Wiley Series on Technologies for the Pharmaceutical Industry
Sean Ekins, Series Editor

Collaborative Computational Technologies for Biomedical Research

Edited by
Sean Ekins
Maggie A.Z. Hupcey
Antony J. Williams

Foreword by Alpheus Bingham

WILEY

COLLABORATIVE COMPUTATIONAL TECHNOLOGIES FOR BIOMEDICAL RESEARCH

Wiley Series on Technologies for the Pharmaceutical Industry
Sean Ekins, Series Editor

Computational Toxicology: Risk Assessment for Pharmaceutical and Environmental Chemicals
Edited by Sean Ekins

Pharmaceutical Applications of Raman Spectroscopy
Edited by Slobodan Šašić

Pathway Analysis for Drug Discovery: Computational Infrastructure and Applications
Edited by Anton Yuryev

Drug Efficacy, Safety, and Biologics Discovery: Enmerging Technologies and Tools
Edited by Sean Ekins and Jinghai J. Xu

The Engines of Hippocrates: From the Dawn of Medicine to Medical and Pharmaceutical Informatics
Barry Robson and O.K. Baek

Pharmaceutical Data Mining: Applications for Drug Discovery
Edited by Konstantin V. Balakin

The Agile Approach to Adaptive Research: Optimizing Efficiency in Clinical Development
Michael J. Rosenberg

Pharmaceutical and Biomedical Project Management in a Changing Global Environment
Scott D. Babler

COLLABORATIVE COMPUTATIONAL TECHNOLOGIES FOR BIOMEDICAL RESEARCH

Edited by

SEAN EKINS
MAGGIE A. Z. HUPCEY
ANTONY J. WILLIAMS

WILEY

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Collaborative computational technologies for biomedical research / edited by Sean Ekins, Maggie A.Z. Hupcey, and Antony J. Williams.
p. cm.
Includes index.
ISBN 978-0-470-63803-3 (cloth)
1. Drug development. 2. Cooperation. 3. Pharmaceutical industry–Data processing. 4. Cloud computing. I. Ekins, Sean. II. Hupcey, Maggie A. (Maggie Anne Zo?), 1972- III. Williams, Antony J.
RM301.25.C65 2011
615'.19–dc22

2010046374

Printed in the United States of America

oBook ISBN: 978-1-11802603-8
ePDF ISBN: 978-1-11802601-4
ePub ISBN: 978-1-11802602-1

10 9 8 7 6 5 4 3 2 1

For Mum and Dad with thanks for letting me follow a route of my own.
Sean Ekins

For Motts, short but loud.
Maggie A. Z. Hupcey

For my twin sons, Taylor and Tyler—two of the best collaborators I know.
Antony J. Williams

In the long history of human kind (and animal kind, too) those who have learned to collaborate and improvise most effectively have prevailed.

Charles Darwin

CONTENTS

FOREWORD

You have in your hands a book on collaboration, more specifically a book on scientific collaboration, and most specifically, a book on collaboration in the science of pharmaceutical development—the discovery of new therapies and medicines—products addressing the, as-yet, unmet medical needs of twenty-first century health. While only a few would take issue with the merits of collaboration, perhaps even *most* fail to appreciate the implications of collaborative technologies in the present day. The ability to fuse ideas—especially ideas that cross disciplines—is a crucial capability responsible for accelerating innovation and progress. Matt Ridley recently gave a TED talk entitled, "When Ideas Have Sex," the salient point being that the fusion of ideas, each bringing its own set of memes, is a powerful way of creating new memetic material.

People have collaborated as long as . . . well . . . as long as there have been people. Often nothing more than self-interest incites us to collaborate, to fill in portions of a solution important to us, portions we were not capable of creating on our own. Unfortunately, modern-day organizational structures very often serve as impediments to collaboration. Collaborating with those outside the walls of an institution may be more than culturally frowned upon, it may even be illegal under legislation written to hinder corporate espionage, or protect trade or national technological capabilities. (I guess if that were the only problem, it could be readily solved by a new set of policies or regulations.)

But institutional boundaries are not the only barriers that impede collaboration. Even *within* an institution—which should be legally, strategically, and financially incented for alignment, and for maximizing the opportunities for internal collaboration—barriers still exist. The subunits of the institution: its departments, its divisions, its components produce collaboration “walls” of varying substantiality. Organizational lore and personal relationships add another layer of “not-invented here” (NIH) culture, and allegiances to local agendas, even to the point of disadvantaging the larger institutional unit. In fact, if we wish to pursue the elimination of collaboration barriers we have to realize that many barriers are not institutional at all. Choices to collaborate or not collaborate are sometimes based not just on current affiliations but on past affiliations, degrees obtained, reputations, and even a less than rational bias as to just who our collaboration partners should be.

A bright spot in recent history has been the open-source movement. It was loosely organized. It was NOT the project management assignment of any large corporate firm filled with project managers looking for substantial development programs like this one. While we acknowledge that there was a component of centralization, that is, Linus Torvald’s role in Linux, the majority of work was exercised in a distributed manner, each module remaining somewhat independent of the constraints often imposed by centralized planning functions. Most importantly, the basis upon which individuals contributed was informed solely by the contribution itself, not perceived qualifications or past reputations.

While the open-source movement has been associated primarily with the development of software, the demonstration that it can compete effectively with the traditional modes of corporate technology development raises the possibility that such collaborative forms will soon move well beyond software and into other arenas of complex development. This is more than mere speculation. In the chapters that follow you’ll see early endeavors to accomplish pharmaceutical development in a much more open manner. While these may still fall short of the phenomenon associated with Linux, they more than hint at a future to come. One barrier to this progression was highlighted in *Harvard Business Review’s* ten best business ideas for 2010; namely, the current lack of a well-accepted and digitized representation of this work. The vast majority of collaborative pharmaceutical development still remains primarily a *local* and classically *social* phenomenon.

While change is still impeded for the reasons described above, the corporate model of the fully integrated pharmaceutical company is under threat for very good reasons. In the past decade, it has shown its inability to create and sustain shareholder value. A closer examination of the business model itself reveals a variety of flaws (or features, if you’d prefer): long monetization cycles, large capital investments with high risks, and a complex union of both information and materials management. We might argue that a typical pharmaceutical company tries to operate, under one roof, three distinctive business entities. It is a high-tech manufacturer, producing exquisitely expensive fine chemicals or complex biotechnical products. It is a purveyor of information to the regulatory and medical communities, information with specifications and demands rarely matched in any other sector. And, finally, it is a high risk research venture, which can only show returns when managed as a portfolio of complex assets demanding constant invention and breakthroughs.

Each of these three business entities would ideally be managed with a distinctive set of overarching strategies and yet such an approach is rarely accommodated. This book addresses, for the most part, only the unique challenge associated with managing large, complex, high-risk research endeavors. But of the three business-entity challenges cited here, a novel new approach to this one could transform the economics of the entire business.

Considering the present state the pharmaceutical industry finds itself in, the promise of innovative medicines for children and our children's children may well depend on finding new collaborative paradigms with attendant business models. The material for this genesis, though nascent, may well be found in these pages.

Alpheus Bingham

April 2011

PREFACE

Biomedical research has become increasingly driven by creating and consuming tremendous volumes of complex data whether biological, genomic, proteomic, metabolomic or molecular in nature. At the same time the pharmaceutical industry is utilizing an extended network of partner organizations of various sorts (CRO's, not-for-profit organizations, clinicians and academics) in order to discover and develop new drugs. Current areas of interest for delivering new technologies or molecules to the industry are Open Innovation, Collaborative Innovation and of course, Open Source. Due to the mounting costs, collaborative research and development is undoubtedly the future of biomedical research. There is currently little if any guidance for managing information and computational resources across collaborations of different types. This represents a large cost as experiments can be repeated inadvertently and the cost and time-savings that could result from precompetitive data sharing have generally been ignored. Improving drug discovery or development technology alone is not the solution and we need intelligent information systems and an understanding of how to use them effectively to create and manage knowledge across these collaborations. This book thoroughly details a real set of problems from the human collaborative and data and informatics aspects and is therefore very relevant to the day-to-day activities of running a laboratory or a collaborative research and development project. The processes, approaches and recommendations provided in this book could be applied to help organizations immediately make critical decisions about managing drug discovery and development partnerships. The chapters provide case histories of biomedical collaborations while the technology specific chapters have effectively balanced technological depth and accessibility for the non-specialist reader. The structure of the book will follow a "*man-methods-machine*" format and the book is divided into four sections:

Part I. Getting People to Collaborate
Part II: Methods and Processes for Collaborations
Part III. Tools for Collaborations
Part IV. The Future of Collaborations

This book may offer the reader a "getting started guide" or instruction on "how to collaborate" for new laboratories, new companies, and new partnerships, as well as a user manual for how to troubleshoot existing collaborations. This book should therefore be of interest to most researchers involved in developing IT systems in the pharmaceutical industry. It should also be particularly pertinent to those leading and participating in collaborative IT consortia for Drug Discovery and Development which are, at the time of writing, increasing in both scope and number.

The book is possible as a result of the contributions of a wide array of authors from pharmaceutical companies, consulting companies, software companies, government institutes, nonprofits, and academia with chapters written by acknowledged pioneers in the field. We have aimed for a complete volume that can be read by all interested in biomedical research and development and with each chapter edited to ensure consistency across the common theme of collaboration and with appropriate explanatory figures and key references. We are confident this book will become a valuable reference work for those interested in collaborative approaches to biomedical research. Certainly this volume represents a point in time for a fast-moving domain of innovation and effort. We hope to revisit this again in the coming years and report on the eventual successes, impacts and shifts in technology as well as cover areas not included in detail.

ACKNOWLEDGMENTS

We are extremely grateful to Jonathan Rose and colleagues at Wiley for their assistance with this book and in particular Bea Roberto for copy editing. Our anonymous proposal reviewers are gratefully acknowledged for their helpful suggestions which, along with other scientists who provided suggestions for additional authors, helped bring this book to fruition.

We are immensely honored that approximately 50 authors agreed to participate sharing their research and ideas and accepting our editorial changes. Clearly this book would have been impossible without their time, effort and input which they provided despite these difficult economic times. This book would have been impossible without their personal sacrifices and collaborations.

We sincerely thank Alph Bingham for the magnificent Foreword and Bryn Williams-Jones for the kind words on the back cover, which they willingly provided at very short notice.

Our authors and ourselves have endeavored to reference as many groups as possible in these chapters but accept and apologize to the many others that may have been unfortunately omitted due to lack of space. We hope we can include you in future volumes!

We acknowledge Tagxedo for the cover image and also made good use of GoogleDocs and its collaborative features when preparing and sharing these chapters. We thank the many scientists that suggested contributors including Dr. Larry Smarr.

Our own research owes a great deal to past, present (and doubtless future) collaborators and we acknowledge them for helping to stimulate this book.

In order to better serve our readers, color versions of selected illustrations from this book can be found at the following ftp address:

ftp://ftp.wiley.com/public/sci_tech_med/collaborative_computational

Finally, we dedicate this book to our families that have followed this project and provided us the time and support to do it.

Sean Ekins
Maggie A. Z. Hupcey
Antony J. Williams

Jenkintown, Pennsylvania
Wake Forest, North Carolina
April 2011

CONTRIBUTORS

Santosh Adayikkoth, Ph.D., Infosys Technologies Limited, Electronic City, Bangalore, India

Renée J. G. Arnold, Pharm.D., R.Ph., Arnold Consultancy & Technology LLC, New York, New York; Master of Public Health Program, Department of Preventive Medicine, Mount Sinai School of Medicine, New York, New York; Division of Social and Administrative Sciences, Arnold and Marie Schwartz College of Pharmacy, Long Island University, Brooklyn, New York

O. K. Baek, IBM Canada Ltd., Markham, Ontario, Canada

Anshu Bhardwaj, Ph.D., Institute of Genomics and Integrative Biology (IGIB), CSIR, Delhi, India

Alpheus Bingham, Ph.D., Cascade Consulting, Carmel, Indiana; InnoCentive, Inc., Waltham, Massachusetts; Monitor Talent, Cambridge, Massachusetts

Jean-Claude Bradley, Ph.D., Department of Chemistry, Drexel University, Philadelphia, Pennsylvania

Samir K. Brahmachari, Ph.D., Council of Scientific and Industrial Research (CSIR), Institute of Genomics and Integrative Biology (IGIB), Delhi, India

Vincent Breton, Ph.D., Laboratory of Corpuscular Physics, Clermont University and University Blaise Pascal, Clermont-Ferrand, France

Barry A. Bunin, Ph.D., Collaborative Drug Discovery, Burlingame, California

Christine Chichester, Ph.D., Netherlands Bioinformatics Center, Nijmegen, The Netherlands

Gabriela Cohen-Freue, Ph.D., PROOF Centre of Excellence, Vancouver, British Columbia, Canada

Ramesh V. Durvasula, Ph.D., Bristol-Myers Squibb Company, Princeton, New Jersey

Sean Ekins, Ph.D., D.Sc., Collaborations In Chemistry, Jenkintown, Pennsylvania; ACT LLC, New York, New York; Collaborative Drug Discovery, Burlingame, California; Department of Pharmaceutical Sciences, University of Maryland, Baltimore, Maryland; Department of Pharmacology, University of Medicine and Dentistry of New Jersey, Robert Wood Johnson Medical School, Piscataway, New Jersey

Rajarshi Guha, Ph.D., NIH Chemical Genomics Center, Rockville, Maryland

Brian D. Halligan Ph.D., Biotechnology and Bioengineering Center, Medical College of Wisconsin, Milwaukee, Wisconsin

Zhiyu He, Ph.D., Graphics, Visualization and Virtual Reality Laboratory (GRAVITY), University of California, San Diego, California

David Hill, Ph.D., Clermont University, University of Blaise Pascal, LIMOS, Clermont-Ferrand, France

Moses M. Hohman, Ph.D., Collaborative Drug Discovery, Burlingame, California

Zsuzsanna Hollander, M.Sc., PMP, PROOF Centre of Excellence, Vancouver, British Columbia, Canada

Victor J. Hruby, Ph.D., Department of Chemistry and Biochemistry, University of Arizona, Tucson, Arizona

Jackie Hunter, Ph.D., OI Pharma Partners, Ltd. Red Sky House, Fairclough Hall Farm, Halls Green, Weston, Hertfordshire, United Kingdom

Maggie A. Z. Hupcey, Ph.D., PA Consulting Group, Princeton, New Jersey

Steve Koch, Ph.D., Center for High Technology Materials, Albuquerque, New Mexico

George A. Komatsoulis, Ph.D., Center for Biomedical Informatics and Information Technology (CBIIT), National Cancer Institute, National Institutes of Health, U.S. Department of Health and Human Services, Rockville, Maryland

Falko Kuester, Ph.D., Graphics, Visualization and Virtual Reality Laboratory (GRAVITY), University of California, San Diego, California

Andrew S. I. D. Lang, Ph.D., Department of Computer Science and Mathematics, Oral Roberts University, Tulsa, Oklahoma

Nick Lynch, Ph.D., AstraZeneca UK Limited, Alderley Park, Macclesfield, United Kingdom

Robert Porter Lynch, Ph.D., The University of Alberta Edmonton, Alberta, Canada and The University of British Columbia, Vancouver, British Columbia, Canada

Lydia Maigne, Ph.D., Laboratory of Corpuscular Physics, Clermont University and University Blaise Pascal, Clermont-Ferrand, France

Shawnmarie Mayrand-Chung, Ph.D., J.D., National Institutes of Health, Public-Private Partnerships Program—Office of Science Policy Analysis, Office of the Director, Bethesda, Maryland

Garrett J. McGowan, Ph.D., Chemistry Department, Alfred University, Alfred, New York

Matthew K. McGowan, Ph.D., Foster College of Business Administration, Peoria, Illinois

Richard J. McGowan, Ph.D., Philosophy and Religion Department, Butler University, Indianapolis, Indiana

Barend Mons, Ph.D., Netherlands Bioinformatics Center, Nijmegen, The Netherlands

Mark A. Musen, Ph.D., Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, California

Cameron Neylon, Ph.D., STFC Rutherford Appleton Laboratory, Harwell Science and Innovation Campus, Didcot, Oxfordshire, United Kingdom

Christina K. Pikas, Doctoral Candidate, College of Information Studies, University of Maryland, College Park, Maryland

Kevin Ponto, Ph.D., Graphics, Visualization and Virtual Reality Laboratory (GRAVITY), University of California, San Diego, California

Brian Pratt, Insilicos LLC, Seattle, Washington

David Sarramia, Ph.D., Laboratory of Corpuscular Physics, Clermont University and University Blaise Pascal, Clermont-Ferrand, France

Vinod Scaria, Ph.D., Institute of Genomics and Integrative Biology (IGIB), CSIR, Delhi, India

Stephan Schürer, Ph.D., Department of Pharmacology, Miller School of Medicine, Center for Computational Science, University of Miami, Miami, Florida

Jeff Shrager, Ph.D., Symbolic Systems Program (consulting), Stanford University, Stanford, California; CollabRx., Inc., Palo Alto, California

Robin W. Spencer, Ph.D., Pfizer Inc. (retired), United States

Ola Spjuth, Ph.D., Department of Pharmaceutical Biosciences, Uppsala University, Uppsala, Sweden

Sándor Szalma, Ph.D., Centocor R&D, Inc. and Johnson & Johnson Corporate Office of Science and Technology, San Diego, California ; Rutgers, The State University of New Jersey, New Brunswick, New Jersey

Keith Taylor, Ph.D., Accelrys, Inc., San Ramon, California

Marty Tenenbaum, Ph.D., CollabRx., Inc., Palo Alto, California

Zakir Thomas, Ph.D., Council of Scientific and Industrial Research (CSIR), Rafi Marg, New Delhi, India

Michael Travers, Ph.D., CollabRx., Inc., Palo Alto, California

Tania Tudorache, Ph.D., Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, California

Chris L. Waller, Ph.D., Pfizer, Inc., Groton, Connecticut

John Wilbanks, Ph.D., Creative Commons, San Francisco, California

Antony J. Williams, Ph.D. F.R.S.C., Royal Society of Chemistry, Wake Forest, North Carolina

Egon Willighagen, Ph.D., Department of Pharmaceutical Biosciences, Uppsala University, Uppsala, Sweden

Edward D. Zanders, Ph.D., BioVillage Ltd., St. John's Innovation Centre, Cambridge, United Kingdom

PART I

GETTING PEOPLE TO COLLABORATE

1

NEED FOR COLLABORATIVE TECHNOLOGIES IN DRUG DISCOVERY

Chris L. Waller, Ramesh V. Durvasula, and Nick Lynch

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.

1.1 INTRODUCTION

From its accidental beginnings in Alexander Fleming's laboratory, pharmaceutical drug discovery and development has emerged as a multi-billion-dollar industry that has revolutionized practically all aspects of human (and animal) life as we know it. Over the past 100 years, serendipitous discovery has been replaced by a structured process that in its current state is highly structured, automated, and regulated. It is also expensive and lengthy and suffers from a 99% failure rate. Industry averages suggest that the cost to bring a new drug to the market under this so-called blockbuster paradigm is in the neighborhood of $1.5–2.0 billion and takes nearly 16 years (Fig. 1.1) [1].

1.1.1 Brief History of Pharmaceutical Industry

The origins of the pharmaceutical industry can be traced back to the 1800s and the dye industry in Switzerland. From the dye industry, specialty chemistry companies emerged with Ciba, Geigy, and Sandoz in Switzerland along with Bayer and Hoechst in Germany evolving into the first pharmaceutical companies. In the early 1900s, the center of pharmaceutical research and development (R&D) migrated to the United States, specifically New Jersey, with companies such as American Home Products, Johnson & Johnson, Warner Lambert, Merck & Co., Pharmacia-Upjohn, Schering-Plough, BASF, Hoechst, Schering AG, Hoffman LaRoche, and Novartis making it the location of choice for their U.S. operations. The late 1900s saw the emergence of North Carolina as a pharmaceutical industry hot spot with Glaxo-Wellcome making its U.S. headquarters there. Also in the late 1900s, the biotechnology industry emerged

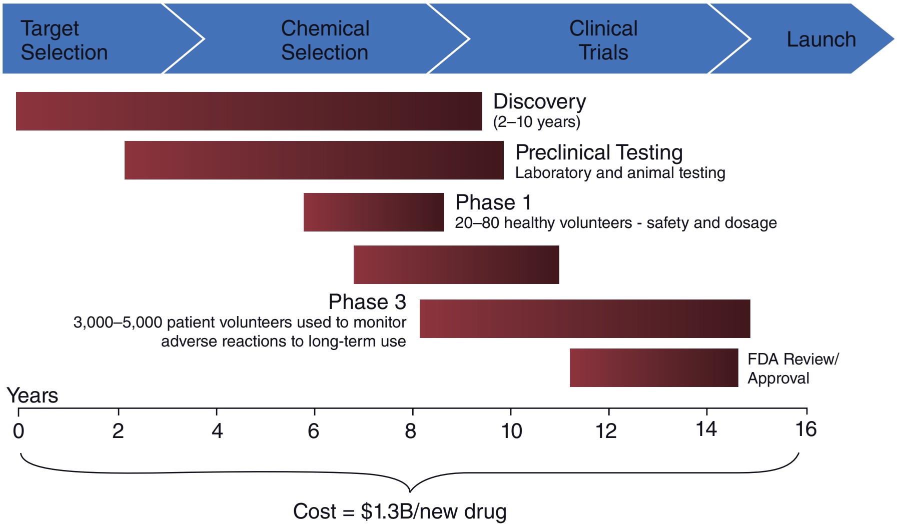


Figure 1.1 Pharmaceutical research and development process.

with companies congregated in the Boston/Cambridge area; the San Francisco Bay Area, San Diego, California; Princeton, New Jersey; Washington, D.C., metro area; as well as Philadelphia. In recent years the economic pressures that forced the pharmaceutical industry to think differently about the sourcing of many operational commodity services has driven a trend toward the emergence of both large pharmaceutical and biotechnology footprints in emerging markets such as Brazil, Russia, India, and China (the traditional BRIC countries) as well as Indonesia [2].

1.1.2 Brief History of Biotechnology

The biotechnology "revolution" began in earnest in 1976 with the founding of Genentech. Inspired by similar movements over the past century in the semiconductor, computer, and advanced materials business, a business model was adopted that would see science evolve from being a tool for the creation of new products and services to being the business itself. Science would move from being "outside" of the business to being the actual business. Genentech was founded as the first of a number of private firms that would monetize the basic research process. Herbert Boyer, an academician, and Robert Swanson, a venture capitalist, invested $500 each into a new business venture that would seek practical uses for the engineered proteins being developed in Boyer's laboratory [3]. Genentech remains one of the largest and most successful of the biotech companies, posting revenues in 2008 in excess of $10 billion, and is now wholly owned by Roche. The Genentech business model continues to be cloned as academicians seek venture capital to advance their ideas and blend science and business.

Despite the business success seen by some of the biotechnology companies, the vast majority of the entrants into this field failed. The business environment imagined (and required) by this new sector was one in which pharmaceutical (R&D) activities were organized through a web of collaborative agreements between the traditional large pharmaceutical and newer biotechnology companies. This collaborative network was envisioned to dramatically alter the industry and transform human health through improved products and services. In reality, while the biotechnology sector has seen exponential growth in revenues over the past 25 years, operational income has been flat or negative, and there has been no discernable difference in research and development productivity as measured by new drug launches. However, the biotechnology sector has contributed to the diversity of treatments in the world's medicine chest. In 2008, 31 new medicines were launched, 10 of biologics (non-small-molecule) origin, the preferred modality of the biotechnology sector [4].

The promise of transformation of the health care industry brought about by the emergence of "science business" biotechnology companies has failed to materialize due to fundamental differences between the pharmaceutical (R&D) business and the organizational models indiscriminately borrowed from the semiconductor industry. Science-based businesses face unique challenges not present in these other industries, and the focus on monetization of intellectual property, rather than products or services, has actually been detrimental to the creation of the collaborative network envisioned by the early pioneers of the biotechnology movement. Specifically, this misaligned focus has led to (1) the creation of numerous information silos and barriers to sharing—a key requirement for collaboration, (2) fragmentation of the industry and duplication of noncompetitive activities, and (3) a proliferation of new firms competing for resources from a limited pool [5].

1.1.3 Brief History of Government-Funded Academic Drug Discovery

In 1980, the Bayh-Dole Act was enacted with the intention to stimulate pharmaceutical research into key disease areas by allowing academic institutions as well as individual researchers to benefit directly from commercialization of their government-funded research efforts. Although greatly criticized as a mechanism that promotes science with no direct market relevance [6], government-funded research spending is significant and increasing. Across the National Institutes of Health (NIH), a number of "center grants" have been awarded over the last several years to build out the necessary infrastructure to power an academic revolution. Examples of the types of work being supported are as follows: (1) Burnham was awarded a $98 million grant to establish one of *four* comprehensive national screening centers as part of the NIH's, Molecular Libraries Probe Production Centers Network (MLPCN); (2) 83 National Center for Research Resources (NCRR)–funded Centers of Biomedical Research Excellence (COBRE) have been awarded two consecutive, five-year, $10 million grants; (3) Northwestern is awarded $11 million to create a Center to Speed Drug Discovery (Northwestern); and (4) a grant from the NIH will help establish the Chicago Tri-Institutional Center for Chemical Methods and Library Development. The NIH will pump $62 million into more than 20 studies focused on using epigenomics to understand how environmental factors, aging, diet, and stress influence human disease.

In 2008, the National Cancer Institute (NCI) alone funded research efforts in excess of $12 billion. More recently, the NCI has been funding efforts that would increase the value of academic research through the creation of public–private partnerships to translate knowledge from academia into new drug treatments. To this end, the NCI has established the Chemical Biology Consortium, which is advertised as an integrated network of chemical biologists, molecular oncologists, and chemical screening centers. Current members of the consortium include. The University of North Carolina in Chapel Hill, North Carolina; Burnham Institute for Medical Research in La Jolla, California; Southern Research Institute in Birmingham, Alabama; Emory University in Atlanta; Georgetown University in Washington, D.C.; the University of Minnesota in St. Paul and Minneapolis; the University of Pittsburgh and the University of Pittsburgh Drug Discovery Institute; Vanderbilt University

Medical Center in Nashville, Tennessee; SRI International in Menlo Park, California; and the University of California at San Francisco.

Like the biotechnology revolution of the late 1970s, the current trend in the creation of networks of public and private institutions, if successfully operationalized, could transform the health care industry. It is important to acknowledge the lessons from the biotechnology revolution as discussed above and plan accordingly to avoid the pitfalls. In order to be successful, the academic institutions must strive to establish truly open and standard data exchange mechanisms and coordinate activities effectively across a highly distributed enterprise that must adopt an integrated business process.

1.2 SETTING THE STAGE FOR COLLABORATIONS

A reorientation of our business models to focus on products and services will be required if the collaborative R&D environment is to be effectively realized. An acknowledgment, by the industry as a whole, must be made that we differentiate ourselves in the marketplace not through our intellectual property but rather through the delivery of products and services that attract and retain consumers. The R&D process, in any industry, is timely, expensive, and, except for those rare instances where true discoveries/inventions are being made, commoditizable across the industry in the sector. A clear understanding and declaration of what differentiates one company from the next in the marketplace must be established and adopted. Only then can we begin to pool our limited resources effectively to solve common problems and focus our specific internal resources on the elements of the R&D process that allow us to transform the health care system and succeed in the marketplace as individual companies.

1.2.1 Current Business, Technical, and Scientific Landscape

The business value of an information technology (IT) system is based on the ability of the system to support and enhance the business process. Fundamentally, open standards are intended to provide resilience to withstand the technical volatility within business processes and their associated systems. If a system and the business process were flawlessly stable over many years, then there would be little value in developing and adopting standards. However, within the pharmaceutical industry, volatility and upheaval abound in every phase of R&D. Perhaps the largest source of upheaval within our industry is the volatility of mergers and acquisitions (M&A) among industry peers as well as business partners, commercial suppliers, and clinical research organizations (CROs) (Fig. 1.2). This M&A volatility—coupled with exponential growth in outsourcing—has placed tremendous pressure on R&D processes to change frequently and dramatically. Common pharmaceutical processes like target identification, compound synthesis, in vivo toxicology, biomarker discovery,

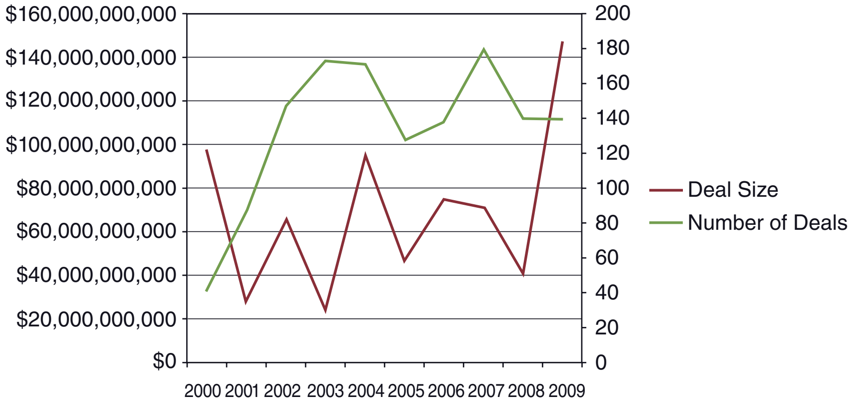


Figure 1.2 Pharmaceutical M&A activity, 2000–2009. (*Source:* http://www.marketwatch.com/story/ten-year-data-on-pharmaceutical-mergers-and-acquisitions-from-dealsearchonlinecom-reveals-top-deals-and-key-companies-2010-03-25. MarketWatch data based on original content from DealSearchOnline.com.)

patent searching, and pharmaceutics are all experiencing revolutions in their processes. The related systems are thus also reacting to this process volatility. This upheaval in the requirements and specifications of R&D IT systems is causing IT budgets to increase, exactly at the moment when all budgets across R&D are sharply decreasing.

We face an unprecedented era of rising process upheaval and constantly evolving business requirements coupled with a cost-conscious environment where chief information officers (CIOs) and R&D executives are looking to simplify their IT architectures and their cost basis. If this trend continues, informatics systems may become a bottleneck to the productivity of pharmaceutical scientists.

1.2.2 Externalization of Research: Collaboration with Partners

The area of greatest process upheaval is the externalization of research processes and the growing collaborations between life science partners throughout the R&D cycle. Originally CROs had been outsource partners, but currently there are outsourcing partners for every phase of the R&D process, from target identification to chemical synthesis to pharmacokinetic studies to clinical supplies, and so on. With this increased opportunity and necessity for outsourcing, samples are constantly getting shipped to and from pharmaceutical laboratories. Every time a sample changes hands, there is a related data exchange as well. Often, for a pharmaceutical company, several CRO partners will be used for a single research project. Also, the CRO will likely have several pharmaceutical clients. In this emerging net-centric industry model, there is a complex graph of data exchange that must be supported (Fig. 1.3).

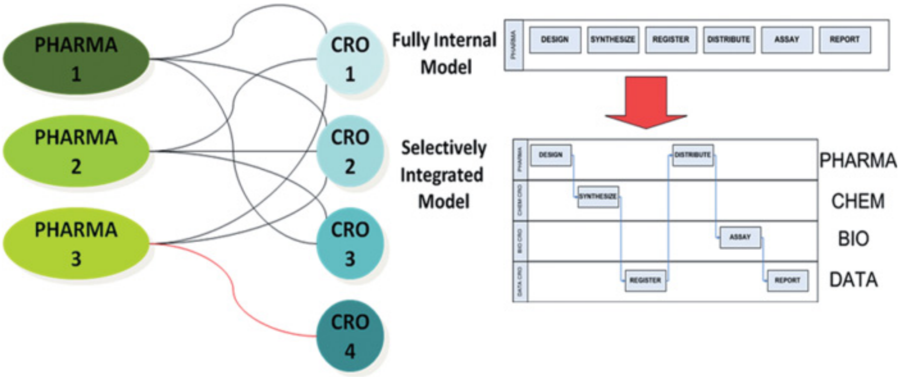


Figure 1.3 Emergence of a selectively integrated drug discovery and development model.

For example, for every pharmaceutical company, there may be two or three chemistry synthesis partners. These partners would likely have their own internal systems for tracking reagents, recording experiments, and registering novel compounds. Since the synthesis is performed on behalf of the pharmaceutical client, a majority of the data from the experiment, from reaction yields to analytical data, must be transmitted to the client along with the synthesized compound in a vial. The challenge is that since the pharmaceutical client has developed mature internal processes, and the synthesis partner has its own internal processes, there is a high likelihood that the processes—and the related IT systems—are different in nature. This leads to the use of different metadata, different vocabularies, and different quality control on the data capture. When an instance of a novel compound is synthesized, the outsource partner may call it a "batch" but the pharmaceutical client may call it a "lot". Also, some compound registration systems assign a different identifier for different salt forms of the compound. One company may handle this by using a suffix of the compound identifier (<compound identifier>–<salt form>), whereas another company may simply assign a completely different base compound identifier to the different salt form. Both of these are legitimate taxonomies to register and identify compounds and their salt forms. The difficulty comes when one company attempts to export its registration data and transmit that to the other company. Reconciling the differences in the semantics and vocabularies of different compound registration systems can be a tedious, error-prone, and often irreconcilable task. Often this reconciliation involves compound registrars and synthetic chemists (and possibly lawyers) from both parties. If the need to transmit compound registration data between business partners was a unique event, then perhaps a manual reconciliation process would suffice. However, since every pharmaceutical company has several synthesis outsourcing partners, and every synthesis CRO has several pharmaceutical clients, this metadata-conflict and reconciliation process is repeated over and over throughout the industry. While this problem of data reconciliation and reformatting is time consuming and error prone in the chemical synthesis domain, this problem is often even more exacerbated in the biological domain.

Often pharmaceutical companies will have outsourcing relationships with contract laboratories that perform assays on compounds owned by the client. These assays could be standard assays that are outsourced for cost efficiencies or proprietary assays that are otherwise not available to the pharmaceutical client. As with compound registration systems, the outsource partner that runs the assays will likely have internal protocol registration and biological assay data management systems to capture the data. These systems will be built to suit the needs of the internal processes within the contract laboratory, so that they can properly manage, interpret, and report on their assay results. However, most pharmaceutical companies like to import the assay results into the pharmaceutical company's internal assay data management system. This would enable the pharmaceutical scientists to interpret the outsourced assay data side by side with all of the other data generated on that proprietary compound. With every partner that generates assay data related to a compound, there is an ongoing, complicated effort to properly format and transmit the data such that the scientists in the pharmaceutical company can understand the nature of the assay and accurately interpret the results. Too often, many days are wasted merely explaining differences between internal and external assay results. Especially with high-throughput or high-content biological assays, there are a significant number of attributes of the experimental design that are important to account for in the data interpretation. For example, which cell line was used? Was it a single-point assay or a dose–response? What was the detection mechanism; fluorescence, phosphorescence, and so on? Furthermore, there are many cases where the proprietary assay platform generates data that have a unique structure.

Perhaps the assay is a high-throughput, low-resolution format, in which case the raw numeric output must be binned into low–medium–high categories and only the binned values are reported to the client, yet the client has stringent data quality, numbers-only rules to which the contract laboratory cannot adhere. Perhaps the assay has a cutoff at a reading threshold, causing the result to be reported as a range instead of an explicit number. Perhaps there is a nonlinear response that requires special curve-fitting software to calculate the half maximal inhibitory concentration (IC_{50}) value. There are many nuances and subtleties to biological assay data, and a large amount of metadata is required to properly describe the experimental method. This must be understood by the scientist who is using that assay data to make design or synthesis decisions for the next molecule. As such, it is important for the contract laboratory to deliver the full experimental description of its data and for the pharmaceutical customer to ingest and report all of that description to its scientists. Again, as with compound synthesis, if this assay data generation was done with a single partner, then a manual process with significant interactions between business partners would be appropriate. However, pharmaceutical companies often send their compounds to many laboratories to be tested in numerous assays, and all of that data must be imported into the assay database of the client, and the data must be interpreted by chemists and biologists who are not the operators of those assays. The further downstream the assay if the assay was an in vivo assay, as opposed to an in vitro assay—the more complicated the experimental design, and thus the harder it is for scientists to interpret the data without being proximal to the biologist who performed the assay.

Both the chemistry and biology examples above highlight the cost and complexity of exchanging data between business partners, and the activities of data exchange and data harmonization are not value-added work for finding drugs. These data tasks are a cost of doing business in life sciences, and as such the industry is looking for ways to reduce these costs without impacting the science. In fact, it could be argued that resources poured into the data activities are actually *diverting* funds away from doing science. So, reducing these costs will actually free up resources to do more science. The challenge of reducing these data-curation costs is that no single entity, neither a pharmaceutical company nor a contract laboratory nor a biotech, can accomplish what is needed to be done, namely to harmonize across the industry. Point-to-point optimizations of data exchange are helpful but only marginally cost effective. For a paradigm shift to occur that would dramatically improve the efficiency of external science, the industry must come together to agree on common methods of exchanging data, delivering services, defining entities, and so on. Thus, a precompetitive collaboration among informatics groups is a natural evolution in our industry. This evolution has already occurred in numerous other industries, from apartments [7] to banking [8] to retail [9].

The nature of every industrywide data standardization effort revolves around defining the terminology, semantics, metadata, entity attributes, and services or functions of the data exchanged between business partners. These definitions and attributes are collaboratively defined by IT or informatics peers who together determine how to harmonize data between disparate systems and processes.

1.3 OVERVIEW OF VALUE OF PRECOMPETITIVE ALLIANCES IN OTHER INDUSTRIES

Other industries have realized the need for precompetitive alliances for some time and have established them over the last two decades. This drive for collaborative alliances has been driven by the same pressures that the life science industry faces today, that of increased pressures on efficiency and the need to divert funding to innovative activities rather than to commodity services. The maturity of the business model for these other industries (telecoms, insurance, automotive, and aerospace) has meant that they have existed prior to work within the early stages of life science and informatics. These other industries realized early on that each company existed as part of an extended ecosystem that relied on the ability to do business with other partners and competitors and hence where the need for interoperable processes and information flows were critical to their mutual success.

1.3.1 Overview of Existing Precompetitive Alliances

Without going into details on all the other industries, some have direct parallels with discovery life science from both other life science areas and financial services. The financial services industry created the VISA processing standards and in creating this concept has led to an explosion in the ways that credit cards are used and their ease of interoperability. Other examples of open approaches include the insurance industry (Polaris) to support data exchange between insurance brokers and the insurance companies offering the policies. In the clinical development workflow of development pharmaceuticals the need to work with multiple partners as part of the delivery of clinical trials and the later delivery of health care services to patients has provided the environment for groups such as the Clinical Data Interchange Standards Consortium (CDISC: www.cdisc.org) and Health Level 7 (www.hl7.org) to be founded and evolve over several years. The drivers here were a need for interoperable standards for information delivery and data markup to support effective and clear communication for submission of clinical trials data and the later management of health care information.

The way these companies do business has changed as the global economy has evolved, but delivering critical information to scientists continues to be the key part of the R&D informatics groups within these pharmaceutical and agrochemical companies and support organizations. There are various ways that the development of software and delivery of information to scientists can be improved through collaboration and open standards. There is evidence from other global businesses where strong open standards have benefited a whole industry sector and delivered improved innovation in the face of cost pressures.

1.3.2 Pistoia Alliance: Construct for Precompetitive Collaborations

There has been a history of organizations working together to promote common standards in the early-stage life science industry over the last decade both as new groups established specifically for life science [Interoperable Information Infrastructures Consortium (I3C: www.i3c.org), Society for Bimolecular Sciences (SBS: www.sbs.org), BioIT Alliance (www.bioitalliance.org)] and those attached to larger groups but wishing to explore and adapt into life science [Object Management Group (OMG: www.omg.org), World Wide Web Consortion (W3C: www.w3c.org)]. The success rate has been variable over the years with various initiatives coming and going and others building a portfolio of activities and evolving. Much of the thinking of setting up the Pistoia

Alliance (www.pistoiaalliance.org) has tried to take the learning from these other groups and understand how they were able to deliver collaborative value.

1.3.3 How Does Pistoia Plan to Differentiate Itself?

There are various factors that we believe make the Pistoia Alliance work slightly differently, including a changing economic environment that is forcing more collaboration and improvements in software design that focus on software services which allow a high level of abstraction and hence more opportunity for cross-company integration. The high-level business processes executed within this sector are very similar between different organizations, and the further appreciation that there is considerable overlap and commonality in the processes executed within the sector has made groups question what is competitive advantage and what are supporting assets that could share some common design (Fig. 1.4).

A key element for the establishment of the Pistoia Alliance was ensuring that the life science business needs were the driving force for the development of common standards and approaches in the group rather than simply a technology/solutions focused view. Hence the projects that have evolved in the first build of the Pistoia Alliance program are intended to show these drivers from developing service requirements (sequence services) and an open

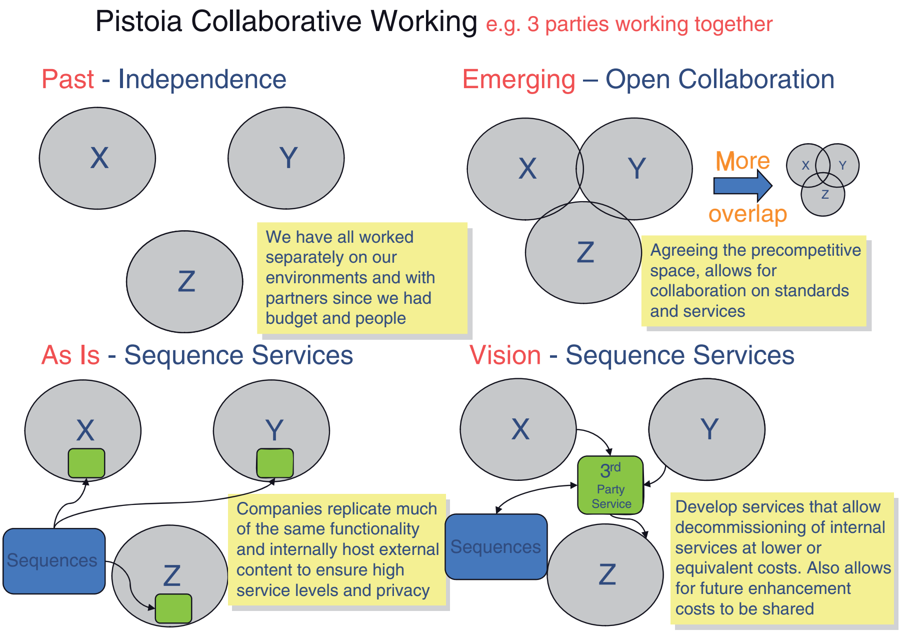

Pistoia Collaborative Working e.g. 3 parties working together
Past - Independence
X
Y
Z
We have all worked separately on our environments and with partners since we had budget and people
Emerging – Open Collaboration
X
Y
Z
More
overlap
X
Y
Z
Agreeing the precompetitive space, allows for collaboration on standards and services
As Is - Sequence Services
X
Y
Z
Sequences
Companies replicate much of the same functionality and internally host external content to ensure high service levels and privacy
Vision - Sequence Services
X
Y
Z
3rd Party Service
Sequences
Develop services that allow decommissioning of internal services at lower or equivalent costs. Also allows for future enhancement costs to be shared

Figure 1.4 Pistoia Alliance collaborative working model.

framework based on existing standards (SESL). The key intention of the Pistoia Alliance was to move beyond standards in their adoption as service requirements and into influencing future business models and be a potential for change in the delivery of information and services in the life science industry. The next-generation business model would ideally shift from products (software programs or databases that need to be installed and maintained) to services (accessing data on Web-based platforms or hosted off-site), eventually maturing to "software as a service," known as SaaS, which would be deployed over the Internet. Standard interfaces, such as those used by Web browsers, would make it easier to simplify IT architectures across the industry, and centralized services would deliver economies in scale and scope. Among the major benefits would be reductions in cost and maintenance as information silos inside company networks are turned off in favor of fewer, more versatile tools. The Alliance has a broad membership because such extensive changes in the business model affect all parts of the supply chain, from life science back to software providers and content providers.

We want to have all parties [suppliers, academics, nongovernmental organizations (NGOs), pharma, and life science companies] actively involved in the Alliance's initiatives, as the intent is to deliver practical pilots and prototypes that demonstrate the collaborative activity. The Pistoia Alliance differentiates itself from groups both past and present through its attempts to embrace and extend the standards and services of these companion groups in technology offerings driven by clear business needs. We wish to adopt existing standards where we can rather than create new ones and also collaborate with existing groups to bring fresh ideas into the value chain. We list a selection of our current portfolio that highlights our current foci and also the wider impact on the information delivery models.

1.3.4 Overview of Current Pistoia Projects

1.3.4.1 SESL—Semantic Enrichment of Scientific Literature The Pistoia Alliance project on biomedical knowledge brokering standards (SESL) is developing a pilot to showcase its key approaches, and its aim is to demonstrate the feasibility of an open knowledge brokering framework which will reduce the costs of integration of disparate data types from several sources. The pilot is focused on the extraction of assertions for type II diabetes mellitus (T2DM) from both the scientific literature, supplied by participating publishers, and structured data resources managed by EMBL-EBI (the European Bioinformatics Institute). The pilot [expected to include an (resource description framework (RDF) triple store] will be published and a prototype demonstrator will be made publicly available to show feasibility (Fig. 1.5).

1.3.4.2 Sequence Services Most major pharmaceutical companies currently host a large number of sequence data and analysis tools within their firewalls. While the genome was still being sequenced, and during the race to patent

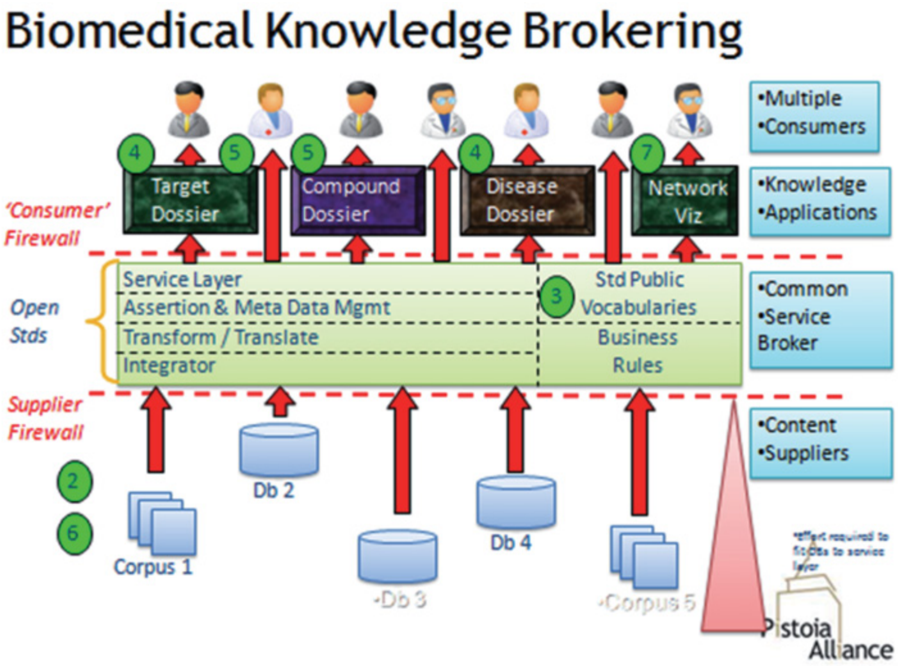


Figure 1.5 Schematic Architecture for SESL project.

genes, these services offered a competitive advantage, and consequently each company built and maintained vast internal systems that both took external public data and merged it with internal private data. However, in the past five years the public domain has caught up (and in many cases surpassed) the expensive, heavily customized commercial and proprietary solutions used by industry.

As a drive to cuts costs, encourage standards, and provide simplification, the Pistoia Alliance is commissioning a pilot set of secure hosted sequence services based on the functional and nonfunctional requirements of its members. These services will provide access to public, private, and commercial data and tools that will enable scientists to search, store, and analyze all their sequence-based data in a single Web interface. Additionally data will be searched and accessed via Web services to allow sophisticated users to flexibly retrieve or pipeline data (Fig. 1.6).

1.3.4.3 ELN Query Services The adoption of an electronic laboratory note book (ELN) within an organization is as much a business change process as it is a technology project, and so the ELNs have traditionally had to focus on the role of the experimental scientist entering new information and ensuring this process is managed and efficient. In areas where ELNs have been used for a few years, such as supporting chemistry synthesis (medicinal chemistry,

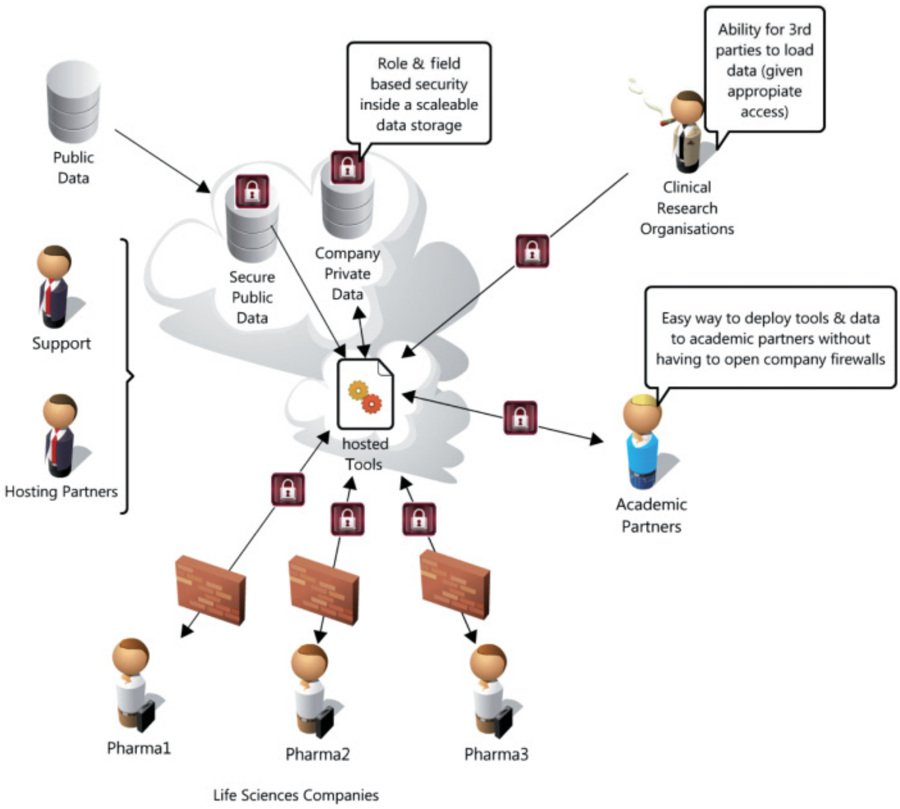

Public Data
Role & field based security inside a scaleable data storage
Ability for 3rd parties to load data (given appropiate access)
Secure Public Data
Company Private Data
Clinical Research Organisations
Support
Hosting Partners
Easy way to deploy tools & data to academic partners without having to open company firewalls
hosted Tools
Academic Partners
Pharma1
Pharma2
Pharma3
Life Sciences Companies

Figure 1.6 Conceptual view for the sequence service project.

process chemistry, operations, and manufacturing), there is a growing demand for enhanced exploitation of the data held within an ELN and the future linking of that data with relevant data held within an organization or further afield. The requirements for knowledge management have grown considerably in the last few years, and this increases the need to query the ELN to extract the high-value information and to build assertions with other data from within an organization or outside (Fig. 1.7).

As the number of ELN installations grows, this requirement becomes more challenging, particularly given the diversity of such ELN implementations (developed commercially, in-house, blended, or as open-source systems). In many companies already a mixture of ELNs have been deployed, either through conscious choice or as a result of mergers and acquisitions. Another key factor is the trend for more business process outsourcing, resulting in the need to be able to work with a CRO partner and share aspects of an ELN knowledge base. So the problem the industry faces is twofold: (1) the need for

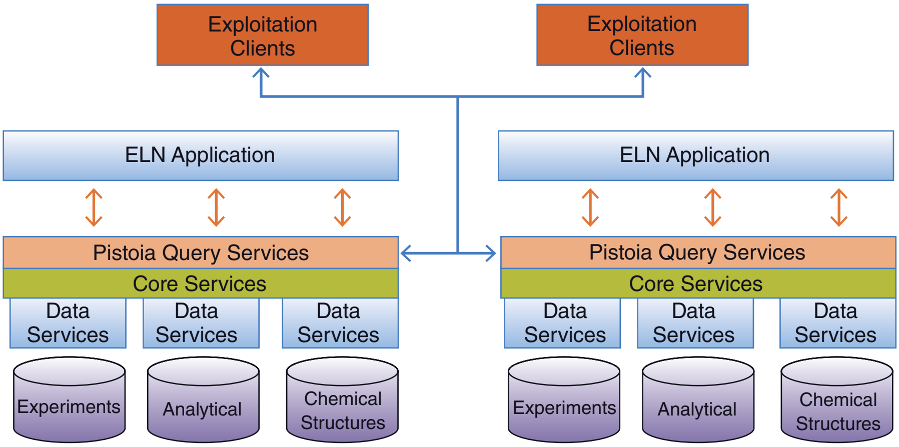


Figure 1.7 Conceptual vision for ELN project.

better exploitation of ELN data and (2) the need to build different ELN implementations using different domain models and designs.

1.4 CONCLUSION

A precompetitive collaboration, the Pistoia Alliance, has been established to provide the foundation of data standards, ontologies, and associated Web services to enable pharmaceutical discovery workflow through common business terms, relationships, and processes. The initial focus has been on chemistry, biological screening, and sample logistics. All pharma companies and software vendors are challenged by the technical interconversion, collation, and interpretation of drug/agrochemical discovery data, and as such, there is a vast amount of duplication, conversion, and testing that could be reduced if a common foundation of data standards, ontologies, and Web services could be promoted and ideally agreed upon within a nonproprietary and noncompetitive framework. This would allow interoperability between a traditionally diverse set of technologies to benefit the health care sector.

REFERENCES

1. *Outlook 2010*. Boston, MA: Tufts University, 2010.
2. Research and development in the pharmaceutical industry. Washington, DC: U.S. Congressional Budget Office, 2006.
3. WGBH. Herbert Boyer. *They Made America*. Boston: PBS, 2004.

4. PhRMA Annual Report 2009. Pharmaceutical Research and Manufacturers of America, 2009.
5. Pisano GP. *Science Business: The Promise, the Reality, and the Future of Biotech*. Boston, MA: Harvard Business Press, 2006.
6. Loewenberg S. The Bayh-Dole Act: A model for promoting research translation? *Mol Oncol* 2009;3:91–92.
7. Multifamily Information and Transactions Standard. Available: http://www.mitsproject.com.
8. Transactions Workflow Innovation Standards Team. Available: http://www.twiststandards.org.
9. The Association for Retail Technology Standards. Available: http://www.nrf-arts.org.

2

COLLABORATIVE INNOVATION: ESSENTIAL FOUNDATION OF SCIENTIFIC DISCOVERY

Robert Porter Lynch

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.

2.1 DAWNING OF ERA OF COLLABORATIVE INNOVATION

As the twentieth century ended, the computer, followed by the explosive growth of the internet, spawned a worldwide "Era of Information." With this profusion of information and data, knowledge itself, for the first time in the history of the human race, has become a commodity. As a commodity, the value of knowledge is not in the information or data; the real value manifests when transformed into how it is (1) applied, (2) integrated, and (3) triggers innovation. Until it is transformed into one of these three areas, knowledge remains data, trivia, or useless information.

Information that used to be proprietary, inaccessible, expensive, or limited to a few elite scholars is now available to virtually everyone and mostly free. Everyone with Internet access has at their fingertips nearly all the world's knowledge. However, it takes more than a grasp of what is known to solve the great problems on the planet: disease, poverty, energy, world peace, or global warming, to name a few.

Knowledge is rooted in what has *already* been learned; thus it is *historic* in nature—the reason Einstein said, "Creativity is more important than knowledge." Creativity, imagination, and inquisitiveness coupled with the ability to cooperate are some of the human being's most endearing characteristics and constitute the foundation of collaborative innovation.

Difficult problems cannot be solved by existing knowledge alone; they require a *collective creativity*, linking the ideas and insights of dozens, scores, hundreds, or thousands of people in collaborative networks focusing their combined imagination, dedication, and understanding on mutual discovery and problem solving.

Neither is what is *known* necessarily imbedded in the context of what is *wise*; wisdom and the ability to innovate—the focus of this chapter—are far higher in the order of human achievements than chronicling, organizing, and managing the profusion of data and knowledge.

Thus the Age of Information will prove to be short-lived, as it is only a brief stepping stone to the dawning of the next era of collaborative innovation—an era based on the creative and cooperative capacities that are natural to nearly every human being. This creative talent is based on our natural curiosity to explore, be curious, and ask innocently outlandish questions. It is this creative drive, when used synergistically with others, that we call "collaborative innovation"; it may be the foundation of all the solutions to the world's greatest problems, as this chapter will describe.

As a reader of this chapter, you may be questioning the veracity of these statements. Traditional thinking has said that it has been the lonesome inventor or experimenter that has created the scientific breakthroughs of the modern age. You may be thinking of the founders of modern scientific inquiry—Leonardo Da Vinci, Isaac Newton, and Louis Pasteur, slaving singly in their laboratories or pouring over textbooks in isolation.

The primary reason individual quests were responsible for most of the historical scientific innovation is because their world was structured neither for ease of collaboration nor for sharing of ideas and data across boundaries. Travel, communication, and information systems were limited and difficult. The structural changes of the latter half of the twentieth century changed all that. Science of the past was isolated and individualistic; science of the present and future will increasingly be (and is rapidly becoming) far more connected and collaborative.

2.2 COLLABORATIVE IMPERATIVE

2.2.1 Driving Forces in Scientific Discovery Today

Technology has not become the great simplifier of our lives, as once predicted. Instead, technology has *enabled* and *accelerated complexity* and *change*. Within our fast-moving, rapidly changing world, innovation has shifted its venue from the individual to the group; almost all innovation today is done collaboratively, in teams, networks, or alliances. This is true not only for scientists but also for those who must commercialize innovations and those who must address the legal complications of bioethical decisions.

To grapple with this complexity, multidisciplinary teams are essential, because, in most cases, it is impossible for one person to grapple with all the intricate information required to create breakthroughs. And most breakthroughs are happening not within a field or specialty but between fields. These multidisciplinary breakthroughs are not just complex, they are also very expensive. Thus it becomes imperative for companies, universities, and laboratories to work in a seamless, synchronistic, and synergistic manner.

The Langer Laboratory at MIT is a perfect example, as Dr. Robert Langer describes[1]:

> My lab has people with 10–12 different disciplines in it—molecular biologists, cell biologists, clinicians, pharmacists, chemical engineers, electrical engineers, materials scientists, physicists, and others. Many of our ideas—such as tissue engineering—require these different disciplines to move from concept to clinical practice. It makes it possible to do nearly anything "discipline wise" in the lab.

2.2.2 Power of Differentials

The value of multidisciplinary teams is founded on the basic principle that all innovation comes from differentials in thinking: If two people think alike, there is no innovation. Innovation occurs when someone decides to think differently—by asking new questions, challenging the status quo, having a vision that there must be a new/better way, or being dissatisfied with the results produced by current solutions.

Harnessing the multidisciplinary power of the differential thinking should be one of the strategic methodologies to generate breakthrough innovation (Table 2.1). Being creative requires *divergent* thinking—generating many unique ideas—and then innovation demands *convergent* thinking—combining those ideas into the best result.

Collaboration triggers the sparks between people that brings out their natural (often suppressed) creativity and enables their differentials in thinking to generate a massive stream of ideas; and then the focus becomes converging, integrating, and aligning those ideas into real innovations. People who innovate collaboratively (as opposed to independently) have a greater chance of learning from others and building the networks that actually enable innovation to become implemented.

For example, one of the best known breakthroughs in biomedicine was the joint insight by Watson and Crick regarding the double-helix structure of DNA. Crick had migrated from the field of physics, and Watson was just a young graduate student. They both came from a place of "not already knowing," an openness to new ideas, rather than thinking of themselves as "experts" in the biomedical profession. They never conducted any experiments, instead looking at the data of others, and interpreted the data from a fresh perspective. Watson and Crick meticulously integrated the work of others in different fields—such as crystallography—and saw unique patterns in the data that enabled them to envision the double helix.

Making collaboration the *central organizing principle* for all research, discovery, development, commercialization, and proliferation for innovative new products, services, and business models will likely result in a far higher chance of producing a breakthrough in thinking and results.

TABLE 2.1 Einstein's Rules for Creating Breakthroughs

1. We cannot solve the problems of today with the same level of thinking that created the problem.
2. Creativity is more important than knowledge.
3. From discord make harmony, from chaos seek order.
4. In the middle of difficulty lies opportunity.
5. There is a simplicity of design behind every level and layer of complexity (if we search for it).

2.3 CREATING CULTURE OF COLLABORATIVE INNOVATION

Nearly every study done on the issue of innovation has concluded that the number one factor in producing innovation depends not upon the quality of the scientists, technicians, and researchers but on the *culture* that supports and reinforces them (Fig. 2.1).

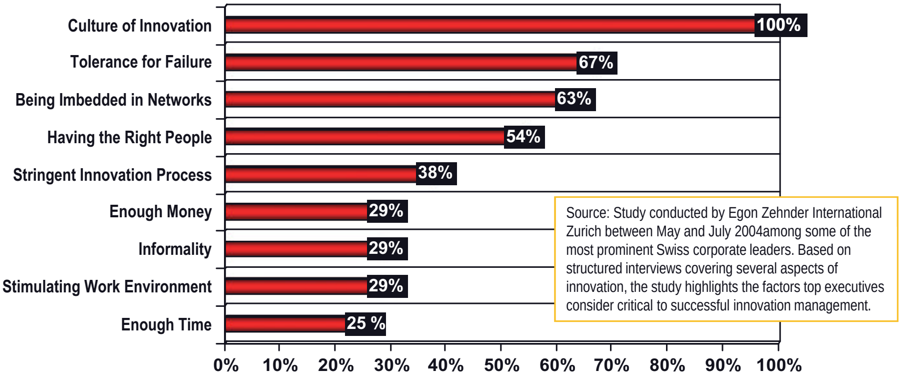
Most scientists, upon deciding they must engage in a collaborative inquiry, will launch the initiative starting with the technological problem. Herein lies the first and biggest trap in collaborative innovation, because it is like learning the words to a song without the music. There are *five key principles* that will create a powerful culture of innovation: select the right people, establish a system of trust, create a spirit of inquiry, eliminate failure, and empower champions. It does not matter where one is located in the innovation process—research, discovery, development, or commercialization—these five principles will always make the difference between success and mediocrity.

2.3.1 Select the Right People

What first characterizes a highly innovative culture is the quality of the people who lead and serve on the innovation team. There are six factors to consider in the choice of people:

1. *Competence* Knowing that the members of the team are highly qualified to conduct research, make modifications to procedures, and thoroughly comprehend the results is the basic standard of excellence.

2. *Character* Individuals with good character are essential to ensuring that team members trust each other and will do the right things for the right reasons. Key characteristics include honesty, good judgment, perseverance under pressure, and a tenacious work ethic. Yet these characteristics alone do not make a great team. More is necessary.



Source: Study conducted by Egon Zehnder International Zurich between May and July 2004among some of the most prominent Swiss corporate leaders. Based on structured interviews covering several aspects of innovation, the study highlights the factors top executives consider critical to successful innovation management.

Figure 2.1 Success factors for innovation (typical example of innovation studies).

3. *Collaboration* Many people who enter the field of scientific research are inherently introspective or shy; others possess minds that are highly logical and analytic. Many scientists were loners in school, perhaps never participating in team activities, such as sports or group governance. This can present difficulties when a large project requires close coordination and human interaction. Teamwork requires communication, sharing information, understanding the human side of research, and mutual support, particularly in times of adversity. People without great collaborative skills may engage in criticism, blame, negativity, and back-biting, often when under high stress. They may horde information for fear it will be used improperly. They may withdraw when others need them most or engage in manipulative behavior to get the attention or credit they yearn for. They many not communicate well, especially listening carefully, and may not understand the human side of technical information.

Collaboration is the enabling force that opens the pathway to group genius:

> When we collaborate, creativity unfolds across people; the sparks fly faster, and the whole is greater than the sum of the parts. Collaboration drives creativity because innovation always emerges from a series of sparks—never a single flash of insight . . . lot's of small ideas . . . each spark lighting the next . . . each critical to the [ultimate] success. [2, pp. 4, 7, 8]
>
> Many stories of innovation, once you get past the smoke and mirrors, reveal a backstage filled with other people, ideas, and objects that were as critical—if not more so—than the one presented onstage. Ultimately, the amount of credit we insist on giving to individuals in the innovation process is absurd.[3, p. 103]

4. *Creativity* Being creative has a massive advantage for a clinical research team. *The quality* of creativity is not limited simply to imagination. It includes a variety of qualities, such as collaborative resourcefulness, inquisitiveness, curiosity, progressive thinking, problem-solving capacity, and even the desire to jump over any obstacle to see ideas carried through to fruition.

Often the most creative people are not necessarily the most academically qualified, because most academia rewards knowledge, having the "right" answers, and analytic skills. Highly creative people often are not primarily analytic but are typically multidisciplined, eclectic, cross-functional, and filled with more questions than answers. Thus they do not always fit into bureaucratic, highly structured environments; they tend to like less structure and thus are often able to live better on the edge of uncertainty because they use a personal set of internal principles to guide themselves rather than external procedures.

What is sought is a "fluency of ideas and flexibility of approach that characterizes scientifically creative individuals working together on a problem" [4, p. 187]. In highly complex environments, Welter and Egmon [1, p. 154; 5, p. 126] point out that collaborative innovation teams will demonstrate five important qualities:

- Freedom to explore beyond the mainstream of conventional thought
- Ability to trust using shared vision and values
- Genuine curiosity and exploration of possibilities and opportunities
- Compelling commitment to make a difference
- Genuine self-awareness of differentials in thinking and learning styles

Some very creative people can lack discipline because they are not easily controlled, preferring to be free spirits. In this case such people may better serve the team in an advisory role.

5. *Courage* Great research teams face many challenges from inception of their idea through to final delivery of a successful product or procedure to a patient. These challenges can often be daunting as the team faces adversity after adversity. The ultimate measure of a successful team is how they face the challenges of difficulty, controversy, and uncertainty while maintaining their honor and integrity. Moving a vision from concept to conclusion requires a championing spirit, a strong commitment to the possibility not yet proven. The championing spirit is focused on both collaboration and innovation. Champions bring a confluence of passion for the vision, strategy for moving forward together, and commitment to the ultimate result [6, p. 82]:

> Ideas do not propel themselves; passion makes them go. Passion is the fuel that generates an intense desire to move forward, smashing through barriers and pushing through to conclusions.

Tenacity and optimism in the face of adversity and unwavering commitment to ideals in spite of the dark nights of the soul are qualities of the true champion. Edison, in his search for an ideal filament for the light bulb, "for eighteen to twenty hours a day experimented with all sorts of materials. . . . He had to find the best type of fiber. . . . He tested more than 6000 materials, and his investigations on this one thing alone cost a small fortune" [7, p. 114]. Edison was courageous and tenacious enough to experience over 6000 failed attempts to get one right solution.

Resilience is another dimension of courage. Resilient people are typically optimists, holding onto their vision and ideals when the skeptic has given up [1, p. 75]:

> Great achievers understand intuitively that the human brain is the most profoundly powerful solution-finding mechanism in the known universe. And they recognize that persistence is the key to keeping that mechanism engaged. . . . Optimists get better results in life; and the main reason is simply because they are less likely to give up. As Dr. Martin Seligman emphasizes, pessimism is self-defeating because it "short-circuits persistence.". . . The real key is . . . to maintain our enthusiasm in the face of seeming failure. Resilience in the face of adversity is the greatest long-term predictor of success for individuals and organizations. Persistence in the process of experimentation, when desired or expected results are elusive, is the way that resilience is expressed.

Resilient people have the ability to flourish on the edge of creative uncertainty, that ambiguous gray area that rigid people perceive as lack of control.

The bottom line is the courage factor that identifies those with a champion spirit, the resilient optimists with the tenacity to produce the persistent actions that get results, not just good intentions.

6. *Cognitive Diversity* All innovation comes from differentials in thinking—people who challenge conventional assumptions, ask uncomfortable questions, and see possibilities in the midst of difficulties. For this reason, cognitive diversity is a fundamental ingredient for success.

An early example of the importance of cognitive diversity spurring innovation comes from Thomas Edison [1, pp. 148–149]:

> Although Edison was an incomparably brilliant independent inventor, he understood and valued the importance of working with others. He knew he needed a trustworthy team of collaborative employees who could illuminate his blind spots and complement his talents. Over the course of his career, Edison cultivated an inner circle of roughly ten core collaborators, each contributing materially to the technologies generated by his laboratories. Edison brought together individuals from diverse disciplines who he would indoctrinate in his methods, then release to freely experiment without his immediate supervision. The diversity of disciplines added tremendous breadth and depth of insight to the laboratory, allowing them to navigate effectively across industry boundaries. . . . they were extensively cross-trained. The teams were bound together by common values of respect and integrity [trust], and a desire to be the best in the world. . . . he placed the value of "team accomplishment" at the heart of his laboratory.

Diversity of thinking, while the stimulus to all innovation, can be a double-edged sword. Many managers are threatened by diversity, desiring instead conformance to a standard set of rules, procedures, and mode of thinking. When organizations are segregated into specialties, such as biology, or marketing, or administration, or any other form of segregation, it is often the case that these specialties become fiefdoms of power and isolation, perhaps isolating themselves because "those others don't think like us." Conflict and competition characterize these groups. They are stuck. Trust will be essential (see next section).

When seeking people for the innovation team, a very useful framework is based on Ned Herrmann's brain dominance patterns [8]. Every human has a preference for how they like to think and learn. In Figure 2.2, the four basic brain patterns are outlined.

While the majority of people tend to be dominant in a single mode, a minority of people will be comfortable in two or even three modes. Very few will have four modes. These are called "multibrain dominant." Many of us are thought of as "left" or "right" brainers, referring to whether we tend to be more analytic (left brain) or more sensitive to people (right brain).

One of the important roles on any diverse team is the role of the "integrator" [9], the person who can translate across boundaries, connecting diverse

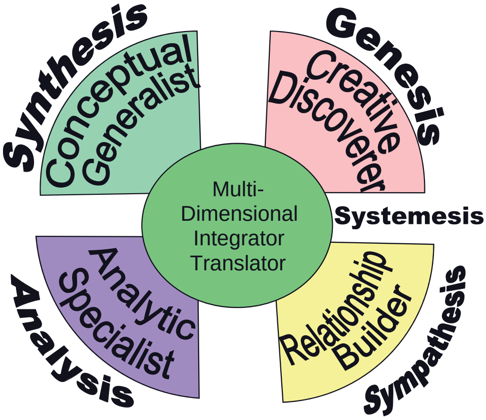


Figure 2.2 Different brain dominance patterns. (*Source:* Adapted from N. Herrmann, *The Creative Brain*, Lake Lure, NC: Brain Books, 1995.)

thinking from one arena to another. This person typically is multibrain dominant, which enables them to see situations and people from a kaleidoscopic perspective, sorting through data, vision, emotions, strategy, and implementation.

2.3.2 Build a System of Synergistic Trust

Ask any person adroit in collaborative innovation about the key factors for a success and you can be assured that trust will be near the top of the list. Trust is a crucial factor for collaborative innovation because it creates the fertile ground for creativity, innovation, and synergy. Without trust, teams disintegrate, and in-fighting predominates. All innovation is, by definition, a force of change; change is destabilizing to most organizational systems and structures, threatening to upend established hierarchies, power structures, procedures, and accepted thinking, preventing the establishment of the linkages of resources and implementation alliances necessary for the innovation to succeed. Thus, without trust, innovation will appear as a threat, fear will overwhelm opportunity, and the organizational immune rejection response will trigger, manifesting as massive resistance to or exclusion of the forces of evolutionary change.

Trust is absolutely essential in generating creativity among innovators. Distrust is the greatest impediment to all innovation. Trust is the essential foundation of synergy—where the innovation team truly becomes greater than the sum of its individuals. Often referred to as “chemistry” (in the psychological sense), trust has unique properties that are more like alchemy: It is simultaneously the *glue* that bonds people together and the *grease* that eliminates interpersonal friction.

Mistrust causes everything to be more complicated, slower, and far more fragmented. In addition, distrust puts a major limitation on collaborative innovation, internal teamwork, and external relationships with suppliers, customers, stockholders, and our community.

Few scientists ever spend the time to create powerful trust-enabled innovation cultures. Often building trust is elusive, filled with platitudes, slogans, and aphorisms such as "trust must be earned," "be skeptical before you trust," "be sure to have an exit strategy," "trust but verify," and so on. Unfortunately none of these approaches really produce any trust.

Highly legalistic attempts to ensure against breaches in trust usually backfire and poison the well before any alliance or collaboration gets started. Often, by trying to protect against distrust, we actually create the conditions we are trying to avoid, which manifests as enormous legal agreements and protracted negotiations that may result in no agreement at all. Trust enables everything to move faster, more effortlessly, and with less conflict. In spite of its importance, trust is too often taken for granted.

It is imperative that innovators today know how to establish a "trust system" that enables collaborators to act honorably with each other, that makes intellectual property safe from incursions, that establishes joint principles of engagement, and that honors the differentials in thinking that stimulates the creative energy so fundamental to all innovation.

To have trust, at a minimum, one must sense that there is a level of *safety* and *security* in the relationship, knowing that I will not be worse off for having this interaction.

Trust, like all disciplines, has an internal "architecture" that can propel the honorable scientist to great heights and weed out the small percentage of "sharks" who would abuse collaborative relationships for their own selfish ends. To understand the nature of trust, it is first important to know the nature of its opposite—distrust.

2.3.2.1 Cause of Distrust What causes distrust? In a word—fear—fear of being taken advantage of, fear of being put in a disadvantageous position, fear of not receiving proper credit, fear of being manipulated or discredited, or fear of one's beliefs and knowledge being subjected to attack.

2.3.2.2 Building Trust Just as the elimination of a disease does not cause health and happiness, neither will the elimination of distrust create solid trust—it just brings everything to "neutral." The lack of ethics will cause distrust, but the presence of honesty and ethics does not necessarily cause trust. Good ethics implies "I won't do something wrong"; it takes the fear out of the picture. But it does not mean "I'll be effective," or "use sound judgment," or "be collaborative," or "be compassionate," or "be spontaneous." Other things are necessary.

The basis for trusting someone is not simply ethics and honesty; it is also how they deal with self-interest. We trust people we can count on to look out after our interests as well as their own—our "mutual" interests, or, put another way, the "greater good." Balancing self-interest with the greater good is the starting point to begin trust.

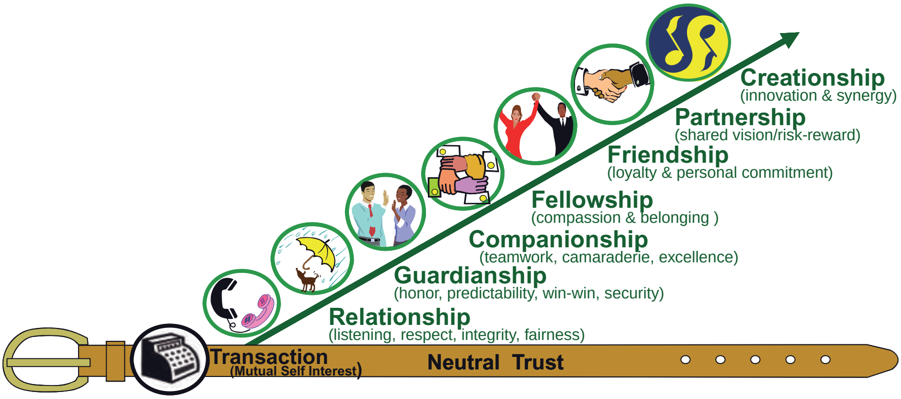
When each person or organization acts to maximize the amount they get from negotiations without consideration of another person's or organization's interests, they are working in their self-interest. Untethered, self-centered decision making creates untenable collaborative situations.

2.3.2.3 Ladder of Trust Traditionally, trust has been rather narrowly defined as *safety, security, reliability,* and *integrity.* This definition should be thought of as the *minimum;* instead think of trust as a spectrum or ladder ranging from neutral trust at the bottom to synergistic trust at the top. As illustrated in Figure 2.3, we refer to "neutral" trust as "transactions."

> The Ladder of Trust is a tool to navigate the journey into a positive world where strong bonds of trust support highly productive collaboration and innovation.

"Below the belt" is the zone of distrust. Here lie the *trust buster behaviors*, such as:

- Acting inconsistently in what they say and do
- Seeking personal gain above shared gain
- Withholding information or cheating
- Lying or telling half truths
- Being closed minded, blaming, personal attacks
- Being disrespectful to anyone, not listening, being uncompassionate
- Withholding support or betraying confidences or breaking promises



Creationship
(innovation & synergy)
Partnership
(shared vision/risk-reward)
Friendship
(loyalty & personal commitment)
Fellowship
(compassion & belonging)
Companionship
(teamwork, camaraderie, excellence)
Guardianship
(honor, predictability, win-win, security)
Relationship
(listening, respect, integrity, fairness)
Transaction
(Mutual Self Interest)
Neutral Trust

Figure 2.3 Ladder of Trust.

The first thing a leader must do is prevent or stop any of these trust buster behaviors from occurring or being rewarded. There must be no tolerance or acceptance of any of these actions which destroy a research team from within.

On the belt line is neutral trust, which manifests as transactions. Transactions happen every day. When shopping, we put enough trust in the "brand" or the store's reputation to complete the exchange of goods or services for money, but not enough trust to engage in any form of deeper relationship.

While the idea of neutral trust may seem benign, there can be some deep downsides to transactionary trust, simply because it may be totally inappropriate for a transactionary relationship to be matched to the circumstances where close teamwork and collaboration are required in solving complex problems that require interactive spontaneity; a transactionary relationship would seem too aloof, distant, and formal.

Above the belt is the zone of trust, where teams can prosper and thrive. Rather than defining trust simply as reliability, security, or integrity (as has been the traditional definition), it is far more useful to define trust on a spectrum ranging from minimal trust to the ultimate forms of trust (see Fig. 2.3). Here are the types of trust in the range above the belt.

2.3.2.4 Relationship The trust journey begins simply with building a relationship with other people by listening. When we listen with compassion, learning, and constructive inquiry, we begin to build trust. People feel like they are receiving *support* because they are heard. When building a trusting relationship, the minimal boundary conditions must be satisfied—both parties must be honored and respected, and both must be counted on to understand each other's personal interests, needs, and concerns, which gives the assurance that ultimately both will be better off from having trusted.

2.3.2.5 Guardianship The next level of trust provides safety and security to the other person. A guardianship can be one way, much like a parent provides to a child, or mutual like soldiers on a battlefield. In a business relationship, *mutual* guardianship means *honor*: We stand guard over each other to defend each other against attacks, lies, dishonesty, and manipulations.

2.3.2.6 Companionship Being a companion means I trust you enough to be in your presence a significant part of my time. In business, this takes the form of working well together in teams. Individuals come to the realization, sometimes painfully, that they win or lose together, that they are on the same team—in the same boat, facing the same storm together.

2.3.2.7 Fellowship This means much more than "membership" to an organization, company, or club; it is more than a company picnic or sales rally. Fellowship implies a powerful attraction, commitment, and buy-in to the values, hearts, and minds of the other members of the community. Because of the weakening of the family structure, for many their workplace has become a surrogate family, and thus the workplace carries with it an additional desire for *fellowship*. Having a powerful set of common values, a sense of purpose, and a unique frame of reference to view the world generates a dedication and energy that are difficult to defeat.

2.3.2.8 Friendship A great friend is always there for me . . . always happy to see me . . . listens to me . . . is loyal, faithful, protective . . . never carries a grudge or the baggage of unfulfilled expectations. When we build trust at the level of friendship, we embrace all the prior levels of trust but add some very energizing and vitality-creating forces.

In a friendship, trust enables our goals and addresses our fears, our deepest yearnings, and our personal limits/failures to be put out in the open with no sense of diminishment. The power of friendship lies not just in the bond of familiarity but also in the mutual commitment to each other's well-being.

2.3.2.9 Partnership A partnership is much more than a friendship; it is an alliance designed to respect and cherish the differentials in thinking and capabilities between two or more people or organizations. It is the synergy between differing strengths and the alignment of common purpose that make a partnership most alluring. Great partnerships rely also on complementary competence and skills, character and integrity, and collaborative behavior.

2.3.2.10 Creationship For this level of trust a new word is needed: A "creationship" implies that we can do something extraordinary—we can co-create. It is at this level that the very best scientific work is done. You do not have to look too far to find wonderful examples of this level of experience (e.g., Watson and Crick, the Wright Brothers, the Manhattan Project team, or the Human Genome teams). A creationship embraces prior elements of trust building, and then, secure in the absence of fear, unleashes a connection between the hearts and minds of the co-creators—new ideas generate like spontaneous combustion.

Building a creationship is extremely rewarding. It can happen between two people or within a research team or in a collaborative alliance. When people engage in a creationship, they seem to abound with an endless source of regenerative energy.

Trust is the most vital thread in the fabric of collaboration. And it is not unusual to find that trust gives work a far deeper sense of meaning and purpose. We neglect the issue of trust at our own peril.

2.4 SPIRIT OF INQUIRY: "CRITICAL PARADOX"

The basis of scientific research is to uncover new insights into the functioning of systems, natural or physical. Inquiry—posing questions—is the essential beginning point of discovery. Scientific research uses a framework of "critical" questions to enhance discovery, much like a trial lawyer or a crime detective, which embrace a strong sense of doubt and skepticism which challenges conventional thinking. To prove one's thesis, it must stand up to a barrage of skepticism supported heavily by evidence. Such is the nature of scientific inquiry. This sounds rather simple, but there is a "catch," often unexpectedly ensnaring research teams, which are the realm of "human" systems.

The paradox is that scientific analysis and human behavior abide by very different operational rules of engagement. The poignant critical and "logical" analysis that facilitates scientific research often destroys human relationships and the ability to co-create, generate synergies, and produce breakthrough thinking.

The way we ask scientific questions, when applied to people, can be accusatory, threatening, distrusting, or even insulting. Seldom are scientists made aware of this important distinction and its corollary: the need to appreciate people while never lowering scientific standards. In Figure 2.4, the different types of questions are charted to help illustrate the distinct differences.

Quadrant I describes questions that qualify as "open inquiry." Questions of this sort tend to let people explore opportunity, possibility, and joint creativity. (A version of this type of question is called "appreciative inquiry.") Human interaction tends to be very positive when faced with questions in this context. Many of these types of questions can be used from a scientific perspective to break deadlocks in thinking or shift paradigms.

Quadrant II works well in forensic work, but it is accusatory in nature. The questioner is not an "inquirer" but rather an "inquisitor." Something's wrong, someone has run afoul, and the inquisitor will find out who is at fault. Similarly

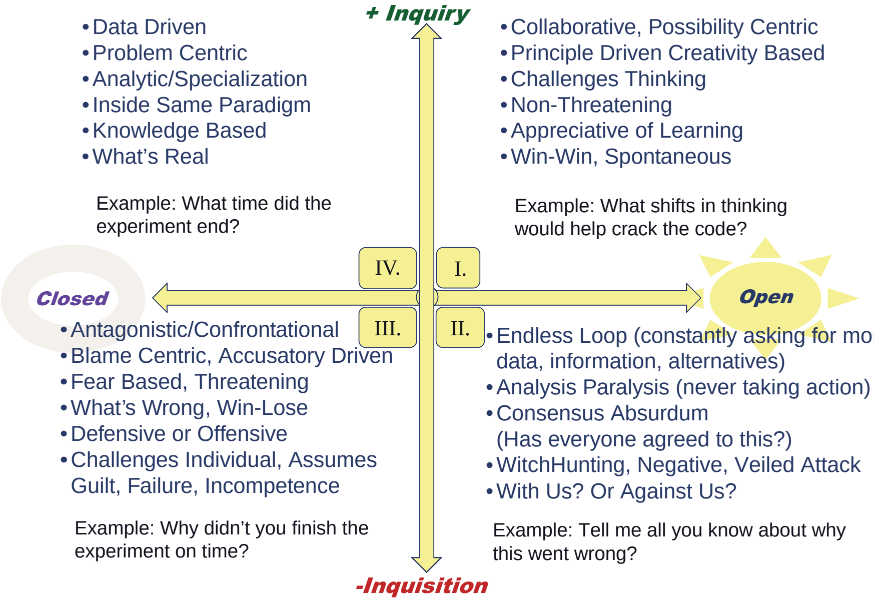

+ Inquiry
• Data Driven
• Problem Centric
• Analytic/Specialization
• Inside Same Paradigm
• Knowledge Based
• What's Real
Example: What time did the experiment end?
• Collaborative, Possibility Centric
• Principle Driven Creativity Based
• Challenges Thinking
• Non-Threatening
• Appreciative of Learning
• Win-Win, Spontaneous
Example: What shifts in thinking would help crack the code?
IV.
I.
Closed
Open
III.
II.
• Antagonistic/Confrontational
• Blame Centric, Accusatory Driven
• Fear Based, Threatening
• What's Wrong, Win-Lose
• Defensive or Offensive
• Challenges Individual, Assumes Guilt, Failure, Incompetence
Example: Why didn't you finish the experiment on time?
• Endless Loop (constantly asking for more data, information, alternatives)
• Analysis Paralysis (never taking action)
• Consensus Absurdum (Has everyone agreed to this?)
• WitchHunting, Negative, Veiled Attack
• With Us? Or Against Us?
Example: Tell me all you know about why this went wrong?
-Inquisition

Figure 2.4 Inquiry versus inquisition—open and closed questions.

quadrant III carries the same inquisitorial context, just asking closed-ended questions that only need a yes or no answer. Any inquisitorial questions will evoke fear, defensiveness, and oftentimes anger and reprisal by the listener. Many research teams have errantly traveled down this path, with less than stellar results as human energy was wasted on protection of status, ego, or honor, instead of focusing on the larger, nobler cause which the research team was trying to achieve.

Quadrant IV describes the types of questions that typically constitute much of scientific research. They tend to be tightly bound, based on evidence, focusing on generating knowledge. While these types of questions can work wonders in the scientific context, they can be very limiting in the human context.

Being aware of these differences can help the leader of any clinical research team shift the content and style of their dialogue to generate a much higher esprit de corps, inspire curiosity, and gain much deeper insight, with an attendant shift in the results produced.

The most transformative creativity results when a group either thinks of a new way to frame a problem or finds a new problem that no one had noticed before. When teams work this way, ideas are often transformed into questions and problems. This is critical, because creativity researchers have discovered that the most creative groups are good at finding new problems rather than simply solving old ones [2].

2.5 ELIMINATE THE WORD: FAILURE

One paramount fear in all scientists is the fear of failure. Studies have shown it to be common to nearly all college graduates. This fear, if used mildly, can motivate people to great heights and long hours of work. But overused or used as a threat, it can paralyze people, causing them to shut down or avoid the possibility of failure, because fear of failure immediately attacks the ego, which never wants to accept the stigma of tragic disappointment. Senior executives have some advice on "failure" (Table 2.2)

Returning to the development of the electric light, Thomas Edison and his research and development (R&D) team provide a superb example of how to

TABLE 2.2 Advice from Senior Executives about "Failure"

- "You only get the ten percent of innovations that succeed if you are ready to accept the ninety percent that fail."
- "If you never failed, you never dared."
- "Relieve failures of their negative aura by calling them 'lessons learned' or 'learning opportunities.'"
- "It's a mistake to punish innovative people for failures, particularly in industries with very short product cycles, where decision making is invariably faster and often based on incomplete knowledge."

deal with the issue of failure versus learning. Edison did not invent the light bulb; it had been created 35 years earlier. His development team in Menlo Park, New Jersey, worked tirelessly to perfect the design of a commercially successful light bulb. It required new technologies to create a vacuum in the bulb, a totally new approach to filaments, and a structure to secure the filament. Edison's team examined and created experiments based on over 3000 theses and conducted over 10,000 experiments [10]. Edison created a database of knowledge coupled with his diverse reading, which ignited his ability to generate a broad range of hypotheses [1].

2.6 EMPOWER INNOVATION CHAMPIONS

2.6.1 Nature of Champions

Scientific research is not easy work. It entails long hours, multiple unknowns, and endless complexity. In the final analysis of success, those who prevail to the end are not the most intelligent (although intelligence doesn't hurt) or the most famous or the most endowed with resources. Rather, success is bestowed upon the most creative, connected, and committed; those who can move from ideas, through strategy, into action. This is the domain of the spirited champion.

2.6.2 Role of Champions

Without champions, the ordinary inertia that plagues most organizations will stifle most innovation, because innovation, by its nature, is change, and change, by its nature, is threatening to most people because it destabilizes the status quo.

To make any innovation occur, three underlying issues must be understood and addressed according to Stanford's Kathleen Eisenhardt [3, pp. viii–ix]:

> First, innovation is the result of synthesizing, or "bridging" ideas from different domains . . . extraordinary innovations are the result of simultaneously thinking in multiple boxes, not of the oft-prescribed "thinking outside the box." In short, extraordinary innovations are often the result of recombinant invention. . . . while it may be appealing to focus on the future, breakthrough innovation depends upon exploiting the *past*. Combining often well-known insights from diverse settings creates novel ideas that can, in turn, evolve into innovations (for example, the Apple iPod used no new technology. Its meteoric sales were due to using existing technology in new ways that improved the user interface).
>
> Second, the organizing *structure* can dominate creativity. Years of academic research suggest that, beyond some fairly low threshold, successful innovators are not really more gifted or creative than the rest of us. Rather, they simply exploit the networked structure of ideas within unique organizational frameworks.

> Third, breakthrough innovations depend on "building" communities. Of course, the substance of the innovation has to be there. But the ideas that go on to become breakthrough innovations rely on fundamentally rearranging established networks of suppliers, buyers, and complementers (people whose skills and abilities "complement" each other to create a synergistic system) into new networks and ecosystems [alliances]. Otherwise, hoped for innovations never develop. The initial innovation is the starting line of the race, not the finish. Innovation is as much *social* as it is technical. Resistance must be met, and alliances forged, because people often cannot understand innovations, or cannot see how they would benefit if the innovations were adopted.

Accomplishing the tasks associated with these three issues is no job for the mundane manager or outsourced technician.

2.6.3 Qualities of Champions

Here are some of the qualities that are found in great champions:

- Passionate visionary who believes there is an innovative or better way
- Seeker and supporter of new ideas, no matter where they come from
- Builder of networks of teams with strong collaborative skills, ethics, and values
- Preserver of trust with unyielding integrity and ethics
- Articulate advocate willing to challenge established thinking
- Persistent networker linking together other supporters and advocates
- Action-oriented mover and shaker intolerant of bureaucratic barriers
- Crusader who will defend an idea or ideal against attack
- Win–win negotiator who sees opportunity in most problem
- Energizer willing to be accountable for reaching powerful objectives

2.7 AVOIDING THE TRAPS

Creating a great collaboration in science does require both discipline and good judgment. Here are a few other key topics that will contribute to supporting and sustaining synergies within the research and development team:

Vision and Value Proposition All members of the initial team should outline a shared vision that will help align their work and the value they believe this will contribute.

Roles and Responsibilities Clarity of knowing who will do what is essential to utilize people's strengths in the most complementary way. It also prevents territoriality from interfering with real work.

Use of Research Data There should be no ambiguity about: How will data be shared? Who owns the output? What publication is expected? What is the authorship sequence? Who owns the patents? What happens to derivative ideas and knowledge? What are the protocols for new people joining?

Joint Operating Principles Bringing diverse groups together means creating a new, hybrid culture based on the norms and values of the many new people that will be engaged. Together they should create a charter or covenant that outlines (on one page) their rules of engagement and the key principles that will ensure trust.

Distant Collaborations Unlike decades past, today many joint investigations occur among scientists stretched far across the globe. Often people have not actually met each other face to face. While social networking technology is getting better and better, it is strongly recommended by the most experienced collaborative innovators to spend some one-on-one time in person with each of the collaborators. (If this cannot be done, a personal telephone call is the next best approach.) During this encounter, be sure to discuss and come to an accord about personal objectives, concerns, trust builders and trust busters, personal mission and style, and quirks.

Misuse of Transactional Emails In an age when electronic communication is fast and pervasive, it is tempting to handle every interaction with an email. Be cautious, as this is only half true. Ordinary transactions, such as setting up meeting times, sending reports, and exchanging information, are perfectly suited for emails. However, emails are a terrible means of managing interpersonal breakdowns, such as conflict, anger, frustration, or disappointment. Do not use emails for this purpose, else you run the risk of massive escalation without resolution. If there is a personal problem, the best method for resolution is a face-to-face conversation where nonverbal communication can be discerned. If this is not possible, using the old-fashioned telephone is far superior to emails.

Poisoning Well of Trust During negotiations to set up the collaboration, very often lawyers, deal makers, and contract managers will be involved in the negotiations. Beware of those who use adversarial methods to wrangle the best terms and conditions for their client. All too often their techniques will “poison the well of trust” for those who later have to make the collaboration work. If you see win–lose techniques being used during the negotiations process, call a halt to that type of action immediately, else a large barrier will be erected between the prospective partners that may never be hurdled later.

2.8 CONCLUSION

Without a powerful commitment that fully embraces collaborative innovation, a research, discovery, or development team risks challenge without inspiration, desire without a dream, drive without destiny, or falling into the abyss between what is real and what is possible.

In the larger perspective, all collaborative innovation in bio-medicine is challenged to utilize complex socio-technical systems to unravel the secrets embedded in intricate biological puzzles.

Complex systems, by their nature, are faced with two paths: either *evolve* utilizing a synergistic set of functional interdependencies or *devolve* because of internal strife, entanglements, and dysfunctionality. Trust is a core determinative factor in the higher evolutionary path.

Complex human systems are uniquely different in nature specifically because humans are the only biological species capable of "inventing" itself—of using our collective intellect to innovate and create, to seek new answers to higher-order questions, and to build upon or interconnect technologies.

The course of history and the destinies of people are dependent on our ability to innovate collaboratively.

REFERENCES

1. Gelb MJ, Caldicott SM. *Innovate Like Edison, The Five-Step System for Breakthrough Business Success*. New York: Plume, 2007.
2. Sawyer K. *Group Genius, The Creative Power of Collaboration*. New York: Basic Books, 2007.
3. Hargadon A. *How Breakthroughs Happen, The Surprising Truth About How Companies Innovate*. Boston: Harvard Business School Press, 2003.
4. John-Steiner V. *Notebooks of the Mind: Explorations of Thinking*. Oxford: Oxford University Press, 1997.
5. Welter B, Egmon J. *The Prepared Mind of a Leader: Eight Skills Leaders Use to Innovate, Make Decisions, and Solve Problems*. San Francisco, CA: Jossey-Bass, 2005.
6. Rosenfeld R. *Making the Invisible Visible, the Human Principles for Sustaining Innovation*. Rochester, NY: Xlibris Corporation 2006.
7. Boyd T. *Prophet of Progress*. New York: E.P. Dutton, 1961.
8. Herrmann N. *The Creative Brain*. Lake Lure, NC: Brain Books, 1995.
9. Lawrence P, Lorsch J. New management Job: The integrator. *Harvard Business Rev* 1967;Nov–Dec.
10. Lathrop GP. Talks with Edison. *Harpers Mag* February 1890: 434.

3

MODELS FOR COLLABORATIONS AND COMPUTATIONAL BIOLOGY

Shawnmarie Mayrand-Chung, Gabriela Cohen-Freue, and Zsuzsanna Hollander

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.

3.1 INTRODUCTION

Over the last decade the biomedical research community has undergone significant changes with respect to the mechanisms and models used to conduct scientific research. These changes have been spurred largely by the shift in the research paradigm that drives drug research and development (R&D) coupled with the accelerated pace of emerging technologies.

For several decades the pharmaceutical industry has employed a "closed model" which was underpinned by two premises: (1) discovering and developing new drugs and (2) patenting those drugs in order to gain a monopoly on the profits of the newly developed drug. Since the development of a monopoly is based on capturing exclusive intellectual property right (e.g., patent exclusivity or exclusive licenses), this traditional model is competitive and inherently discourages the use of collaborative mechanisms.

While historically yielding high rates of return on investment, this closed model of R&D has been encumbered by increasing development costs and regulatory hurdles. For example, the average cost of developing a new drug in 2002 was $800 million, while the 2006 study from the U.S. Federal Trade Commission estimated costs of new drug development ranging from $500 million to $2 billion [1].

The decline in the U.S. economic status coupled with the escalating cost of drug development has forced the pharmaceutical industry, as well as the biotechnology and diagnostic industries, to reevaluate the traditional "solo" model of conducting research—and has led to the emergence of nontraditional partnerships and collaborative models. This chapter will introduce the concepts of modern partnerships and will describe some of the current partnership models that are being used to conduct collaborative scientific research.

3.2 IMPORTANCE OF PARTNERSHIPS

The concept of working collaboratively to advance scientific research efforts is not a new idea. However, the notion of collaborating within the scientific research arena through the twentieth century had been limited to the sharing of reagents and/or tools—mice, cell lines, antibodies, and so on. These traditional "partnerships" were generally the result of scientists sharing their curiosity for a particular scientific question and recognition that there may be some value in sharing controlled amounts of information to advance that shared research interest.

While traditional partnerships were successful in bringing together researchers sharing common research goals, they were limited in both scope and outcomes. Additionally, these early efforts to collaborate were limited to academic researchers and rarely, if ever, involved the industry sector.

Multisector partnerships are relatively new and are still largely competitive versus precompetitive (see Section 3.3.1) in their structure. However, many of the multisector consortiums developed over the last five years are precompetitive in their design, and many of the consortium outcomes and results are aimed at developing public resources and/or placing information into the public domain.

3.3 CONSORTIA MODEL

A consortium can be envisioned as a multistakeholder effort developed in order to undertake a large-scale initiative that no single entity could achieve alone. In general, a consortium can be distinguished from a partnership based on the scope and size of stakeholders. However, the underlying concepts of leveraging resources, sharing cost and risk, and increasing intellectual input exist in both the basic partnership model and the consortia model.

One distinguishable element of any collaborative effort is whether or not the initiative will be structured as a competitive or precompetitive initiative. While there is no exact formula for building a precompetitive collaboration, a hallmark shared by all precompetitive efforts is that the outcomes of the initiative will benefit the scientific community at large—as opposed to merely benefiting the participants to the effort.

A second distinguishing element of the consortia model is that a more diverse stakeholder group is involved. Increasingly, U.S. government agencies such as the National Institutes of Health (NIH) and the Food and Drug Administration (FDA), nonprofit organizations, and others are joining with academic and industry researchers to further their individual mandates and missions.

The decision for a government agency or any other entity to join a consortium goes much further than whether the effort will be competitive or precompetitive (e.g., will public resources be generated as a result of the consortium's work). When the government agency decides to participate in a consortium or other partnership, there must be careful consideration of all aspects of its involvement, including (1) whether the mission of the consortium aligns with the mission of the agency, (2) whether the use of the agency's resources in the consortium is justified, (3) whether the agency's involvement in the consortium will be the subject of public controversy, for example.

Regardless of whether or not a consortium is a public–private partnership (e.g., government is a participant) or a private–private partnership or a public–public partnership (e.g., only government participants), a successful consortium will always set forth at its inception the mission, policies, governance structure, and expectation of partners, at a minimum. However, as with any large-scale collaboration, perceived and inherent barriers exist that must be overcome. Examples of perceived barriers include:

- Concerns that by sharing data and information a company might give away information that could someday generate profits (e.g., loss of intellectual property)

- Concern that antitrust laws might prohibit sharing and collaboration or that lack of an adequate antitrust policy might put the companies in legal jeopardy
- Concern about the loss of intellectual property rights to drug development programs, for example, by sharing data on the safety of a compound, a company inherently discloses its commitment to commercialize a related class of compounds
- Concern about liability, for example, who is responsible if the work is done by another company and outside the control of their organization (e.g., who will manage the work)
- Who will be responsible for governance issues
- Who will perform the work and how the work will be divided
- Who will fund the collaborative work
- Who will benefit from the collaborative work

One successful mechanism for addressing many of the above concerns is to establish a complete set of policies and procedures in advance of commencing the partnership. This preemptive approach allows potential consortium partners to understand both the rewards and expectations associated with their involvement. A unified understanding of the expectations, outcomes, policies and procedures, governance structure, and so on, affords the consortium the greatest opportunity for success.

While there are often many good reasons for a stakeholder to join a consortium, the decision to join or not join ultimately rests with an analysis of several factors and a determination of whether or not there is adequate alignment of their interests with the other stakeholders to justify participation.

3.3.1 Precompetitive Models

Precompetitive collaborations can be defined in several ways, for example, as "pertaining to the time during research and development in which there is collaboration but no competition" [2] or "open collaborations between companies that usually are intellectual property (IP) competitors" [3] or "early stages of research that benefit all [stakeholders]" [4]. More particularly, Webster's defines the term *precompetitive research and development* as "noncompetitive, cooperative research and development which leads the way to full scale competitive development in the future by addressing key requirements of new technology for the low-cost realization of [independent business concern] IBC equipments and services" [4].

Regardless of the precise definition employed the general thinking is the same—*to work collaboratively in the first instance in order to expedite the generation of resources (e.g., tools, data, specimens) that each stakeholder (e.g., academia, industry, nonprofits) can use to drive the success of that organization's commercial products.*

For the drug development industry, the competitive space is focused on the marketing of proprietary therapeutic agents that meet with regulatory approval for on-label treatment, as such the opportunity to share the cost and risk of upstream research affords the pharmaceutical companies a precompetive opportunity. Of course, the IP constraints associated with the precompetitive effort must be carefully considered so as to not create a barrier for downstream commercialization.

Since the commercial benefit of diagnostic and biotechnology companies is further upstream than the pharmaceutical companies, the ability to create precompetitive opportunities is more challenging. However, the goal of determining relevant precompetitive opportunities for any company is to identify mechanisms for discovering or developing fundamental knowledge and outcomes that drive a net positive return on investment for the company's competitive products.

The concept of precompetitive collaboration has been recognized as a successful strategy for accelerating drug development by U.S. and European regulatory agencies through their participation in several initiatives. For example, the FDA is a founding member and active participant in several precompetitive consortia, including the Critical Path Institute (C-Path), Predictive Safety Testing Consortium (PSTC), and the Biomarkers Consortium (BC) (see Section 3.4 for the details of each initiative). In May 2010, the FDA and the European Medicines Agency (EMA) concluded the first joint precompetitive qualification process for biomarkers.

More particularly, this FDA–EMA initiative involved a joint effort to consider use of renal toxicity markers proposed by the PSTC and involved the participation of 16 pharmaceutical companies. This unprecedented sharing of data by multiple pharmaceutical companies has served to test a joint FDA–EMA data submission process to receive, review, and approve new methods as qualified for use in drug development. Through the joint process, the PSTC submitted a single (preclinical) biomarker data application to both regulatory agencies and then met jointly with scientists from both the FDA and EMA to discuss the details of the submission and to address any scientific questions posed by the regulators. Each regulatory agency then reviewed the application and made independent decisions on use of the new biomarkers.

The leveraging of contributions and risks with appropriate partners improves the chances for positive outcomes and reduces the risk of investment. As such, there is likely an opportunity for most commercial stakeholders to find benefits in the precompetitive space, though it may take some looking.

3.4 EXAMPLES OF SUCCESSFUL LARGE-SCALE PARTNERSHIPS

The emergence of large-scale partnerships exemplifies the industry's recognition that the "old model" for conducting scientific research needs to be reconsidered. Whether done in a precompetitive fashion or using a controlled competitive model, the need to collaborate is being embraced by the scientific community.

As evidence of acceptance, several large-scale partnerships have emerged over the last five years, which demonstrate the utility and sensibility of leveraging resources to (1) share cost and risk among the research community, (2) develop standards to achieve more accurate measures of research, and (3) expedite the advancement of scientific knowledge, for example.

Below are examples of large-scale partnerships that have proven successful to advance research efforts. Some of these exemplar partnerships are private–private (involve only privately held companies/organizations), while others include one or more government entities—thus termed public–private partnerships. The PSTC and BC are two examples of U.S.-led large-scale consortia. Examples of large-scale European initiatives include the Innovative Medicines Initiative Undertaking (IMI) (see Chapter 4), the European Personalized Medicine Association (EPEMED), and the Pistoia Alliance (see Chapter 1, http://www.pistoiaalliance.org/overview). Sections 3.4.1–3.4.4 provide detailed descriptions of existing large-scale consortia.

3.4.1 Predictive Safety Testing Consortium

The PSTC is a unique public–private partnership led by the nonprofit C-Path. The mission of the PSTC is to bring together pharmaceutical companies to share and validate each other's safety testing methods under advisement of the FDA and its European counterpart, the EMA. The goal of the PSTC is depicted in Figure 3.1 and can be summarized as an attempt to "create a productive environment among private sector competitors while balancing their needs with those of the FDA, academic scientists, and the public health" (http://www.c-path.org/consortia.cfm).

The 16 industry members of the PSTC share internally developed preclinical safety biomarkers in five workgroups: carcinogenicity, kidney, liver, muscle, and vascular injury. For more detail see http://www.c-path.org/pstc.cfm.

3.4.2 Biomarkers Consortium

The BC is a precompetitive public–private partnership managed by The Foundation for the NIH (FNIH) and the three founding members, the NIH, the FDA, and PhRMA (the trade organization for the U.S. pharmaceutical industry). In addition to the founding members of the consortium other partners in the consortium include the Centers for Medicare & Medicaid Services and the Biotechnology Industry Organization. The genesis for the BC was initially with PhRMA, which in discussions with the NIH and the FDA realized the need for robust and meaningful biomarkers, well characterized for use and widely available to the research community [7].

The BC is the largest public–private consortium to date with participation from a variety of stakeholders, including government, industry, academia,

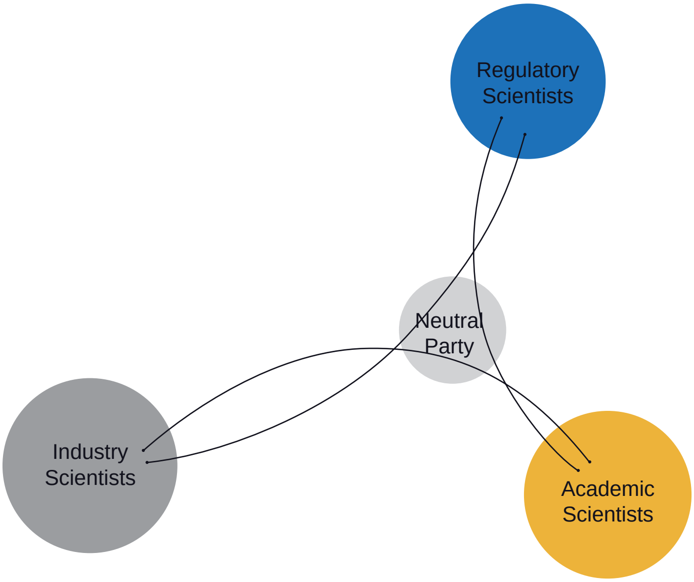


Figure 3.1 Schemata of relationships of stakeholder groups for PSTC.

patient advocacy, and other nonprofit organizations. This multistakeholder consortium was launched in October 2006 with the mission of joint discovery, development, and qualification of biomarkers. More specifically, the BC's products and outcomes include (1) identification and execution of cross-sectoral biomarker projects, (2) publications, and (3) cross-sector familiarization, increased trust, new approaches to collaboration, and improved cultural competency among the more than 60 participating organizations, agencies, and companies.

One hallmark of the BC is that there is an absolute requirement that the founding members participate in *all* activities and at every level (e.g., executive and steering committees, project teams and subteams, and work groups) of the BC.

To date, the consortium has implemented 10 projects in areas such as Alzheimer's disease, cardiovascular disease, and cancer imaging; a number of other promising projects are also moving forward for implementation. The BC completed its first project, titled "Evaluate the Utility of Adiponectin as a Biomarker Predictive of Glycemic Efficacy by Pooling Existing Clinical Trial Data from Previously Conducted Studies" (adiponectin study) in 2009.

The results from the adiponectin study were published in June 2009 [4]. Conducted entirely via in-kind contributions from F. Hoffman LaRoche, GlaxoSmithKline, Merck & Co, and Quintiles Translational Corporation, the project involved aggregating data from clinical trials of peroxisome proliferator-activated receptor (PPAR) agonists at GlaxoSmithKline, Eli Lilly, Merck, and Roche. These pooled data were then subjected to analysis by statisticians at

Quintiles and at the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Among the project's results was evidence that adiponectin is a robust predictor of glycemic response to PPAR agonists in type II diabetes patients and that adiponectin has potential utility across the spectrum of glucose tolerance. Despite the challenges overcome by this project, the most important lesson learned is that cross-company precompetitive collaboration is a feasible robust approach to biomarker qualification [4]. Additionally—and equally important—this project served as a positive demonstration that cross-company collaboration is a feasible and powerful approach to biomarker qualification. For more detail see www.biomarkersconsortium.org.

3.4.3 Alzheimer's Disease Neuroimaging Initiative

Launched on October 1, 2004, the Alzheimer's Disease Neuroimaging Initiative (ADNI) is the NIH's largest public–private partnership focused on brain research and designed to gather and analyze thousands of brain scans, genetic profiles, and biomarkers in blood and cerebrospinal fluid (CSF).

This large-scale initiative was initially designed as a five-year research project aimed at defining biomarkers for use in clinical trials to determine the best way to measure treatment effects of Alzheimer's disease (AD), but the goal has been expanded to using biomarkers to identify AD at a predementia stage. ADNI involves scientists at 59 research centers, 54 in the United States and 5 in Canada. There are over 800 participants comprised of 200 with AD, 400 with mild cognitive impairment (MCI), and 200 with normal cognition. The success of this initial study has led to the launch of a *phase 2* of ADNI in the spring of 2010.

The overall goal of ADNI is to define the rate of progress of mild cognitive impairment and AD, to develop improved methods for clinical trials in this area, and to provide a large database which will improve design of treatment trials. A secondary, long-term expectation is that the results of this project will provide information and methods which will help lead to effective treatments for AD, leading to effective prevention.

The ADNI project was originally funded by the government for $60 million, with $40 million from the National Institute on Aging (NIA) and National Institute of Bioimaging and Bioengineering (NIBIB), which was leveraged with $20 million from the pharmaceutical industry and several foundations.

A hallmark of ADNI is that there is public access of the clinical and imaging data through the ADNI website and a parallel website at the Laboratory of Neuroimaging, making this a true "precompetitive" consortium.

A second important aspect of ADNI is it is governed by a steering committee comprised of the principal investigators (PIs), all funded core leaders, all site PIs, representatives of the NIH and FDA, and representatives of the companies contributing funding (observers only). Together with the Executive Committee and the Industry Scientific Advisory Board, these bodies ensure that the ADNI project adheres to the study design and methodology laid out

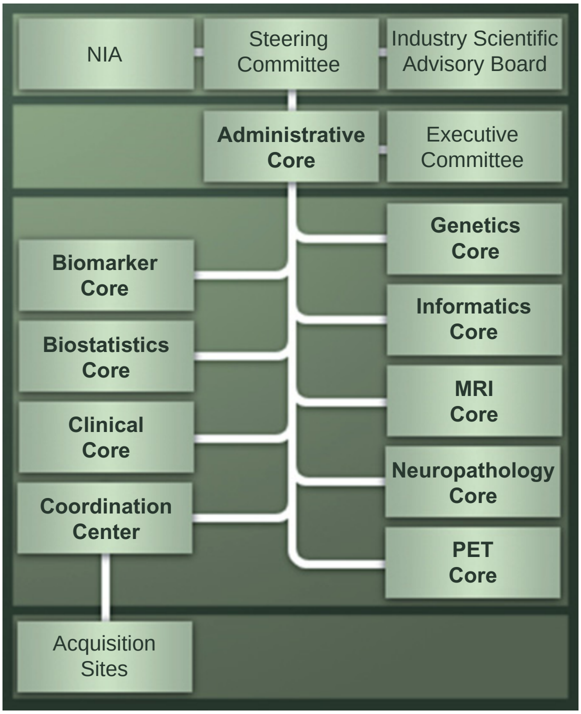


Figure 3.2 Schemata of organizational structure of Alzheimer's Disease Neuroimaging Initiative (ADNI).

in the grant submission. The highly developed structure of ADNI has helped to assure the success and productivity of this landmark public–private partnership (PPP). Figure 3.2 depicts the multiple levels of oversight and the elaborate organization that has enabled this PPP to be regarded as a successful example of a large-scale consortium. For more detail see www.loni.ucla.edu/ADNI and www.adni-info.org.

3.4.4 Osteoarthritis Initiative

The Osteoarthritis Initiative (OAI) is a multicenter, longitudinal, prospective observational study of knee osteoarthritis. The initiative is coordinated by the NIH's National Institute of Arthritis and Musculoskeletal and Skin Disease (NIAMS) and the NIA. The overall goal of the OAI is to develop and put into the public domain a research resource to expedite the evaluation of biomarkers for osteoarthritis as potential surrogate endpoints for disease onset and progression.

To support the initial efforts of NIAMS and NIA, a public–private consortium was established in 2002 to allow industry an opportunity to participate in this initiative. The OAI consortium leverages public funding from the NIH with private funding from several pharmaceutical company partners (GlaxoSmithKline, Merck & Co., Novartis, and Pfizer) and is managed by the FNIH.

The overall goal of the OAI is to generate an unparalleled, state-of-the-art database that is capable of explaining the natural progression of osteoarthritis and providing relevant information on imaging and biochemical biomarkers as well as outcome measures for osteoarthritis. Originally designed as a seven-year project, the success of this consortium led to a second generation of the OAI, which is presently under development. For more detail see http://oai.epi-ucsf.org/datarelease/ and http://www.niams.nih.gov/Funding/Funded_Research/Osteoarthritis_Initiative/default.asp.

3.5 OPPORTUNITIES FOR COMPUTATIONAL BIOLOGY RESEARCH PARTNERSHIPS

While some areas of research have been quicker to embrace the use of large-scale partnerships to advance individual research interests, the field of computational biology seems to be slow in warming up to this new model. As a result, to date the most successful large collaborations in computational biology have been run as "divide-and-conquer" strategies, as opposed to a "too-many-cooks-in-the-kitchen" scenario of shared analyses. As such, existing collaborative tools in our experience do not go much beyond teleconferencing, screen sharing/multicasting (e.g., via Virtual Network Computing [VNC]), and document editing (e.g., wikis). Over the last two decades, the accelerating pace of technological progress has generated a massive volume of biomedical data, opening a new era of life science investigation. For example, recent technical advances in the field of microarrays have produced considerable quantities of gene expression and metadata associated with various human diseases and conditions. The ongoing evolution of new technologies required computational methodologies to evolve in parallel.

As a result of this evolution, collaborative initiatives in computational biology have started to emerge to address the unique challenges arising in the field. In particular, an area that received special attention was the establishment of standards for data storage and data management. An illustrious example is given by the Minimum Information About a Microarray Experiment (MIAME) consortium, which outlines the content and structure of the necessary information required for recording and reporting microarray-based gene expression data [8]. The guidelines provided by MIAME include the standards for the raw data (e.g., Affymetrix CEL or GenePix GPR files), the processed (normalized) data, the sample annotations, the experimental design, the annotation of the array, and the laboratory and data processing protocols [8].

Similar examples of collaborative efforts to develop and agree on common vocabularies and standards include the International Health Terminology Standards Development Organization, which developed, maintains, and promotes suitable standardized clinical terminologies, notably SNOMED [9].

Furthermore, the Human Proteome Organization Proteomics Standards Initiative (HPO-SI) has helped enable the development of unified standards for proteomics. Through a collaborative model, HPO-SI has had an integral role in the development, publication, and adoption of several new interchange formats, commonly accepted terminologies, and data standardization, and presents a view on developments and policy. Additionally, repositories to support the HPO-SI formats are readily being established, while minimum reporting requirements have been developed and submitted for journal publication after prolonged exposure to community input via the PSI website [11].

Building off the standards developed by the PSI-MI, several public interaction databases came together to establish the International Molecular Interaction Exchange (IMEx) consortium (http://imex.sf.net). The IMEx consortium is an international collaboration among several public interaction data providers all agreeing to share curation efforts and to:

- Develop and work to a single set of curation rules when capturing data from both directly deposited interaction data and publications in peer-reviewed journals
- Capture full details of an interaction in a “deep” curation model
- Perform a complete curation of all protein–protein interactions experimentally demonstrated within a publication
- Make interaction available in a single search interface on a common website
- Provide the data in standards compliant download formats
- Make all IMEx records freely accessible under the Creative Commons Attribution License

One additional example of the standardization efforts is the Minimum Information About a Proteomics Experiment (MIAPE) guidelines, which are aimed at the establishment of “MIAPE-compliant” reporting [12].

3.6 CHALLENGES AND OPPORTUNITIES IN COMPUTATIONAL BIOLOGY

3.6.1 Open-Access Repositories

Due to the large amount of genomic data accumulated, the need for publicly available data repositories arose for collecting the data generated by the various research groups. This need, for example, is filled in North America by the Gene Expression Omnibus (GEO), developed by the National Center for Biotechnology Information (NCBI) [13]; in Europe by ArrayExpress, developed by the European Bioinformatics Institute (EBI) (http://www.ebi.ac.uk/microarray-as/ae); and in Japan by the Center for Information Biology Gene Expression database (CIBEX), developed by the DNA Data Bank of Japan (DDBJ) (http://cibex.nig.ac.jp). These repositories were established to store

MIAME-compliant gene expression data. Proteomics data repositories have also been developed, including the Global Proteome Machine Database (GPMDB) (http://gpmdb.thegpm.org), the Proteomics Identifications Database (PRIDE) [14] and PeptideAtlas [15].

Due to these efforts, investigators from various fields, including biomedical sciences and computational biology, can access public data repositories to enhance their research. For example, Lukk and colleagues constructed a global gene expression map based on data corresponding to over 5000 samples from 206 studies from 163 laboratories they obtained from GEO and ArrayExpress. These large-scale studies would have required huge collaboration efforts and most likely would not have been possible without public data repositories. Other benefits also include the possibility of validating results in an external cohort of subjects that were made public in one of these databases. More importantly, smaller laboratories that do not have the resources to collect their own samples and generate gene expression data now have the capability of using publicly available data to make important discoveries.

Many journals, including the *New England Journal of Medicine*, require the posting of the data before a manuscript is even considered for review to be accepted for publication. The benefit of this is that other researchers can replicate published results and build on the published work. The drawback is that when collaborating with pharmaceutical or other companies/institutions that would like to further mine the data before releasing it to the world, it might make it impossible to submit the manuscripts to the journals of choice.

3.6.2 Data Storage and Management

One of the requirements of collaborating within the field of computational biology is the need to merge data from the different groups. This means not only the storage of the data but also proper management such that data can be merged and queried seamlessly. Storing data from collaborators might only require extra hard drive space, but in most cases it entails far more. Usually data security involving multitier login-based access to the data is needed. If data are on human subjects, data anonymity is also a must. Although there are guidelines for data storage and management of transcriptomics and proteomics data, metadata might need to be addressed in a different matter. In most cases, metadata collected by different research groups cannot be merged very easily. It might require building a new database and finding relations between the different sources in such a way that joining of the tables in the different schemas or databases can be performed.

The establishment of standards enabled the development of new data management and data analysis tools to support collaborative and multidisciplinary studies. For example, MicroGen is a Web system used to store, manage, and exchange data characterizing spotted microarray experiments according to the MIAME standards [17]. Similar examples are EDGE(3) for Agilent two-color microarray experiments [18], MARS (Microarray Analysis and Retrieval

System) for multicolor microarray data [19], MiMiR for Affymetrix microarray and potentially other -omics technologies [20], EMMA 2 for spotted arrays and synthesized oligo arrays [21], and RAD [22].

Overall, the development of these new tools demonstrated the increasing need in the field to establish a shared and fluid dialogue among multidisciplinary actors from the beginning of a study and to collect and exchange efficiently detailed information on the experimental and computational procedures related to collaborative studies.

3.7 TOOLS FOR INNOVATION IN COMPUTATIONAL BIOLOGY: BIOCONDUCTOR AND R SOFTWARE

Bioconductor is an open-source software developed specifically for the mining of genomic data (http://bioconductor.org), but it now contains methods that can be applied for the analysis of other “omic” data too. It has a large collection and selection of statistical and bioinformatics methods, thus providing great flexibility in data analyses. Bioconductor is also an open development software project for computational biology and bioinformatics [23], and many researchers who develop new data analysis, data mining, and other related methods can create a package within this environment which is then made publicly available. These packages contain the methods (functions) and a description of what they do and how they can be used. The Bioconductor packages run under R, a free software environment for statistical analysis, computing, and graphics (http://www.r-project.org/).

Bioconductor aids computational biology collaboration in three ways. First, it supports virtual collaborations since the new methods developed for data mining are freely available in this environment by the method creators and are used by statisticians, bioinformaticians, computational biologists, and others. Second, different statistical technique developers can work together in building a comprehensive tool for data analysts. Third, Bioconductor allows different computational biology groups to use the same methods for free and analyze data in exactly the same way, share analysis tasks, and/or reproduce each other’s work.

Bioconductor is one example of the tools that will help spur the continued efforts for collaborative research in computational biology.

3.8 DISCUSSION

The material covered in this chapter sets forth and demonstrates the important role that collaborative efforts can, and should, play in biomedical research. As demonstrated by large-scale initiatives such as the BC, the consortia model is a valuable mechanism for leveraging resource, sharing cost and risk, and increasing the intellectual capacity of a scientific research effort. As individual stakeholders learn more about how these structures can provide "value-added" benefits for their own personal mission—and continue to see demonstrations of the success of research consortia—the interest in and acceptance of large-scale partnerships should continue to grow.

As with any successful partnership, there must be a "meeting of the minds," or at the very least a clear understanding of: Why each partner is involved. What each partner looks to get out of the collaboration. What policies and procedures must be in place to assure that each partner's mission and goals are achieved through the partnership's activities.

Like other areas of biomedical research, the field of computational biology has much to benefit from breaking out of the traditional model of collaboration and exploring the possibilities that large-scale collaborations can provide each stakeholder.

REFERENCES

1. Adams CP, Brantner VV. Estimating the cost of new drug development: Is it really $802 million? *Health Affairs* 2006;25(2):420–428.
2. Dictionary.com's 21st century lexicon. Available: http://dictionary.reference.com/browse/precompetitive.
3. Vargas G, Boutouyrie B, Ostrowitzki S, Santarelli L. Arguments against precompetitive collaboration. *Clin Pharmacol Ther* 2010;87(5):527–529.
4. Wagner JA, et al. The Biomarkers Consortium: Practice and pitfalls of open-source precompetitive collaboration. *Clin Pharmacol Ther* 2010;87(5):539–542.
5. Food and Drug Administration. Critical path opportunities report and list. http://www.fda.gov/oc/initiatives/criticalpath/reports/opp_list.pdf. 3-16-2006.
6. Woodcock J, Woosley R. The FDA Critical Path Initiative and its influence on new drug development. *Annu Rev Med* 2008;59:1–12.
7. Mayrand-Chung S. The Biomarkers Consortium: Advancing biomarkers research. Paper presented at Pharma Focus Asia 2009;9. Available at: http://www.pharmafocusasia.com/clinical_trials/biomarkers_consortium_advancing_research.htm.
8. Brazma A, et al. Minimum information about a microarray experiment (MIAME)—Toward standards for microarray data. *Nat Genet* 2001;29(4):365–371.
9. Cornet R, de Keizer N. Forty years of SNOMED: A literature review. *BMC Med Inform Decis Mak* 2008;8(Suppl 1):S2.
10. Martens L, Orchard S, Apweiler R, Hermjakob H. Human Proteome Organization Proteomics Standards Initiative: Data standardization, a view on developments and policy. *Mol Cell Proteom* 2007;6(9):1666–1667.
11. Orchard S, et al. Entering the implementation era: A report on the HUPO-PSI Fall workshop 25–27 September 2006, Washington DC, USA. *Proteomics* 2007;7(3):337–339.
12. Taylor CF, et al. The minimum information about a proteomics experiment (MIAPE). *Nat Biotechnol* 2007;25(8):887–893.

13. Barrett T, et al. NCBI GEO: Archive for high-throughput functional genomic data. *Nucleic Acids Res* 2009;37(database issue):D885–890.

14. Vizcaíno JA, et al. A guide to the proteomics identifications database proteomics data repository. *Proteomics* 2009;9(18):4276–4283.

15. Desiere F, et al. The PeptideAtlas Project. *Nucleic Acids Res* 2006;34(database issue):D655–658.

16. Lukk M, et al. A global map of human gene expression. *Nat Biotechnol* 2010;28(4):322–324.

17. Burgarella S, et al. MicroGen: A MIAME compliant web system for microarray experiment information and workflow management. *BMC Bioinform* 2005;6 (Suppl 4):S6.

18. Vollrath AL, et al. EDGE(3): A web-based solution for management and analysis of Agilent two color microarray experiments. *BMC Bioinformatics* 2009;10:280.

19. Maurer, et al. MARS: Microarray analysis, retrieval and storage system. *BMC Bioinform* 2005;6:101.

20. Tomlinson, et al. MiMiR: An integrated platform for microarray data sharing, mining and analysis. *BMC Bioinform* 2008;9:379.

21. Dondrup, et al. EMMA 2: A MAGE-compliant system for the collaborative analysis and integration of microarray data. *BMC Bioinform* 2009;10:50.

22. Manduchi E, et al. RAD and the RAD Study-Annotator: an approach to collection, organization and exchange of all relevant information for high-throughput gene expression studies. *Bioinform* 2004;20(4):452–459.

23. Gentleman RC, et al. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol* 2004;5:R80.

4

PRECOMPETITIVE COLLABORATIONS IN PHARMACEUTICAL INDUSTRY

Jackie Hunter

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.

4.1 INTRODUCTION

Over the past decade there has been an enormous increase in the number and range of precompetitive collaborations between pharmaceutical companies, biotechnology companies, non-for-profit organizations, and academia. Yet precompetitive research still represents only a small fraction of the research effort and funding that are expended by pharmaceutical and biotechnology companies. In part this is because traditionally pharmaceutical companies were fiercely competitive and many of these precompetitive collaborations are relatively recent. It is timely to assess what the drivers for change were, what obstacles may still be present, what lessons can be learnt from existing precompetitive efforts, and what the future of precompetitive research might be.

4.1.1 Definition of Precompetitive Research

It is useful to have an agreed-upon definition of precompetitive research together with an appreciation of how it differs from other approaches in a more open pharmaceutical research and development (R&D) framework. Although precompetitive research has been defined as competitors sharing early stages of research that benefit all [1], Janet Woodcock has more recently defined precompetitive research as "a subset of translational research that is focused on improving the tools and techniques needed for successful translation, and not on development of a specific product" [2].

A comparison of various forms of collaborative sourcing efforts is shown in Table 4.1. It can be seen from this that there are clear distinctions between these approaches and that the working definition proposed above for precompetitive research is a useful one that distinguishes it from some of the other types of collaborative activity, such as crowdsourcing. It also follows from this that some areas, such as safety science, can adopt a precompetitive agenda more readily than others, for example, developing new targets in a particular disease area. However, in looking to the future, the areas for precompetitive collaboration could be greatly expanded (see Section 4.4).

4.1.2 Drivers for Change

Ten years ago there were far fewer precompetitive collaborations than at present. Pharmaceutical companies were still yielding high profits and growth for their investors, the biotechnology bubble had yet to burst, and academic biomedical research was, in general, well funded in the United States and Europe. In the last decade, there has been a dramatic change in the finances of large pharmaceutical companies and biotechnology companies. The need for increased spending on pharmaceutical R&D, aligned with a lack of concomitant success in terms of new product approvals, has been well documented and discussed earlier in this volume. The cost of developing a new drug, includ-

TABLE 4.1 Comparison of Mechanisms to Explore More Open Approach to R&D

A patent pool is an agreement between two or more patent owners to license one or more of their patents to one another or third parties. A patent pool allows interested parties to gather all the necessary tools to practice a certain technology in one place, e.g., "one-stop shopping," rather than obtaining licenses from each patent owner individually.

Precompetitive research consists of basic or applied research and can include part of the development phase. At the precompetitive stage, research results are not immediately marketable even thought they are the basic tools for creating new products and processes.

Crowdsourcing is the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call.

Creative commons—ownership is retained by the originator, but they allow others to use the ideas/concepts/patents on the conditions the originator specifies. In many cases, there will be no rights and the information will be placed for free access in the public domain.

Open source refers mainly to software that is distributed with its source code so that end-user organizations and vendors can modify it for their own purposes. Most open-source licences allow the software to be redistributed without restriction under the same terms as the licence.

Open innovation is the use of targeted inflows and outflows of knowledge across organizational boundaries to accelerate internal innovation and to expand markets for the external use of innovation.

ing the cost of failed compounds, is now in excess of $1 billion [3, 4]. For the industry to remain financially viable, either more compounds must make it to the market or the industry has to find ways to reduce the cost of failure. Both of these approaches can benefit from precompetitive collaboration. The major causes of failure of compounds today are unexpected toxicity in animals and humans and lack of efficacy in early- or late-phase clinical studies. In order to tackle these issues, the industry needs more predictive biomarkers for early detection of toxicity as well as developing better technologies to ensure the efficient translation of early research into clinical efficacy. In most cases no one company has sufficient data to be able to address these questions or expend the time and financial or other resources in developing these new technologies and tools. Additionally, although companies still fund academic centers, they are expecting a greater degree of value creation from such collaborations. Governments are also seeking greater signs of economic return on their biomedical research funding, providing further stimuli to industry–academic interaction. This academic–industry interface is the subject of a subsequent chapter, but there are lessons to be learnt from precompetitive collaborations which can apply to many types of collaborations. It is clear that the more traditional model of companies doling out large sums of money to academics and remaining distant from the subsequent science that the company's money funds is unlikely to be more successful than it has in the past. New ways of working with the intellectual input, expertise, and resources of both sides being valued should have a better chance of creating innovative medicines through increasing efficiency and better harnessing of emerging scientific knowledge.

The public sector can also provide an impetus for change. Over the past decade there has also been a realization by governments and funding agencies that the failure of the industry to address some of the most serious health care challenges such as the diseases of old age (e.g., dementia, cardiovascular disease) will have serious economic consequences for society both in the developed and the developing world. Therefore governments and agencies have, in different ways, sought to encourage the precompetitive agenda. Examples include the Innovative Medicines Initiative (IMI) in Europe, the Critical Path Institute (CPI) in the United States, and precompetitive funding [e.g., The Structural Genomics Consortium (SGC) http://www.thesgc.com] by organizations such as the Wellcome Trust [5–7].

Another driver that is not often explicitly mentioned is that of transparency. The industry needs to build confidence and trust with a variety of stakeholders, including the general public. Precompetitive consortia, if publicized appropriately, can do much to enable dialogue with patients, regulators, and others, which should lead to increased understanding and communication as well as increased trust [5].

4.2 EXAMPLES OF PRECOMPETITIVE CONSORTIA

Many different models exist for precompetitive consortia and a number of consortia exist (see the Appendix). Two established consortia are described below, but a number of recent initiatives have been announced in both the developed and the developing world. In February 2010 Lilly, Merck, and Pfizer announced that they would share data, through an independent organization, the Asian Cancer Research Group, on pharmacogenetic data related to lung and gastric cancers. GlaxoSmithKline (GSK) set up a patent pool in 2009 to facilitate access to intellectual property (IP), industrial expertise, and technologies to stimulate research into neglected tropical diseases. This pool was joined by Alnylam Pharmaceuticals and administered by BIO Ventures for Global Health. The Massachusetts Institute of Technology and South Africa's Technology Innovation Agency joined the initiative, now known as the Pool for Open Innovation against Neglected Tropical Diseases, in 2010. In order to support researchers, GSK also set up laboratories for their use at its research center in Tres Cantos, Spain. More recently the data from screens for inhibitors of *Plasmodium falciparum*, the protozoa that is responsible for malaria, were put into the public domain [8–10].

The two initiatives described below are both public–private partnerships (PPPs) where industry is collaborating precompetitively. They differ in scale, scope, and location, but they demonstrate how such large collaborations can be successfully established.

4.2.1 Innovative Medicines Initiative

The IMI was officially established by the European Commission and EFPIA, the European Federation of Pharmaceutical Industry Associations, in 2007 and is therefore relatively new. The dialogue for this initiative started in 2004 when the commission approached the Research Directors Group (RDG) of the EFPIA and asked them what the bottlenecks were in the drug discovery and development process and how the commission could stimulate research in Europe to address these. The RDG worked with the commission and various stakeholder groups, including academia, biotechnology companies, and patients, to draw up a strategic research agenda from which the topics for research funding would be drawn. A pilot project bringing companies and researchers together in two areas, safety pharmacology and neurodegeneration, was also funded under the Framework VI Programme to identify some best practices and issues that might arise in collaborations of this nature (http://www.innomed-addneuromed.com).

Collaborations of this scale required a new funding mechanism—one that could be used across a range of industries, not just the pharmaceutical industry. Such a new funding vehicle was established by the European Commission—the Joint Technology Initiatives (JTIs). The JTIs were a new way of realizing PPPs in research at the European level that were released as part of the Framework VII Programme for Research, Development and Demonstration. They were intended to support transnational cooperation in fields of key importance for industrial research [11].

The overarching objective of the IMI was to promote Europe as the most attractive place for pharmaceutical R&D, thereby enhancing access to innovative medicines for patients with a key deliverable of the provision for new tools and methodologies to remove major bottlenecks in drug development. It is one of the largest biomedical research initiatives in the world with a contribution of €1 billion from the European Commission which is being matched by a contribution of €1 billion from the EFPIA and its member companies. What makes it unique is that the industry contribution is primarily given as "in-kind contribution." This can be in the form of reagents, personnel, clinical data and samples, preclinical data and samples, and so on. The money from the commission goes to fund the work of participating academic groups, small- to medium-sized biotech companies (SMEs), and other participants such as patient groups. The main focus of the research is on developing and validating new techniques and methods to enhance the prediction of safety and efficacy of new medicines. This is underpinned by better knowledge management that will provide the necessary data pooling and data processing. Education and training programs will ensure a workforce in Europe that is more skilled for the future needs of this sector [12]. This is shown schematically in Figure 4.1.

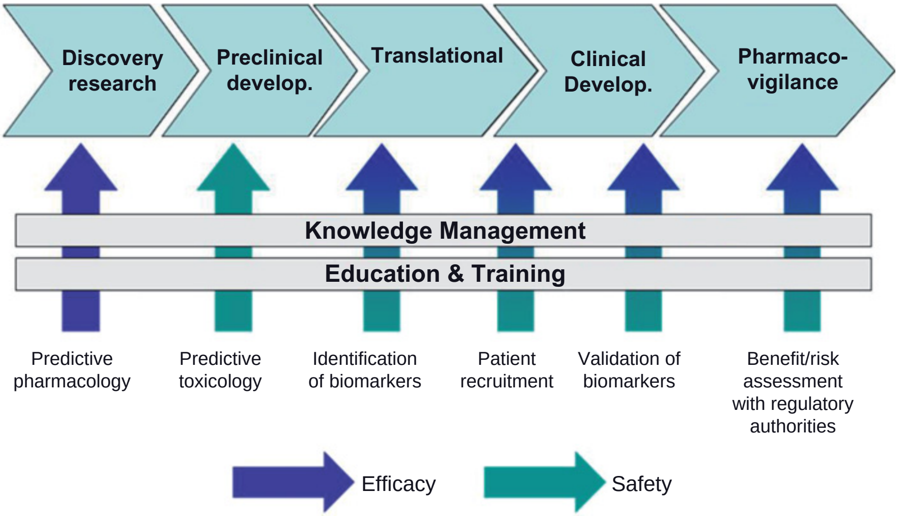


Figure 4.1 Schematic showing key areas where IMI activity will impact R&D process.

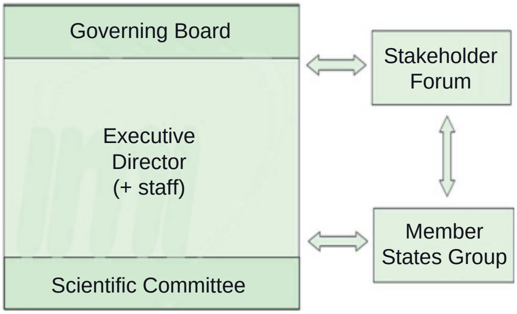


Figure 4.2 Governance structure of IMI.

The initial strategic research agenda was approved by the IMI governing board meeting on March 3, 2008, and is available on the IMI website.

4.2.1.1 Governance The governing and management structure of the IMI consist of several bodies and committees, as shown in Figure 4.2. The official legal entity formed by the EFPIA and the EC is called the IMI Joint Undertaking (JU) and is responsible for implementation of the IMI program. It is comprised of the IMI executive office, the IMI governing board, and the IMI scientific committee.

The executive office is really a neutral third party with three main functions:

- Implementation of all programs and activities in the best interest (common interest) of all stakeholders
- Financial accountability for the use of the funds from both industry and the commission
- To ensure fair and reasonable conditions for optimal knowledge exploitation and dissemination

The governing board is the main decision-making body of the IMI JU and has overall responsibility for the operations of the undertaking and oversight of the implementation of its activities, ensuring their alignment with IMI objectives. The governing board is composed of 10 board members representing equally the two founding members of the IMI JU: 5 from the European Commission, representing the European Community, and 5 from EFPIA, representing the pharmaceutical industry in Europe.

The scientific committee provides scientific advice to the governing board. It is currently composed of 15 members and gives strategic science-based recommendations to the IMI JU, advises on the scientific priorities which form the basis for call topics, and in 2010 led the revision of the strategic research agenda.

In September 2009, Michel Goldman became the first executive director of the IMI and has played a critical role in moving the IMI agenda forward. Two other committees also play an important role in the dissemination of IMI to all stakeholders and conduits for feedback to the IMI—these are the Member States Group and the Stakeholder Forum. The aim is very much to use these bodies both as advocates across Europe for the IMI and to gain feedback on ways in which IMI procedures and processes could be improved for the future.

4.2.1.2 How IMI Research Process Works The EFPIA companies discuss together which topics are important for the next call for research proposals. These topics are then validated and agreed in discussions with the European Commission and the scientific committee and members of the EFPIA form consortia to support each call topic. The role of the executive office at this point is to attract the best partners for the agreed-upon call topics in Europe by launching a call for proposals for the agreed-upon topics. At this stage the proposals are known as expressions of interest (EOIs), which are less than 10 pages in length and come from the academic and other non-EFPIA partners—the aim here is to minimize the amount of work involved at this stage as only one EOI will be selected for each call topic. The submitted EOIs are then evaluated by a panel of experts and the best EOI is selected for each topic. The consortium of academics that submitted this successful EOI then meets with the respective industrial consortium and the merged consortium works on the final project proposal. It is at this point that the details of the in-kind contribution, project management, timelines, deliverables, and any IP considerations are negotiated and written down in a formal project proposal. This is again evaluated by an independent scientific committee before it can be approved.

The first call was launched in 2008 and 15 of the 18 call topics resulted in consortia projects that made it all the way through to approval. Most consortia are led by a coordinator from the EFPIA with a deputy coordinator from academia, although this is not always the case, with PROTECT being led by the European Medicines Agency (EMA). These 15 consortia involve some 395 teams across Europe and cover the efficacy, safety, and education and training themes of the IMI. In terms of efficacy there are projects in diabetes, asthma, chronic obstructive pulmonary disease (COPD), cognitive disorders, schizophrenia, and pain.

Safety Projects

- Non–genotoxic carcinogenesis (MARCAR) aims to provide validated reliable early biomarkers for the prediction of cancer development.
- Expert systems for *in silico* toxicity prediction (eTOX) will develop a pharmaceutical toxicity database and *in silico* expert systems for the computational prediction of drug secondary pharmacology and direct drug-induced toxicity.
- Qualification of translational safety biomarkers (SAFE-T) aims to develop new specific and sensitive safety biomarkers and their respective assays for human samples to improve predictivity between nonclinical and early clinical studies.
- Strengthening the monitoring of the benefit/risk of medicines (PROTECT) will develop new methodologies in pharmacovigilance and pharmacoepidemiology. It is important that these safety projects communicate regularly with other similar projects both within Europe and without and where possible avoid duplication of effort. For example, the clinical IMI projects are complementary to those of the Predictive Safety Testing Consortium (PSTC), which is focusing on preclinical safety and toxicity.

Efficacy Projects

- Islet cell research (IMIDIA) will gain a better understanding of β-cell function and survival.
- The aim of surrogate markers for vascular endpoints (SUMMIT) is to validate agency acceptable surrogate markers for micro- and macrovascular diabetic complications, thereby enhancing the efficiency of drug development studies and shorten clinical trials.
- In pain research, EUROPAIN aims to improve the understanding of pathways and mechanisms mediating different kinds of pain and develop markers for patient stratification and quantitative pain assessment for efficient testing of new analgesics. This is desperately needed in light of

the recent failure of a number of potential pain therapies in phase II [e.g., Radopril (RGH-896)] in neuropathic pain [13].

- New tools for the development of novel therapies in psychiatric disorders (NEWMEDS) will validate blood and CSF markers as well as dynamic physiological and structural measures suitable for both clinical and preclinical clinical assessments to enable better progression and prediction for new drugs for psychiatric disorders.
- In neurodegenerative disorders (PHARMACOG) this initiative will develop better translatable animal and human volunteer models for increased predictivity of translation of efficacy in these models for new therapies in patients with Alzheimer's disease.
- Understanding severe asthma (U-BIOPRED) aims to create a large longitudinal patient cohort enabling validation of novel biomarkers and development of diagnostic criteria for mechanistic and therapeutic trials.
- A COPD patient-recorded outcomes project (PROACTIVE) will develop a comprehensive framework for better understanding of patients' physical activity in COPD in dimensions considered relevant by the patients and leading to developing strategies for measuring clinical trials outcomes.

Education and Training Programs

- The European Medicines Research Training Network (EMTRAIN) aims to develop a European biopharmaceutical research training platform to provide a sustainable academia–industry cross-disciplinary approach to efficient organization of training courses on emerging science and technologies across Europe.
- The safety sciences for medicines training program (SAFESCIMET) will produce a training program that will integrate all safety-relevant disciplines linking animal and human/patient safety data, thereby facilitating a more holistic evaluation of new medicines.
- The pharmaceutical medicine training program (PHARMTRAIN) will establish a network of academic centers that delivers postgraduate training programs in pharmaceutical medicine, including quality management of the processes and outcomes.
- The pharmacovigilance training program (EU2P) will develop a pan-European training and education network platform in pharmacovigilance and pharmacoepidemiology to train professionals for the pharmaceutical industry, regulatory authorities, and health care organizations. The long-term objective is to improve the understanding and effectiveness of risk communication.

As can be seen from the above, the scope and range of topics are extremely large. The second call research priorities were announced in 2009 and focused on knowledge management and efficacy areas with the efficacy projects being focused on oncology, inflammation, and infectious diseases. As with the first call there were a large number (124) of expressions of interest and the winning applicant consortia have submitted full project proposals. The third-call priorities have also been announced and include projects in safety, efficacy, and education and training.

4.2.1.3 IP Policy The IMI has a clear, published (http://imi.europa.eu) IP policy that is distinct from the IP policy in other Framework VII–funded programs. The IP policy was designed to be aligned with objectives of the IMI as a whole but to retain the flexibility to be tailored to the needs of each individual project. It was put together by a highly experienced group of experts and will be monitored by a working group of IP experts and representatives of the IMI founding members (EFPIA and the European Commission) and the member states group. The initial policy together with a clarification note issued by the working group (http://imi.europa.eu) very clearly state, what is meant by background, foreground, and sideground IP and is a good model for other PPP to use.

Although each project has clear objectives as stated above, there are shorter term, broader benefits that the IMI might be expected to deliver. One of the more immediate benefits should be a shared understanding across all stakeholders of the challenges and opportunities in drug discovery and development. For example, a number of regulators are involved in IMI projects including the EMA and the agencies of the United Kingdom (MHRA), Denmark (DKMA), Spain (AEMPS), Switzerland (SwissMedic), and France (AFSSAPS).

There is also a growing involvement of SMEs in ongoing projects—24 companies received €13.9 million in the first calls. Initially SMEs, especially those developing biomarkers, viewed the IMI with suspicion, but the positive experiences of those involved in the first call will hopefully encourage further SME involvement in subsequent calls. Patient groups are also involved in a number of these first calls, including Asthma UK, European Lung Foundation, International Alliances of Patients Organisations, and Alzheimer's Europe. The voice of the patient will become more important as the industry and health care providers move further toward a personalized medicine agenda.

4.2.1.4 Learnings from the IMI The size and scope of the project were such that initially many academics and industry people did not believe it would be workable to bring so many partners together. However, as can be seen above, this has been achieved and projects have commenced. The enthusiasm and commitment of scientists from both industry and academia is something that the IMI office has commented on at various meetings. It will be important going forward that this level of involvement, enthusiasm, and commitment is maintained, and incentivized, by participating companies. The need for the projects to obtain in-kind contribution from industry has meant that they are very focused on things that could make a real difference to the drug discovery and development process. However, this did cause some difficulties early on in the review process as reviewers, primarily from academic backgrounds, were more used to reviewing very early, more blue-sky projects. Likewise in the discussions around the IP policy, it was much easier to define the boundaries in safety projects than in efficacy projects, although there were no IP issues that precluded projects from the first call progressing. Efficacy projects also had challenges as they are very much dependent on the long-term commitment of a company to the particular therapeutic area in question. In 2010 a number of companies exited aspects of neuroscience research, which meant that there had to be some changes to consortia involved in the neuroscience efficacy projects.

Another lesson from the IMI was the importance of providing subject matter experts within companies to drive forward the IMI centrally. This included people from finance, legal, IP, and project management functions. With such a large program of work there is clearly a need for dedicated tools for data sharing and knowledge management and the IMI office will have to make sure that there is sufficient cross talk across programs and projects to ensure data can be transferred and stored in an accessible way with little duplication. The IMI office is also working with groups globally, such as the Biomarkers Consortium, to reduce overlap and duplication and ensure efforts are synergistic where possible. Indeed there is a clear requirement going forward for the IMI to complement rather than compete with other initiatives globally. Resources within the industry will continue to be constrained, and as large companies are global, they will seek to avoid duplication of effort in precompetitive collaborations.

Leadership from the industry participants was also important in moving the IMI agenda forward. The RDG developed a good working relationship that allowed the industry to agree on the bottlenecks that needed to be solved and show a clear commitment to the precompetitive agenda. Such open and honest communication takes time to develop but is necessary to tackle the inevitable issues that arise.

There do remain a number of challenges going forward. There needs to be greater involvement of patients and lay persons to maximize their input and ensure the aims and achievements of the IMI are communicated widely. The need to consider contributions from outside Europe and from pharmaceutical companies not in the EFPIA or from related industries (e.g., diagnostics) is also a challenge that remains despite being flagged as an issue very early on in the process. There are also other financial challenges such as the differences between the IMI and other Framework VII funding programs in funding parameters such as overheads. However, the IMI has clearly shown that precompetitive research can happen on a grand scale and has largely been successful in its original aim of stimulating more pharmaceutical R&D in Europe.

4.2.2 Structural Genomics Consortium

There are a number of initiatives aimed at solving protein structures [14]. The SGC is a PPP that, like the IMI, aims to facilitate the development of new medicine, but in this case it does so by carrying out the basic science of relevance to drug discovery through increasing basic knowledge of protein structures. It was established in 2003 and operations commenced in mid-2004, and therefore the SGC is a relatively mature precompetitive effort. The SGC is based in three centers—the University of Toronto, the Karolinska Institute and the, University of Oxford—and thus is globally more diverse than the IMI. It originally was comprised of approximately 200 scientists funded by three different types of funding sources. Government funds were from Canada (Genome Canada, Canada Fund for Innovation, Canadian Institutes of Health Research, Ontario Genomics Institute, Ontario Ministry for Research and Innovation) and Sweden (Swedish Foundation for Strategic Research, Vinnova). Charity funding came from the Wellcome Trust and the Knut and Alice Wallenberg Foundation, and industrial funding was provided by GSK initially, with Merck and Novartis joining phase II in 2007. It differs from other precompetitive consortia in this area because industry is a key sponsor of this effort, which ensures that many of the targets are directly relevant to drug discovery. For example, the Targeted Proteins Research Programme in Japan (www.tanpaku.org) is wholly publically funded, although it is studying proteins of relevance to both academia and industry. The original objective of the SGC was clearly to promote drug discovery by substantially increasing the number of medically relevant human protein structures, as well as related reagents and protocols, available in the public domain. In this regard it was also different from the Protein Structure Initiative (www.nigms.nih.gov/Initiative/PSI) in the United States, which had not previously focused on human proteins. In the SGC the targets are selected by the funders and the first phase of the SGC aimed to generate 386 protein structures. At the end of phase I, 450 structures had actually been solved. The target for the second phase of operations (2007–2011) is 660. By mid-June 2010, nearly 1000 human protein structures (cumulative for phases I and II) had been solved and 2000 human proteins purified, and therefore the consortium is performing ahead of the original expectations for phase II. An example of this is shown in Figure 4.3, where the SGC has been responsible for solving nearly 40% of all kinase structures. One of the aims of the SGC was to provide the research community with reagents and tools to maximize the scientific impact of the SGC, and as an example of this hundreds of cDNA clones are distributed to academic and industrial collaborators each year. The SGC also publishes protocols for the expression and purification of proteins that have been successfully purified in the SGC laboratories and has developed a number of software tools to aid researchers in structural biology. An additional aim was to publish extensively in high-impact journals, and the track record of the consortium to date shows that this has been achieved (http://www.thesgc.com).

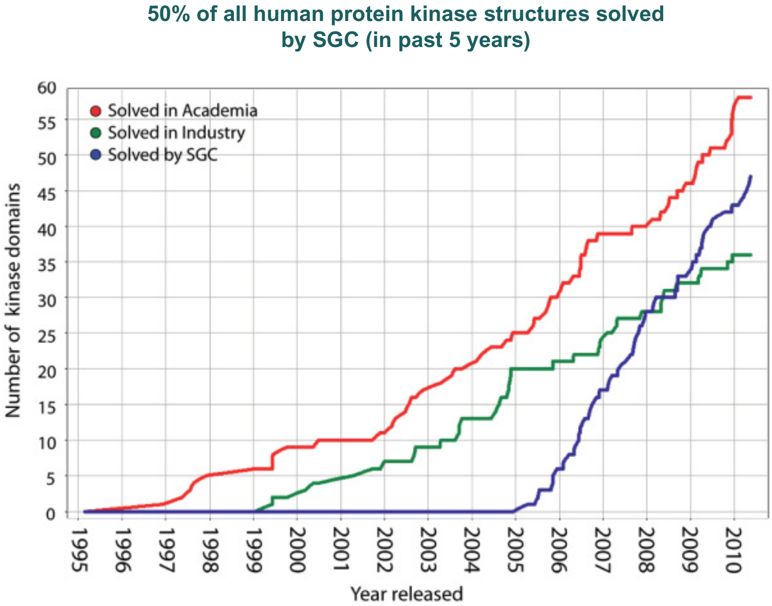


Figure 4.3 Contribution of SGC to solving protein kinase structures.

4.2.2.1 IP Policy Importantly, the SGC operates in precompetitive space and does not generate any IP; thus the consortium is committed to placing any results promptly into the public domain and does not file for patent protection on any of its research outputs [15]. It also means that the SGC seeks the same commitment from any of its research collaborators; despite this, the policy does appear to confer significant advantages, primarily through allowing speed of interaction with potential collaborators. It also means that the consortium can work with multiple private partners on the same project. The SGC currently has over 250 collaborators in 19 countries, and it is easy to access the relevant individuals through the consortium website.

4.2.2.2 Governance The SGC is run by a chief executive and oversight is provided by a board of directors (primarily nominated by the funders) and a scientific advisory board. Operations at each of the three sites are managed by a chief scientist and these individuals along with the CEO sit on the board of directors.

The scientific committee approves the list of protein targets for the SGC to work on and also those targets which are deprioritized. Reasons for the SGC ceasing to work on a target include the structure being solved elsewhere, intractability of the protein to solubilization or crystallization, or realignment of strategic focus.

In 2009, the SGC added an additional consortium project—the International Epigenetics Chemical Probes Consortium, which aims to stimulate epigenetics research via the discovery of "open-access" reagents. The consortium will produce chemical probes that are optimized for potency, epigenetic target selectivity, and cellular activity and hence will be able to stimulate or block the activity of proteins under epigenetic control. This was financed initially by a grant from the Wellcome Trust, but in-kind contributions came from the Chemical Genomics Centre based in Baltimore, which agreed to run 20 high-throughput screens, and GSK, which gave eight full time employees (FTEs) to the project. Subsequently the consortium has been able to attract more funding and partners with Pfizer and Novartis (each contributing eight FTEs) and other organizations contributing FTEs and funding from Canada (Ontario Ministry of Research and Innovation) and Sweden (Swedish Foundation for Strategic Research). In keeping with SGC policy, the structure and function (including physicochemical properties and screening data) of each probe will be made available. Each probe is declared as such by an independent scientific committee comprising world leaders in epigenetics, cellular assays, medicinal chemistry, and drug discovery.

The SGC model demonstrates how an open-source model can be applied to precompetitive drug discovery when both the academic and industrial partners obtain clear benefits.

4.2.2.3 Lessons from the SGC Clearly the policy of making everything freely available with no IP constraints has promoted a wealth of collaboration and data sharing that otherwise might not have been possible and removed a lot of potential bottlenecks. Going forward, as the consortium moves into chemical probe space, there may be more pressure from both academic and industry partners on the IP policy, and it will be interesting to observe how this develops. Even today the "no IP" policy means there are some limitations to what the consortium can do.

The chemical molecules that meet certain predetermined criteria in terms of selectivity will be made available through a commercial supplier's catalogue as probes for academics and industry. However the structure activity relationship (SAR) that underpinned them will not necessarily be available in the public domain, and this may be perceived as against the precompetitive spirit of the collaboration.

Working with many partners also brings its challenges and puts significant onus on the investigators to organize their collaborative network and ensure that the consortium is fully aware of all collaborative activities. The clear governance structure with its representation from all the sponsors means that issues around transparency are minimized.

4.3 IMPLEMENTATION AND MANAGEMENT OF PRECOMPETITIVE CONSORTIA

Given the large number of precompetitive consortia that have been in existence for some years, there is now a wealth of evidence about factors that increase the effectiveness of such collaborations. These lessons can be distilled into a few clear principles which, if addressed early on in collaborative discussions, can save much time and energy in the formation of any new consortia. One thing is clear, however—consortia such as the IMI, SGC, the Biomarkers Consortium, and Asian Cancer research group have shown that cross-company precompetitive collaboration is feasible and provides a new route to tackling some of the major health care challenges that exist today.

4.3.1 Best Practice Principles

1. *Agree on Goals and Objectives Up Front* It sounds obvious that all parties' objectives and goals are aligned and clearly articulated when the consortium is formed. However, this is not always the case. Frequently it is due to tacit assumptions about the goals and objectives of other parties, but these really need to be tested and agreed-upon explicitly. Ideally the aims and objectives should be captured as part of any consortium agreement. While this is important in any collaboration, it becomes even more so where multiple partners from different sectors are involved. The deliverables have to have value for all parties to retain the interest of the sponsoring organizations. Lack of alignment at this early stage has the potential to cause the consortium to fail very early on. For example, before the launch of the Biomarkers Consortium, the founding members had multiple discussions on what specific questions in the therapeutic areas of focus might be of interest to all stakeholders in the consortium [6]. For the Biomarkers Consortium this preproject consensus building was seen as critically important to the project. The single nucleotide polymorphism (SNP) consortium had a very clear objective which was easy to describe and identify with—other initiatives with more diverse, longer term objectives, such as the Biomarkers Consortium and IMI, have had to work much harder to communicate their objectives to stakeholders.

2. *Strong Leadership* There needs to be a key person in each organization that will act as the internal champion for the project to promote the project internally and externally and act as a key conduit of information. It is important that each organization provides the right people for these roles and gives them the right amount of resources and authority to contribute optimally to the consortium. This is facilitated significantly when there is senior sponsorship within the organization.

It is also important to have a single point of contact within each organization for all the legal issues whether it is for the drawing up of consortium contracts, discussion of data sharing issues, or other legal matters (e.g., freedom to use materials from third parties such as transgenic animals, software, or reagents and antitrust considerations). This was the case within the IMI and facilitated the construction of common templates for grant agreements, for example, and this was also one of the early lessons from the Biomarker Consortium [6].

3. *Have an Agreed-Upon and Clearly Articulated IP Policy* With precompetitive consortia it becomes very important to have very clear IP guidelines that all members of the consortia sign up to—even if it is only to agree to a "no IP" policy as with the SGC. As part of the IP discussion, a range of likely scenarios should be covered identifying the likely associated ownership of the IP that would result from each scenario. Other legal issues should also be specified where appropriate.

4. *Importance of Good Project Management and Governance* The funders of most large-scale collaborations now insist that project management tools and milestones be used to plan and monitor the progression of a project. It is also important to ensure that appropriately skilled individuals are given the time and resources to implement the project plans. Any private company spending millions of pounds (euros or dollars) would ensure that there were individuals responsible for operational (as opposed to the technical/scientific) management—the sums involved in large-scale precompetitive projects are no different and have the same operational demands. Part of this is clearly defining what each party will bring to the consortium, what they are accountable for delivering and when it will be delivered, and monitoring this regularly. As can be seen from the IMI and SGC examples, clear governance structures were put in place which helped the smooth running of the consortium, and this is also the case for the consortia listed in the Appendix.

One thing noted with the IMI example is the importance of defining where clinical governance will reside for clinical studies undertaken by the consortium.

5. *Standards* It is also important that standards for experiments, data storage, and analysis and clear expectations on the level of recordkeeping are agreed upon early on. Pharmaceutical companies operate in a highly regulated environment where standards are clearly laid out whereas academic laboratories do not have to operate to the same level of standard operating procedures. However, if data obtained in the collaboration are to be used later for regulatory submissions, accurate and comprehensive recordkeeping will be required. Alongside this resources should be explicitly requested for data mining, analysis, storage, and so on—this can often be a major issue and yet consortia such as the Coalition Against Major Diseases (CAMD) depend on having these functions adequately resourced.

6. *Have Clear Exit Strategy* All projects should have a clearly defined exit strategy—whether this is at the end of the natural life of the project or a reflection of changing information or scientific priorities of one or more of the individual parties involved. This is often hard to contemplate—just as no one enters a marriage planning for divorce, no collaborator wants to entertain the option of failure. But just as prenuptial agreements can save a lot of time, angst, and lawyer's fees, so can agreed-upon exit strategies for large-scale collaborations. The strategy could also allow for flexibility on behalf of the participants to change the direction of the project as new data emerge.

7. *Communication* Internally it is important to ensure that there are clear lines of communication between participants, and the use of collaborative resources such as websites and social networking tools is recommended. It is important early on in any precompetitive consortia for the key members to meet face to face to build trust between participants and facilitate efficient consensus building, for example, around goals and objectives.

Externally many precompetitive consortia would be able to do more to enhance the reputation of the pharmaceutical industry if the pharmaceutical participants were able to devote more time and resources to publicizing their progress. On the initiation of a collaboration, the participants are all usually involved in press releases, but it is usually the academic partners that then maintain the flow of information around progress (e.g., SGC).

4.3.2 Capturing Value Added in a Collaboration

Given the increase in precompetitive collaborations across the whole of biomedicine, it will be much more important going forward to be able to assess their value and to manage precompetitive collaborations as a portfolio within one company. This means adopting a strategic approach so that the maximum benefit can be obtained from such collaborations and so that they fill gaps so as to deliver real value to the company. However, measuring the value is something that has not received much attention.

From an individual company, a good starting point is to consider why the company entered into the collaboration in the first place:

- What were the tangible deliverables—impact on time, cost, novel endpoints for clinical trials?
- What were the intangible benefits—access to talent, networks, knowledge, key stakeholders such as patient groups?

The next step is to map these against the current landscape, a snap shot in time against which future progress can be made. Finally, a framework should be established that allows progress against the relevant metrics to be captured as the project progresses. This can provide an important set of data for communicating the value of a project both internally and externally with other stakeholders. Recently such a framework has been developed which allows both quantitative and qualitative capture of expected and actual value [16]. Many companies have failed to reap the benefits of their investment in precompetitive consortia because they have not paid attention to capturing the value both directly and indirectly. For example, a project in the IMI could provide access to key thinkers across Europe as well as opportunities to explain some of the challenges of drug discovery and development to the lay people, potentially via member states countries as well as via the commission. These “softer” outcomes are often ignored or not captured and yet they could provide some very early real benefits, which is important given that many of these projects will take years to deliver the more tangible, directly related project outputs.

4.3.3 Potential obstacles to precompetitive collaboration

Vargas et al. [17] have recently outlined some of the challenges to precompetitive collaboration, including definitions and standards, organizational complexity, and IP. In the case of IP they rightly point out that many of the issues around data sharing are with data collected in the past and that these should be reduced by basing precompetitive collaborations on prospectively designed experiments. They conclude that "it is good to see that PPPs have successfully formed and seem to have successfully tackled IP challenges." However, it will be interesting to see whether the IP issues will become more significant and precompetitive efforts seek to access tool compounds or share information around novel targets. Already the deposition of the compounds in the Pool for Open Innovation against Neglected Tropical Diseases has caused some comments that these compounds are not "druglike," as claimed by GSK [18]. This highlights the need for very clear communication about the value and nature of what is being shared as any perceived lack of transparency could easily create a negative impression for the consortia.

Decreasing size of pharmaceutical R&D is also a challenge—this may seem paradoxical as this should be an impetus to precompetitive data sharing. However, as companies shrink their internal R&D efforts, their R&D employees may be increasingly reluctant to become more externally facing as they feel the need to focus more on maintaining their internal impact. This is why it is important, if external collaboration is seen as a strategic imperative, to reward and recognize those individuals playing significant roles in precompetitive efforts. Additionally, if the personnel responsible for initiating and leading the precompetitive consortia then leave a company, it is important to ensure a suitable replacement it found. This reinforces the need for senior-level endorsement and support of the project as mentioned above. There is also the potential for conflict more generally in terms of the way large pharmaceutical companies are organized—while there has been a growing acceptance of networked R&D models in the industry, networks alone cannot increase innovation [19]. There still needs to be a significant cultural shift at all levels (especially in the more traditional middle management strata) within most pharmaceutical companies to take full advantage of new, precompetitive ways of working.

A further potential obstacle is a lack of coordination globally between funders, researchers, and companies on precompetitive funding initiatives. Although current PPPs such as the Biomarkers Consortium and the IMI have begun to communicate and coordinate activities, there still remains a lot of potential overlap across all the PPP space. In addition, it is key that industry agrees where it needs to focus resources and efforts in the precompetitive arena. In that respect the EFPIA Research Directors Group has made considerable progress, but this is primarily focused around research based in Europe and really needs to be extended globally and to ensure that companies not part of the EFPIA are brought into the debate. There are other fora where industrial partners work together strategically in the precompetitive arena.

One example is the industry program at the European Bioinformatics Institute (EBI, www.ebi.ac.uk/industry/ind-prog-index.html) where the 16 member companies meet on a quarterly basis to discuss issues of relevance to bioinformatics [20]. Such fora have highlighted the need for more global standards for data collection, annotation, and storage, which when agreed upon and implemented would aid cross talk between projects and collaborations and speed up data integration and analysis. Going forward ways of maintaining data warehouses and other central resources will have to be found and new funding mechanisms established to enable this.

Although there has been considerable progress made in this area, there still appears to be a need for a cultural change among some academics and funders to appreciate the different nature of the questions addressed by precompetitive research collaborations. As with the IMI example, these are frequently more applied questions than pure blue-sky research and as such may need to be evaluated by different research criteria to those of the more normal, speculative, basic research grant proposals. This does not mean that they should not be underpinned by excellent science but rather that the other criteria for funding may be different. The need for real collaboration between the pharmaceutical company partners and others also means that the more traditional view of pharmaceutical companies as a source of funds must change [17].

Although some precompetitive collaborations have clearly delivered on their objectives for all stakeholders (e.g., the Dundee Kinase Consortium, the SGC, and the Serious Adverse Events Consortium [SAEC]), others have not been in place for long enough to have delivered significant value. In light of this, there remains the threat that such collaborations will not be able to deliver value in a time frame that will ensure their survival.

4.4 FUTURE TRENDS

There is no doubt that the drivers that stimulated precompetitive initiatives will continue to be a key feature of the industry. Whether precompetitive collaborations increase or whether some other forms of collaboration shown in Table 4.1, such as open innovation [21] or more open access models [22], come more to the fore will depend on several factors. One of the most important questions is where the boundaries of precompetitive activity will be drawn in the future. Given the definition proposed by Woodcock at the start of this chapter, precompetitive research has hitherto been focused mostly on tools and technologies, but if we draw the precompetitive boundary after target identification and validation, this could stimulate more collaborations at this critical stage of drug discovery and development. This in turn could allow the industry to more rapidly identify which targets are the most suitable for drug discovery and development efforts and hence improve efficiency and reduce unnecessary duplication and cost. Such an approach has been advocated by Barnes et al. [20], who argue that this could provide precompetitive opportunities in the biology and mechanisms of diseases. The picture is complicated because although initiatives like the IMI and SNP consortium are clearly seen as precompetitive by the large pharmaceutical companies, small biomedical companies with novel biomarkers or data-mining algorithms might set the precompetitive boundaries differently to protect their IP. A recent meeting sponsored by the Wellcome Trust on the precompetitive boundaries highlighted some of these issues and identified some areas for future focus (Precompetitive Boundaries and Open Innovation in Drug Discovery and Development meeting, June 2010).

There will also be an increasing need for intermediaries such as the CPI or the EBI. Such neutral third parties can reduce the bureaucracy of intercompany collaboration through the use of standard agreements [7, 20] and allows an interface with members of the scientific community. In addition, the number of other non-for-profit organizations such as Sage Bionetworks (www.sagebase.org) is likely to increase, and one of the challenges will be making sure that these repositories allow the cross talk mentioned above, through the common implementation of standards.

Some of the breadth and depth of the existing precompetitive space can be seen in the summaries of some of the major partnerships in the Appendix as well as the preceding descriptions of the IMI and SGC. Ten years ago such a wealth of precompetitive activity would have been unthinkable. This chapter began with the reasons for change in the industry and the recognition that the current process is not financially sustainable. The precompetitive agenda will continue to evolve and drive changes in the drug discovery and development process—the skill will be in fully leveraging the opportunities presented to really accelerate new medicines development. If this can be achieved, it will have benefits for both industry and society as a whole.

APPENDIX SUMMARY OF PRECOMPETITIVE CONSORTIA

The Biomarkers Consortium: www.biomarkersconsortium.org

Members/ participants	In 2010 there were 56 contributing members, including: Centers for Medicare & Medicaid Services, the National Institutes of Health (NIH), U.S. Food and Drug Administration (FDA), Althea Technologies, AstraZeneca, Avalon Pharmaceuticals, BG Medicine, Boehringer-Ingelheim, Bristol-Myers Squibb, Digilab Biovision GmbH, EMD Serono, Roche, Genstruct, GlaxoSmithKline (GSK), GVK BioSciences, Ingenuity Systems, J&J, Lilly, Luminex, Lundbeck, Merck, Novartis, Novo Nordisk, Pfizer, Rules-Based Medicine, Pfizer, and a multitude of nonprofit and trade organizations

Objectives and deliverables

To search for and validate new biomarkers to accelerate the competitive delivery of successful new technologies, medicines, and therapies for prevention, early detection, diagnosis, and treatment of disease. Efforts are focussed in the areas of oncology, inflammation and immunology, metabolic disorders, and neurosciences:

- Facilitate the discovery and development of biomarkers using new and existing technologies
- Validate their ability to diagnose disease, predict therapeutic response, and modify medical practice
- Identify patient groups who are better suited for a particular intervention
- Speed the approval of new therapeutic entities
- Make consortium project results broadly available

Description of consortium

The Biomarker Consortium is not a funding body; rather it encourages the submission of biomarker project concepts to the appropriate scientific steering committee, composed of experts in each therapeutic area. When approved, concepts are resubmitted as formal project proposals to the executive committee. This group directs the Biomarkers Consortium and is composed of the founding partners and other stakeholders. Approved projects receive the attention of consortium fundraising activities within the Foundation for NIH (FNIH).

Time frame

Started October 2006, no specified end date. One project completed in 2009—adiponectin as a marker of glycemic control

Current status

Projects currently funded through FNIH partnerships:

- Fluorodeoxyglucose positron emission tomography (FDG-PET) Lung and Lymphoma Projects interim analysis due in 2010
- Carotid magnetic resonance imaging (MRI) reproducibility study will complete in 2010
- Sarcopenia consensus summit to report in 2011
- Alzheimer's disease (AD) proteomics project with results end 2010
- Analysis of placebo data from AD and mild cognitive impairment
- I-SPY2 trial launched in breast cancer

Cardiac Safety Research Consortium (CSRC): www.cardiac-safety.org

Members

FDA, Health Canada, Duke University Clinical Research Institute, New York Presbyterian Research Hospital, GSK, Merck, Pfizer, Abbott, Lilly, Quintiles, Mortara Instrument, NeoCardio, Roche, Genentech

Objectives and deliverables

To advance scientific knowledge on cardiac safety for new and existing medical products the CSRC will:

- Facilitate focused, pragmatic nonproprietary research to inform regulatory processes
- Create common nomenclature and standards for cardiac safety evaluation
- Develop knowledge and improve the evaluative sciences related to cardiac safety and product development, specifically addressing the use and qualification of biomarkers for assessing cardiac safety
- Establish infrastructure and operational processes to allow effective sharing of knowledge and resources while preserving patient's rights and proprietary interest
- Develop white papers to address gaps in cardiovascular safety for drugs and devices
- Develop a forum for "think tanks" to address specific cardiovascular safety issues

Description of consortium

Investigators are invited to submit and develop research proposals in alignment with CSRC objectives. The scientific oversight committee then assesses and approves proposals, establishing teams to conduct the research, and monitors progress. One example is analysis of the FDA database of more than 2 million electrocardiograms (ECGs) from the clinical trial data submitted as part of new drug applications. The initial focus will be on proarrhythmic risks, with longer term development of evaluative tools, standards, validated tests, and cardiovascular biomarkers related to broader aspects of cardiac safety, including but not limited to arrhythmia, thrombosis, myocardial infarction, and heart failure.

Time frame

Started 2006, no specified end date

Current status

White papers in progress for: evaluation of ventricular arrhythmias in early-phase drug development; use of cardiac troponins in drug development; evaluation of QT for biologics; thorough blood pressure studies; developing compounds with cardiac safety signal; autonomic effects on the QT interval; pharmacokinetic/pharmacodynamic (PK/PD) analysis of QT effects; QT evaluation for oncology compounds; evaluation of non-QT ECG safety. Current think tanks: Atrial Fibrillation; CSRC-HESI; dual antiplatelet therapy duration

Dundee Kinase Consortium: www.lifesci.dundee.ac.uk/research/dstt

Members AstraZeneca, Boehringer Ingelheim, GSK, Merck KGaA/Serono, Pfizer fund research in the MRC Protein Phosphorylation Unit (eight groups), College of Life Sciences (four groups) and the Medical School (one group) of the University of Dundee

Objectives and deliverables The consortium provides access to world-leading expertise in functional analysis of proteins in cell-signaling pathways, to understand better the members of this family currently of interest for drug discovery, and to identify new drug targets. Reagent supply, assay definition, and compound profiling are also available to members.

Description of consortium Research is directed by Professors Cohen, Alessi, and Downes, with regular oversight meetings with industrial partners. The companies share access to all unpublished results, technology, know-how, and reagents and have first rights to licence the intellectual property generated. Compound profiling data are blinded and remain strictly confidential to each participating company.

Time frame Initiated in 1998 and funding secured until 2012

Current status The consortium is expanding to utilize better the pathway expansion capabilities of Dundee scientists and establish a lipid kinase assay profiling service. Currently, the Division of Signal Transduction Therapy (DSTT) makes 166 kinases and phosphatases, 200 antibodies and 4000 DNA constructs per annum and screens 80,000–100,000 data points per month. Reagent and assay provision remains a core component. The compound selectivity profiling service against a protein kinase assay panel is currently 85 and will be building to 104 by the end of 2010.

Observational Medicines Outcomes Partnership (OMOP): http://omop.fnih.org

Members FDA, FNIH, PhRMA, Abbott, Amgen, AstraZeneca, Bayer, Bristol-Myers Squibb, Eli Lilly, GSK, Roche, Johnson & Johnson, Lundbeck, Merck, Novartis, Pfizer, Sanofi-Aventis, Schering-Plough

Objectives and deliverables	OMOP is testing whether observational data can be used to improve understanding of drug safety and benefit outcomes. The three main objectives are: • Conduct a series of experiments to assess the value, feasibility, and utility of observational data (claims and electronic health record data) to identify and evaluate the safety risks and potential benefits of prescription drugs using a range of analytical methods and multiple observational data sources, beyond the currently available tools and data sources • Test approaches for creating the infrastructure for accessing and managing the required data, including multiple claims and electronic health records data sources • Establish and evaluate a suitable governance structure for a public–private partnership for these tasks and to inform future efforts to monitor drug safety and benefit outcomes systematically
Description of consortium	The core research team has designed and developed tools and technologies for assessing data and databases. Collaborators will be selected via an open, competitive application and award process managed by the principle investigators (PIs) and the OMOP executive director. The research protocols, data models, database evaluation and quality assurance tools, analytical programs, and findings generated by OMOP are being published and made available, allowing other researchers the opportunity to run the protocols on their own data and develop parallel or complementary tools and approaches. OMOP aims to encourage sharing of results and tools developed in this way for the public benefit; external institutions whose aims are in alignment with those of OMOP, who can credibly perform relevant research, and who are willing to share tools, approaches, findings, and other intellectual property developed as a result may be additionally recognized as members of the OMOP Extended Research Consortium as announced in April 2010.
Time frame	2008–2010
Current status	Phase 1: completed in 2009 • Established a consistent framework to use across disparate observational data sources and an OMOP research community

Phase 2: completed in 2009

- Developed and tested analysis methods within the OMOP research lab and other data environments
- Established standard data characterization procedures
- Implemented health outcomes of interest definitions and facilitated comparisons across databases

Phases 3 and 4 will complete in 2010

Predictive Safety Testing Consortium (PSTC): http://www.c-path.org/pstc.cfm

Members
Abbott, Amgen, AstraZeneca, Boehringer Ingelheim, Bristol-Myers Squibb, ClinXus, Daiichi Sankyo, Lilly, GSK, Johnson & Johnson, Merck, Mitsubishi, Novartis, Pfizer, Roche, Sanofi-Aventis, Schering Plough, Critical Path Institute

Objectives and deliverables

- To identify and cross qualify new and improved preclinical safety testing methods through a collaboration of scientists from the pharmaceutical industry, FDA, European Medicines Agency (EMEA), and academia. This involves validating predictive, preclinical animal model biomarkers aimed at reducing the cost and time of preclinical safety studies and providing potential early indicators of clinical safety in drug development and postmarketing surveillance.
- To facilitate the development of new FDA processes for approving such testing methods and applying these processes for approvals and guidances.

Description of consortium
Led by the Critical Path Institute, the PSTC brings together pharmaceutical companies to share and validate safety testing methods. The corporate members of the consortium share internally developed preclinical safety biomarkers for examination and cross-validation in five work groups: carcinogenicity, kidney, liver, muscle, and vascular injury. This should enable the FDA and EMEA to write new guidances that identify more accurate methods to predict drug safety. Notably, the FDA and EMEA scientists are not acting as regulators but provide assistance and advice to the consortium.

Time frame
Commenced 2006, no end date

Current status
In 2008 the FDA and EMEA published jointly seven new kidney biomarkers. Working groups running for renal toxicity, myopathy, hepatotoxicity, vascular injury, and carcinogenicity.

Serious Adverse Events Consortium (SAEC): www.saeconsortium.org

Members	Abbott, Daiichi Sankyo, GSK, J&J, Novartis, Pfizer, Roche, Sanofi-Aventis, Takeda, Wellcome Trust, Wyeth, Cerner, Columbia University, Dundee University, DILIGEN, EUDRAGENE, Expression Analysis, Malaga University, Illumina, U.S. VA Center for Drug Safety
Objectives and deliverables	The SAEC aims to identify genes that can influence adverse drug reactions. This will be achieved by comparing genetic sequences of individuals with well-characterized adverse events with control individuals to identify genetic variants that may be risk factors for these events. During the pilot phase, the focus has been on serious skin reactions and drug-induced liver injury.
Description of consortium	Sponsors of the SAEC contribute genetic and phenotype data from individuals who have had an adverse reaction to a drug together with data from controls. These are assigned to scientific collaborators for analysis to determine if the adverse events can be correlated with genetic variation. After full analysis the results are made available to all qualified researchers on a nondiscriminatory basis to maximize the speed of follow-on discoveries and published as appropriate. The SAEC is governed by a board of directors which has management control over the property, activities, and funds of the corporation. It is made up of a director from each sponsoring member and a chief executive officer. The board structure allows for involvement of other nonprofit research and governmental organizations via "associate membership" and it makes decisions using a "majority rules" model. All key research collaborations are formed via a rigorous "RFP process," with final selection determined by the scientific management committee. The committee outlines research strategies, well-defined research programs, and advice on study design and methodology and data release policies
Time frame	2007–2012

Top Institute Pharma Consortium: www.tipharma.com

Members Agendia, Nycomed, Astra Zeneca, Bio Detection Systems, Centocor/Johnson & Johnson, DNAge, Eli Lilly, GSK, Hal Allergy, IQCorporation ISA, Lundbeck, Merck, Netherlands Vaccine Institute, Nobilon, NOTOX, Novartis, Numico, Octoplus, Organon, Pamgene, Pepscan systems, Pfizer, PRA International, Prosensa, PROXY laboratories, Pyxis Discovery, Solvay Pharmaceuticals, Winap, Xendo; plus 25 universities and medical centres in Holland

Objectives and deliverables TI Pharma is focused on the development of tools required to shorten drug development timelines and a reduction of the major risks of clinical failure of potential new medicines. TI Pharma aims to attain a position as a leading pharmaceutical research and training institute in Europe and hence a coordinating partner in European research networks.

Description of consortium TI Pharma is a collaborative research structure consisting of pharma and academic groups. Individual projects are managed centrally from a funding perspective; partners also contribute in kind (with data and time). Ph.D. students and postdoctoral fellows are responsible for the daily research activities. Research is conducted at the institutes brought together by TI Pharma, focused on six main themes: autoimmune diseases, cardiovascular diseases, cancer, infections/vaccines, brain diseases, efficiency analysis of the process of drug discovery and development. There also seven technological disciplines: therapeutic target finding; validation and animal Models; lead selection and in silico and PK/PD modeling; predictive drug disposition and toxicology; biomarkers and biosensors; drug formulation, delivery, and targeting; pharmaceutical production technologies; molecular informatics.

Time frame TI Pharma started mid-2006

Current status Forthy-eight research projects have been initiated by end 2009 with 340 postdoctoral researchers and students. One hundred twenty-six papers were accepted for publication in 2009 and five patents filed with seven pending. In 2009 the consortium underwent a scientific review that recommended continuation.

Other Public–Private Consortia

The Coalition Against Major Diseases (CAMD), www.c-path.org/CAMD.cfm

Members include the Critical Path Institute, the Engelberg Center for Health Care Reform at the Brookings Institution, 6 nonprofit groups representing patients' interests, 15 leading pharmaceutical companies, the US Food and Drug Administration (FDA), the European Medicines Agency (EMEA), 2 institutes of the National Institutes of Health (NIH)—the National Institute on Aging (NIA) and the National Institute of Neurological Disorders and Stroke (NINDS)—and representatives from academia. The coalition's purpose is to transform the drug development paradigm for neurodegenerative diseases and serve as a model for other major diseases

Center for Translational Molecular Medicine (CTMM), www.ctmm.nl

CTMM is a public–private consortium of 105 partners including universities, academic medical centers, medical technology enterprises, and chemical and pharmaceutical companies. It is dedicated to the development of medical technologies that enable the design of new and "personalized" treatments for the main causes of mortality and diminished quality of life (cancer and cardiovascular diseases and neurodegenerative and infectious/autoimmune diseases) and the rapid translation of these treatments to the patient.

HRP Initiative, www.hrpinitiative.com

The HRP Initiative is a precompetitive industry collaboration that is focused on discovering and developing novel blood tests and imaging methods to find individuals with high-risk plaque disease before the occurrence of the first cardiovascular event. AZ, Merck, Abbott, Takeda, Phillips, and BGMedicine are industrial sponsors.

Quebec Pain Research Network, www.qprn.ca

Founded in 2001, now has 50 regular and 26 associate members. It is a multidisciplinary research program to tackle the challenge of pain in its multiple dimensions, ranging from increasing our understanding of the basic mechanisms of pain transmission to improving the assessment and treatment of pain in humans.

Québec Consortium for Drug Discovery (CQDM), www.cqdm.org	CQDM is a nonprofit organization whose mission is to identify, fund, and support research projects carried out in partnership between the academic and hospital milieus in the public sector and the biotechnology and contract research organizations in the private sector. Research projects funded by CQDM aim at developing tools or enabling technologies that facilitate and accelerate the drug discovery process. AZ, Merck, and Pfizer are the pharma partners.
Stem Cells for Safer Medicines, www.sc4sm.org	Its aim is to enable the creation of a bank of stem cells, open protocols, and standardized systems in stem cell technology to enable consistent differentiation of stem cells into stable homogenous populations of particular cell types, with physiologically relevant phenotypes suitable for toxicology testing in high-throughput platforms.

REFERENCES

1. Weber S. *The Success of Open Source*. Cambridge, MA: Harvard University Press, 2004.
2. Woodcock J. Precompetitive research: A new prescription for drug development? *Clin Pharmacol ther* 2010;87(5):521–523.
3. Rawlins M. Cutting the cost of drug development? *Nature Drug Discov* 2004. 3:361–364.
4. DiMasi J, Hansen RW, Grabowski HG. The price of innovation: New estimates of drug development costs. *J. Health Econ* 2003;22:151–185.
5. Wagner JA. Open minded to open innovation and precompetitive collaboration. *Clin Pharmacol Thera* 2010;87(5):511–515.
6. Wagner JA, et al. The Biomarkers Consortium: Practice and pitfalls of open-source precompetitive collaboration. *Clin Pharmacol Ther* 2010;87(5):539–542.
7. Woosley R, Myers RT, Goodsaid F. The Critical Path Institute's approach to sharing and advancing regulatory science. *Clin Pharmacol Ther* 2010;87(5):530–533.
8. Gamo F-J, et al. Thousands of chemical starting points for antimicrobial lead identification. *Nature* 2010;465:305–310.
9. Guiguemde WA, et al. Chemical genetics of *Plasmodium falciparium*. *Nature* 2010;465:311–315.
10. Strauss S. Pharma embraces open source models. *Nature Biotechnol* 2010;28: 631–634.
11. Joint technology initiatives: Background, state-of-play and main features. Commission Staff Working Document SEC(2007) 692, Brussels; CEC, 2007.
12. IMI fact sheet. Available: ftp://ftp.cordis.europa.eu/pub/fp7/docs/factsheet2_imi_en.pdf.

13. Radiprodil fails Ph II study for diabetic neuropathic pain. *Scrip World Pharmaceutical News* July 9, 2010, p. 15.
14. Ledford H. Protein mapping gains a human focus. *Nature* 2010;466:544.
15. Edwards AM. Open source to enable drug discovery. *Drug Discov Today* 2008; 13:731.
16. Pardoe DA, Barrett P, Hunter AJ, Cooke RM. Assessing the value of R&D research partnerships. *Drug Discovery World* 2010; 11:9–17.
17. Vargas G, Boutouyrie B, Ostrowitzki S, Santarelli L. Arguements against precompetitive collaboration. *Clin Pharmacol Ther* 2010;87(5):527–529.
18. Ekins S, Williams EJ. When pharmaceutical companies publish large datasets: An abundance of riches or fool's gold. *Drug Discov Today*, Doi.101016/j.drudis. 2011.02.016.
19. Munos B. Can open source drug R&D repower pharmaceutical innovation. *Clin Pharmacol Ther* 2010;87(5):534–536.
20. Barnes MR, et al. Lowering industry firewalls: Precompetitive informatics initiatives in drug discovery. *Nature Rev Drug Discov* 2009;8:701–708.
21. Melese T, Lin SM, Chang J, Cohen NH. Open innovation networks between academia and industry: An imperative for breakthrough therapies. *Nature Med* 2009;15:502–507.
22. Edwards AM, Bountra C, Kerr DJ, Wilson TM. Open access chemical and clinical probes to support drug discovery. *Nature Chem Biol* 2009;5:436–440.

5

COLLABORATIONS IN CHEMISTRY

Sean Ekins, Antony J. Williams, and Christina K. Pikas

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.

5.1 INTRODUCTION

Sometimes it seems like the pharmaceutical industry has been a "dedicated follower of fashion" on its quest for reinvention. In a decade we have sequentially gone through major trends of combinatorial chemistry, high-throughput screening, genomics, systems biology, and biomarkers. Now it seems that outsourcing, virtual drug discovery, and collaborations are the new trends in parallel. The aim of this chapter is to describe some of the technologies available for collaborations in chemistry and examine some of the initiatives that are already underway. But first we will provide some historical context, examples of how such collaborative tools are used in big science, and our thoughts on communication trends.

5.1.1 Historical Context

During the late 1990s, a hot topic was how small biotechnology companies could partner successfully with pharmaceutical companies, as both were seen as having divergent cultures and needs [1]. Both needed each other: The small companies needed money and their new technologies and molecules were needed by the industry to boost pipelines. Now the need for alliances and collaborations within and between companies, universities, and other organizations could be imagined as complex networks that lead to knowledge and innovation [2]. Of course you can look outside the industry for inspiration. In the same decade we have seen an increase in cooperation in software development [think Linux operating system and a myriad of open-source development projects, the majority hosted on Sourceforge (http://sourceforge.net/)], and we are also seeing nonprofits and virtual drug companies come into their own to fund and drive neglected and orphan disease research.

5.1.2 Collaborations in Big Science

Big projects require big funding and undoubtedly drug development is costly and could be tackled in the same way *big science* projects are; genome sequencing, particle physics, and the international space station are just some examples. The National Academy of Public Administration in the United States recognized that (1) optimal use of collaborative technology is for technical transfer, support, and education; (2) such technology allows teams to form around data sets for more sophisticated analyses; (3) Web-based social software allows the public to participate in peer review; (4) much of the data are unstructured; and (5) such technology could unleash innovation and process improvements [3]. Examples from within the U.S. government include the Department of Energy using a blog, the Environmental Protection Agency using a wiki, the U.S. patent and trademark office allowing the public to peer review patents, and a government thrust to release terabytes of previously guarded data into the public domain [3]. The term "public domain" in this context is problematic: In the United States, unlike European countries, government data are automatically in the public domain. Data.gov just makes it (a) easier to find and (b) more usable. On the academic side, the increased competition for funding has already seen some great success in Europe with a framework funding mechanism that supports collaborative research.

5.1.3 Current and Future Trends in Communication

So where does this put collaborative chemistry? We are seeing an increasing use of crowdsourcing, databases, wikis, and networking tools (Table 5.1), and these in many ways are supplementing or changing the traditional ways that research is done, shared, and communicated. Tapping into the global

TABLE 5.1 Examples of Crowdsourcing, Databases and Networking Resources

Name	Website	Function
myExperiment	http://www.myexperiment.org/	Workflows, communities
DIYbio	http://diybio.org/	Community for do-it-yourself biologists
Protocol online	http://protocol-online.org/	Biology protocols
Open wetware	http://openwetware.org/wiki/Main_Page	Materials, protocols, and resources
Open-notebook science challenge	http://onschallenge.wikispaces.com/	Crowdsourced science challenge, initially on solubility measurement
UsefulChem project	http://usefulchem.wikispaces.com/	Example of one scientist's open notebook
Laboratree	http://laboratree.org/pages/home	Science networking site
Science Commons	http://sciencecommons.org/	Strategies and tools too faster, efficient Web-enabled scientific research
WikiPathways	http://www.wikipathways.org/index.php/WikiPathways	Curated biological pathways
Open Source Drug Discovery	http://www.osdd.net/home	Collaboration around genomics and computational technologies
Wikipedia	http://www.wikipedia.org/	Crowdsourced encyclopedia of knowledge, includes chemical pages (http://en.wikipedia.org/wiki/Wikipedia%3ACHEMISTRY)
BioSpace	http://www.biospace.com/	News and jobs in pharmaceutical and biotech industry

(*Continued*)

TABLE 5.1 (***Continued***)

Name	Website	Function
BioPortfolio	http://www.bioportfolio.com/	Information aggregator of news and biological and chemical entities
ACS Member Network	http://portal.acs.org/portal/acs/corg/networkLanding?_nfpb=true&_pageLabel=PP_MNLANDING&node_id=2127&use_sec=false&__uuid=f1b95b5e-fc29-450b-8b1e-695b8cf2e321	Networking website
LabMeeting	http://www.labmeeting.com/	Website for organizing and sharing papers
Nature Network	http://network.nature.com/	Networking website
Innocentive	http://www.innocentive.com/	Uses crowdsoursing to solve science challenges
PD^2	https://pd2.lilly.com/pd2Web/	Crowdsourcing site to bring molecules to Lilly for phenotypic testing
ChemSpider	www.chemspider.com	Chemisty aggregator and database
Pubchem	http://pubchem.ncbi.nlm.nih.gov/	Molecule structure and bioassay data
DrugBank	http://www.drugbank.ca/	Detailed drug and target information
eMolecules	http://www.emolecules.com/	Chemistry structure search engine and supplier information
CDD	www.collaborativedrug.com	Collaborative database enabling secure selective sharing
ZINC	http://zinc.docking.org/	Over 13 million commercially available molecules for virtual screening

community of chemistry research expands resources beyond a single laboratory. The ability to do this in real time via blogs [4, 5] or wikis, commonly used to host open notebooks [6] (see Chapter 25), rather than having to wait for a journal article to publish findings, obviously could greatly increase the speed (by months) with which chemistry syntheses could be made available, thereby preventing unnecessary and costly repetition in other laboratories. The work can be hosted not only in these common environments but via exemplar efforts such as the crowdsourced ChemSpider SyntheticPages (http://cssp.chemspider.com/) where the community can publish and share reaction syntheses data on the same day that they do the work.

The complete process of scientific research (including chemistry) can be supported by a myriad of tools and technologies online. As yet, compounds cannot be synthesized automatically by a simple request online but the command can be sent to a far-off laboratory and the results shipped to you. Your research can easily be posted online for others to see and comment on using a multitude of Web 2.0 platforms (see Table 5.1). At the time of writing there are generally no associated costs with sharing the data other than the time and effort associated with posting the information (data upload, checking, and publishing). Although only a tiny minority of chemists do this today, in the future this is likely to grow as a Web presence for a scientist is measured not only by peer-reviewed publications online but rather by the contributions of a scientist to the scientific commons. We predict that with this change will come dramatic improvements in scientific communication and, one can envisage, improved data validation, engaging feedback between scientists and the initiation of new collaborations between previously unconnected scientists. What is preventing this today? One roadblock is the delay in uptake of new communication technologies in the chemistry field which has been termed "latency" [7]. This naturally leads to more questions around what can be done to foster communication (and embrace new technologies) between chemists and other groups in pharma, biotech, and academia. What is needed for the collaborators to successfully accomplish their goals?

5.2 CROWDSOURCING

Crowdsourcing uses the wisdom of the many (the "crowd") and their varied perspectives to benefit community-based efforts. A loose definition of crowdsourcing is "outsourcing a task to a group or community of people in an open call"—a phenomenon, culture, or movement best summarized in the book *Wikinomics, How Mass Collaboration Changes Everything* [8]. Crowdsourcing approaches have contributed enormous societal benefit as well as created new businesses. One of the greatest success stories of crowdsourced approaches is the phenomenon known as Wikipedia. The open-source operating system, Linux, is the result of the programming efforts of thousands of people around the world contributing to a free code base. The business value to organizations that have adopted Linux is enormous and IBM alone has estimated savings in the hundreds of millions of dollars by adopting the Linux platform. Wikipedia is a global success story and, one would assume, has coerced the hearts and minds of the masses to assist in the creation of the world's foremost encyclopedia, a free resource where even the data can be reused and repurposed under appropriate licenses. The reality is a little different, however: While many thousands of contributors have helped to shape the multilingual articles, the number is a tiny fraction of the number of people who access Wikipedia. The reality of crowdsourcing is that there are only a small number of contributors relative to the number of consumers. In general, studies have shown that the 90-10-1 rule holds (http://www.90-9-1.com/); only 10% of the visitors will edit or comment and only 1% of the visitors will contribute new content. The same is likely true of collaborative chemistry resources on the Internet.

5.2.1 Crowdsourcing Platforms

A number of platforms enabling collaborative contributions to chemistry are already available. Certainly Wikipedia is one of these platforms and many thousands of encyclopedic articles regarding chemical compounds, materials, and synthetic reactions have been compiled by contributors. There is concentrated effort—a Wiki Project—to curate and correct chemistry information on Wikipedia (http://en.wikipedia.org/wiki/Wikipedia%3ACHEMISTRY). In addition to chemical information, there are contributions from biologists and medical experts to add details about diseases, genes, and proteins, and this has already created an incredibly rich resource of information.

Other examples related to chemistry include Innocentive (www.innocentive.com), a website which posts challenges which can be anything from proposing syntheses of specific materials to projects to identify molecules binding to a particular receptor. The successful individual or team selected is rewarded financially and also with publicity if desired. The Nature Publishing Group also recently teamed up with Innocentive to form the Open Innovation Pavilion such that nature.com readers could be directed to Innocentive challenges (http://www.nature.com/openinnovation/index.html). Another initiative from Eli Lilly is phenotypic drug discovery (PD^2, https://pd2.lilly.com/pd2Web/), whereby scientists can submit their molecular structures via a secure portal where they are evaluated for novelty and drug likeness. If a molecule is selected, it is screened in phenotypic assays for diabetes, cancer, Alzheimers', and osteoporosis. The goal of such an approach is to bring compounds from academia and companies that might never have the potential to be tested against these diseases. Obviously, Eli Lilly is then in a position to license compounds it finds that are active. These types of e-science initiatives are remarkable in bringing solutions to the companies. While Innocentive is very specific to well-defined challenges (in general), PD^2 is sampling academic or biotech company compound space in a less well defined manner. Initial filtering is performed computationally and followed by various whole-cell biology and secondary assays to perform further filtering. The advantage of both such approaches are that they do not need to employ the scientific participants full time or pay for the time and effort that went into the initial synthesis of compounds tested. In both cases, significant parts of the research and development (R&D) process are essentially outsourced and the companies involved do not have to pay for that (Eli Lilly in the case of PD^2). Such an approach should lead to a reduction in the costs of R&D. It will be of interest to see whether other companies will adopt similar initiatives to PD^2 because currently one company has a monopoly on this. The PD^2 approach could potentially be applied to other parts of the R&D pipeline even as a means to pull in later stage candidates for licensing. Lilly have indicated that it will develop a similar approach for drug targets.

A final example, while not specific to chemistry, can reveal the value of collaboration. It involves the Alzheimer's Disease Neuroimaging Initiative (ADNI, http://www.adni-info.org/), which started with government organizations, academic, nonprofit, and industry members that formed a public–private partnership to find biomarkers for the disease from clinical studies. The National Institutes Health (NIH) serves as the coordinator between all the organizations and the data are shared and open to all for analysis. To date it appears that several hundred publications have been generated from the data and many more studies are underway. The collaborative nature in this case is helping drive the field of Alzheimer's biomarkers forward.

5.3 COLLABORATORIES

The National Academies recently issued a report entitled "A New Biology for the 21st Century" that describes science as more complex and more global in nature than ever before and focused specifically on the biological sciences [9]. Scientific software and hardware are expensive and, for many, sharing of such technologies is essential for their progress. Software in particular can be readily distributed within or between institutes or collaborators via collaboratories. Collaboratories are a Web-based infrastructure for collaboration that allow the sharing of computational tools and data and enable distributed research by providing access to resources for research using the off-the-shelf tools. Examples include BIRN (http://www.birncommunity.org/) for biomedical imaging and genetics [10] and BioCORE (http://www.ks.uiuc.edu/Research/biocore/) for bioinformatics and computational chemistry [11]. The national e-Science Centre at the University of Glasgow (http://www.nesc.ac.uk/hub/) supports research collaborations that are interdisciplinary and include bioinformatics, clinical trials using high-performance computing, and grid-based technologies [12–14]. The academy for medical development and collaboration developed shared facilities for genomics. Other scientific efforts could also be seen, for example, collaboratories such as the open-source drug discovery network (http://www.osdd.net/) described in Chapter 20. Collaboratories could become a critical component of future drug discovery efforts, particularly those in countries with limited scientific resources or those between academic groups.

5.4 DATABASES

One way to share scientific data is via public domain databases available on the Web (Table 5.1). Astronomy, physics, biology, and chemistry have all contributed enormous amounts of data to the public domain and, when available, have been the basis of platforms to allow crowdsourced analysis, validation, and annotation of the data. Examples from the world of astronomy are GalaxyZoo (http://www.galaxyzoo.org/) and MoonZoo (http://www.moonzoo.org/) while in chemistry the ChemSpider database, coincidentally established by ChemZoo (http://www.chemspider.com) (see Chapter 22), is the preeminent example. In regards to chemistry the past five years has seen an explosion in the availability of databases hosting chemical compound collections and generally accessible via a cheminformatics platform allowing searching by molecular structure. As a result of these efforts, chemistry information on the Internet is increasingly becoming much more widely accessible, with numerous chemical compound databases on the Web providing free access to molecular structures and related data [15, 16]. However, there are multiple issues: As previously described [17], these databases generally contain the chemical identifiers in the form of chemical names (systematic and trade) and registry numbers and, due to their assembly in a heterogeneous manner, the data can be plagued with quality issues and these can impact downstream uses such as computational modeling. We are aware of many databases that curate all manner of information that might be of relevance to chemists involved in biomedical research, from chemical vendor catalogs, to patents, to spectra of various kinds. A recent article describes the public and commercial databases of bioactive compounds [18] and concludes that the commercial efforts are ahead of the public ones at this point in time, yet both are complementary.

5.4.1 PubChem

PubChem, a molecule database, launched in 2004 to support the "New Pathways to Discovery" component of the Roadmap for Medical Research [19] (http://pubchem.ncbi.nlm.nih.gov/), is probably the most widely known and yet it covers only a small fraction of the chemical universe. At present PubChem is the informatics backbone for the Molecular Libraries and Imaging Initiative, which is part of the NIH Roadmap [19]. PubChem presently contains almost 31 million unique structures with biological property information provided for a fraction of the compounds. Although it is authoritative and built on an excellent informatics platform with a well-resourced infrastructure, there are a number of constraints and issues with PubChem. Specifically, it is a repository of data and information and does not make any special effort toward curating the data depending instead on the whims of the depositors to ensure the quality and validity of the data. As a result any errors in the data deposited into PubChem may be, and already have been, transferred into other online databases that treat PubChem as an authority. This in turn can impact the research of others using computational models. The issues are not limited only to the validity of the chemical structures but, more generally, to the structure–identifier relationships and resulting dictionaries that have been derived from the data. As a simple example of structure–identifier errors, examination of the list of identifiers associated with the simplest organic molecule in PubChem, methane, lists associated identifiers including carbon, diamond, soot, fullerene, and many other more complex organic molecules. There are multiple hits for well-defined compounds such as vancomycin and taxol and other much more simple organic molecules. While useful, PubChem is, quite simply, inappropriate to treat as an authority.

5.4.2 Other Molecule Databases

Another interesting database of relevance to biomedical researchers is the Chemical Entities of Biological Interest, or ChEBI database (http://www.ebi.ac.uk/chebi/). The data are curated on an ongoing basis and as of this writing ChEBI release 69 is available, with 584,456 total entities, of which 21,369 are fully annotated and curated. ChEBI includes an ontology which identifies the relationships between molecular entities or classes of entities and their "parents" and/or "children". Another database of primary interest to biomedical researchers is DrugBank (http://www.drugbank.ca/), a manually curated database [20] linking out to other public domain databases [e.g., the Kyoto Encyclopedia of Genes and Genomes, KEGG, (http://www.genome.jp/kegg/) [21], PubChem, ChEBI, the protein databank (PDB, http://www.pdb.org/pdb/home/home.do), Swiss-Prot (http://www.expasy.ch/sprot/), and GenBank (http://www.ncbi.nlm.nih.gov/genbank/)] and additional data from the laboratories of the hosts. The database aggregates both bioinformatics, cheminformatics data, detailed drug data, and comprehensive drug target information and contains U.S. Food and Drug Administration (FDA)–approved small-molecule and biotech drugs representing nearly 5000 molecules [22]. For those interested in vendor libraries for use for docking studies, ZINC (http://zinc.docking.org/index.shtml) represents a free, searchable database of over 20 million molecules commercially available compounds for virtual screening available as three-dimensional structures [23, 24]. Importantly all the molecules are assigned biologically relevant protonation states and are annotated with other molecular properties which may be of interest for hit filtering.

5.4.3 ChemSpider

Another example of a freely available database (also described in Chapter 22) is ChemSpider (http://www.chemspider.com/) [15, 16], a community resource for chemists provided by the Royal Society of Chemistry. It currently contains almost 25 million unique chemical entities aggregated from almost 400 diverse data sources, including government databases, chemical vendors, commercial database vendors, publishers, as well as all of the prior described databases and individual chemists. The unique capabilities of ChemSpider relative to other public chemistry databases include the real-time curation of the data by the community and annotation of the data. It is also possible to deposit New data to the database: New chemical compounds can be deposited to the database as singletons or as large collections with up to half a million compounds already having been deposited in a single day; analytical data and activity data can also be deposited against existing chemical compounds.

As ChemSpider has grown in popularity and scope, numerous websites have started to link to the databases. Wikipedia commonly includes links from its chemical compound articles to ChemSpider, ZINC has included links in its databases, Nature Publishing Group links from its chemical compound pages from both its *Nature Chemistry* and *Nature Chemical Biology* journals, as does the Royal Society of Chemistry for its *prospected* articles (http://www.rsc.org/publishing/journals/projectprospect/faq.asp). ChemSpider also provides access to a series of web services to allow querying of the data. For example, four of the primary analytical instrumentation vendors (Bruker, Waters, Thermo, and Agilent) have established or are presently pursuing integration of the ChemSpider data into their mass spectrometry data processing software packages. The Web services are also used by other public and private databases in either academia or industry. For example, Collaborative Drug Discovery (CDD, www.collaborativedrug.com, also described in Chapter 21), provides links to ChemSpider for molecules in their CDD database [25]. CDD itself is a highly secure, commercial collaborative drug discovery informatics platform with both a Vault for proprietary data, technologies to enable selective sharing, and over 50 publically accessible data sets available upon registration.

Recently pharmaceutical companies such as GlaxoSmithKline and Novartis have shared some sets of active compounds for malaria and made them available in CDD, ChEBI, and PubChem. This deposition of large numbers of compounds raises some issues relating to how the data will be used and accessibility of the compounds for follow-up evaluation [26].

5.5 BLOGS

The tools described thus far facilitate collaborations in which the partners work in a shared space online or pull information from a community online. Another mode of collaboration involves working asynchronously and then sharing the product of your work. Blogs and other types of social software tools like image-sharing sites and presentation-sharing sites support this type of collaboration. Blogs are defined by their format because the content varies greatly. They are typically lists of individual posts, each with a permanent link, a place for comments, can be tagged with a topic, and are posted in reverse chronological order. Collaboration software which many companies already have, such as Microsoft SharePoint Server (http://sharepoint.microsoft.com/en-us/Pages/default.aspx) or Atlassian Confluence (http://www.atlassian.com/software/confluence/), support blogging to some extent, but specialized blogging software products such as MovableType (http://movabletype.com/) and Wordpress (http://wordpress.org/) have more features and are more common on the Web.

5.5.1 Why Blog?

Scientists typically do not blog on public sites about their work in progress or work that has not been published. Instead, they blog about the process of science, interesting articles they have read, tutorials on things they have figured out or the basics for a broader audience, and ideas and information they have that is not enough for a separate publication.

The process of preparing and writing a blog post as well as receiving comments and feedback helps the scientists learn the topic better and clarify their thinking on it. Blogging tutorials on the basics of a topic or on how to do something in the laboratory can serve as a distributed apprenticeship [27]. Some scientists find that they have more content than they can fit in the page limits of a publication and so use their blogs to provide additional information. Similarly, if they have an idea that they do not have time to pursue or that is not enough for a separate publication, the blog is a way to explore and share. Blogging makes the information posted findable by the author later, provides a time and date stamp, and also makes the information findable by other scientists when they need it.

Another advantage of blogging is the speed in that someone can write something which is published without peer review and their opinion is instantly out in the open and joins the scientific discourse. Of course, the anonymity also allows people to vent their feelings in a manner impossible in scientific papers, letters to the editor, and so on. A great source of chemistry blogs as well as examples of "scientific venting" is Derek Lowe's in the pipeline (http://pipeline.corante.com/).

5.5.2 To Blog Publicly or on the Intranet?

It may be tempting to blog on an intranet or to only allow collaboration partners to read the posts. Unfortunately, it is difficult to get a critical mass of readers to get the full value with an intranet blog. Intranet blogs generally do not get as many comments and so may not be as rewarding to the blogger. On the other hand, the blogger might feel more free to share details of his or her current work and the blog will still be useful for refinding information. Also there may be a real sense of fear blogging internally, in case management disagrees with your opinion and effectively sanitizes it.

5.5.3 Using Computational Methods to Make Use of Information and Knowledge in Blogs

A nascent use of blogs is to gather opinions on scholarly articles. Publishers are beginning to search, mine, and aggregate blog posts that discuss articles. Public opinion researchers also mine blog posts and use sentiment analysis to understand how their product is viewed. This could be extended to chemistry blogs to crowdsource new ideas and approaches.

5.6 WHERE WILL COLLABORATIVE TECHNOLOGIES TAKE CHEMISTRY?

Improving efficiency is necessary since squeezed budgets and costcutting in both academia and industry are impacting pharma R&D. Collaborative software for chemistry can help in several ways. First, sharing data or molecular structures either openly or securely can enable a chemist in one location, say in a drug company in the United States or Europe, to suggest a synthesis route for a molecule being made in China, India, or Russia. Drug companies can tap into the global chemistry community to work on problems that may be beyond the capabilities of their own staff via collaboration networks such as Innocentive. Alternatively, they can attract molecules or technologies to them for potential testing via initiatives like PD^2. Free Web-based databases of molecules and their properties, patents, and reactions can reduce the dependency on commercial databases as well as provide links to many other initiatives that such closed commercial systems cannot. Collaborative technologies have the capability to facilitate virtual laboratories that can be located in multiple locations while at the same time potentially bringing together different disciplines like biologists and chemists. This may be ideal for neglected diseases or rare orphan diseases in which there is either limited funding, scientific capability, or scientific interest to have critical mass.

While there are significant precompetitive informatics efforts in the pharmaceutical industry [28] that cover both chemistry and biology, to date there has not been a focus on creating standards or even setting requirements for collaborative technologies that could be used between or within companies and between companies and researchers or external contractors. There is therefore significant opportunity for collaborative technologies to expand and impact chemistry in general from teaching [29–31] through to research and development, and it is equally likely that success may come outside of biomedical fields.

REFERENCES

1. Sapienza AM, Stork D. *Leading Biotechnology Alliances Right from the Start*. Hoboken, NJ.: Wiley, 2001.
2. Oliver AL. *Networks for Learning and Knowledge Creation in Biotechnology*. Cambridge: Cambridge University Press, 2009.
3. Clark MA, et al. More than just a slick website: The use of collaborative technology to solve organizational challenges in federal agencies. USDA Graduate School executive potential program Team #1, 2009. Available: http://www.collaborationproject.org/wp-content/uploads/2011/03/More_Than_Just_A_Slick_Website.pdf.
4. Pikas CK. How and why physicists and chemists use blogs. Nat Precedings posted 8 May 2010. Available: http://precedings.nature.com/documents/4429/version/1.

5. Rowe A. The new water cooler. *Chemic Heritage* 2009;Fall:40–41.
6. Bradley J-C, Guha R, Lang A, Lindenbaum P, Neylon C, Williams A. Beautifying data in the real world. In Segaran T, Hammerbacher J, Editors. *Beautiful Data*. Sebastopol: O'Reilly Media, 2009.
7. Velden T, Lagoze C. Communicating chemistry. *Nat Chem* 2009;1:673–678.
8. Tapscott D, Williams AJ. *Wikinomics: How Mass Collaboration Changes Everything*. New York: Portfolio, 2006.
9. Academies TN. *A New Biology for the 21st Century*. Washington, DC: National Research Council of the National Academies, 2009.
10. Keator DB, Grethe JS, Marcus D, Ozyurt B, Gadde S, Murphy S. A national human neuroimaging collaboratory enabled by the Biomedical Informatics Research Network (BIRN). *IEEE Trans Inf Technol Biomed* 2008;12:162–172.
11. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E. Scalable molecular dynamics with NAMD. *J Comput Chem* 2005;26:1781–1802.
12. Sinnott RO, Stell AJ, Ajayi O. Supporting grid-based clinical trials in Scotland. *Health Inform J* 2008;14:79–93.
13. Sinnott R, Bayliss C, Jiang J. Security-oriented data grids for microarray expression profiles. *Studies Health Technol and inform* 2007;126:67–76.
14. Sinnott R, Stell A, Ajayi O. Development of grid frameworks for clinical trials and epidemiological studies. *Studies Health Technol and Inform* 2006;120:117–130.
15. Williams AJ. A perspective of publicly accessible/open-access chemistry databases. *Drug Discov Today* 2008;13:495–501.
16. Williams AJ. Internet-based tools for communication and collaboration in chemistry. *Drug Discov Today* 2008;13:502–506.
17. Williams AJ, Tkachenko V, Lipinski C, Tropsha A, Ekins S. Free online resources enabling crowdsourced drug discovery. *Drug Discov World* 2009;10(Winter): 33–38.
18. Southan C, Varkonyi P, Muresan S. Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds. *J Cheminform* 2009;1:10.
19. The NIH Roadmap Initiative. Washington, DC: Office of Portfolio Analysis and Strategic Initiatives NIH, 2008.
20. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P. DrugBank: A comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* 2006;34:D668–D672.
21. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;28:27–30.
22. Wishart DS, et al. DrugBank: A knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* 2008;36:D901–D906.
23. Irwin JJ, Shoichet BK. ZINC—A free database of commercially available compounds for virtual screening. *J Chem Inf Model* 2005;45:177–182.
24. Irwin JJ, Raushel FM, Shoichet BK. Virtual screening against metalloenzymes for inhibitors and substrates. *Biochemistry* 2005;44:12316–12328.
25. Hohman M, Gregory K, Chibale K, Smith PJ, Ekins S, Bunin B. Novel web-based tools combining chemistry informatics, biology and social networks for drug discovery. *Drug Disc Today* 2009;14:261–270.

26. Ekins S, Williams AJ. When pharmaceutical companies publish large datasets: An abundance of riches or fool's gold? *Drug Disc Today* 2010;15:812–815.
27. Efimova L, Fiedler S, Verwijs C, Boyd A. Legitimised theft: Distributed apprenticeship in weblog networks. Paper presented at I-KNOW 04, Graz, Austria, 2004.
28. Barnes MR, Harland L, Foord SM, Hall MD, Dix I, Thomas S. Lowering industry firewalls: Pre-competitive informatics initiatives in drug discovery. *Nat Rev Drug Discov* 2009;8:701–708.
29. Tsovaltzi D, Rummel N, B.M. M, Pinkwart N, Scheuer O, Harrer A. Extending a virtual chemistry laboratory with a collaboration script to promote conceptual learning. *Int J Technol Enhanced Learning* 2010;2:91–110.
30. McLaren BM, Rummel N, Pinkwart N, Tsovaltzi D, Harrer A, Scheuer O. Learning chemistry through collaboration: A Wizard-of-Oz study of adaptive collaboration support. Paper presented at the Third European Conference on Technology Enhanced Learning (EC-TEL 2008), Maastricht, The Netherlands, 2008.
31. Glaser RE, Poole MJ. Organic chemistry online: Building collaborative learning communities through electronic communication tools. *J Chem Ed* 1999;76: 699–703.

6

CONSISTENT PATTERNS IN LARGE-SCALE COLLABORATION

Robin W. Spencer

6.1 INTRODUCTION

No one will dispute the motivation for this book: Without significant change in the rate of new drug approvals over decades [1] despite steeply rising expenditure and knowledge, no stone should be left unturned as we seek higher productivity in health care. Now that the Internet and storage technologies have reduced the costs of moving and storing information by orders of

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.

magnitude, surely we are poised for a positive revolution based on facile collaboration? Yet we are scientists, and a good dose of data is always healthy to calibrate our expectations and help us set strategy. In this chapter we will see that large-scale voluntary collaboration systems show remarkably consistent patterns in contributors' behavior within scientific collaborations in the pharmaceutical industry and extending to every other company and industry examined. This behavior has all the signatures of a power law, driven by positive feedback from the individual and the group, and with a "long tail" such that approximately half of all contributions come from people who contribute only once to any given campaign. Interestingly the evidence also suggests that networks of acquaintanceship are not an essential driving force, which makes us revise our concept of "community." Finally we review the data, not just for collaborative idea generation, but for collaborative evaluation and decision making, and see that the most popular methods are prone to strong bias by minority views.

6.2 BACKGROUND

From late 2005 to 2010 I created and then managed the "Idea Farm," an online collaborative problem-solving system within Pfizer, the world's largest pharmaceutical firm. The underlying model was the campaign, or challenge, in which a business need is identified with a specific sponsor, the problem or opportunity is reframed for the online medium, then in the "diverge" phase broadcast (usually via e-mail) to a large diverse audience who then may contribute using an easy-to-use system designed to support the challenge model [2]. In the subsequent "converge" phase, the entries are collected, organized, built upon (by the crowd and/or an assigned review team), evaluated, and trimmed and decisions made on implementation. The challenge model also underpins Innocentive, DARPA challenges (e.g., robot vehicles crossing the desert), X-Prizes, the Netflix Prize, and many more.* Arguably the first and most successful challenge was the longitude problem, in which the late-eighteenth-century Parliament and the Admiralty sponsored an apparently impossible problem to which John Harrison, an unknown clockmaker from the north of England, dedicated his life; he won by inventing the marine chronometer [3]. If we take innovation to consist of inspiration (the stereotypical "aha"), followed by invention (the proof of concept), followed by implementation (the scaling of the invention and dissemination to its customers), there is no question that implementation is the most lengthy, costly, and likely to fail [4]. The challenge model succeeds because it addresses this at the outset by selecting only those challenges where a serious need is matched by a serious and specific individual who already has the mandate

*See http://www.darpa.mil/grandchallenge/index.asp, http://www.xprize.org/, and http://www.netflixprize.com/ for examples.

and resources to address the problem and who will accept input from a large and diverse audience to get a better and faster solution [5]. While accessible throughout the corporation (secure behind its firewall), the sponsorship of the Idea Farm in Pfizer research and development (R&D) resulted in a majority of the campaigns involving scientific subjects. The baseline data consist of over 200 campaigns and 3000 separate authors of over 12,000 ideas, supplemented by anonymized data sets from colleagues in similar roles at other large corporations.

6.3 THE LONG TAIL OF COLLABORATION

"What was the average number of ideas per contributor last year?" is an innocent and reasonable question that has no good answer. It has an answer (typically around 2) , but it is not a good answer because the question implies, incorrectly, that the distribution is somewhat like a bell curve: If the average is 2 ideas per person, there probably were fewer people that put in 0 or 1 and fewer that put in 5 or 10, right? Just like if the average height is 5 foot 7, there should be fewer people 3 feet tall or 10 feet tall? Very wrong. Figure 6.1 is a rank–frequency plot of over four years' worth of ideas and comments entered into the Idea Farm, where the leftmost author (rank 1) put in about 700 entries

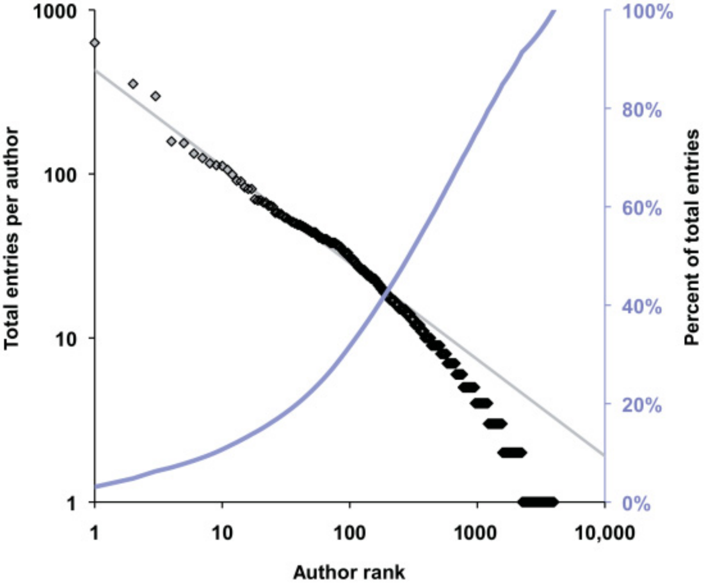


Figure 6.1 Rank–frequency plot of all ideas and comments submitted to Pfizer Idea Farm, 2006–2010: 4004 authors, 20,505 entries. Gray line: power law with α = 2.7. Overlay curve and right axis: cumulative percent of all entries.

and authors ranked 2000–4000 put in one each. A straight line on a log–log plot is a power law; power law distributions are prevalent in natural and human situations [6] but nearly ignored in statistics courses and textbooks. For power law distributions, "average" makes no sense (there is no peak) and the range of values can be enormous. In general, events which are mutually independent (the flipping of a "coin with no memory") will produce Gaussian, or normal, distributions, while events which are mutually dependent will produce power laws. Avalanches, earthquakes, salaries, and network connectivities all follow power laws, and a strong case can be made that the 2008–2009 financial collapse was due in part to our financial systems' underappreciation of the long tail of this distribution [7]. Figure 6.1 shows just what a "long tail" of a power law consists of: those 3000 people at the lower right (rank 1000–4000) who put in just one, two, or three entries each.

The importance of the tail in a power law phenomenon is the subject of Chris Anderson's eponymous book [8], where he describes how Internet technologies, by reducing transaction costs nearly to zero compared to brick-and-mortar stores, enabled Amazon and iTunes to extend the reach of book and music retail to orders-of-magnitude more content and consumers than had been previously feasible—to their notable profit and market dominance benefit.

Figure 6.1 is not unique to Pfizer or even pharmaceuticals or scientific problem solving; Figure 6.2 shows Pfizer's data (just ideas, not ideas and comments) with data from Cargill, a huge multinational agribusiness corporation. Both companies have secure firewalls with no contact or commonality of people, business needs, challenges, demographics, or cultures, yet their statistics

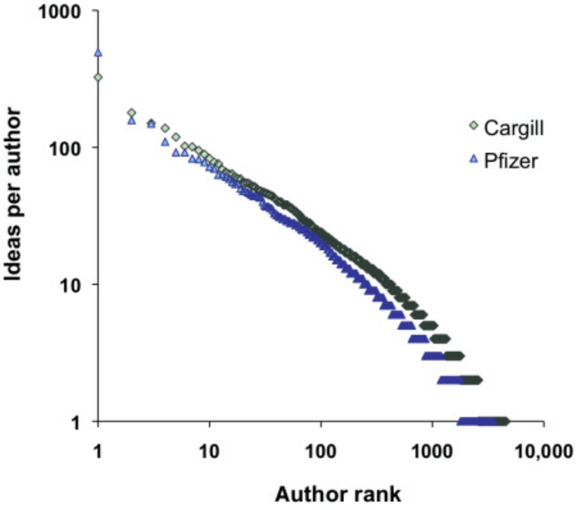


Figure 6.2 Rank–frequency plot of ideas submitted to Pfizer Idea Farm (triangles) and similar system in large Cargill business unit (diamonds).

of participation are indistinguishable. This follows for every case we have examined, and at all scales: Individual challenges give the same plots with the same slope [9]. This is strong support for an underlying power law mechanism since it is the only distribution that is scale free [6].

There is every reason to expect these power law properties to extend to every type of large online collaboration, in part because of the diversity and number of our private-sector data sets [9] and because the contributions to Wikipedia, perhaps the largest open collaborative intellectual effort of all time, follow the same pattern [10].

6.4 VALUE OF AN IDEA

Whether or not the power law property matters depends on the value of what we are measuring, specifically whether the ideas from the "head" (those relatively few people who put in many ideas each) are more or less valuable than those from the "tail" (those many people who put in very few ideas each). Figure 6.3 suggests three general possibilities, where the ideas are counted in the same order as the participation level of their authors, that is, ideas from the most prolific authors at the left and from the occasional authors to the right.

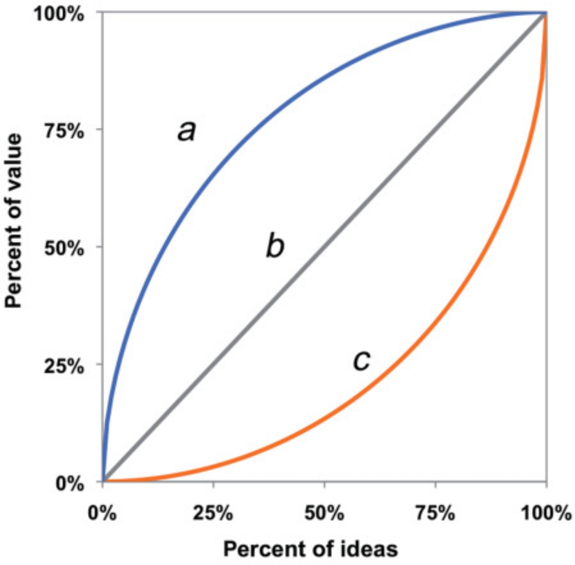


Figure 6.3 Models of cumulative value vs. cumulative quantity, where quantity (horizontal axis) is ordered by author rank (as in Figs. 6.1 and 6.2): (*a*) "head" participants have better ideas; (*b*) all ideas are (probabilistically) equal; (*c*) "tail" participants have better ideas.

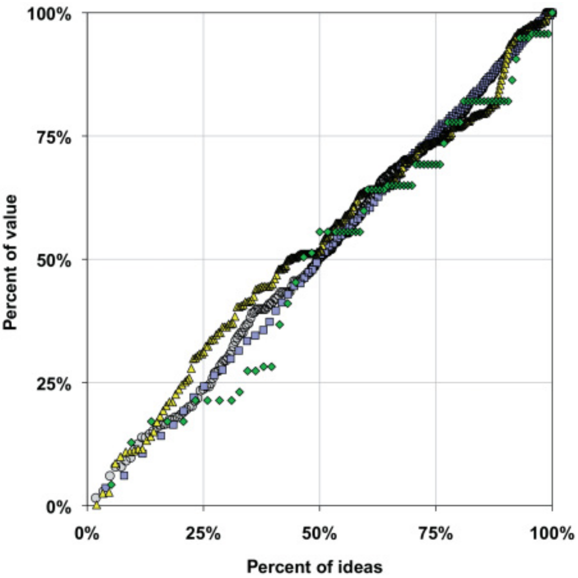


Figure 6.4 Cumulative value vs. cumulative quantity from four large Pfizer campaigns: (triangles) an all-R&D campaign seeking nontraditional, marketable IP; (circles) a challenge to reduce operating costs for a mobile sales force; (squares) a process improvement challenge to reduce time and complexity in clinical document preparation; (diamonds) a scientific–medical challenge for additional indications for an existing drug.

In case (*a*), not only are the "idea people" prolific, their ideas are better. If there is some sort of expertise or talent for idea generation, or if people with more talent also have more confidence resulting in higher participation, this is what we might expect. It would be a case of an "80–20" rule where a minority dominates in quantity and quality. On the other hand, a case can be made for (*c*), where the ideas from the rare participants should be better: There is good evidence that teams get tired and less effective over time and need external stimuli [11], and Gary Hamel makes a strong case that value comes by "listening to the periphery," those voices seldom heard by dint of corporate culture, geography, or generational deafness [12].

Our data, while not as complete as for the power law itself, are consistent and provocative. Figure 6.4 shows the results from four large Pfizer challenges in which semiquantitative estimates of idea value were available. Importantly, in all cases entry value was assigned by the review team established by the campaign sponsor and judged by criteria agreed on in advance [typically along dimensions of technical feasibility, potential market value or cost or time reduction, competitive advantage, and intellectual property (IP) risk]; in other words, value was estimated by those who would benefit by success and be involved in implementation. Ideas rated low by such a team have essentially no chance to be implemented and so, however intrinsically brilliant, have no value: a harsh pragmatic reality. Such teams usually begin with a high–medium–low binning before any ideas are excluded; Figure 6.4 shows the results with high = 10, medium = 4, and low = 0 points, though the weighting factors made no qualitative difference to the results.

This is a very interesting result, supporting neither above hypothesis but rather suggesting that idea value is independent of whether the author is prolific or occasional. In other words, if 1 in 100 ideas is valuable (for example), then we might expect one valuable idea from the three people who put in 50, 30, and 20 each and also might expect one valuable idea from the 100 people who only put in one each. Note how accurately this parallels the value proposition in iTunes or Amazon's Kindle bookstore, where songs cost about 99 cents and books cost $10, roughly constant from best sellers to the most obscure titles.

Now we can return to the overlay of Figure 6.1, the cumulative area under the log–log graph. This represent not only the cumulative number of entries but also the cumulative value. About half the value is contributed by authors 1–300, and the other half by authors 301–4000. If your reaction is to think "I want those top 300!" you are missing the opportunity of large-scale collaboration in three important ways. First, exactly which 300 you need is going to change for every given business problem. Innovation must be specific and purposeful; calls for "we only want big game-changing ideas" are guaranteed to fail [4], and so successful campaigns are quite content specific and useful contributions draw on deep personal expertise and experience. Second, traditional teams become dysfunctional beyond a dozen or so participants [13]. Even scheduling meetings for a team of 20 becomes infeasible. If you fall back to the idea of "top 10" for a team, Figure 6.1 tells us that you will knowingly miss 90% of the value you could have had. If your organization does not have 4000 people in it, that is still true, they just do not all work for you. Third, and optimistically, recall the lessons of Chris Anderson: Do not run from the long tail, exploit it. Systems designed to facilitate, run campaigns, and manage evaluation and next steps are readily available [2], and the cost is not the system but the opportunity lost by ignoring the value in the tail (which you now know how to predict).

In fact, the tail value is greater for individual campaigns than Figure 6.1 suggests. The only exceptions to power law behavior seen to date are from nonvoluntary campaigns. The electronic tools for mass collaboration work equally well in a "command" situation; for example, it is very productive to have a half-day meeting with several background presentations followed by a "flash event" in which every member of the audience is exhorted to spend the next 15 minutes writing down four ideas for how their work team could support the presented proposal [14]. In these cases the result is not a power law distribution [9] but closer to a Gaussian; on a rank–frequency plot the tail drops quickly. The Pfizer data are an aggregate of many campaigns including large involuntary ones of this type, which probably accounts for the deviation from power law at the bottom right. When large voluntary individual campaigns alone are considered, the tail extends farther [9, Fig. 5a], with the consequence that fully half of the total entries (and therefore value) come from people who only ever contribute once [9, Fig. 8].

6.5 COMMUNITIES?

We have seen that voluntary, large-scale, collaborative challenges on scientific topics are feasible, sustainable, and technically well understood and that a great deal of the value derived comes from the occasional contributors. But are these really "communities," or is that word becoming overworked, in the same way that calling a stranger who accesses your blog or photos a "friend" does not make them one. It is more than a semantic quibble if our beliefs affect our strategies for attracting new participants or rewarding and recognizing past contributors.

We have one relevant data set, but it is objective and large scale. For years Pfizer, like many companies, has had a link on its public website, saying in effect, "send us your ideas to improve our offerings." Figure 6.5 shows the familiar rank–frequency plot for several years of this activity; again, it is an excellent power law. What this data set has in common with the others is that it is from a large-scale voluntary process, seeking new ideas and concepts for business purposes. Where it differs is that for intellectual property and legal reasons the process has been implemented as a "drop box," in which contribu-

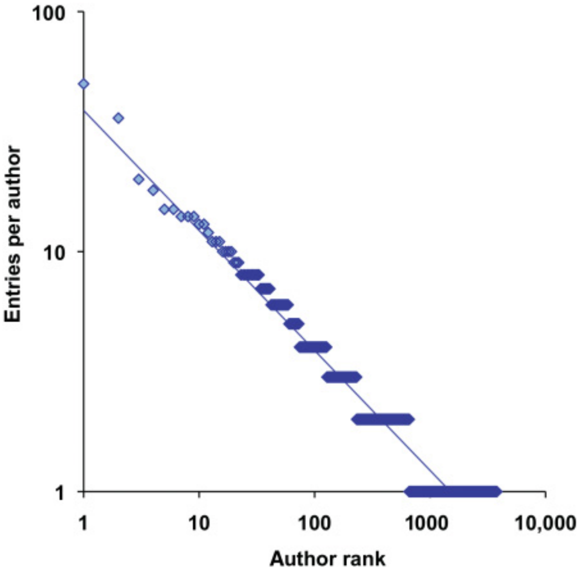


Figure 6.5 Rank–frequency plot for unsolicited open website suggestions. The line has exponent α = 3.

tor's identities are not accessible to each other and there is no possible commenting or cross-contributor collaboration (a hallmark of the internal challenges). These contributors cannot by any stretch be called a community, because they cannot know or communicate with each other. And yet we have a power law signature, including the same exponent ($\alpha = 2.7 \pm 0.3$, [9]) as all others observed.

Figure 6.5 refutes the hypothesis that our power laws might derive from a network effect. It is well established that human networks show power law statistics in their connectivity [15], and it would be reasonable to suppose that our observations somehow derive their statistics from a driving force dependent on a network: For example, I am more likely to contribute if people in my social network contribute. It remains perfectly possible that such network effects could amplify the contributions to a challenge, but Figure 6.5 shows that something more intrinsic, more local is going on. A source of positive feedback is the most likely origin of the power laws [6], and simulations suggest that feedback comes approximately half from one's own behavior ("I put in an idea last week, it wasn't hard or scary, I'll probably do it again") and half from general observations of others ("Other people are doing this, I'll give it a try") [9]. The difference between general ("other people") and specific ("people I know and trust") is arguably the difference between a collaboration process and a community-driven process.

With "community" now shown to be a tenuous concept, we have to consider how to advertize our campaigns and induce people to contribute. We can certainly hope that people will tell all their friends but cannot rely on it and have data suggesting instead that contribution is more likely a private choice. As manager and facilitator for hundreds of challenges, this is not surprising. Almost without exception, announcements of a new broad challenge that depend on propagated e-mails will fail: first, because the e-mails simply do not get sent and, second, because they are sent with a generic title ("Please Read" or "On Behalf Of"), which does nothing to convey the content or opportunity. In a world of spam and information overload, this is a bucket of cold water.

6.6 MOTIVATION AND SUSTAINABILITY

If we cannot expect true community behavior, and if the specificity of our business needs is such that each campaign will uniquely interest different people, how can we make large-scale collaboration work? My role at Pfizer brought me into a true community of peers from other companies, brought together face to face in vendor-sponsored user groups. There are definitely best and worst practices, learned over and over; for example:

- *Do Not Offer Tangible Prizes or Rewards* Especially for cutting-edge scientific challenges, the participants you need are probably well paid and not particularly enthused by another tee shirt, coffee cup, or $100 voucher.

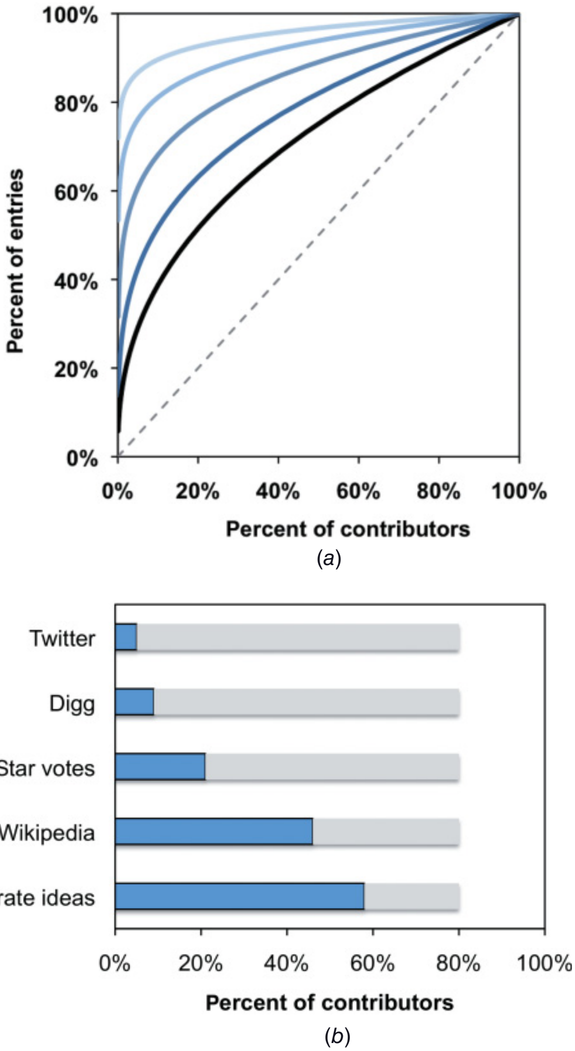
There is intriguing literature that, in fact, monetizing an otherwise altruistic bargain will decrease participation [16]. If that is not enough, offering tangible rewards comes at significant cost: Who will get the prize? (Let's call a meeting …). Who has the prize budget? Just do not do it unless you are not prepared to make a full business of it (e.g., Innocentive's prizes, which may typically be in the $5000–$40,000 range).

- *Do Offer Recognition But Watch for Overload* Absolutely recognize contributors when campaign results are known; every organization has appropriate newsletters for this. But beware of the cynicism that follows from too many employee-of-the-month-type programs [17]. It is not kindergarten; not everyone gets a star.
- *Highlight Based on Quality, Not Quantity* Since most of your campaign value will come from people who only ever put in one, two, or three contributions, do not cut off the tail by hyping the high contributors and implicity offending the rare ones. Do not set up a "reputation" system based on mouse clicks rather than serious content. Do highlight contributors, but based on quality, not quantity.
- *Remember Herzberg* A generation ago, Herzberg studied employees' motivators and demotivators; his article [18] has been the most-requested reprint from the *Harvard Business Review*. Even more important than recognition is to make the task serious and real; people seek achievement and responsibility. In other words, the challenges you pose must *matter*, and it must be clear how they matter and to whom. Never pose toy challenges or ones that address minor issues; it devalues the entire program. Equally important is to assure that your collaboration system avoids the Herzberg demotivators (or "hygiene factors"), principal of which is the perception of unfair or inappropriate policies and bureaucracy. In other words, if you want voluntary help, do not make the contributor suffer through three pages of legal caveats or a picky survey, and make the challenge about something known to be important to the sponsor and, perhaps altruistically, to the contributor.

6.7 COLLABORATIVE EVALUATION

Soliciting and collecting ideas are only the divergent half of a campaign; the convergence process of evaluation and decision must follow if there is to be implementation. For typical departmental-scale campaigns in which entered ideas number in the dozens to hundreds, a review team appointed by the original project sponsor is very effective, because it taps directly into the organization's norms for project responsibility and funding. However, when entries approach the thousands, it may be useful to enlist the "crowd" to assist in their evaluation.

But the data suggest caution. Figure 6.6 illustrates how we need to be aware of the possibility that crowd evaluations, however democratic and open in



(*a*)

(*b*)

Figure 6.6 Exponent of power laws reflects the distribution of participation. (*a*) Curves are Newman's equation 28 [6], for (top to bottom) $\alpha = 2.05, 2.1, 2.2, 2.4, 2.7$, which are approximately the exponents for Twitter entries [19], Digg promote–demote voting [10], five-star voting (this work and [9]), Wikipedia edits [10], and corporate ideas (this work and [9]). Dashed gray line, if all participants contributed equally. (*b*) Slice through part (*a*) illustrating how many people contribute 80% of the content.

intent, may be driven by small minorities. The types of data in Figure 6.6 appear to be in order of difficulty or knowledge required (i.e., a five-star vote takes less effort or know-how than typing in an original contribution), which suggests an important possibility: that the easier or less content rich the task, the more it is likely to be driven by an active minority of participants. This is perhaps counterintuitive: "Make it easy" would seem to encourage a more democratic, representative outcome. But the data sets behind Figure 6.6 are so large that we must take very seriously the possibility that "make it hard and specific" is the better way to assure a broader source of input.

Of equal concern are the very large scale observations of five-star voting at Amazon [20], namely that it is biased, compressed (with an average of 4.4 out of five stars), and prone to follower behavior that drives to extreme opinions rather than balance. Consistent with the observation of Figure 6.6, the authors recommend making the online book review process more difficult, rather than less, to achieve better quality and balance.

6.8 CONCLUSIONS

Multiple large data sets from diverse private and public sources show that contributions to large-scale voluntary collaboration campaigns (including scientific challenges) generally follow a power law and with an exponent considerably higher ($\alpha = 2.7 \pm 0.3$) than "easier" tasks (Twitter, Digg, five-star rating; $\alpha \simeq 2$–2.2). The consequence is that these campaigns depend for a majority of their content on a "long tail" of people who contribute only a couple of ideas each. Because power laws are scale free, this generalization applies to small as well as global-scale campaigns. The phenomenon may benefit from, but does not depend on, social networks because blinded "drop box" challenges have the same signature. Thus, rather than speak of "communities," we would be more accurate to refer to "personal responses to a particular challenge." To encourage participation, we should respect our contributors as individuals, recognize quality over quantity, and remember the strong motivation of contributing to real work that makes a difference. Large-scale collaborative evaluation of options is more problematic, since the data for popular techniques like promote–demote and five-star voting reveal a potential for considerable bias and dominance of minority opinions.

ACKNOWLEDGMENTS

The author, recently retired as Senior Research Fellow at Pfizer, wishes to thank Steve Street for his support and courage to embrace change as well as Tim Woods at Imaginatik, Doug Phillips at Pfizer, and Anne Rogers and Kurt Detloff at Cargill for sharing anonymized data.

REFERENCES

1. Summary of NDA Approvals & Receipts, 1938 to the present. Available: http://www.fda.gov/AboutFDA/WhatWeDo/History/ProductRegulation/SummaryofNDAApprovalsReceipts1938tothepresent/default.htm. Accessed July 5. 2010.

2. The system behind the Pfizer Idea Farm is Idea Central® from Imaginatik PLC. Available: http://www.imaginatik.com.
3. Sobel D. *Longitude*. London: Fourth Estate, 1995.
4. Drucker P. *Innovation and Entrepreneurship*. New York: HarperBusiness, 1985, Ch. I.2.
5. Suroweicki J. *The Wisdom of Crowds*. New York: Doubleday, 2004.
6. Newman MEJ. Power laws, Pareto distributions and Zipf's law. http://arxiv.org/abs/cond-mat/0412004v3, 2004.
7. Olmerod P. *Why Most Things Fail*. New York: Pantheon, 2005, pp. 173–179.
8. Anderson C. *The Long Tail*. New York: Hyperion, 2006.
9. Spencer R, Woods T. The long tail of idea generation. *Int J Innovation Sci* 2010;2(2):53–63.
10. Wilkinson D. Strong regularities in online peer production. In: *Proceedings of the 2008 ACM Conference on E-Commerce*, Chicago, IL, July 2008. Available: http://www.hpl.hp.com/research/scl/papers/regularities/.
11. Katz R, Allen TJ. Investigating the NIH syndrome. In Tushman M, Moore M, Eds. *Readings in the Management of Innovation*, 2nd ed. New York: HarperBusiness, 1988, pp. 293–309.
12. Hamel G. *Leading the Revolution*. Cambridge: Harvard Business School Press, 2000, pp. 261–264.
13. Brooks F. *The Mythical Man-Month*. New York: Addison-Wesley, 1995.
14. Spencer R. Innovation by the side door. *Res-Technol Manag* 2007;50(5):10–12.
15. Barabasi A-L, Albert R. Emergence of scaling in random networks. *Science* 1999;286:509–512.
16. Bowles S. When economic incentives backfire. Available: http://hbr.org/2009/03/when-economic-incentives-backfire/ar/1.
17. Deming WE. *Out of the Crisis*. Cambridge: MIT Press, 2000.
18. Herzberg F. One more time: How do you motivate employees? *Harvard Business Rev* 1987;65(5):109–120.
19. Piskorski M. Networks as covers: Evidence from an on-line social network. Working Paper, Harvard Business School. Available: http://www.iq.harvard.edu/blog/netgov/2009/06/hbs_research_twitter_oligarchy.html.
20. Wu F, Huberman BA. How public opinion forms, Social Computing Lab, HP Labs, Palo Alto. Available: http://www.hpl.hp.com/research/scl/papers/howopinions/wine.pdf. Kostakos V. Is the crowd's wisdom biased? A quantitative assessment of three online communities. Available: http://arxiv.org/pdf/0909.0237.

7

COLLABORATIONS BETWEEN CHEMISTS AND BIOLOGISTS

Victor J. Hruby

7.1 INTRODUCTION

Biology is the science of life. More specifically, it is the study of the organization of matter that can reproduce itself and maintain its specific functional properties over many generations as life. Chemistry is the study of matter which composes the universe, their combinations, and the manner in which they interact with each other. Since the molecular biology revolution of the past 50 years, it has become increasingly clear that the most fundamental biological problems, whether health or disease, are essentially chemical problems. Unfortunately, most chemists are ill prepared to do state-of-the-art biology and, indeed, often have very negative attitudes about biologists. On the other hand, biologists are generally ill prepared to do state-of-the-art chemistry and generally avoid interacting with chemists.

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.

To bridge the gap, scientific "opinion leaders" have suggested that biologists and chemists should collaborate to help solve the problems of human health and disease. But in practice these "leaders" generally act to punish true collaborations by reducing funding and other amenities that biologists and chemists have if they worked independently (e.g., a top biologist and a top chemist can each generally readily get a $250,000-per-year grant independently doing innovative research, but when working together reviewers, administrators, etc., balk at giving the two working together a $500,000-per-year grant). I know this because I have done the experiment. So, not surprisingly, what has happened in the past 20 years or so is biologists and chemical biologists (chemists working in biology) have resorted to data collecting. Genomics, proteomics, metabolomics, structural biology [X-ray and nuclear magnetic resonance (NMR)], chemical libraries, high-throughput assays, and so on, have been essentially data-collecting exercises. The "exciting discoveries" are made by robots, machines, and computers which collect enormous amounts of data. In the process human thought often seems to have become of secondary importance. At the same time, creative collaborations between chemists and biologists are often marginalized and starved for the resources they need to scientifically investigate the very difficult problems of understanding life processes and how disease and other dysfunctions arise and what might be done to fully understand these processes. In many ways, it is a tragedy for both chemistry and biology. I certainly realize that the cult of the individual dominates our society and its award systems, and our power structures, especially in science, enforce the myths which form the basis for our scientific culture. Clearly a less arrogant and more collaborative scientific culture will happen only slowly and incrementally. Nonetheless, it seems clear that we will never solve the problems of human health and disease until we chemists and biologists (and other scientists) work together as equals in true collaboration, without arrogance of fields, to solve these complex problems.

I believe that it is possible to develop such collaborative interactions and have spent most of my 45 years in science as a chemical biologist trying to do that. Using a few examples taken from my own efforts with biologists, I will try and illustrate how daily collaborations with biologists have led to novel insights into biology using chemistry and the way in which the chemistry of biology is manifested. Some aspects of these efforts have already been discussed in the literature [1–3].

7.2 ORGANIZING SUCCESSFUL COLLABORATIONS BETWEEN CHEMISTS AND BIOLOGISTS TO SOLVE IMPORTANT PROBLEMS IN CHEMICAL BIOLOGY AND MEDICINE

It goes without saying (it seems so obvious) that any successful collaboration between chemists and biologists comes with the recognition that the scientific problem in biology, or drug design, or some aspect of medicine requires that both the chemist involved and the biologist involved recognize that the problem will not be solved without highly integrated creative efforts by both chemist and biologist. Life, good health, and disease are chemical processes in the context of the complexities of life, in the case of human health and disease multicellular life. Thus in entering a collaboration both chemist and biologist will need to recognize that collaboration will require that they both will be doing research that they had not previously considered, and thus they must be prepared to be completely open to new thinking about the problem they are trying to solve together. Indeed, in my successful collaborations, generally very quickly, often within a few months or a year, the research has taken on new directions not initially envisioned. I would suggest that you know you have a good collaboration when both parties involved are soon doing research that they had not initially planned to do.

On the other hand, I have not had as much success in collaboration when the potential collaborator is convinced that what they had in mind to do is all they will do. They are convinced they know where they are going and do not need to consider alternate thinking. Alternatively, they see their collaborator as simply a means to their ends. No doubt this can be and often is very useful. Such cooperations are very important in science and often lead to new and useful results. However, they are not collaborations, and the possibilities for creativity and novel insights and directions are lost or more generally either explicitly or implicitly suppressed. Even more to the point, most research these days, whatever the source of support, is "hypothesis" driven. Such an approach to research is very congenial for the bureaucracy and bureaucrat who can defend his or her support for the research on very practical grounds without any real knowledge of the research area. In a good collaboration "hypotheses" are viewed as temporary starting points with the understanding that modifying the working hypothesis is not only a possible but also a desired likelihood as the research progresses. One other critical aspect of creative and productive collaboration is that all participants must take ownership and thus responsibility for success while at the same time share in both the success and failure. In this regard, doing high-risk, high-reward research generally has failures. Indeed, failure often is a critical part for ultimate success. Knowing when to reexamine and change your most cherished ideas is always difficult whether or not the research involves collaboration, but doing so when collaboration is involved is especially difficult. Fingerpointing never solves any problems whether in life or in research, but here congenial critical examination of the problem in a group environment is critical. If done properly, it often can lead to the most creative solutions moving forward. For this purpose I have found that when such research problems occur it is best if all participants—the principal investigators, students, postdocs, and technicians—discuss the problem together. Such discussions can often lead to the most creative solutions, especially as it provides those who have experienced failure an opportunity to fully discuss their efforts and what has been learned and gives them permission from the entire group to move in new directions with enthusiasm and confidence. In this regard, a key to success in any collaborative effort is the necessity that all lines of communication remain open. Each research group involved in a collaboration has its own culture and modes of communication and cooperation. Since many different types of expertise are often involved, and often quite different knowledge is necessary to address the problem, it is critical that all participants be aware of and committed to the multidisciplinary requirements of the project and recognize that success will depend on the success of all components of the research. In other words, making everyone around you successful is the key to your success. To ensure this commitment and understanding, I have found that meetings of the entire group in which research progress is formally discussed, including related recently published literature research, be held often (weekly or biweekly). Not only do these group meetings provide open communication channels and shared goals, but equally importantly they provide everyone the opportunity to know who is doing what and how it all fits into the ultimate goals. From these discussions it often also is clear who should write which papers and who will be first author. I have found this minimizes any conflicts down the road about who gets credit for what and why and whether and what aspects of the research should be published in chemical/biophysical journals and which should be published in biological/medical journals.

Another key issue that is increasingly critical for successful collaborative research at the interface of chemistry and biology (including medicine) is the availability of state-of-the-art infrastructure. Generally the biological chemists will need access to outstanding X-ray, NMR, mass spectrometry, sequencing, cell scanning and imaging, cell development and cell growth facilities, and screening/assay facilities, not to mention ultracentrifuges, high-performance liquid chromatographs (analytical and preparative), and so on. For their part the biologist needs outstanding facilities for obtaining genomics, proteomics and other critical biological data, animal models and the facilities to develop new animal models, knockout and knockin animals, cellular and animal imaging equipment, facilities for a wide variety of behavioral and other whole-animal studies, and much more. The beauty of good collaboration, of course, is that various diverse physical, chemical, and biological tools can be brought to bear on the problem, but of course this requires careful division of labor and maximal communication so that all involved can maximize creativity and productivity. Simply wishing or hoping that such efficacy will be obtained in collaborations will not lead to creative accomplishments. Everyone must be committed to success of the larger goals and be able and willing to commit their time and creativity to the overall goals.

Finally, and this is every bit as critical to successful collaboration, the administration of the departments involved, the colleges involved, the business and research offices involved, and the upper administration (vice presidents, provosts, presidents) must be committed to the success of others, especially others whom they cannot and should not control. From my experience this is where collaborative efforts are often stymied or destroyed, often with utter disregard for the critical science lost, but especially the human carnage created by well-intentioned but often ill-conceived requirements for the "approval" of the grants, of grants management, and intercollege or interdepartment "agreements." Interference in scientific, management, and fiscal affairs related to the research by administrators is often the most difficult barrier to doing collaborative research. For successful collaborations it is essential that the principal investigators organize and decide who will do the science, how the financial and related resources will be distributed, and who will be the principal investigator in grant applications. Why department heads, deans, provosts, vice presidents for research, facility directors, and such bureaucrats think that they have the knowledge and ability to dictate or determine such matters has always astonished this observer. The damage these arrogant, power-hungry bureaucrats can cause and have caused cannot be overestimated, and the human carnage which follows is even more devastating. On the other hand, when these same people act as facilitators, problem solvers, and organizers of the proper infrastructure channels, business office coordinations, and related administrative requirements, they can greatly facilitate success both short term but especially long term, especially if their facilitation is constant and consistent. Of course problems will arise. Human beings are imperfect, make mistakes, do stupid things, and so on. In these cases wise and prudent principle investigators (PIs) and administrators will play crucial roles for long-term success.

In this regard it is most useful, indeed essential, to have an administrative secretary who can oversee and facilitate the daily financial, personnel, and technical issues that arise. Administration and granting agencies should not only provide support for such persons but also require their presence.

Unfortunately, as is often noted, "new brooms sweep clean," and often they sweep out the good with the bad. I have had long-term collaborations with some of my collaborators for 30 or more years. They have been highly productive and creative in the biological areas of neuroscience, pain, addiction, feeding behavior, sexual behavior, diabetes, cancer, and so on. Despite these successes and the tens of millions of dollars that we have brought to the university, the limitations of what we could accomplish have often been due to administrative decisions. Agreements that had been reached among collaborators and agreed upon earlier are not honored or are overturned by unthinking, unscrupulous, and arrogant new administrators who take what they want, ignore what has been done and agreed on, and modify what is to be because they can. As Lord Acton said almost two centuries ago, "Power corrupts, absolute power corrupts absolutely." Thus it is and always will be with human beings. If it happens at the university level, as it often has, we have just gone on with our collaboration, but often at a diminished level compared to what could have been done. Things become more problematic when it happens at the national level, when bureaucratic decisions simply stop ongoing science. I will only give two examples from my own career. For the first 20 plus years of my independent academic career I had a grant from the National Science Foundation (NSF). This grant, though never large, was the major grant for my group to pursue the highest risk research in peptide and peptidomimetic science which would be applicable to biological problems. Its major goals evolved around the design of novel constrained peptide and peptidomimetic structures, new structural templates, and the asymmetric synthesis of novel amino acids and other "templates" which could be used to explore the importance of chi space in peptide and protein structural and biological function [4, 5]. In other words, its goal was to make synthetic and structural organic chemistry compatible with the chemistry of peptides, proteins, and other biological compounds. These novel structures and synthetic methods were designed for applications to the peptide hormones and neurotransmitters we were investigating with our biological collaborators. As would be expected, these novel templates and structures failed from time to time, but they also led to numerous successful innovations and novel biological functions and were beginning to provide new and useful insight into what kind of structural constraints in phi/psi space and also in chi space could lead to structures with unique biology. However, just as this grant got the best reviews of my career from the NSF, the NSF decided not to continue to support this research. Though I protested and asked for a rationale for why it would no longer support my research in this area of ligand design and its relationship to biological function, I never received an explanation. Much of what I proposed still awaits effort to determine its potential in peptide and peptide mimetic design and synthesis.

A similar fate occurred several years later in a grant that I had for over 20 years from the National Institutes of Heath (NIH) in diabetes research and the involvement of glucagon in diabetes among its goals. We were the first to design and prepare a glucagon receptor antagonist and to demonstrate with our collaborator David Johnson that it lowers blood glucose levels in diabetic animals [6]. However, it also had partial agonist activity, and in other animal models it was not as effective. We continued to develop more potent glucagon receptor antagonist and using these and other glucagon analogues were able to demonstrate with Miles Houslay that glucagon stimulates more than one signaling pathway, which at that time was a revolutionary discovery [7], that is that signaling through G-protein coupled receptors (GPCRs) could be mediated by multiple signaling pathway [cyclic adenosine monophosphate (cAMP) Ca^{2+}, phosphoinositol, etc.]. As a result of these and other novel discoveries we were awarded an NIH MERIT Award. However, obtaining a highly efficacious glucagon antagonist that could be used for treating diabetics proved elusive in our research group and in others in academia and industry. Using a highly sensitive assay, we had found that these potent in vitro antagonists had weak agonist activity in vivo, which significantly reduced their efficacy in reducing glucose levels in most animal diabetic models. Eventually we obtained pure glucagon antagonists/inverse agonists, but just as we were testing the efficacy of one of these analogues in animal models (dogs and rats), our NIH support was terminated. The possible utilization of pure glucagon receptor antagonists and inverse agonists in the treatment of diabetes remains to be determined.

Of course, it is possible, perhaps likely, that the judgment of the administrators for these scientific and medical problems was correct and those of us working on these collaborative projects needed to move on or quit doing this research. Only time and further research will answer these questions. Nonetheless, when such decisions are made in the midst of novel productive collaborative research, it has an immediate and lasting impact on the future aspirations, especially of the more junior people involved. From my experience the young graduate and medical students and postdocs involved in these "failed" projects move away from research and take other career paths. In a few cases, they completely leave science. Is this relevant to the future of collaborative research? Who can say?

7.3 CONCLUDING DISCUSSION

Virtually all important and lasting science is a group and collective enterprise. Great advances in the understanding of our universe have been the result of efforts of many individuals often from diverse areas of science. If this is the case, then it would seem obvious that we would make greater progress in science, especially for highly complex scientific problems such as the underlying mechanism of human health and disease and how we might most effectively promote the former and be able to prevent and, if not prevent, treat the latter, if we more often and more effectively collaborate with each other. This clearly will require some fundamental changes in our personal behavior but even more so in our cultural, institutional, and power structures. Psychologists and other scientists who study human behavior have found that most humans find greatest happiness and fulfillments when they do things and accomplish things with their fellow human beings. It seems increasingly clear that promoting and rewarding collaborative research result in a win–win situation for both the scientists who do the research and the society which benefits from it.

ACKNOWLEDGMENTS

I thank my many scientific collaborators who have made my scientific career both productive and satisfying. I am grateful for their friendship and kindness, but especially for their criticisms and challenges to my scientific ideas and hypotheses. These especially have led to new ideas, new approaches, and new insights into the chemistry and biology of life for which I am eternally grateful.

REFERENCES

1. Hruby VJ. Peptide science: Exploring the use of chemical principles and interdisciplinary collaboration for understanding life processes. *J Med Chem* 2003;46: 4215–4231.

2. Hruby VJ. Organic chemistry and biology: Chemical biology through the eyes of collaboration. *J Org Chem* 2009;74:9245–9264.
3. Hruby VJ. Design of peptide and peptidomimetic ligands with novel pharmacological activity profiles. *Annu Rev Pharmacol Toxicol* 2011, in press.
4. Hruby VJ, Al-Obeidi F, Kazmierski WM. Emerging approaches in the molecular design of receptor selective peptide ligands: Conformational, topographical and dynamic considerations. *Biochem J* 1990;268:249–262.
5. Hruby VJ, Li G, Haskell-Luevano C, Shenderovich MD. Design of peptides, proteins, and peptidomimetics in chi space. *Biopolym (Peptide Sci)* 1997;43:219–266.
6. Johnson DG, Goebel CU, Hruby VJ, Bregman MD, Trivedi DB. Decrease in hyperglycemia of diabetic rats by a glucagon receptor antagonist. *Science* 1982;215:1115–1116.
7. Wakelam MJO, Murphy GJ, Hruby VJ, Houslay MD. Activation of two signal-transduction systems in hepatocytes by glucagon. *Nature* 1986;323:68–71.

8

ETHICS OF COLLABORATION

Richard J. McGowan, Matthew K. McGowan, and
Garrett J. McGowan

8.1 INTRODUCTION

Perhaps a chapter on the ethics of collaboration ought to begin with a remark attributed to Isaac Newton, one of the greatest scientists who ever lived: "If I have seen further, it is by standing on the shoulders of giants." Newton realized that much of his success was owed to the work of scientists who came before him. He understood that success in scientific investigation depended on the

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.

success of other scientists and that scientific advance was not a function of a lone scientist working in isolated conditions. Newton had the matter correct, that scientific collaboration is conducive to scientific progress, especially as science has been conducted over the last two centuries.

Rosenberg and Birdzell [1] argued that the "Western miracle," the process of growth and development in the nineteenth and twentieth centuries, was a consequence of the growth of scientific knowledge and the rise of technology in free-market economies. Rosenberg and Birdzell [1] believe that "Western science has made a better organized attack on the secrets of nature and used greater resources in the assault than science in other cultures." The organized attack involved the situating of scientists under one roof. Rosenberg and Birdzell noted [1] that "although the idea of bringing scientists together for directed research in an institute equipped with laboratory instruments and a suitable library was tried successfully in the first half of the 15th century by Prince Henry the Navigator of Portugal, it came into common practice only in the 19th century."

In short, the Western miracle, the tremendous advances in scientific knowledge and in the subsequent standard of living, has deep roots in cooperative, mutual endeavors guided by scientific inquiry and goals. Today's scientific laboratories, both academic and nonacademic, rely on the contributions of many people working together to achieve a better understanding of the natural world. McGowan and McGowan [2] suggested that the history of scientific publication shows growing awareness of collaboration within the scientific disciplines. They reviewed the history of attribution in the journal *Science* and discovered that over the years fewer and fewer articles were published under the name of a single author. We may safely conclude that collaboration and cooperation are hallmarks of modern science.

The apparent necessity of collaboration and cooperation in the sciences, in an ideal world, would have scientists undertake collaborative endeavors efficiently and agreeably. In the real world, problems arise both systemically and locally. One example of a systemic problem that has rendered scientific collaboration less efficient is the problem of exclusion. Indeed, much has been said and written by feminists alleging the lack of a female presence in the scientific community.

Local problems exist, too, as case studies show. We begin, though, with a brief examination of collaboration itself.

8.2 TEAMWORK, COOPERATION, AND COLLABORATION

Not every collective endeavor, that is, an endeavor undertaken by several people, can be called a collaboration. Griesel [3] suggests that the word *collaboration* sometimes produces confusion because of its elasticity. Griesel suggests this word could mean teamwork, partnership, cooperation between two or more people, or a more restrictive form of organizational structure.

Griesel warns of the confusion in terminology when collaboration is used [3]. At least we can observe that a collaboration represents organized behavior of a certain sort.

If the various terms above, such as *team effort* or *partnership*, are examined, a more precise understanding of collaboration can be acquired. For instance, "team effort implies competition" [3]. However, many collaborative efforts, especially in science, are designed around cooperation without a sense of vanquishing or "beating" others. Also, Griesel observes that collaboration is "operational" in that it is has a fluidity of process and governance and does not ask the individual to place the team first [3]. The popular sports directive, "take one for the team," suggests that a team player is somehow required to set aside his or her interests for the good of the team. Collaborative activity does not seem to have that rigidity if Griesel is correct.

Kagan [4] believed that cooperation was the first step toward collaboration, followed by the second step of coordinated action, and then ending in the most complex level of collective organization, namely collaboration. Kagan said that "collaborations are defined as organizational and interorganizational structures where resources, power, and authority are shared and where people are brought together to achieve common goals that could not be accomplished by a single individual or organization" [4]. Kagan's view reformulates a long-standing account of collective human structures. Years ago, Schein [5] observed that "an organization is the rational coordination of the activities of a number of people for the achievement of some common explicit purpose or goal, through a division of labor and function and through a hierarchy of authority and responsibility."

However, Kagan's view suggests shared power and authority [5]. Griesel observes that the more complex stages of collaboration require participation of members in the organization of ethical guidelines for the collaborative tasks, in the selection of instruments, methods, and scientific procedures, and in the development of the "participatory design" [3]. Collaboration built on Griesel's observations requires shared decision making and permits the possibility of redefining or reworking the project, mission, or activity in response to changing conditions. Implicit in this sort of collaboration is the notion that the "common ground" of the collaborators be continually reviewed and reshaped so as to serve the overarching goal or goals of the participants in the collaboration.

Schein's view [5] is certainly more traditional in that it stresses a hierarchy. Schein's position identifies a "hierarchy of authority and responsibility." Implicit in Schein's notion of collaboration is that participants do not necessarily meet as equals to achieve the tasks Griesel identifies as necessary to a collaboration, for example, organize together the ethical guidelines that will determine conduct among the participants. As such, the analysis by Thagard [6] is helpful. He argues "that there are at least four different levels of collaboration, reflecting the different backgrounds and roles of the collaborators." He identifies the employer–employee collaboration, which he argues is the weakest form of collaboration; the teacher–apprentice form of collaboration, where the apprentice learns from and emulates the teacher; a "peer-similar" collaboration, where individuals with similar knowledge base, interest, and status work together toward a common end; and the "peer-different" collaboration, in which individuals with different knowledge base and background work together to achieve some goal or end [6].

Drake and Schacter [7] draw a distinction relevant to the sorts of collaborations Thagard identifies. Drake and Schacter suggest that collaboration can be dictatorial or sustainable [7]. In their understanding, dictatorial collaboration does not allow equal power sharing or shared authority. Instead, some participants in a collaborative endeavor are forced into certain behaviors. In sustainable collaboration, the participants in the collaboration meet as equals and are enjoined in the process of government for the collaboration. Thagard [8] argues, therefore, that collaboration in science demands that potential collaborators have not only a substantial knowledge of science but also considerable procedural knowledge of how to collaborate.

8.3 THE IDEAL COLLABORATOR

The ideal collaborator will have knowledge of procedural fairness if Thagard [6] is correct. The knowledge must be built into the collaboration itself and must be such that each collaborator is an effective participant in the government of the collaboration. The ideal collaborator not only has the knowledge of procedural fairness but also has the aptitude to use that knowledge. The key, perhaps, for a successful collaboration is trust among the collaborators, which procedural know-how will ensure.

Rorty [9] argues that, from the broadest perspective, collaboration presupposes trust and requires a commitment to the common good. Rocha and Miles [10] make a similar point in arguing for an Aristotelian–Thomistic approach to collaboration. They argue that the Aristotelian–Thomistic approach treats "self-regarding" and "other-regarding' preferences as ends in themselves. The upshot of this approach is that people in the collaboration are not used as instruments to serve another's end. The collaboration, then, is what Thagard would call "peer different" or "peer similar." When people are peers, they are likely to share power and authority inasmuch as they are equal participants in decision making. Thus, the collaboration is more likely to be sustainable as opposed to dictatorial.

Raza [11] also identifies trust as the most important ingredient in collaborations and believes that moral virtues are necessary to any collaborative enterprise. Raza lists altruism, empathy, and individual commitment among the necessary components of the collaborator's moral character [11]. Essential ingredients of a sustainable collaboration, according to Raza, include effective communication, establishment of minimum goals and objectives, shared and assigned responsibilities, rules and norms for sharing and handling data as well as other information, shared responsibility for writing and publishing together, and disclosure and settling of financial interests [11]. Of course, Raza is describing the ideal colleague or, from a superior's perspective, an ideal subordinate. The person or persons who have the traits and capabilities Raza enumerates need to be present before collaborative work can be or should be undertaken.

Another important ingredient of successful collaboration is conflict resolution. In the real world, collaborations do not always go smoothly, sailing along in a problem-free manner. Also, in the real world, a need for team leaders exists or, minimally, third parties that would aid in dispute resolution. Raza suggests that team leaders resolve disputes [11] and that they do so by talking directly with each other. However, if some sort of agreement has not been reached beforehand about how to resolve conflict, it may not matter that team leaders have the responsibility of settling disputes. Griesel makes many of the same practical points as Raza.

Griesel's "Guidelines for Ethics of Collaboration Checklist" [3] also has pragmatic suggestion. Griesel says potential collaborators ask these sorts of questions [3]: "Do the guidelines promote the overall mission? Do the guidelines allow for a positive conflict resolution plan? Are the guidelines for personal, professional, and public responsibility clearly stated? Do the guidelines encourage freedom of choice? Do the guidelines allow for change and further development? Do the guidelines encourage goodwill, cooperation, and responsibility? Are the guidelines democratic?" Were a potential collaborator to ask these questions, the likelihood of a successful collaboration would be enhanced. In other words, despite the intentional goodwill, collaborations do fail.

One way that collaborations fail is by exclusion. The modern women's movement has pointed out what would be a fatal flaw for the notion that scientific inquiry is a collaborative activity. Many feminist thinkers, particularly Harding [12], have argued or implied that the activity of science excludes women, can hardly be called inclusive, and, to that extent, is not collaborative or cooperative at all. Harding suggests, for example, that feminists "would have to reinvent both science and theorizing itself in order to make sense of women's social experience" [13, p. 251].

At the heart of this sort of criticism is that science and the scientific community have not been collaborative inasmuch as they have not been democratic and open to all. While data may suggest otherwise, for instance, that by 2008–2009 women earned 60% of the doctorates in the social sciences, 70% of the doctorates in the health sciences, and 51% of the doctorates in the biological and agricultural sciences, as reported by Bell [14], the criticism shows what could be a genuine problem for the ethics of collaboration, namely, exclusion. Discounting an individual for a collaborative activity on the basis of sex is simply wrong.

Regardless of the merits of the feminist arguments "against" science, the feminist critiques do point to the necessity of stepping outside the scientific community and appraising it for inclusion. The requirement demands an ethical appraisal and constant assessment of the scientific community and the scientific project. Or, as Thagard [8] argued, people in the sciences, or any discipline, who wish to collaborate need to have some procedural knowledge of how to collaborate.

The procedural knowledge should certainly involve cultural awareness. If communication is necessary for full accountability, visibility, and transparency, then individuals involved in a collaborative activity should be culturally knowledgeable of differences in communication and communicative practices. The demand for this knowledge is reciprocal in that each individual assumes the responsibility for acquisition of knowledge related to other participant's culture.

Disciplinary communication can also hinder collaborative activity. If we look at the peer-different collaboration proposed by Thagard [6], we see that often the difference is in education and discipline-specific knowledge. For instance, different disciplines have different ideas about what constitutes a standard of proof, what laboratory practices are customary, or how observations are expressed. Again, it is imperative that the assorted difficulties that might arise in a peer-different collaboration be addressed at the start of the collaborative activity.

Furthermore, individuals themselves communicate differently. People speak with different inflection, different vocabulary, different gestures, and so on. If the heart of any collaboration is trust, as several people noted above, then each individual has the responsibility of encountering other individuals with good will and commitment. The responsibility means that the individual must care about the other people so that care about what they say becomes natural.

Within the field of chemistry, knowledge and advancement were classically (150 years ago) shared through society meetings wherein a single presenter would invite peers and nobility to an exhibition/seminar on his or her particular research. The lecturer would expound on findings within his or her (generally) laboratory and its implications to society. While the audience would include laypersons and scientists alike, so few sufficient laboratories were in existence during these times that to duplicate a presenter's findings would not be trivial and attribution of the findings would indeed directly go to the presenter.

In today's culture, achievements and advances in science are also shared via lectureships and conferences; the primary difference is that there is a greater likelihood that members of the audience include scientists whose laboratories could quickly reproduce the presented findings. To present cutting-edge research that has yet to be published requires a modicum of trust in the audience and a moral minimalism in the behavior of the audience. Discussions among like-minded research groups that begin at a conference can often lead to future collaboration provided cooperation and attribution are appropriate.

One consequence of trust and the impulse toward greater awareness of cultural, disciplinary, and personal communication is a more inclusive team of collaborators. People will not be discounted for mission-irrelevant criteria. The result is a higher level of procedural competence and the application of procedural knowledge.

Of course, most collective, organized activities do not reach the level of collaboration or at least the most complex arrangements, that is, peer-similar or peer-different, sustainable collaborations. Most organizations have elements of teamwork and are to some extent cooperative. Most organizations produce coordinated action. However, most organizations do have hierarchical arrangements where power and authority are not nearly equal among the organization's members. To that extent, the organizations may be dictatorial, as Drake and Schacter [7] put it. Yet, collaborations, the most complex level of collective activity, appear to be the most productive—or so Rozenberg and Birdzell [1] suggest. For collaborations to exist, though, the laundry list of virtues identified by Raza [11] must be present in the character of the participants. Only then can the conditions allowing trust be possible. The "ground floor" of the steps from teamwork to collaboration begins with some commitment to cooperate with others.

While the reality of the typical organization is hierarchical, the literature points to the necessity of shared power and authority where the members of a potential collaboration have a strong voice in the guidelines governing the group. It would seem, therefore, that organizational leaders would be wise to share their power and authority. Raza [11] states that team leaders need to build and maintain trust, promote respect, accommodate needs, respond to the needs of the collaborators, empower team members to discharge responsibilities, appreciate individuals at whatever level they might be, respect disagreement, and considerately provide feedback when there is conflict [11]. The likely result, according to Raza, is that the team leader will elevate what might only be coordinated activity to become a sustainable collaboration [11].

Raza's attributes of a team leader are, of course, the attributes of an ideal "boss." The ideal boss, or manager, is virtuous and respects the dignity and person of his or her workers. Again, the character of the manager must be virtuous in that the leader approaches others with good will, accords respect, and is "other-regarding" and "self-regarding" as well as "organization-regarding." Only then can trust flourish in a hierarchical organization and that organization move toward a collaborative enterprise. The basic building block of trust and the ensuing cooperation it enables lead to greater efficiency and productivity characteristic of collaborations.

Nonetheless, even with trust present, certain aspects of collaboration are more prone to problems.

Any form of collaborative effort is faced with ethical issues. However, when a collaborative effort involves the use of information systems and information technologies, there are additional considerations that collaborators must address. The following section outlines the issues that arise when using information systems and information technologies.

8.4 INFORMATION TECHNOLOGY ISSUES

There are four broad categories of ethical issues related to information technology and information systems. Mason's [15] seminal work identified the following four categories: privacy, accuracy, property, and accessibility. Subsequent work, including that by Johnson [16], has elaborated on these themes and sometimes used different terminology. The following sections examine each of these categories and how they relate to collaborative efforts in pharmaceutical research.

8.4.1 Privacy

Privacy can mean several things, but in the context of pharmaceutical research, the aspect of privacy dealing with information about individuals is most relevant. In collecting data for pharmaceutical research, particularly for clinical trials, personal information may be collected and stored in a database. A major problem in collecting data is ignoring human subject requirements [17]. Research collaborations need to ensure that data are kept secure; many systems do not provide security beyond a user name and password.

8.4.2 Accessibility

The issue of accessibility is related to the idea of privacy and considers what information is accessible to whom, under which conditions, and with what safeguards [15]. By its nature, collaboration involves information sharing. The study by Martinson et al. [17] identified two problematic behaviors that relate to information access: unauthorized use of information for one's own research and failing to present data that are inconsistent with one's own research. Unauthorized use of information could happen when one researcher collects information for the collaborative effort and uses it for another purpose. Similarly, a researcher could collect data for the collaborative effort and then withhold it from the group. A researcher could do this to try to keep the data for his or her own research or because it conflicts with his or her research.

8.4.3 Accuracy

Accuracy refers to the authenticity, fidelity, and correctness of the data [15]. Martinson et al. [17] found the most often reported problem in scientific research is falsifying or manipulating research data. They also found that overlooking other researchers' use of flawed data was a common problem. A researcher might do this to make sure that the collaborative effort results in significant findings. A researcher might choose not to inform collaborators that the data or the technique used to interpret the data are flawed. Techniques used to collect data and even the sample population may be inappropriate.

8.4.4 Property

The question of property concerns ownership of the information and what constitutes fair exchange [15]. Collaborative research agreements should spell out who owns the research data and whether it may be used for any other purposes. GlaxoSmithKline (GSK) was sued about its patent claim on a drug (see following case) because it received some federal funding [18]. Large pharmaceutical companies are partnering with various external organizations, including academic institutions, which can lead to confusion regarding ownership rights [19]. Traditionally, pharmaceutical companies have contracted with individuals to perform specific tasks and retained ownership since these were works for hire [19]. Now teams of academic researchers are working with teams of company employees, and issues of ownership are unclear.

8.4.4.1 Case Study: GlaxoSmithKline and AZT Intellectual Property Rights This case illustrates two possible issues related to collaborative pharmaceutical research: (1) ownership of intellectual property rights and (2) jurisdictional issues of international research efforts.

GlaxoSmithKline owns the intellectual property rights to azidothymidine (AZT). AZT was developed to help treat the symptoms of HIV/AIDS. It has been effective in increasing the life expectancy of infected persons and is often prescribed as part of a "drug cocktail." GSK owns the rights to use AZT as an anti-HIV treatment and also has a process patent protecting the technique by which AZT is produced [18].

In 1994, the AIDS Healthcare Foundation argued that GSK scientists were not the sole inventors of the drug. Some arguments against GSK's ownership contended that GSK should not have exclusive rights over AZT because the drug was developed with the help of government-funded research [18]. The U.S. Court of Appeals ruled that GSK employees were the sole inventors of the drug and the government had no right to share in the credit for developing the drug [18].

The same issue was also heard by the Supreme Court in Canada. The Canadian court reached the same conclusion as the U.S. courts. However, the fact that GSK had to defend its rights in multiple countries demonstrates the problems of international pharmaceutical research. What if the ruling had been different in another jurisdiction?

8.5 CONCLUSIONS

Sustainable collaborations, where power and authority are shared, require virtuous character traits among the participants in the collaboration. The individuals should be other regarding as well as self-regarding, empathetic, and committed to the collaboration as well as the goal or goals of the collaboration. Essential ingredients of a sustainable collaboration include effective communication, especially in international settings; clear goals and objectives mutually agreed upon; shared and explicitly defined responsibilities; rules and norms for sharing and handling data and other information; shared responsibility for writing, publishing, and other tasks associated with collaboration; and disclosure and settling of financial interests.

Recognizing that collaborations at times run into problems, a method of resolution must be in place prior to any dispute, large or small. The optimum resolution to any problem would, of course, be found among the collaborators themselves, for then trust would be enhanced. Also, the conflict or problem would be solved at the lowest level.

Organizations that wish to have collaborative activities and engagement must also take action or adopt policy that fosters trust. The key for such organizations is to value relationships while being other regarding and not merely concerned with the bottom line (in whatever form an important or final goal may take). The conditions of trust are likely to be established.

The leaders in an organization that desires collaborative activity, internal or external, must be willing to share power and authority, trusting subordinates to discharge their collaborative responsibilities. Organizational leaders need to create the conditions that enable trust, respect the dignity of subordinates, respond to the needs of the collaborators, provide resources for team members to engage the work of the collaboration successfully, appreciate individuals at whatever level they might be, allow disagreement without being disagreeable, and considerately provide guidance as needed.

Were organizations, scientific or otherwise, to establish sustainable collaborations, not only would the participants in the collaboration have enhanced opportunities for individual growth but the organization would likely experience a growth-enhanced bottom line [20–22].

REFERENCES

1. Rosenberg N, Birdzell LE, Jr. Science, technology, and the western miracle. *Sci Am* 1990;263:42–54.
2. McGowan GJ, McGowan RJ. Attribution, cooperation, girls, and science. *Bull Sic Technol Soci* 1999;19:547–552.
3. Griesel P. Ethics of collaboration: A quest for guidelines. 1992. Available: www.eric.ed.gov/ERICWebPortal/recordDetail?accno=ED360235.
4. Kagan SL. *United We Stand: Collaboration for Child Care and Early Education Services*. New York: Teachers College Press, 1991.
5. Schein EH. *Organizational Psychology*. Englewood Cliffs, NJ: Prentice-Hall, 1965.
6. Thagard P. Collaborative knowledge. *NOUS* 1997;31:242–261.
7. Drake MJ, Schacter JT. A virtue-ethics analysis of supply chain collaboration. *J Business Ethics* 2008;82:851–864.
8. Thagard P. How to collaborate: Procedural knowledge in the cooperative development of science. *Southern J Philoso* 2006;44(Suppl):177–196.

9. Rorty A. On being rational. *Ratio* 2009;22:350–358.
10. Rocha H, Miles R. A model of collaborative entrepreneurship for a more humanistic management. *J Business Ethics* 2009;88:445–462.
11. Raza M. Collaborative healthcare research: Some ethical considerations. *Sci Eng Ethics* 2005;11:177–186.
12. Harding S. *Whose Science? Whose Knowledge?* Ithaca, NY: Cornell University Press, 1991.
13. Harding S. *The Science Question in Feminism*. Ithaca, NY: Cornell University Press, 1986.
14. Bell N. *Graduate Enrollment and Degrees: 1999 to 2009*. Washington, DC: Council of Graduate Schools, 2010.
15. Mason RO. Four ethical issues of the information age. *MIS Q* 1986;10:5–12.
16. Johnson D. *Computer Ethics*. Upper Saddle River, NJ: Prentice-Hall, 1994.
17. Martinson BC, Anderson MS, de Vries R. Scientists behaving badly. *Nature* 2005;423:737–738.
18. Gewertz NM, Amado R. Intellectual property and the pharmaceutical industry: a moral crossroads between health and property. *J Business Ethics* 2004;55: 295–308.
19. Clyde-Watson Z. Pharma deals prompt ownership questions. *Manag Intellect Property* 2008;184:63–65.
20. White TI. Business, ethics, and Carol Gilligan's "Two Voices". *Business Ethics Q* 1992;2:51–61.
21. Gilligan C. *In a Different Voice*. Cambridge, MA: Harvard University Press, 1982.
22. Bartlett C, Ghosal S. Changing the role of top management: Beyond strategy to purpose. *Harvard Business Revi* 1994;Nov/Dec:9–88.

9

INTELLECTUAL PROPERTY ASPECTS OF COLLABORATION

JOHN WILBANKS

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.

9.1 BACKGROUND ON INTELLECTUAL PROPERTY RIGHTS

Intellectual property rights (IPRs) play a central role in modern biology and its related businesses. Negotiating the disposition of IPRs in research is an essential element of most collaborations and funded projects and is the subject of an extensive literature. However, "intellectual property" (IP) is in fact a wide variety of disparate forms of protection and exclusive rights which apply in different ways at different points in the scientific research cycle and the business value creation and capture cycle.

This chapter begins with an introduction to the most common forms of IP—copyright, patent, trademark, and trade secret—as well as a brief treatment of the relationship of how IP affects data and databases. The second section of the chapter looks at the key transactional elements of a collaboration, including materials transfer, patent licensing, and the way that those elements can affect a negotiation. The third section of the chapter provides some pragmatic resources for simplifying negotiations, reducing transaction costs, and amplifying discoverability for materials and inventions created in the course of a collaborative research arrangement.

9.2 SECTION I: INTELLECTUAL PROPERTY RIGHTS

Intellectual property rights are inspired by traditional property rights—the idea is that, just as one can own a hectare of land and maintain exclusive rights to live on or develop that land, one can also own something less tangible, like a discovery, or a method, or an expression, or a symbol, or a piece of music. IPRs are based, at least in part, on the economic principle that the provision of exclusive rights in these intangible assets creates a financial incentive to create more of the assets or develop them more completely through research and development [1].

IPRs cut through the life sciences at multiple points in the research cycle. When a scientist takes notes in a laboratory notebook, he or she is creating a copyrighted work—but is also fixing some key elements of work that might be used later to prove an invention as part of a patent application. The scientist may treat that lab notebook as a trade secret until he or she is ready to publish and the data on which the research rests are subject to a complex and internationally patchy set of laws and regulations. This section makes a brief introduction to the major elements of IPR and attempts to place them in the context of collaborative biomedical research.

9.2.1 Copyright

Copyright is the set of exclusive rights granted to the author of a new creative work, such as a song, photograph, blog post, or software. The primary rights held by the author are the right to copy, distribute, and adapt what they have created. Copyright gives the creator the legal right to control copying—to prevent others from using those rights to copy, distribute, and adapt. The copyright is an exclusive right that both creates the author's power to copy and distribute and allows the author the power to prevent anyone else from doing so [2].

In the life sciences, we see copyrights most clearly in the scholarly publishing industry. Whether writing a journal article or a textbook, the text emerging from biomedical research is a clear example of a copyrighted work. The importance of copyright here is primarily in the transaction between scientist and publisher, as traditional scholarly publishers have developed business models that depend on the transfer of copyrights from authors to journals. The journals then use the exclusive right to the articles to sell copies of the journals and to prevent anyone from copying those journals without permission [3].

But copyrights cover far more content than the articles. Laboratory notebooks, e-mail, meeting notes, journal club reports, powerpoint presentations, conference posters, abstracts, and more, all carry an automatic copyright, as do many expressions of underlying data (especially data rendered in photographic or video forms).

Due to the continuing expansion of the reach and lifespan of copyright [4], combined with the explosion of digital communications, life scientists create copyrighted materials at a remarkable pace. Though there is tremendous potential to publish and share these materials, the default position of copyright makes the reuse of these materials an infringement in the absence of a license—a positive grant of rights to users to make and distribute copies. Thus, at the very moment we have the technical capacity to capture, store, and publish the intermediate literature that postal delivery rendered inefficient economically, the defaults of copyright make achieving that goal complex. This default position is one that lasts a very long time.

Although copyrights are not permanent, their lifespan is tied to the date of their creation, and after a time (defined by national laws and differing country by country) the copyright expires and the underlying work passes into the "public domain." Copyright lasts for 50–100 years after the death of the author. In science, that is the time difference between a world in which we have the core theory of DNA and one in which we do not.

Copyright is subject to what is known as the "idea–expression dichotomy," which is a complicated way of saying that copyrights govern the expression of an idea, not the idea itself [5]. Thus, Watson and Crick own the copyrights over the words they chose to use to explain the structure of the DNA molecule, but not of the idea or the facts they described. A user might come along and take the core ideas of an article and use them without proper citation or attribution and not infringe copyright as long as that user did not copy the actual phrasing. The user would be a plagiarist, but that is a matter for science and not the courts.

The author is automatically the initial owner of the copyright in an original work of authorship as soon as the work has been fixed in a tangible medium of expression [6]. Since facts and ideas are not copyrightable, the results and underlying data reported in an article are facts that are not subject to copyright. Similarly, the insight or idea leading to an experiment is also not subject to copyright. In the case of journal articles, the copyright applies to the author's creative expression, such as the choice of text to describe materials and methods, an experiment, or its result. Tables, figures, charts, or other accompanying material are copyrightable only if some minimally creative decisions were required in their design.

Once the copyright vests in the author, he or she can authorize others to use the work in one of four ways: (1) assign the entire copyright, (2) grant an exclusive license, (3) grant a nonexclusive license, or (4) dedicate the copyright to the public domain. An author must sign a written document to effectively assign the copyright or grant an exclusive license. In contrast, a nonexclusive license or permission can be granted quite casually. A verbal okay or even conduct, such as posting a work on a publicly accessible Web server, can be deemed to be the grant of a nonexclusive license [7].

9.2.2 Patent

The patent is a totally different property right from copyright. A patent is a set of exclusive rights granted to a person who "invents or discovers any new and useful process, machine, article of manufacture, or composition of matter, or any new and useful improvement thereof" (http://www.uspto.gov/web/offices/pac/doc/general/what.htm), but there is a catch. The inventor has to apply for the patent; it is not automatically granted, and the rights are granted as a trade, in exchange for the inventor publicly describing the invention in enough detail that an ordinary expert in the field could make use of it [8].

Like copyrights, patents are not permanent. They expire after a much shorter period of time than copyrights, 20 years in the United States. Unlike copyrights, they are nationally based, which means that a patent granted in the United States is of no value in China as the inventor must file patent applications in multiple national jurisdictions to exercise the exclusive rights.

Patents are subject to statutory requirements as well. A patented invention must be novel. One cannot repatent an invention already patented or one that was disclosed publicly before by another party. A patented invention must be "nonobvious" (this clause, predictably, creates a lot of debate in the life sciences). A patented invention must also have some sort of use or application. If an inventor can satisfy these three requirements, then he or she receives the right to prevent anyone else from making, using, or selling the invention [8].

This is important in the life sciences, because the exclusionary nature that the patent right carries for the inventor is innately different from the positive rights copyright gives their creators. A scientist may acquire a patent which is the right to exclude someone from some activity but not actually possess the rights to practice the patent because of other preexisting patents that "block" implementation. Very few patents are untouched by other patents, as the vast majority of innovation is incremental.

Richard Jefferson of Cambia (http://www.cambia.org/daisy/cambia/home.html), a thought leader in open biology, explains with a simple example. If Richard has patented a four-leg chair as well as the idea for chairs, it does not block me from patenting a three-leg chair, but it does prevent me from manufacturing and selling that three-leg chair, because I do not have the right to sell chairs at all without a license to Richard's patent. I in turn can block him from selling a three-leg chair because I own the improvement patent. This reality means that patents exist in a world of complicated claims, competing ownership, and lengthy negotiations, one in which transparency in patent landscapes is almost impossible [9]. We will return to this issue in the licensing section, as it has real effects on the transactions around patents.

9.2.3 Trademark

A trademark is perhaps the least relevant aspect of IPR to the life sciences, as it serves primarily to identify that some form of goods, products, or services actually come from the owner of the mark. It is a quality indicator or an indicator of brand differentiation; for example, think of the Nike swoosh, the Coca-Cola bottle, or iconic Apple logo [10].

Trademarks might enter into collaborative discussions in the life sciences in the naming of new entities or in the branding of drugs versus their generic equivalents. Ambien™ may compete against generic zolpidem, but some consumers prefer to trust the quality control processes that are associated with the Ambien trademark. Trademarks can be words, logos, sentences, drawings, and other elements.

9.2.4 Trade Secret

Some kinds of property are not amenable to the patent or the copyright system and are thus well suited to trade secret regimes. A good example of this is the recipe for Coca-Cola or Kentucky Fried Chicken. Both have been kept secret for decades, protecting the respective company's interest.

Examples of the kinds of property subject to trade secret are formulas, practices, processes, designs, instruments, patterns, and information. Trade secrets typically carry three core qualities: First, they are unknown to the general population; second, they give the owner of the secret an economic advantage (and importantly, the advantage must be connected to the secrecy); and third, the owner has to try to keep it secret [11].

Trade secrets are a major class of IPRs, although they rarely enter the conversation in the "open" IPR context, because a trade secret exists, by definition, only if it is kept secret. Trade secrets and sharing, at least open sharing on the Internet, do not coexist. Trade secrets in the life sciences could include a database of compounds used for investigatory purposes (whose compositions of matter would enter the public domain if published before patent claims were filed) or preliminary laboratory findings ahead of publication.

9.3 INTELLECTUAL PROPERTY RIGHTS AND DATA

Data do not naturally fall under the four classes of IPR previously listed, especially the sorts of data now emerging from contemporary high-throughput life sciences laboratories. Measurements of cell activity, lists of gene sequences, and protein crystal structures are not "expressions of creativity" but instead purport to give us facts about the real world. They are not inventions or processes. Data in the United States have a long tradition of being in the public domain and of being protected at the database level by only a very thin layer of copyright. A database may carry a copyright only on the elements of the database related to the "selection and arrangement" of the data contained therein and not on the data itself [12].

However, the U.S. approach is not an international standard. There are varying laws in various countries that create protections similar to IP for elements, like data, that do not fit the traditional regime, and indeed the European Union (EU) does precisely this for databases. To be fair, the U.S. law protects ship hulls, the French protect fashion, and the "mask works" used to design semiconductor chips receive protection as well.

But the EU database directive, as it is known, is the most relevant to the life sciences discussion. The directive was passed in 1996 in the goal of creating a regional advantage for databases built in the EU. The holder of the database rights can prohibit the extraction or reutilization of all the data in the database or of a "substantial" subset of the data. Substantial can be calculated either objectively or subjectively, and the owner is not allowed to restrict the users who make "insubstantial" uses of the database [13].

EU database rights last for 15 years, dated from the publication of the database, but can be renewed through substantial new investments in the database. This means that any interesting database, one that is growing, is legally allowed to have perpetual database rights under the EU system. To make things more complex, database rights are orthogonal to and independent of copyright, making it possible for different owners to assert copyrights and database rights to the same product.

9.4 LICENSING AND CONTRACTS

9.4.1 Material Transfer Agreement

A material transfer agreement (MTA) is a contract covering the transfer of physical research materials in which the transfer is intended to facilitate research purposes at the receiving laboratory. MTAs are commonly associated with biological materials, especially recombinant ones like stem cells, plasmids, or model organisms, and the contract serves to delineate the rights and obligations of both parties to the transfer. MTAs can also be used for nonrecombinant materials like spinal fluid, tumor, blood, and other tissue samples as well as for chemicals and other technologies that may assist in drug discovery [14].

Negotiating an MTA can be a lengthy process. In the university-to-university setting, estimates range from delays of over a month for between 11 and 16% of requests, "a substantial delay in a fast-moving field," to estimates that there are routine delays of over six months for 20% of requests and over two months for 42% of requests. Studies also show increasing rates of outright denial of requests and abandonment of "promising research projects" because materials are not received. In the commercial–university arena, with no standardized agreements at all, most observers believe the situation is worse. Commercial–academic denial rates are estimated to be nearly twice those in the academic–academic context (33 vs. 18%) [15].

Scholarly and empirical research shows a range of estimates. The lowest estimate is that one in six suffer a delay of more than one month—"a substantial delay in a fast-moving research field" [16]. The higher estimates (from a large survey conducted by the American Association for the Advancement of Science [AAAS]) are that 25% of those researchers seeking materials through MTAs suffer delays of one to two months, 23% two to six months, and 19% over six months. Most of the literature falls between those extremes, but closer to the second estimate. Articles published by technology transfer officers and interviews with scientists offer support for negotiation delays being significant and widespread [15].

Separating the purely legal negotiating delays from other delays in the MTA process is difficult, however. The lowest estimate on this factor was 11% suffering delays of more than a month—this amount of time was assessed as "a substantial delay." Technology Transfer Office (TTO) estimates and other surveys here suggest that legal/licensing delays of two to six months are routine. The AAAS figures also support this conclusion. Finally, some requests are not simply delayed. They are denied. Forty-seven percent of all academic geneticists who had asked "other faculty for additional information, data, or materials regarding published research reported that at least one of their requests had been denied in the preceding 3 years" [17]. Systems attempting to address these delays have been put into place by universities, funders, and nonprofit organizations and will be addressed in the final section of this chapter [15].

MTAs take three forms—academic to academic, academic to industry, and industry to academic. The differences in delays noted in the literature above come from the kind of elements found in the relevant MTAs to each transfer, with industry to academic often being the most difficult to negotiate.

Academic-to-academic transfers are in many cases quite simple, with some major research institutions (such as Stanford) and some biobanks (such as the Jackson Laboratories mouse facility) doing away with "outbound" MTAs altogether in lieu of norms or website terms of use. And for those not quite ready to do away with MTAs for academic-to-academic transfer, the Uniform

Biological Material Transfer Agreement (UBMTA) and Simple Letter Agreement for the Transfer of Non-Proprietary Biological Material are widely available and widely used standardized tools. For institutions that have signed the UBMTA master agreement, materials can be transferred under the terms of the UBMTA upon execution of an implementing letter for the particular transfer. Often the problem in academic-to-academic transfer is not legal, but instead one of the incentives, opportunity costs, and not-infrequent issue of competitive withholding [18].

Many interesting collaborations, however, involve a company and an academic laboratory, necessitating the existence of more complex MTAs depending on whether industry is the source of the material or its recipient. Industry-to-academic MTAs are widely reported to be the most complex and time-consuming contracts to negotiate by technology transfer offices and regularly contain requests that "reach through" the research to future rights on technologies that might be invented using the material or requests to create embargo periods on publication while the research is reviewed for commercial value and/or patentability. Other frequent elements of industrial MTAs are nondisclosure agreements, restrictions on redistribution of the material outside the specified laboratory, or restrictions on allowed fields of use.

Not surprisingly, many inventions are thus encumbered by rights imposed earlier in the research life cycle by MTAs. A classic example is the so-called Golden Rice, a genetically engineered rice variety aimed at ending vitamin deficiency disorders in the global South. After all the parties involved in creating Golden Rice decided to give away the variety, studies indicated that 44 patented products or processes and at least 15 materials, many of which were governed by MTAs, were potentially used in its development. The intellectual and technical property landscape surrounding Golden Rice has reported that the unfair use of one MTA had been particularly problematic [18].

9.4.2 Patent Licensing

Licensing of patents is often a major part of negotiating collaborations, or, if not licensing the patents, then deciding on key elements of how any patents emerging from the collaboration will be handled. A patent license grants certain rights to practice a patented invention from the owner to the recipient. The license lays out the rights granted, the freedoms given, and the requirements imposed and is a conditional grant, that is, if the licensee does not comply with the requirements and obligations, then the right to practice is nullified [19].

Collaborative research in the life sciences often starts at an early stage, before any patents are filed. Thus the negotiation will often be over ownership—who will own the patent? The university or the company and what rights will be licensed from the owner to the other parties and under what terms?

Two of the most common moving pieces in patent licensing are revenue elements and field-of-use restrictions. Revenues can be dealt with in multiple ways, from one-time fees to annual fees to annual royalties based on a percentage of revenues received by the licensee. If a licensee failed to pay revenues as promised, then the license could be terminated by the patent owner. The conditional license is a powerful tool. Patent licenses can also contain performance obligations, but these can be harder to negotiate at very early stages before the patent itself is filed.

9.4.3 Tools and Systems

The slowdowns imposed by materials transfer negotiation and by complex patent licensing negotiation have created some demand for standardized systems. However, those systems remain in the early stages. Three examples are presented here: a legal system covering both MTAs and patent licenses intended to facilitate industry–academic collaboration, an e-commerce system intended to address discoverability and cataloging of innovations, and a software system to provide a "wizard" for negotiating university–industry collaboration.

9.4.3.1 Creative Commons—MTAs and Patent Licenses Creative Commons, a nonprofit organization that is famous for providing copyright licenses that facilitate voluntary sharing, has developed a set of new MTAs that use modular contract options to promote the development and evolution of standard MTAs for transfers between academia and industry. The new MTAs published by Creative Commons provide for a more flexible range of options while at the same time adhering to the core guidelines of rapid transfer, low transaction costs, and increased research use advocated by the National Institutes of Health (NIH) guidelines on access to research tools [20].

Creative Commons MTAs distinguish between activities for internal use and commercialization, and they do not provide options that restrict publication or that contain reach-through royalties, grant backs, commercialization options, or other obligations with regard to downstream inventions made by the recipient. Creative Commons has developed a simple, Web-based user interface that guides a user through key considerations and options associated with selecting a particular MTA [21] and has already accomplished the integration of the MTAs into online systems such as the iBridge Network for use on life sciences research tools. These material transfer agreements are already available and in use for research materials such as stem cells, for uses ranging from the narrow (neurodegeneration field-of-use restraints) to the wide (commercial and clinical use rights granted in advance).

The patent licenses developed by Creative Commons are the second component of the legal system. First, to promote basic research, the patent owner commits to nonenforcement of patents against users engaged in basic nonprofit research ("the research nonassertion pledge"). Second, the patent owner may provide a public license offer to enable use of specific patents chosen by the patent owner for applications beyond nonprofit research. In return, users may be asked to provide usage metrics, patent marking, and attribution and pay a fee or royalty as determined by the patent owner [22].

The research nonassertion pledge is a commitment by a company not to assert its patents against anyone (whether an individual scientist or an institution) for engaging in activities related only to qualified nonprofit research. This is intended to enable the kind of research that takes place at universities and other academic, government, and nonprofit institutions. Patent owners are encouraged to consider adopting this policy widely to cover all their patents. However, a patent owner has ultimate control over what patents to dedicate to the research nonassertion pledge. Therefore, a patent owner may pledge all or some of its patents under the research nonassertion pledge. It can do so by providing a list of patents to be included or conversely by stating an overall policy but listing patents excluded from the pledge.

To promote use of sustainable technologies requires adoption and application of the technology in the real world, which goes beyond basic research. Therefore, a license to use the technologies in real-world applications, including some commercial ones, is needed. Companies wishing to provide such licenses may either adopt the Creative Commons model patent license that is available through the innovation exchange or offer their own. The criteria for inclusion of a patent license offer in the exchange, whether based on the model patent license or a custom one, is that the terms and conditions of the license must be fully specified and made publicly available and the offer must be a valid offer that can be accepted by anyone without the need for further negotiation. These two criteria constitute what Creative Commons calls a "public license offer," the primary benefits of which are its transparency and low transaction costs.

The transparency and nondiscriminatory terms permit entrepreneurs and potential adopters to plan and make early decisions about technology adoption, while the low transaction costs allow a patent owner to offer a license widely without incurring large administrative and negotiation costs. These two factors can bridge the otherwise difficult and unpredictable license negotiation process, which can deter technology adopters or trigger economically wasteful design-around attempts. The history of innovation highlights the value of the unanticipated uses of a technology. Both the research nonassertion pledge and the public license offers are options for not only promoting and widely disseminating technologies that can solve global sustainability challenges but also stimulating interest in broader or unanticipated applications of existing technologies.

Owners do not abandon their patents under the exchange; they retain the ownership of the patent rights, with full ability to enforce the terms of the license or nonassertion. If the user is found to violate the conditions of the license, the patent owner retains the full right to pursue remedies and damages under the law.

9.4.3.2 Kauffman Foundation—iBridge Network The iBridge Network is a Web-based mechanism for the dissemination of innovations such as research results and reports, computer software and other copyrighted works, biological research materials, and patented inventions. It is implemented as a database with Web interfaces and electronic commerce capabilities. Providers are predominantly universities and their individual researchers, although federal laboratories and for-profit organizations were also encouraged to contribute. Intended adopters are other researchers and entrepreneurial individuals, groups, and organizations [23].

The model underlying the iBridge Network is that aggregating research results and inventions from multiple institutions and establishing simple ways to search and transact will increase the flow of innovation to entrepreneurial actors for further development, application, and delivery to society. In operation, researchers and their institutions post descriptions of their results and inventions on the iBridge Network site and set terms of transfer. Prospective adopters search for and review nonconfidential summaries of available innovations, agree to terms of transfer, and download an item directly from the site (in the case of electronic media such as software, data, or reports) or arrange for delivery from the provider (in the case of biological research materials or other tangibles). At the option of the provider, innovations may be acquired without special terms or may require acceptance of an electronic license and may be without charge or fee based. Any provider charges are payable by e-commerce mechanisms, but the Network itself is a nonprofit enterprise.

The iBridge Network goes beyond earlier experiments in expediting technology transfer in several ways:

- Greater emphasis on one-to-many transfers (nonexclusive licenses)
- Aggregation of innovations from multiple research institutions
- Permission and functionality for transactions directly between an innovation provider and the network and directly between the network and an adopter (disintermediation)
- Options for fee-based and electronic license-based transactions
- Design principles and functionality determined by input from traditional university technology transfer offices.
- Management as a not for profit

The iBridge Network's website, www.iBridgeNetwork.org, facilitates linkages between universities, industry, and entrepreneurs. On the website, users browse research or use the search tools to quickly and accurately find research related to specific topics. Tools include bookmarking, note taking, tagging, and smart searches. Users contact universities directly about the assets or, if the e-commerce function is available, they can purchase and license the product right from the site. Existing university online metrics and data systems can be easily integrated, allowing participation without more work. Universities post their research using an online form or "plug-and-play" integration with existing database tools, enabling them to post more research.

9.4.3.3 University–Industry Demonstration Project—TurboNegotiator The University–Industry Demonstration Project, or UIDP, is a group of university and industry representatives brought together under the aegis of the National Academies of Science and Engineering. UIDP looks for ways to lower the transaction costs and effort of collaborative research agreement negotiations:

> One university estimates that it requires approximately 5 times the staff effort (per dollar received) to negotiate an industry contract/grant as compared to a government one. University and company data are consistent in painting contract negotiation as a process that takes, on average, approximately 70 days—but can range up to years. Regardless of which side of the negotiating table one sits, over 2 months to reach an agreement is unacceptable [24].

As a first project, the UIDP began developing the TurboNegotiator tool, which is available in pilot form as of this writing. Inspired by the popular TurboTax software, TurboNegotiator (TN) is designed to be a step-by-step "wizard" that walks collaborating parties through the key elements of a collaborative research arrangement, drawing on a database of standard legal clauses, filling in variable blanks such as revenue targets or royalty terms or fields of use, and compiling a complete agreement. TN is designed to cover IPRs, indemnification, publication rights, and more [25].

The benefits of a system like TN are easy to imagine—faster turnaround on negotiations due to rapid information exchange, harmonized expectations, and shared understanding of needs. It also helps to build capacity in negotiations where one party may be significantly less skilled or funded than the other by creating a level playing field to engage in complex issues of property.

9.5 CONCLUSION

Collaboration in the life sciences touches on international IPRs at almost every point of the knowledge creation cycle and is at the heart of the majority of contemporary business models for life sciences. Making sure that IPRs, especially patent rights, have been negotiated in advance when possible is the easiest way to make long-term commercial use possible.

However, the relentless pursuit of IPRs—again, especially patents—comes at a price. The pressure to know where patents do and do not exist, who owns them, and what must be licensed to get to market exacts a high cost on all parties, but none more so than those attempting to take drugs or foods to market that face a small market. Adopting a licensing strategy in negotiation that reserves safe harbors for rare diseases, nonmarket uses, humanitarian uses, and other uses is a good practice to implement, and one can use standardized tools such as those provided by Creative Commons for this purpose.

More good practices include making inventions, tools, and materials available via e-commerce methodologies. This increases transaction flow without increasing transaction costs, which can increase the overall likelihood of a discovery that leverages the invention, tool, or material.

It is also important to deal with copyrights and data issues in collaboration. Making the research literature emerging from a project available increases the impact of the research, but doing so often requires ensuring that the rights necessary to distribute the articles are reserved at the beginning of a project. Similarly, data and database rights can differ so much from country to country that it is easier to reserve rights to distribute internationally via public domain systems at the beginning of a project rather than at the end.

The use of standard legal and technical systems creates a more even negotiating ground between parties that might otherwise be disadvantaged due to economics, expertise, or time pressure. These systems can make it easy to reserve broad rights to the public for socially responsible uses while reserving the right to negotiate bespoke agreements and licenses for valuable inventions and discoveries. The emergence of standardized negotiating frameworks is beginning to transform those bespoke agreements as well. But nothing can replace the value of a good lawyer, in the end, when there are high stakes such as those in the life sciences.

REFERENCES

1. Boyle J. The second enclosure movement. *Law and Contemporary Problems* 2003;66:33–74.
2. Jones H, Benson C. *Publishing law*. Routledge, 2002:12–13.
3. MIT. Available: http://info-libraries.mit.edu/scholarly/faculty-and-researchers/, 2010.
4. Lessig L. Copyright's First Amendment. *48 UCLA Law Rev* 2001:1057–1074.
5. Landes WM, Posner RA. An economic analysis of copyright law. *J Legal Studies* 1989;18:325–363.
6. Copyright Law of the United States of America and Related Laws Contained in Title 17 of the United States Code, Circular 92, Chapter 1.
7. Complying with the National Institutes of Health Public Access Policy: Copyright considerations and options by Michael Carroll. Jointly released by SPARC, Science Commons, and ARL. Available: http://sciencecommons.org/wp-content/uploads/nih_copyright_v1.pdf.
8. Patents: Frequently asked questions. World Intellectual Property Organization.
9. FAQs, Patent Lens, Cambia. http://www.patentlens.net/daisy/patentlens/search/faq-plens.html.
10. United States Patent and Trademark Office. Trademark basics. Available: http://www.uspto.gov/trademarks/basics/index.jsp.

11. Rockman HB. *Nature of a Trade Secret: Intellectual Property Law for Engineers and Scientists*, Wiley-IEEE, 2004.
12. Gasaway L. Databases and the law. Cyberspace law course. University of North Carolina at Chapel Hill, Spring 2006.
13. Directive 96/9/EC of the European Parliament and of the Council of March 11, 1996, on the legal protection of databases. 1996. Available: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31996L0009:EN:HTML.
14. Quick guide to material transfer agreements at UC Berkeley. Available: http://www.spo.berkeley.edu/guide/mtaquick.html.
15. Commons S. Empirical data about materials transfer problems. Available: http://sciencecommons.org/projects/licensing/empirical-data-about-materials-transfer/.
16. Walsh JP, Cho C, Cohen WM. View from the bench: patents and material transfers. *Science* 2005;23:2002–2003.
17. Campbell EG, Clarridge BR, Gokhale M, Birenbaum L, Hilgartner S, Holtzman NA, Blumenthal D. Data withholding in academic genetics: evidence from a national survey. *JAMA* 2002;287(4):473–480.
18. Nguyen T. Science commons: Material transfer agreement project. *Innovations: Technology, Governance, Globalization* 2007;2:137–143.
19. Organization WIP. To license a patent—Or, to assign it: Factors influencing the choice. Available: http://www.wipo.int/sme/en/documents/license_assign_patent.htm.
20. Creative Commons Materials Transfer Project. Available: http://sciencecommons.org/projects/licensing.
21. Chooser CCMT. Available: http://mta.sciencecommons.org.
22. Project CCP. Available: http://sciencecommons.org/projects/patents.
23. iBridge Network, a Kauffman Innovation Network project. Available: http://ibridgenetwork.org.
24. Mayo MJ. University-industry demonstration partnership. Available: http://sites.nationalacademies.org/PGA/uidp/PGA_049074.
25. University-industry demonstration partnership. Available: http://sites.nationalacademies.org/pga/uidp/index.htm.

PART II

METHODS AND PROCESSES FOR COLLABORATIONS

10

SCIENTIFIC NETWORKING AND COLLABORATIONS

Edward D. Zanders

10.1 INTRODUCTION

Worldwide investment in biological research, although slowed down by recent economic pressures, continues its upward growth trajectory. This is due in large part to the need for better understanding and treatment of disease as well as new environmental challenges such as the development of biofuels for energy

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.

production. The result of this investment is an ever-increasing number of scientists who publish in an expanding number of journals and online media. Furthermore, countries such as Brazil, China, and India are now becoming significant players in the biomedical research arena, challenging the predominance of the United States, Europe, and Japan. The astonishing increase in efficiency in communication brought about by the Internet has opened up new possibilities for networking across this new global landscape. This chapter will examine how biomedical networking has evolved from scientific academies and personal correspondence to networking websites and data resources on the Internet. It will also provide some examples of how recent initiatives in computer-aided networking in the drug discovery arena may herald changes in the management and dissemination of pharmaceutical discoveries.

10.2 HISTORY AND BACKGROUND OF SCIENTIFIC NETWORKS

The year 2010 marks the 350th anniversary of the founding of the Royal Society, the world's oldest scientific networking organization. Now every country with a significant commitment to scientific research has its own national academy operating along the same lines as the London institution. The noticeable feature of these academies is their hierarchical structure built from carefully selected individuals who make up an elite cadre of scientists. Networking in face-to-face meetings was (and is) not always feasible, so those people who want to communicate results, share or request samples, and coordinate multinational projects would have to use whatever tools were available to them. Until recently these tools consisted of just the letter and the telephone. Charles Darwin is an example of someone who acquired vast amounts of biological data from his correspondence with fellow naturalists and others with intimate knowledge of the natural world. Although the Victorian postal service that he used was comparatively efficient compared with today's "snail mail" (and Darwin's correspondence was prodigious), the need to communicate by letter writing inevitably slowed the development of his ideas on evolution. A modern Darwin alive today would be able to condense decades of networking into a few years, but it is interesting to speculate whether this speed would come at the expense of deep critical thought undertaken at a more leisurely pace. With a vast increase in competition through the globalization of science, quiet contemplation is a luxury we can no longer afford; the challenge therefore is how to maximize the efficiency of all the traditional and electronic networking tools that are now available to the biomedical scientist who is suffering from information overload.

10.3 ONLINE NETWORKS

Many readers of this book will have little or no recollection of the time when computers were primitive devices with no connectivity to the outside world.

This of course changed with the development of the Internet, which is now taken for granted as a communication and networking tool for transmitting e-mails, data sets, and online publications. The rollout of fast broadband connections in the early part of the new century has increased connection speeds and data volumes so that new opportunities have arisen for social and professional interactions between individuals regardless of their geographical location and time zone. We have previously discussed this in the context of social networking and drug discovery [1]. In our paper we concluded that the common theme of different networking sites (such as Facebook and LinkedIn) is one of interactive participation through contributions to discussion boards, blogs, or wikis. Personal interaction has always been recognized as vital for seeding and developing scientific ideas, but this is necessarily restricted to a limited number of people. Web-based collaboration dramatically expands the number of people that can participate in essential professional activities such as job hunting, opinion seeking, and sharing and collaboration on projects.

In the short time since the article was published, the demographic of Internet collaboration has begun to change, so that people of all ages actively contribute to, for example, Facebook, rather than just the younger generation. The time is now ripe for a further evaluation of computer-aided networking, particularly in light of new models of data sharing and scientific publication.

10.3.1 Definitions

A reasonable answer to the question “How many computers are there in the world with Internet connections?” would be “somewhere in the millions.” It has been argued, however, that the real number is one. It is the Internet itself that is the dominant operating system [2]. This concept is being developed in the form of “cloud computing,” where programs and data are all stored on remote servers to be accessed by computers that are essentially just devices for connecting to the Internet [3]. Modern netbook computers, mobile devices, and new-generation tablet computers are heading that way already. So the Internet has evolved from a World Wide Web of information that flows from provider to user to one where that flow is in multiple directions. Internet networking has spawned its own vocabulary, much of which is now in common use. The following section focuses on the new manifestations of the World Wide Web that have the potential to transform the way that scientific research is undertaken.

10.3.1.1 Web 1.0 and 2.0 The first client–server interaction over the World Wide Web occurred in December 1990 [4]. During the rest of the decade up until the early twenty-first century, the Web (retrospectively named Web 1.0) became the de facto system for electronic networking. Its performance was, however, limited by a number of factors. First, the connection speeds were limited by telephone modem technology to around 56 kps and, second, Web pages were static repositories of information with limited interactivity. The situation changed through the introduction of fast broadband connections and software that would allow interactivity between different computers. The concept of Web 2.0 was born [5].

Interactivity was made possible through the introduction of programming languages such as JavaScript and Extensible Markup Language (XML) that allow websites to execute programs and transmit data across all hardware and software platforms. Asynchronous JavaScript and XML (AJAX) were introduced in 2005 by Garrett. This system allows website users to interact with a site while at the same time allowing communication between the client and server machines to occur in the background [6]. The resulting ease of communication and transfer of information between different computers has led to the scientific networking revolution that is the subject of increasing study [7] and of course is the subject of this chapter.

10.3.1.2 Web 3.0: The Semantic Web At present, the content of most Web pages can only be interpreted by a human being rather than being interpretable by machine. In order to overcome this serious limitation, Tim Berners-Lee proposed the creation of a machine-readable Semantic Web to automatically interpret meaning and connections between disparate sets of information [8]. As with the development of Web 2.0, the Semantic Web requires the implementation of new software systems, one of which is the resource description framework (RDF) standard [9]. This software enables the Web to automatically store, exchange, and use machine-readable information. The word "automatically" is key here and is the feature that drives a far greater level of interaction than can be undertaken by human operators alone.

A schematic diagram of the three versions of the Web based is shown in Figure 10.1, based on Deus et al. [10].

10.3.1.3 e-Science and the Fourth Paradigm The Web technology described above is, of course, used for a vast range of human activities, of which e-commerce and social networking are among the most prominent. Of equal or greater importance to human development is the scientific enterprise, with its global reach and dependence on acquiring and analyzing increasingly complex data sets. It is no surprise therefore that advances in Internet technology have attracted the attention of those who have responsibility for managing science as well as those with visions of how the technology can be harnessed.

The term "e-science" was coined in 2000 by John Taylor, Director General of Research Councils Office of Science and Technology in the United Kingdom, and is defined as follows: "e-Science is about global collaboration in key areas of science, and the next generation of infrastructure that will enable it—e-Science will change the dynamic of the way science is undertaken" (http://www.nesc.ac.uk/nesc/define.html). In keeping with this vision, a number of e-science centers have been created in the United Kingdom (http://www.nesc.ac.uk/centres/). A common theme running through these institutions is the development of high-performance grid computing to create the

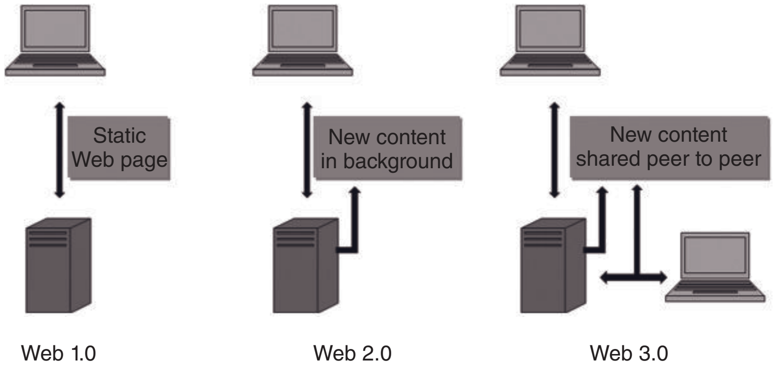


Figure 10.1 Evolution of client–server configurations from Web 1.0 to Semantic Web. The static links on a Web page give way to a two-way communication process between the client (laptop) and server that feeds in information through AJAX and RDF systems.

physical infrastructure that allows data transfer and manipulation at the petabyte level (see Section 10.3.2). Incidentally, the name eScience (unhyphenated) has been adopted by the Chemical Abstracts Service as part of its information portal and is a registered trademark of the American Chemical Society (http://www.escience.org/).

A wide-ranging and visionary description of e-science has been offered by the late Jim Gray of Microsoft Research, who coined the term "the fourth paradigm" [11]. The word *paradigm*, originally used by Thomas Kuhn in his seminal book *The Structure of Scientific Revolutions* [12], is widely used and abused in scientific (and other) circles. Kuhn's definition is based on the idea that scientific progress occurs in a nonlinear fashion, with infrequent paradigm shifts that challenge and supersede existing dogma. Thus the paradigm of an earth-centered universe was overthrown as a result of the discoveries of Galileo, Copernicus, Kepler, and others. The paradigm described by Gray is not really the same, since he refers to a *process* or *theme* central to the way science works. Nevertheless, it offers a useful description of how scientific research has evolved in distinct phases. The first paradigm is the bedrock of the discipline, experimental science, followed by the second, theory. The third paradigm is simulation, an area that has only been possible through the development of computer science over the last few decades. Finally, the fourth paradigm can be summed up as data-intensive science which involves capture, curation, and analysis of experimental data, from the output of stars to signals from gene chips. Here, enormous data sets generated from many different types of experiment are distributed globally and analyzed using the mature computational technologies brought about by Web 2.0 and Web 3.0. The required exponential increase in computing power is delivered by new hardware. Some examples of e-science/the fourth paradigm in action are to be seen in astronomy, oceanography, and high-energy physics as well as the life sciences to be covered in this chapter [13]. The astronomy community, for example, already has a formidable array of telescopes operating at different wavelengths on earth and in space. These instruments are streaming enormous amounts of data to astronomers worldwide that can in turn spend years analyzing the data at a cost that is greater than the cost of the telescopes themselves. A good example of the implementation of e-science is provided by the "virtual observatory" project where observational data from different sources are collated for analysis by astronomers at different geographical locations (http://www.us-vo.org/). Ambitious future plans for massive astronomical surveys plus ongoing scientific activities such as particle physics and climate science are creating huge demands for hardware and software that can cope with the deluge of data.

Until quite recently, data-intensive science in biology was limited to taxonomy. Genome sequencing, expression analysis, and pharmaceutical research have changed all that. Thousands of DNA sequencing runs are being performed worldwide using templates from many different forms of life. Metagenomics is a powerful but data-intensive approach of reconstructing individual sequences from heterogeneous biological samples containing many different organisms. A recent example is the sequencing of the human gut microbiome which contains 10 times more bacterial cells than there are cells making up the human body [14].

Despite the ongoing efforts to sequence as much plant, animal, and microbial life as possible, the main focus of biomedical interest remains the human genome. As new technologies allow the cost of sequencing to approach the figure of $1000 for a 3-gigabase genome, the output of digital data will be easily on a par with the output from the particle physics laboratories and astronomical surveys and there will be some tough choices that will have to be made regarding how much raw experimental data should be archived and how much permanently deleted.

To conclude this section, interactive Web technologies are providing the physical and virtual infrastructures required for collaboration and networking in many scientific activities, the life sciences included. Some (but not all) of this activity involves large-scale sharing and analysis of data, the fourth paradigm of Jim Gray. Before considering networking in the life sciences in more detail, we will examine some of the issues that are being addressed in order to turn these visions into reality.

10.3.2 Computing Infrastructures

Modern computers are orders of magnitude more powerful than the primitive machines that were developed at the end of the World War II. This is a statement of the obvious to all readers of this book, but it is worth reflecting on just how big this magnitude difference actually is. As noted by Nielsen in a recent *Nature* review [15], the electronic numerical integrator and computer (ENIAC) was a state-of-the-art machine for modeling complex phenomena in 1946. Now, the large hadron collider at the European Organization for

Nuclear Research (CERN) will produce enough data in a single second to occupy the memory capacity of six million ENIAC machines. Given the exponential rise in data output and scientific networking, it is clear that enormous strain will be put upon current hardware and communication infrastructures in the very near future. Modern computers are themselves based on processors that have reached the maximum speed at which surplus heat can be efficiently removed. The solution for increasing performance is to turn computers into massively parallel processing machines, as either purpose-built supercomputers or relatively inexpensive PC clusters. A further development lies in the form of cloud computing, in which all data processing is undertaken not by individual PCs but by remote networks. A good analogy is that of the replacement of individual power generators by an electricity grid. Electricity thus became a utility in the nineteenth century and computing has moved this way in the twenty-first century [16]. Computing grids for scientific collaboration are being established by consortia such as The Enabling Grids for e-Science project (EGEE) originating from CERN (http://www.eu-egee.org/). Quoting from its website: "The Production Service infrastructure is a large multi-science Grid infrastructure, federating some 250 resource centres world-wide, providing some 40,000 CPUs and several Petabytes of storage. This infrastructure is used on a daily basis by several thousands of scientists federated in over 200 Virtual Organizations on a daily basis."

While the above specifications are impressive, these organisations depend upon the reliable performance of an Internet that is reaching saturation. Broadband transmission speeds across networks vary significantly according to the physical nature of the links, be they fiber-optic cable, wireless transmission, or copper phone lines. While construction of fast fiber-optic links is being undertaken across the world, there are still limitations in the way the Internet itself operates. This is in large part due to the equal weight given to data packets regardless of importance, a legacy of the communication protocols [Transmission Control Protocol Internet Protocol (TCP/IP) system] devised by the Internet's founders. Given that current Internet bandwidth is reaching saturation, new approaches to data transfer and management are called for. Some ideas about how to achieve this have been highlighted in a recent article [17].

10.4 LIFE SCIENCES AND THE INTERNET

Three important aspects of electronic collaboration for life sciences and medicine are discussed below.

10.4.1 Bioinformatics

Online analysis of nucleotide and amino acid sequence data is now decades old and is a major part of the bioinformatics enterprise. The storage and curation of the data is the responsibility of the International Nucleotide

Sequence Database Collaboration (INSDC, http://www.insdc.org/). This is comprised of data repositories from Europe (EMBL-Bank), Japan (DDBJ), and the United States (GenBank) that exchange sequence data on a daily basis. There are of course many different areas of biology that have (and are) benefiting from access to these publically available data sets. A good example of this is provided by Southan and Cameron from the European Bioinformatics Institute. They note that the nucleotide sequence of the 2009 H1N1 influenza virus was produced within days of detection in patients and the information was rapidly disseminated to those who could make use of it [18].

10.4.2 Biomedical Images

Images are central to biomedical sciences ranging in scale from scans of whole organisms down to images at the scale of single molecules. Now that most images are in digital form, they can be readily shared across the Internet. The potential of image sharing for improving medical diagnosis has been recognized in the form of networked resources. Examples include the Spine Pathology & Image Retrieval System (SPIRS) [19] and the National Cancer Institute (NCI)–sponsored cancer Biomedical Informatics Grid (caBIG) program [20] (see also Chapter 17). Digital pathology is beginning to have an impact upon the operations of histopathology laboratories. Companies such as Aperio Technologies (http://www.aperio.com) produce the hardware to digitize slides so that electronic images can be distributed to pathologists in many different sites (e.g., see [21]). A whole range of desktop communication tools (e.g., WebEx, www.webex.com) are now commercially available that provide sophisticated videoconferencing systems to allow display of images and other data as well as discussion in real time. The above considerations apply to video images, although the bandwidth required may be considerably higher than with static images. Video is commonplace in the form of embedded clips in websites for journals, news, and entertainment and it serves a useful function in training, for example, in laboratory procedures. Videos available on YouTube are providing an unlikely source of medical information to researchers interested in how patients view their disease and what sources of information they see as being useful. A number of examples are available in the recent literature [e.g., 22].

10.4.3 New Publication Models

Van de Sompel et al. recently stated, “The current scholarly communication system is nothing but a scanned copy of the paper-based system” [23]. Despite the extensive use of online publication for scientific communication, most electronic papers are still copies of a printed manuscript, albeit with online supplementary methods and hyperlinked references. It is now impossible for the average scientist to digest all the information about their field that is being published at an ever-increasing rate and which shows no sign of slowing down. This is why the concept of the Semantic Web is so important for the creative use of Internet communications. There is a need to objectify knowledge so that new hypotheses can be generated by using the results of automated literature analysis.

The following examples illustrate the application of semi- or fully automated text mining to specific biomedical applications. The National Center for Biotechnology Information (NCBI) PubMed Central journal collection (http://www.ncbi.nlm.nih.gov/pmc/) contains extensive cross referencing via hyperlinks, but this requires some degree of manual curation. The same is true for the journal *Molecular BioSystems* (http://www.rsc.org/Publishing/Journals/MB/), although the links of chemical terms to online databases is very much in the spirit of a publication format that is much more than that used by a traditional journal. A fully automatic text-mining system is illustrated by Reflect, a system devised at EMBL in Germany that automatically links the names of genes, proteins, or small molecules to online databases (http://reflect.embl.de/). Reflect can be used as a plug-in for Web browsers or as a stand-alone Web page. Finally, the Semantic Web Applications in Neuromedicine (SWAN) Web tool (http://swan.mindinformatics.org/) allows researchers to access the most relevant papers in their field of neuroscience and to make hypotheses based on connections that would not have been obvious just by browsing the literature in the usual way.

10.5 NETWORKING AND OPEN-SOURCE DRUG DISCOVERY

Large biopharmaceutical companies have until recently been notoriously reluctant to openly collaborate with their peers because of (actually understandable) concerns about the security of their proprietary data. One major threat to this outlook is the emerging view that the process of drug discovery can and indeed should be restructured to take account of the realities of patent expiries and depleted pipelines. This restructuring has to involve more openness and collaboration to encourage those with good ideas but limited resources to make a contribution to top-level drug discovery as practiced by the traditional pharmaceutical companies. This is even more important given the emergence of BRIC countries (Brazil, Russia, India, and China) as rising stars in pharmaceutical research and development (R&D). There are no certainties any more about large companies like GlaxoSmithKline (GSK), which are diversifying into new markets, changing their portfolio of medicines (more biologicals) and also growing consumer health products that are totally unrelated to their pharmaceutical cousins. Perhaps given these strategic changes, it should not be a surprise that GSK has recently announced a major initiative in so-called open-source drug discovery by freely releasing data derived from the screening of nearly two million compounds against the malaria parasite *Plasmodium falciparum* [24]. The structures and biological activities of over 13,000 active compounds are now available to any organization that wishes to exploit them. This is a significant extension of the public–private partnerships (PPPs) such as the Medicines for Malaria Venture, TB Alliance, Institute for

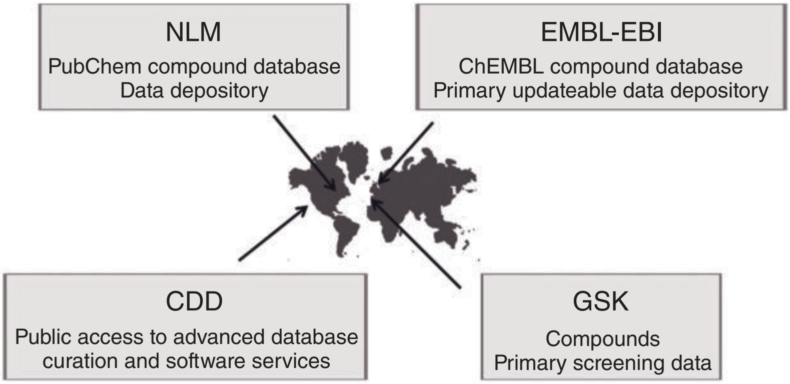

NLM
PubChem compound database
Data depository
EMBL-EBI
ChEMBL compound database
Primary updateable data depository
CDD
Public access to advanced database curation and software services
GSK
Compounds
Primary screening data

Figure 10.2 Global network of open-source drug discovery data. Primary screening data for hits identified by GSK in Tres Cantos Spain are deposited in the EMBL-EBI database near Cambridge, England. The data are also sent to the National Library of Medicine PubChem database in Bethesda, Maryland, and to Collaborative Drug Discovery in Burlingame, California. All of the organizations apart from GSK can be accessed globally by anyone wishing to use the data for their research.

One World Health, and Drugs for Neglected Diseases Initiative (DNDi) that are focused on neglected developing world diseases.* The GSK collaboration used a network of organizations to process and distribute the data in a way that may set a precedent for future initiatives of this type. The key organizations involved in this open-source network are shown diagrammatically in Figure 10.2.

The GSK laboratories in Tres Cantos, Spain, performed the initial compound selection and wet laboratory testing against the malaria parasite. Biological data such as the concentration of compounds that produce 50% growth inhibition (XC_{50}) are then sent with compound data to the primary data silo in the ChEMBL database at the European Bioinformatics Institute near Cambridge, United Kingdom. The data are held in a database called ChEMBL-neglected tropical diseases (NTDs), which also contains data from other initiatives, including compound data from Novartis [25]. It is important to note that, in addition to downloading data, scientists can also upload their own experimental results, thus making the system a kind of virtual laboratory notebook that anyone can access. The data are also distributed to the PubChem database hosted by the National Library of Medicine in Bethesda, Maryland [26]. Finally a commercial organization, Collaborative Drug Discovery (CDD), in Burlingame, California, has access to the data for deposition in its CDD public database [27]. The company is an example of how such enablers of drug discovery networking may evolve over the next few years through the provi-

* http://www.mmv.org/; http://www.tballiance.org/home/home.php; http://www.oneworldhealth.org/; http://www.dndi.org/.

sion of free open-source data to the research community while at the same time maintaining commercial viability through proprietary database services. All of the organizations comprising the GSK initiative are offering valuable information to academic and other groups that simply cannot afford to establish their own chemoinformatics and related infrastructures. Of course, while the GSK example represents a significant milestone in sharing preclinical data, it is clear that malaria is not high on the list of priorities for large pharmaceutical companies. It is really too early to say how (or if) the above model can be applied to major competitive programs like neurodegeneration or oncology, so it will be a matter of "wait and see." In the meantime, there is an interesting trend toward open-source software for pharmaceutical development that will go some way toward leveling the playing field between large pharma, small biotechs, and academia. Drug discovery is a highly complex business and requires many different software tools to manage biological, chemical, and clinical data. As many of these tools are proprietary, there is a call for greater uptake of open-source systems to allow all parties access to the same software. This will have the obvious effect of facilitating networking between different organizations since they can then use common software to seamlessly transfer and analyze multiple data types. An example of open-source software for drug development is the OpenClinica® system developed by Akaza Research in the United States [27]. As with CDD, the business model involves free access to data and software while providing enhancements paid for by subscription.

10.6 CONCLUSION

It is clear that electronic networking is evolving rapidly, not just in terms of the actual hardware and software, but also in terms of how barriers to access are being torn down. The free spirit of open-source software, file downloading, and free access that has characterized telecommunications, entertainment, and social networking is now clearly being applied to pharmaceutical and other biomedical research. This will be important for anyone working inside a large pharmaceutical company on IT and related systems as they will find that the choice of resources is widening far beyond the traditional "off-the-shelf" commercial products. Despite the many unanswered questions about how networking will impact upon mainstream pharmaceutical research in the near future, one thing is certain—there is plenty of excitement to come.

REFERENCES

1. Bailey D, Zanders E. Drug discovery in the era of Facebook—New tools for scientific networking. *Drug Discov Today* 2008;13:863–868.
2. Hannay T. From Web 2.0 to the global database. In Hey T, Tansley S, Tolle K, Eds. *The Fourth Paradigm*. Redmond, WA: Microsoft Research, 2009, pp. 215–219.

3. http://searchcloudcomputing.techtarget.com/sDefinition/0,,sid201_gci1287881,00.html.
4. Pre-W3C web and Internet background. http://www.w3.org/2004/Talks/w3c10-HowItAllStarted/?n=15.
5. What Is Web 2.0: Design patterns and business models for the next generation of software. http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html.
6. Ajax: A new approach to web applications. http://www.adaptivepath.com/ideas/essays/archives/000385.php.
7. Sagotsky JA, et al. Life sciences and the web: A new era for collaboration. *Mol Syst Biol* 2008;4:1–10.
8. Berners-Lee T, Hendler J. Publishing on the semantic web. *Nature* 2001;410: 1023–1024.
9. RDF/XML syntax specification (revised) W3C recommendation, 10 February 2004. http://www.w3.org/TR/REC-rdf-syntax/.
10. Deus HF, et al. A Semantic Web management model for integrative biomedical informatics. *PLoS ONE* 2008;3:e2946.
11. eScience: A transformed scientific method. http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt.
12. Kuhn TS. *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press, 1962.
13. Bell G, Hey T, Szalay A. Beyond the data deluge. *Science* 2009;323:1297–1298.
14. Qin J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 2010;464:59–65.
15. Nielsen MA. A guide to the day of big data. *Nature* 2009;462:722–723.
16. Carr N. *The Big Switch Our New Digital Destiny*. New York: W.W. Norton, 2008.
17. Gammon K. Four ways to reinvent the Internet. *Nature* 2010;463:602–604.
18. Southan C, Cameron G. Beyond the tsunami: Developing the infrastructure to deal with life sciences data. In Hey T, Tansley S, Tolle K. Eds. *The Fourth Paradigm*. Redmond, WA: Microsoft Research, 2009, pp. 117–123.
19. Hsu W, et al. Web-based image retrieval system for large biomedical databases. *Int J Med Inform* 2009;78(Suppl 1):S13–24.
20. Oster S, et al. caGrid 1.0: An enterprise Grid infrastructure for biomedical research. *J Am Med Inform Assoc* 2008;15:138–149.
21. Perkel JM. Digitizing pathology. Available: http://www.biosciencetechnology.com/Articles/2010/02/Imaging-Digitizing-Pathology/.
22. Fernandez-Luque L, Elahl N, Grajales I. An analysis of personal medical information disclosed on YouTube videos created by patients with multiple sclerosis. *Studies Health Technol Inform*. 2009;150:292–296.
23. Van de Sompel H, et al. Rethinking scholarly communication: Building the system that scholars deserve. *D-Lib Mag*. doi:10.1045/september2004-vandesompel.
24. Gamo F-J, et al. Thousands of chemical starting points for antimalarial lead identification. *Nature* 2010;465:305–310.
25. CHEMBL-NTD. http://www.ebi.ac.uk/chemblntd.
26. https://www.collaborativedrug.com/pages/public_access.
27. http://www.openclinica.org/.

11

CANCER COMMONS: BIOMEDICINE IN THE INTERNET AGE

Jeff Shrager, Jay M. Tenenbaum, and Michael Travers

11.1 INTRODUCTION

Despite the enormous strides medicine has made in the understanding and treatment of human disease, millions of people each year succumb to cancer. This chapter describes Cancer Commons, an approach to the treatment of cancer that echoes the eloquent words of Alexander Dumas in his 1844 masterpiece, *The Three Musketeers: Un pour tous, tous pour un* ("One for all, all for one"). The central idea that drives Cancer Commons is that biomedicine is essentially a huge ongoing experiment and the data resulting from every

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.

patient encounter should contribute as efficiently as possible to improved treatment for each subsequent patient. In theory, to make this work, one need only treat each patient based upon the best available knowledge, gather all possible data from every such encounter, combine and analyze all of the data instantaneously, update the knowledge base accordingly, and then disseminate the updated knowledge base to inform treatment decisions for every subsequent patient encounter. Although difficult in practice, Cancer Commons is bringing this vision to reality through Internet technologies and advanced laboratory and computational techniques for dissecting cancer at the molecular level.

In its quest to create a closed-loop "rapid learning" system for advancing the treatment of cancer, Cancer Commons employs a number of tightly interlocking methods and technologies:

1. A *Web-based platform* that records the relevant genomic and clinical data from patients, produces actionable information regarding potential treatments (including applicable clinical trials), captures outcomes data, and makes these data available for researchers to improve knowledge about the disease.
2. The treatment information produced by the platform is based upon a computationally accessible *molecular disease model* that integrates the latest available knowledge regarding genomic subtypes of cancer.
3. The disease model needs to be updated when (a) it does not contain any actionable treatments relevant to a particular patient, (b) a standard treatment fails or (c) a nonstandard one succeeds in a particular patient, or (d) new results are published. Cases (a), (b), and (c) are the most interesting:
    - **a.** When no information regarding potential treatments exists in the disease model for the particular genomic and clinical pattern presented by the patient, one can deploy a *personalized virtual biotech* project with the goal of finding such a treatment. This is accomplished by enabling ad hoc teams of researchers to utilize outsourced services and workflow management facilities offered by the Cancer Commons platform to undertake small-scale, highly targeted drug development. Findings from such projects are curated back into the disease model, thereby reaching every subsequent patient seeking treatment recommendations through Cancer Commons.
    - **b.** When a treatment does not work, there may be an error in the disease model. There are many possible sources of such an error, but the general response is similar to that of personalized virtual biotech [case (a)]. However, rather than drug discovery, the goal is to understand what went wrong in this particular case, if possible, and update the model accordingly. For example, this may be a new genomic subtype of cancer that was not previously recognized.

c. Whenever a patient responds unexpectedly to a novel therapy, it is critical to try to understand what subtype of cancer that patient had so that the results can be generalized to benefit others. Such serendipitous findings can result from formal clinical trials (e.g., isolated responders in a large clinical trial that failed overall) or from the "*N*-of-1" experiments that oncologists perform everyday in their practices. Again, a personal virtual biotech can be used to determine the molecular subtype of cancer for which the putative therapy applies and then to organize an adaptive study that attempts to validate the finding in patients with that phenotype.

All patients whose treatment is being managed within Cancer Commons are, in effect, participating in a huge, continuously running, adaptive clinical trial that is constantly testing and refining both the molecular disease model and potential treatments in a continuously improving loop. At each point in time, patients are treated with the best available therapies for their tumors' molecular subtype. Subtypes are then split, corresponding to responders and nonresponders, and new subtypes are added to accommodate previously unseen tumor types. Each patient encounter thus becomes an opportunity to update knowledge (represented by the molecular disease model) and have it immediately affect subsequent patient encounters. Over time, subtypes in the disease model will be defined with greater and greater specificity and will be linked to increasingly efficacious therapies.

This closed-loop, rapid learning approach to translational cancer research and the platform that supports it are described in the sections that follow.

11.2 GENOME-BASED CANCER TREATMENT, CANCER COMMONS, AND THE MOLECULAR DISEASE MODEL

Individual patient genomes—more precisely, the genomes of their tumors—are already being used to hypothesize treatments for particular cancers. In a recent segment on National Public Radio's *Talk of the Nation* program [1], Dr. Harold Varmus, newly appointed Director of the National Cancer Institute (NCI) put it this way [1] (emphasis added):

> Over the [past] 30 or 40 years . . . it's become apparent that we can identify the set of genes that play a major role in cancer. That role is played when those genes are damaged by mutations or rejoined or amplified or inappropriately expressed in some way. . . . we're now, one by one, picking apart the insides of a cancer cell, understanding how cancers grow, how they invade their environments, how they metastasize. . . . We have some successes over the last 10 years of drugs that are precisely targeted to the damage in the cancer cell that result in dramatic remissions, in some cases, sustained absence of tumors. . . . while there are commonalities, *every cancer is different in some way*. Cancers have lots of changes, and the task for modern cancer biologists is to understand which of those [changes] is

> most significant with respect to the likelihood of the cancer actually leading to death, and more importantly, *for the patient, offering a different set of prospects for treatment*. . . . [T]he genetic testing that's done on a cancer [involves actually comparing] the cancer to normal tissue from the same individual to look for mutations that occurred during life that gave rise to that cancer.

Because an individual patient is being analyzed, and the genomic profile is created from that particular person's tumor, this process is sometimes referred to as "personalized genomic medicine."

Cancer Commons is Web-based paradigm whereby patients with advanced cancers (along with their physicians*) can enter the relevant parts of their clinical and genomic profiles, to the extent these are known, and receive information regarding potentially useful treatments based upon the latest science. This information comes from the Cancer Commons *molecular disease model* which is, in effect, a *living review paper* regarding genomic subtypes of cancer and their management. The molecular disease model is created by experts in the field of genomic cancer treatment, usually in the specific type of cancer with which the patient has been diagnosed. It encodes mappings among these aspects of genomic medicine: genomic profiles, biochemical mechanisms of the disease (commonly called "pathways"), and treatment options, including drugs and combinations of drugs, clinical trials, and other treatment options. Together these constitute a set of "actionable subtypes" of cancer, each representing a functionally different molecular etiology driving the disease, and links it to distinct therapeutic approaches. The molecular disease model also contains citations to the relevant literature, including case studies that support the mappings.

The molecular disease model is a "living" document in at least two senses. First, it is represented in both human-readable and computationally accessible formats and can thus be used directly by the algorithms that enable the Web-based Cancer Commons platform to produce information regarding potential treatments or trials appropriate for particular genomic and clinical profiles. Second, it is constantly updated by expert curators based on the latest findings and opinions in the field, as well as by information resulting from the case flow that takes place within Cancer Commons itself. An initial disease model may be created top down, for example, by a panel of recognized disease experts, or bottom up, for example, via a statistical analysis of clinically annotated genomic data. Regardless of how it is created, the initial model is likely to be incomplete and contain errors and so will need to be refined, as described in the next section.

*We speak in a number of places of what patients do, as though there are no physicians involved. In reality, the target users of Cancer Commons are physicians (usually oncologists), physician/patient dyads, or what we call "superpatients" who can operate nearly at the level of oncologists but with a focus on their own disease.

11.3 UPDATING THE MOLECULAR DISEASE MODEL

For personalized genomic medicine to be successful, one must be able to correlate specific genomic and clinical profiles with potentially effective therapies. Unfortunately, there are so many possible genomic-level variations that every patient could potentially represent an entirely new disease. For example, Lee et al. [2, p. 473] reported ">50,000 high-confidence single nucleotide variants" and "530 somatic single nucleotide variants in [one lung cancer tumor], including [392] in coding regions, as well as 43 large-scale structural variations"; *All this from a single patient's tumor*! One hopes that the huge number of possible mutations can be categorized into a handful of functional subtypes based upon the molecular drivers of the disease (i.e., biochemical pathways), and so into a handful of therapeutic approaches, but we cannot know at the outset, of course, what the subtypes are and how to treat them.

Thus, to carry this program forward, it is necessary to obtain genomic information from a wide range of individuals—indeed, potentially a huge number of them and probably multiple samples from each patient in order to assess intratumor variability. Moreover, in order to determine treatment efficacy, one needs to correlate the details of this massive body of genomic information with information about how individuals with different genomic profiles are treated and how they respond.

The molecular disease model summarizes some of this complexity into actionable subtypes in terms of tumor profiles, pathways, and so on, but it is only useful when it has something to offer and when it is correct. Determining when the disease model is incomplete is easier than determining when it is wrong. It is *incomplete* when no actionable treatments can be computed from the model. This deficit is addressed by the *personalized virtual biotech* methodology described in the next section. The model is *wrong* when information is forthcoming regarding potential treatments but when the patient fails to respond to a treatment selected from this set (i.e., their cancer progresses). Whether a particular patient's disease is progressing more under one treatment than under another, and taking into account the near infinitude of variables that may cause particular individuals with otherwise genomically (and functionally) identical tumors to respond differently to treatment, is a very difficult problem which we address through a *macroscale N-of-1 adaptive trial strategy*, described in the following section.

11.3.1 Personalized Virtual Biotech

When no effective treatment hypotheses can be found in the disease model, the personalized virtual biotech methodology provides a means to efficiently carry out ad hoc research to discover potential novel approaches to a particular patient's disease. The idea of virtualizing therapy discovery is not new; large pharmaceutical companies often operate projects that are nearly completely outsourced, but these usually require extensive (and expensive) management and information infrastructures. The Internet has upended the economics of this sort of virtualization, bringing enterprise management tools such as SalesForce.com (http://www.salesforce.com/), computational tools such as BioBike (http://www.biobike.org/ and [3]), databases such as those found at the National Center for Biotechnology Information (NCBI, http://www.nbci.nlm.nih.gov/), and enterprise information infrastructure such as Google Apps (http://apps.google.com/), well within the range of price and usability of small distributed teams. Moreover, the technologies required for drug development, such as high-throughput sequencing, drug screening, computational modeling, and so forth, have traditionally been available only inside large pharmaceutical and biotech companies. However, rapid technological advances have made such services readily available at reasonable cost. Indeed, noncorporate virtual treatment discovery projects are already emerging, funded by nonprofit disease foundations like the Cure Huntington's Disease Initiative (http://www.highqfoundation.org/), the Myelin Repair Foundation (http://www.myelinrepair.org/), the Michael J. Fox Foundation for Parkinson's Research (http://www.michaeljfox.org/), and others.

There are several unique aspects of the way that virtual biotech is deployed in the setting of Cancer Commons. First, *personalized* virtual biotechs can be created out of Cancer Commons when no specific treatment hypotheses can be found in the molecular disease model for a particular patient's disease. The molecular disease model is in this sense a cache of previously solved targeted drug discovery activities: A particular patient seeking information regarding potential treatments queries the model using his or her clinical and genomic profile. If there is information regarding one or more potential treatments already cached in the model, they are returned as treatment *hypotheses*, along with an explanation of why they were chosen. Significantly, these treatment hypotheses can include available therapies that proved effective on patients with other types of cancer whose tumors exhibited similar genomic or pathway aberrations. If, on the other hand, no such hypotheses are forthcoming, then a personalized virtual biotech project can be created to try to find a treatment.

The second unique feature of personalized virtual biotech, as it operates in the Cancer Commons setting, is that the Cancer Commons platform supports the process via an integrated services and workflow architecture, integrated state-of-the-art bioinformatics, and, most importantly, the ability to run real-time live in-patient experiments where *real patients* are treated by real physicians and their genomic and clinical profiles and their response to treatment are tracked. These factors enable researchers involved in a personalized virtual biotech to hypothesize and test novel treatments in settings ranging from *in silico* to in patient. This will be explained in more detail in the next section.

Historically, good leads are often found by using existing drugs off-label or combining them in cocktails. This strategy has the advantage that these leads can be used immediately in patients. The Cancer Commons platform can enable multiple personalized virtual biotechs, working on the same or related diseases, to collaborate and share data, knowledge, and resources. By coordinating their efforts, a network of personalized virtual biotechs can systematically explore the opportunity space of targets and leads and avoid unnecessary replication of experiments.

One can conceptualize the ongoing activity of a given personalized virtual biotech in terms of requests of various sorts ranging in complexity. Examples of requests may include a call for opinions on a particular scientific question (say, by asking for votes), a request for some data set (or reference to a paper) that addresses some specific scientific question, a request for help with some complex bioinformatic calculation, and so on. Note that the requests form a hierarchy: At the top a physician might request a drug to treat a particular patient given the patient's clinical and genomic profile. This is, of course, a very broad request, and hard to satisfy as stated, but must be broken down into a series of simpler requests: a request from pathology for a tissue sample, a request for a laboratory to culture the sample, a request for a high-throughput screening facility, such as the National Institutes of Health (NIH) Chemical Genomics Center (NCGC) (http://www.ncgc.nih.gov/) to run assays on the tissue model using existing drugs, a request for the gathering together of those results, a request for experts to interpret those analyses and rank the results to create a pipeline, a request for a mouse model of the disease, a request for someone to run the highest ranked molecules through the mouse model, and so on. Of course, each of these will usually be broken down further into subrequests until one reaches small doable tasks or small answerable questions. This recursive request structure is commonplace in science, and indeed in any structured problem solving, and can work across a community [4]. What is unique about operating such a hierarchy of goals in Cancer Commons is that such requests may be efficiently fanned out across a wide-ranging distributed community. This ability to efficiently leverage distributed resources in a call-and-response manner is a unique capability of Internet-based technologies [4]. Moreover, the Internet offers methods for distributed decision analysis, which can be utilized to prioritize opportunities such as targets and leads, to collaboratively analyze and interpret data, and to make other complex decisions.

The Cancer Commons platform facilitates this process of distributed call and response by employing a series of unique technologies. Specifically, the platform includes a registry of services (such as those exemplified just above) and a hub-and-spoke architecture where the hubs can automatically and securely communicate with one another through the Web. Each hub corresponds to a community of clinicians and/or researchers organized around a project, an institution, a disease, or a discipline. Each hub advertises the services that are available within that hub's community (i.e., the knowledge, tools, and expertise of those in that particular community). When requests are made within a given hub (either by researchers or in the process of operating semi-automated workflows), if the services required to respond to that request are not locally available, the hub architecture will automatically call out to hubs that advertise relevant services. By tightly integrating knowledge, decisions, and action, the personalized virtual biotech methodology can accelerate the research cycle, driving it on behalf of particular patents' disease, focusing resources on the most promising opportunities and minimizing redundant work.

To be clear, personalized virtual biotech is a *methodology*. That is, it is a particular way of employing the services and technologies provided by the Web-based platform that operates Cancer Commons, *not* a set of tools separate from that platform. This methodology also creates a novel economic model for drug discovery and development. It can reposition drugs that worked for other cancers or diseases or that failed to demonstrate efficacy against their original indication, or it can develop new molecules. However, because it is not usually practical to develop a new molecule for a novel target on the timescale of an individual patient, the most common case is likely to be the off-label use of drugs or new combinations of existing therapies. Physicians do both of these routinely in practice, but the results of these *N*-of-1 experiments are seldom disseminated.

Beyond off-label and combinational use of approved or investigational drugs, there are the many potential drugs languishing on the shelves of pharma companies as a result of failed trials for particular indications, or those many hundreds that showed some activity in the target indication, but not enough to be worthwhile for the company to pursue them. Collaborative Drug Development (CDD), for example, has spearheaded the effort to get pharmaceutical developers to open some of these libraries for use in diseases of low profitability but large worldwide impact, such as malaria (Chapter 21). Finally, beyond existing treatment combinations and orphan pharma libraries there are the many promising treatments being studied in academic laboratories with no path to market through the traditional channels. The traditional path for academics to translate their research into therapies is either to start a biotech company or license their intellectual property (IP) to an existing company. Unfortunately, the costs of a startup are high ($5million-80million) and the odds of success dismal. Moreover, potential licensees of IP require extensive validation data that are hard to generate in an academic setting. As a result, many promising treatments are left languishing in the so-called Valley of Death between the academic laboratory and the clinic. Personalized virtual biotech offers a radically different approach, driven directly by the needs and financial resources of patients. This approach leverages research collaborations that academics are used to but provides them with previously unavailable technical services, access to highly motivated patients for small-scale clinical studies, and potentially novel funding sources such as Internet fundraising campaigns. Moreover, the products of personalized virtual biotech are rapidly disseminated through Cancer Commons to all patients who might potentially benefit from them.

The Cancer Commons platform enables multiple personalized virtual biotechs working on the same or related diseases to collaborate and share data, knowledge, and resources. By coordinating their efforts, a network of personalized virtual biotechs can systematically explore the opportunity space of targets and leads and avoid unnecessary replication of experiments. By virtue of such sharing, a network of personalized virtual biotechs can achieve unprecedented economies of scale and acceleration by leveraging knowledge about targets and leads across diseases.

11.3.2 Cancer Commons as a Macroscale *N*-of-1 Adaptive Trial

All patients whose treatment is being managed within Cancer Commons are, in effect, participating in a huge adaptive clinical trial testing the molecular disease model and leading to changes in the model resulting from sources such as personalized virtual biotech. One can think of this as a macroscale (across the whole Cancer Commons community), *N*-of-1 (each patient is being treated through personalized genomic medicine), adaptive trial. Let us see how and why this is needed and how it works.

Treating cancer (as is the case for most diseases) is, in computational parlance, a very high dimensionality search problem. The number of potential hypotheses about the causes and associated treatments of the disease is huge, especially when complete genomic profiles and combinational therapies are considered—exponentially larger than the number of patients that could represent each possible combination of factors.

Classical large-scale clinical trials are inappropriate for systematically exploring very high dimensional spaces where, as we have seen, there are potentially a huge number of differences at the genomic level and a huge space of possible treatments. One reason for this is that a large-scale controlled trial is very likely to have a heterogeneous population of patients with functionally different genomic tumor profiles. As a result, it is possible that a treatment that failed statistically in a large-scale controlled trial actually helped a small subset of the patients who have a specific genomic profile, but these cases were mixed together with many others who were not helped by the treatment. To make matters (much!) worse, the most effective treatments will probably involve combinational therapies and dosages, resulting in a nearly infinite number of potential treatments, even if we only consider approved drugs. Computer scientists call the situation "the curse of dimensionality" [5]: There are nowhere near enough patients to explore a space of this size using classical clinical trials.

Even though the raw dimensionality of genomic profiles crossed with treatments is so high that there cannot be enough patients to sort out genomic-level treatment effectiveness precisely, a great deal is known about cancer treatment from large-scale controlled trials, from research using *in vitro*, *in vivo, and in silico* models, and from the experiential knowledge of the treatment community. The molecular disease model approach taken by Cancer Commons posits that a great deal is actually already known about how to treat many genomic subtypes of cancer, but this knowledge is not gathered together or disseminated efficiently because medicine relies upon the publication of review papers or guidelines which are only occasionally updated, and their results are not computationally accessible and so cannot be efficiently utilized in clinical settings. Moreover, there is enormous variability in treatments and outcomes, even given the guidelines. What is needed in all of these cases is to capture and integrate data and evidence not only from large-scale controlled trials but also from the thousands of ad hoc *N*-of-1 experiments that occur every day in the practice of oncology. These experiments test hypotheses based on the creativity, knowledge, and experience of practioners seeking the best possible outcomes for their patients, not pharmaceutical companies seeking regulatory approval for a particular drug.

The concepts of adaptive trials and of *N*-of-1 medicine are, independently, fairly well understood. Indeed, there are entire books dedicated to the design of adaptive trials [e.g., 6], although they remain somewhat controversial and must be designed with great care [7]. There is also a long tradition in medicine of publishing case histories as a way of communicating treatment experiments to others.

Technically, an *N*-of-1 study is a specific sort of experimental design where the subject, for example, an individual patient, acts as his or her own control. For example, a baseline measurement may be made, serving as a control, and then a treatment applied and a response observed. The treatment may then be removed, the patient observed returning to baseline, and so forth, in accord with a protocol that can be replicated to get the required sample. In diseases like cancer such a design could sometimes be applied, but usually what one means by *N*-of-1 in these settings is not such a specific experimental design, but rather just the assumption that each patient's disease is different from every other patient's disease. Although in a trivial sense this is always true, it is clear that in some cases it is more practically relevant than in others. For example, if your friend were to acquire a bacterial infection from you, it is likely that you *functionally* have the same disease—that is, an antibiotic that works on you is very likely to work on your friend as well. Of course, there is a small chance that the organism could have mutated along the way or that your friend is allergic to the antibiotic, so adjustments will always have to be made from one person to the next, but generally speaking bacterially transmitted diseases do not require a great deal of intricate tailoring of treatments to each individual. This is not the case for many other diseases, especially ones where there is a great deal of variability at the genomic level, the paradigmatic example of this being HIV/AIDS, which is known to mutate rapidly. Cancer is a particularly complex case because, whereas we know that all cancers are due to genomic mutations, we do not know what the practical number of effective carcinogenic mutations is; and the number of functional subtypes requiring different treatments could range from tens to hundreds or even thousands. As a practical matter, in the case of cancer *N*-of-1 treatment has come to mean profiling the specific genomic mutations that drive an individual's cancer and then choosing a treatment based upon that specific genomic profile. Of course, this is much more easily said than done. The profiling step is relatively simple in these days of microarrays that measure everything from gene expression to chromosomal rearrangements, but choosing a treatment based upon such profiles has not even begun to become practical and is unlikely to progress very rapidly because of the high dimensionality of the problem.

The Cancer Commons approach to this problem combines *N*-of-1 treatment (as defined above in cancer) with adaptive trials in a particular way that begins to address the high dimensionality of the problem. We call this approach *N-of-1 adaptive trials*. The idea is to genomically profile (to the degree possible) every patient and then treat each of them with the best possible treatment. Of course, it is often not possible to determine the best possible treatment, either because there are multiple treatment options available that have not been tested head to head, or because no effective treatment is currently known. The latter case is dealt with by personalized virtual biotech, described above. Here we focus on the former case, where there are a range of potential treatments. In this case, we would like to essentially run an adaptive experiment over a set of patients, trying multiple treatments, collecting and analyzing the response data as efficiently as possible, and then integrating the results into the molecular disease model to tune the treatment regimens for these patients (if possible) and for subsequent patients. It is important to note that, on an individual basis and leaving the genomics out of the picture, treating a particular patient, seeing how things go, and adjusting the treatment accordingly are what physicians do all the time and have done through time immemorial. What is new here is efficiently collecting the results of a large number of such *N*-of-1 experiments and using them in a rapid learning loop to adjust the treatment regimens for the patients being treated and for subsequent patients.

As plausible as this might seem, the *N*-of-1 adaptive approach holds hidden dangers [7] and must be carefully thought out to avoid inappropriate channelization into a less effective treatment based on an apparent early success due, say, to a misclassification of the disease or perhaps noisy data [8]. To take an extreme case, say that, when one drug appeared slightly better in one patient all other patients were immediately switched to it, it would be the end of the trial. One should, in theory, wait until one has statistically strong evidence before switching everyone onto another treatment, but because there is a large statistical price to pay for our "peeking" at the data at every step, it may well be that such a trial requires many more subjects than a classical trial. Moreover, if we continue treating everyone with one of the two treatments at random (more precisely, in accord with randomized trial protocols), then if it requires more patients to see an effect in the adaptive trial, we may have harmed the excess number of patients assigned to the poorer treatment over the classical trial by virtue of not having treated them with the best available treatment as soon as would have been possible in the classical model. Of course, in powering the study to begin with, one can tell how many subjects one is likely to need, but the adaptive approach will always require, in the worst case, more subjects than a classical trial. ("Whopping" effects in either direction are, of course, also considered as grounds for termination of a classical trial, so there is no advantage to adaptive trials when the effect is very large.)

The way out of this conundrum is through the second type of virtual biotech, mentioned above: using personalized virtual biotech to examine, at a deep molecular level, why a treatment failed in a particular individual. Note that this is very similar to the problem of finding a new treatment for the case, except that we now have an example of a treatment that was supposed to work but did not. What was different about this particular disease process that differentiates it from the cases where this treatment worked such that it was in the molecular disease model to begin with? If we learn this—which, if it can be learned at all, requires more or less the same machinery as creating a new treatment—then we have learned a new distinction between subtypes of the disease, and the molecular disease model should be updated accordingly (regardless of whether or not we have in hand a new treatment for the new subtype).

In sum, at each point in time, patients will be treated with the best available therapies for their tumors' molecular subtype based on the associated (clinically indicated or proposed) treatment guidelines. The power in the Cancer Commons approach lies in the rapid feedback loop that is generated when patients are rationally treated based on their molecular subtype, and then if they do not respond to the experimental agent, researchers attempt to uncover why they failed to respond and apply these findings to the next patient with a similar molecular subtype.

Studying response outliers armed with a full genomic panel also allows one to pose interesting questions. For example, for the subset of patients who may have responded in a trial that failed to meet its clinical endpoints: "What disease did these people have, which was previously lumped into a broader subclass (e.g., triple negative breast cancer) for which there might now be an effective therapy?" "How many others have that disease?" Or, for those patients who failed to respond in a successful trial, one can go back to the translational scientists on whose research the molecular disease model and proposed guidelines were based and ask: "What didn't you understand about the disease, drug, or molecular target now that we have thousands of new molecular data points from real cancer patients?" These questions are not academic; they pertain to patients who have just failed therapy and are urgently seeking new options. Answers to questions like these will lead to splitting existing subclasses, corresponding to responders and nonresponders, or adding new ones to accommodate previously unseen tumor types. Over time, molecular subtypes in the disease model will be defined with greater and greater specificity and linked to increasingly efficacious therapies.

11.4 DETAILS OF THE CANCER COMMONS PLATFORM

The goal of Cancer Commons is to provide patients and physicians with the latest information, tools, and resources they need to obtain the best possible outcome and to capture and aggregate the results over all studied patients to improve cancer treatment generally. Operating Cancer Commons, personalized virtual biotech, and macroscale adaptive *N*-of-1 trials efficiently requires a computational platform that is both deep in biocomputation and broad in its ability to support collaboration across a wide range of different types of users: patients, physicians, and researchers. The platform must also support the rapid creation, publication, and computational use of living, computationally accessible documents such as the molecular disease model, and case reports created directly from patient clinical and genomic data and must enable the community to efficiently provide commentary on these. Finally, the platform must provide the facilities required to create and operate personalized virtual biotechs, including access to numerous commercial and noncommercial services, state-of-the-art bioinformatics, integrated knowledge bases, workflow facilities, and management tools. The following paragraphs describe these facilities in more detail.

All Cancer Commons users interact with one another and with community resources through personalized Web-based portals. Researchers use portals to design and run experiments and collaborate on data analysis and interpretation. Ecosystem partners use portals to advertise and provide services. Clinicians use portals to report clinical results and experiences and to track the latest research relevant to their patients, and patients themselves may use portals to report on their personal experiences and to track state-of-the-art developments relevant to them.

Each patient's status is described by a continuity of care record (CCR), which is essentially the portion of their electronic health record (EHR) relating to the particular disease process under consideration. The CCR serves as a dynamic record of the patient's treatment and response/outcomes and acts as a query to the molecular disease model (via computational machinery underlying the platform, described below) to produce information regarding potential treatments, relevant literature, and so forth.

The platform employs a variety of biocomputation algorithms, operated by the BioBike platform [3], to combine the disease model with a large integrated knowledge base and generate information regarding relevant treatments. These algorithms usually cannot make a specific treatment recommendation, so there is usually the need to compare nearly equivalent choices. For this case the platform includes a variety of collaborative decision support tools based upon the CACHE system [9] that can be used by patients and physicians to help make treatment decisions and by researchers to help in considering hypotheses.

One of the central features that makes Cancer Commons a *commons* is a rapid communitywide interactive publication model for two-way communication and dissemination of data, knowledge, case details, commentary, and so forth. This model is utilized throughout the commons. Private forum facilities allow patients and physicians to create "Virtual Tumor Boards" by inviting specific other members of the community to collaborate in the analysis of a particular case. CCRs can be deidentified and opened up as case studies published within Cancer Commons, and these can be discussed as well. In addition to displaying and computing with the molecular disease model, the platform provides a means for members of the community to discuss and suggest refinements to the model. These are fed back to the curators for consideration and potential revision of the model and can be read by any member of the community.

The rapid communitywide interactive publication model of Cancer Commons overcomes three major issues that weigh down traditional biomedical communications. First, while formal, peer-reviewed, professionally copy-edited research articles are integral to medical research, they are not an efficient means of disseminating timely knowledge that may have immediate clinical utility for patients. Second, with tens of thousands of peer-reviewed publications, abstracts, conference proceedings, and the like, there is far too much information for anyone to absorb and meaningfully apply in making treatment decisions. Finally, in the process of preparing formal papers, much information is delayed or, worse, lost. Interim data and results are typically discarded, especially the results of failed experiments and clinical trials, dooming others to waste time rediscovering them again—a tragedy when patients are involved. Although journals and funding agencies are committed, in principle, to requiring that data associated with publications be made available, in practice, this only succeeds in the few cases for which community-endorsed repositories exist, for example, those maintained by the NCBI. Beyond access to data, there is the deeper issue of making the conclusions conveyed in a scientific paper available in a structured form that can be understood and manipulated by computers as well as by human scientists.

In Cancer Commons, the equivalent of peer review occurs continuously through discussion forums. The experts who curate the molecular disease model aggregate this information and make it actionable for patients with specific molecular subtypes of cancer, and revisions are disseminated through the Career Commons rapid communitywide interactive publication model. As such, the molecular disease model becomes, in effect, a dynamic review article subject to continuous review and revision based on the latest clinical and laboratory findings, with all conclusions and supporting data available in a structured format suitable for computation.

All contents of Cancer Commons are available for semantic annotation so that they can be brought immediately to the attention of all researchers, clinicians, and patients for whom that information may be relevant. Beyond ensuring timely access to knowledge by humans, semantic annotation is also the key to making that knowledge machine understandable.

The Cancer Commons platform includes applications that enable participants to not just annotate papers but also comment on the annotations of others and weave them into structured arguments that support and refute hypotheses [9]. Applications can use this structured knowledge, for example, to track and prioritize targets, leads, and trials or suggest experiments to validate clinically observed responses. Capturing and using this structured discourse adds a new dimension to knowledge sharing not available in traditional publishing media.

The platform also provides access to an ecosystem and catalog of core services and capabilities. Individual researchers, institutions, and companies can publish information about their core competencies and resources, from clinical trial design to molecular profiling, and (where possible) computational interfaces to these services (commonly known as application programming interfaces, or APIs). The operators of personalized virtual biotechs can readily discover and use these APIs through the platform, thus offering participating researchers unprecedented access to industrial-scale, high-quality services. Workflow planning, execution, and tracking, mentioned above (although not yet implemented in the current Cancer Commons platform), go hand-in-hand with this service's ecosystem. When workflow facilities become available, they will enable operators of personalized virtual biotechs to automate and share best practices.

A typical complex workflow might continuously search connected specimen banks for relevant newly deposited materials, automatically dispatch them for molecular profiling, gather and reintegrate the resulting data, run it through statistical methods that correlate the molecular with clinical data related to the specimens, and make the results available for collaborative interpretation and decision making. When a decision is made regarding hypotheses to move forward, an adaptive *N*-of-1 experiment, such as described in the preceding section, can be planned and operated through the platform, and the data can be gathered and analyzed in the same setting.

The sheer volume and complexity of available information—genes, proteins, pathways, disease, mechanisms, drugs, patient clinical, genomic, and response profiles, and so forth—far exceed the synthetic and analytic capacities of any individual or human analysts. To create understanding from information, we need biocomputational tools that can manipulate, check, and use the formal representations to make predictions and form explanations. Such tools for hypothesis generation, knowledge capture, and so forth, use formal representation and computational methods to enable scientists to work efficiently with complex models [10]. By virtue of having been built on top of BioBike, a powerful cloud-based semantic biocomputation platform [3], the Cancer Commons platform provides access to state-of-the-art computational biology and machine learning services and integrated access to a variety of data and knowledge from databases that contain curated representations of clinical trials, known regulatory pathways, drug targeting information, genomic, proteomic, and biochemical knowledge, and so forth, as needed by the operations of the various computations carried out through the platform.

Taken in combination, the facilities of Cancer Commons offer its community of patients, physicians, and researchers unprecedented power to collaboratively manipulate and interpret the wealth of data and knowledge contained within the commons, and from other (mostly public) resources, to efficiently search for treatments and cures.

11.5 DISCUSSION

The goal of cancer commons is to enable physicians and researchers to provide each and every cancer patient with the best possible treatment options based upon the most up-to-date science, and to aggregate and analyze the results of all such activities to improve cancer treatment for every subsequent patient. By virtue of patients, clinicians, and researchers sharing a common platform, the lag from treatment to data to analysis to proposed treatment guidelines to implementation in individual patients can be dramatically reduced.

This new collaborative model for real-time translational research and personalized oncology is powered by a transformational cloud-based platform that employs state-of-the-art computational technologies and communication paradigms, enabling individuals and organizations to build on each other's services, creating new services and linking them into an industry-transforming, network-centric model for delivering personalized medical care.

CollabRx is building Cancer Commons incrementally, one cancer type at a time, beginning with melanoma, one of the least tractable and most rapidly increasing types of cancer. We are developing each new commons in partnership with leading professional and patient advocacy organizations in that cancer. Additionally, we encourage large institutions such as universities and pharmaceutical companies to create private commons to coordinate their cancer initiatives and supplement their own data, selectively sharing results with the public commons as they feel ready to release them. We envision eventually linking these individual commons through the Cancer Commons platform to exploit cross-learnings and economies of scale across cancers and institutions. Because the Cancer Commons model leverages Internet communications speed, flexibility, and scale, it can ultimately range over diverse diseases, disciplines, institutions, geography, and timescales.

The Cancer Commons collaboration model and underlying computational infrastructure enable a new paradigm for translational biomedicine in which each and every patient's experience, each and every physician's knowledge, and each and every researcher's best efforts are brought to bear on ensuring that each and every cancer patient receives the best possible personalized therapy and obtains the best possible outcome.

Un pour tous, tous pour un!

ACKNOWLEDGMENTS

The Cancer Commons platform was built by the CollabRx engineering team led by Mike Travers, Randy Gobbel, and Jay Bartels. Smruti Vidwans developed the first molecular disease model, in melanoma. We also gratefully acknowledge the intellectual contributions of John Curd, George Lundberg, Rob Rodin, and Allan Schiffman. Steve Bagley, Sean Ekins, Gavin Gordon, and George Lundberg provided helpful comments on various drafts of this chapter.

REFERENCES

1. Harold Varmus returns to politics. http://www.npr.org/templates/story/story.php?storyId=128568241. Accessed.
2. Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* 2010;465:473–477.
3. Elhai J, Taton A, Massar JP, Myers JK, Travers M, Casey J, BioBIKE: A Web-based, programmable, integrated biological knowledge base. *Nucl Acids Res* 2009; 37(Web Server issue):W28–W32.
4. Tenenbaum JM. AI meets Web 2.0: Building the Web of tomorrow, today. *AI Mag* 2006;27(4):47–68.
5. Curse of dimensionality. http://en.wikipedia.org/wiki/Curse_of_dimensionality; accessed 201007014.
6. Berry SM, Carlin BP, Lee JJ, Muller P. *Bayesian Adaptive Methods for Clinical Trials*. Biostatistics Series. Boca Raton, FL: Chapman & Hall/CRC, 2010.
7. Scott CT, Baker M. Overhauling clinical trials. *Nature Biotechnol* 2007; 25: 287–292.
8. Shrager J. The promise and perils of pre-publication review: A multi-agent simulation of biomedical discovery under varying levels of review stringency. *PLoS ONE* 2010;5(5):e10782.
9. Shrager J, Billman DO, Convertino G, Massar JP, Pirolli PL, Soccer science and the Bayes community: Exploring the cognitive implications of modern scientific communication. *Topics Cognitive Sci* 2009;2:53–72.
10. Fedoroff N, Racunas S, Shrager J. Tools for thought in the age of biological knowledge. *Scientist* 2005;19:20–21.

12

COLLABORATIVE DEVELOPMENT OF LARGE-SCALE BIOMEDICAL ONTOLOGIES

TANIA TUDORACHE AND MARK A. MUSEN

12.1 ONTOLOGIES IN BIOMEDICINE

The use of formal systems to define biomedical concepts and to represent and store biomedical knowledge has never been more important. From the

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.

implementation of hospital information systems to the organization of experimental data for bioinformatics research, developers now identify the key issue to be the manner in which salient concepts are labeled and defined and ultimately used computationally. In the past decade, the notion of ontology has become a well-recognized substrate for research in informatics and computer science. Ontologies are formal descriptions of the entities that exist in an application area, the properties of those entities, and the relationships among those entities that can be referenced by both humans and computers.

In biology and medicine, ontologies have become central to the construction of intelligent decision support systems, simulation systems, information retrieval systems, and natural language systems [1, 2]. The World Wide Web Consortium (W3C) has developed the resource description framework (RDF) [3] and Web Ontology Language (OWL) [4], standard languages for representing ontologies on the Semantic Web. The introduction of these W3C standards has become an impetus to an even wider use of ontologies and knowledge-based resources. W3C has an extremely active Semantic Web for Health Care and Life Sciences Interest Group [5]. A fast growing collection of well-known biomedical ontologies and terminologies [e.g., BioPAX, the National Cancer Institute (NCI) Thesaurus, the Ontology of Biomedical Investigations] now embraces W3C's OWL standard.

With this adoption of ontologies by the broad biomedical community, however, comes a new challenge: Ontologies and terminologies can become so large, diverse, and specialized that it is often impossible for any single centralized group to develop them effectively. Indeed, several prominent ontology engineering projects are promoting community participation as a key element in their development work. Below, we describe briefly three of the most visible collaborative ontology projects now ongoing: The *Gene Ontology* (GO), The 11th revision of the *International Classification of Diseases* (ICD), and the NCI *Thesaurus*. GO is one of the more prominent examples of an ontology that is a product of a collaborative process [6]. GO provides terminology for consistent description of gene products in different model–organism databases in terms of their associated biological processes, cellular components, and molecular functions in a species-independent manner. Members of the GO community constantly suggest new terms for this ontology. Three full-time curators examine the suggestions and incorporate them into GO as appropriate on a continual basis.

The ICD (http://www.who.int/classifications/icd/en/) is a public global standard to organize and classify information about diseases and related health problems. The World Health Organization (WHO) plans three major shifts for the upcoming 11th revision of ICD (ICD-11) [7, 8]: First, the ICD will represent clinical knowledge explicitly in machine-processable form using OWL. The WHO plans to use an ontological approach, formalizing the definitions of each clinical entity and organizing the terms in a semantically meaningful way. Second, the WHO will open the process of ICD revision to a wide community of experts. Topic advisory groups (TAGs) will serve as planning and coordinating bodies for specific areas of medicine, such as oncology, mental health, and communicable diseases. Each TAG will support several international working groups and an additional corps of field testers who will use online tools to evaluate the evolving ontology and to generate proposals for revisions and enhancements. Third, ICD-11 will include direct linkages to terms in other standardized terminologies, such as SNOMED-CT. The WHO plans to open the development of ICD-11 to the broader community in a social process similar to that supported by Wikipedia.

The NCI Thesaurus is a biomedical reference ontology that covers areas of basic cancer biology, translational science, and clinical oncology developed at the NCI Center for Bioinformatics and Information Technology [9, 10]. The content of the NCI Thesaurus is contributed by internal editors who work separately. Every month or so, a curator goes over the editors' proposed changes and approves or rejects them. Once the curation process is completed, a new version of the NCI Thesaurus is published, and the entire development process starts again.

Each of these projects relies on both a well-defined social process as well as a set of specialized interactive tools to support collaborative ontology engineering—each in a very different way. At the U.S. National Center for Biomedical Ontology (NCBO), we have been working to engineer both processes and tools to enable alternative forms of collaborative ontology engineering.

In this chapter, we will present methods and tools to support the community-driven authoring of large biomedical ontologies. In Section 12.2, we present Collaborative Protégé—a tool that supports collaboration as an integral part of the ontology editing process, and that has features for tracking provenance and changes, as well as discussions in the context of the ontology. In Section 12.3, we describe WebProtégé—a Web-based client for Collaborative Protégé that we built as a highly customizable platform which can be adapted as a knowledge acquisition tool for domain experts. Section 12.4 presents the architecture of the collaboration framework that both Collaborative Protégé and WebProtégé are using. In Section 12.5, we give an overview of BioPortal—a repository of biomedical terminologies and ontologies on the Web. BioPortal provides Web services for publishing, storing, and retrieving ontologies from the repository, supports a proposal and discussion mechanism, and has facilities for versioning and mapping between ontologies. Work on Collaborative Protégé and WebProtégé is becoming tightly linked, as the NCBO explores coordinated methods of community-based ontology authoring, dissemination, and peer review. We discuss our experience with these technologies and talk about future plans in Section 12.6.

12.2 COLLABORATIVE PROTÉGÉ

Our laboratory has developed Protégé—probably the most widely used open-source ontology and knowledge base editor [11]. At the time of this writing, Protégé has more than 150,000 registered users. Users can build ontologies in

Protégé using different representation formalisms, ranging from frames to RDF(S) and to OWL, and can store their ontologies in file or database backends. Protégé is both robust and scalable and is being used in production environments by many government and industrial groups. Protégé supports a plugin architecture that allows other developers to implement their own custom extensions that can be used either in the Protégé user interface or as part of other applications that use Protégé services.

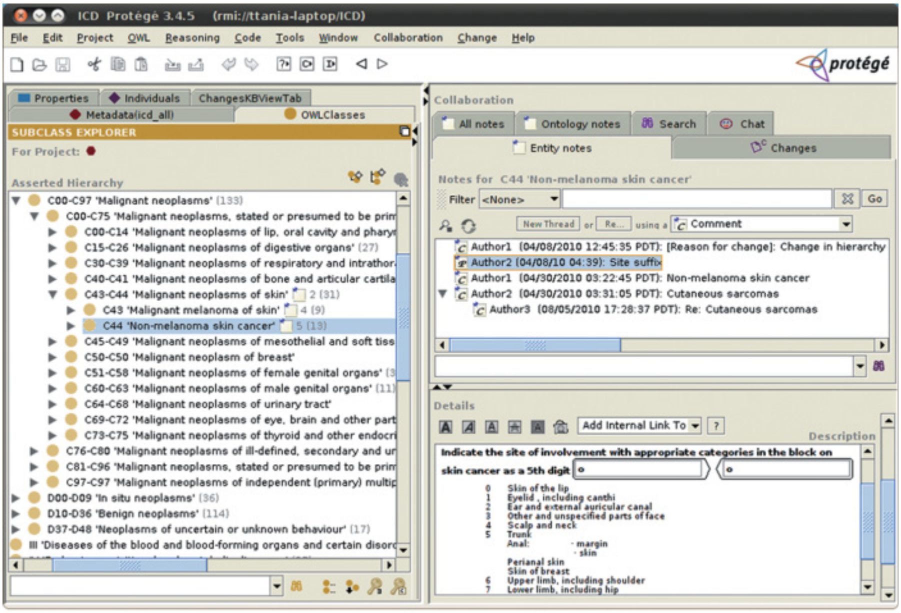
Protégé also offers a client–server mode, in which multiple users can browse and edit a shared ontology at the same time. If a user makes a change to an ontology, then other users will see that change right away. This immediate synchronization of changes minimizes the number of editing conflicts in a collaborative setting, because each of the distributed users receive updates from the other clients in a matter of seconds. The immediate updates make editing conflicts (e.g., two users editing the same property value at the same time) less likely. A different approach for synchronizing changes is found in source-control repositories, such as SVN or CVS. In these settings, users check out a copy of the shared ontology and perform changes in the local copy. When the user is done making changes, he or she has to commit the local changes back to the shared copy stored in the repository. In this setting, it is much more likely that editing conflicts will occur, because other users may already have changed the shared copy since the time that the user checked out his or her own copy. The user has to solve the conflicting changes manually, and this is a very effort-intensive and error-prone task. Both synchronization modes, the immediate mode available in Protégé and the check-out/check-in approach used in SVN, are appropriate for different collaboration scenarios. We have chosen to use the immediate synchronization mode because it is preferred when distributed users may need to make changes that may have ramifications that are dispersed within the edited content and when the users cannot tolerate being locked out from making certain edits for long periods of time.

Although enabling distributed users to browse and edit a shared copy of an ontology brings a lot of advantages, there are still many features that are needed to support real collaboration among ontology builders. Our group has developed Collaborative Protégé (http://protegewiki.stanford.edu/wiki/Collaborative_Protege) as an extension of the client–server version of Protégé to support collaboration as an integral part of the ontology development process [12, 13]. Collaborative Protégé has features for adding comments and notes and discussion threads to ontology classes, properties, and individuals. All changes made by users are tracked and stored together with the related provenance information. It is also possible to add different types of proposals for changes to entities in the ontology and then later to ask users to vote on these proposals. An access policy mechanism allows an administrator to set different privileges for different user groups and hence to restrict access based on particular accounts. A chat feature enables users who are connected at the same time to a Protégé server to exchange text messages and to send references to entities in an ontology. In the remainder of this section, we present some of the most salient features of Collaborative Protégé.

12.2.1 Marginal Notes and Support for Discussions

When an ontology is developed by a large community in a collaborative manner, it is only natural that there will be discussion related to specific modeling decisions or other issues. Usually, such discussions take place in discussion forums, teleconferences, or personal conversations, and often the design rationale or the motivation for certain decisions is lost and it is hard to retrieve at a later time.

Collaborative Protégé offers facilities for carrying out such discussion as an integral part of ontology development. Figure 12.1 shows the user interface of Collaborative Protégé. Collaborative Protégé is an extension of the Protégé user interface: The "regular" user interface showing the class tree and properties has been extended with an additional window that manages the collaboration-related information. The right-hand window shows the collaborative panels, such as Entity notes, Changes, Ontology notes, and so on. In the example in Figure 12.1, users have added five notes on the class selected



C00-C97 'Malignant neoplasms' (133)
C00-C75 'Malignant neoplasms, stated or presumed to be prim
C00-C14 'Malignant neoplasms of lip, oral cavity and phary
C15-C26 'Malignant neoplasms of digestive organs' (27)
C30-C39 'Malignant neoplasms of respiratory and intrathor
C40-C41 'Malignant neoplasms of bone and articular cartila
C43-C44 'Malignant neoplasms of skin' 2 (31)
C43 'Malignant melanoma of skin' 4 (9)
C44 'Non-melanoma skin cancer' 5 (13)
C45-C49 'Malignant neoplasms of mesothelial and soft tiss
C50-C50 'Malignant neoplasm of breast'
C51-C58 'Malignant neoplasms of female genital organs' (3
C60-C63 'Malignant neoplasms of male genital organs' (11)
C64-C68 'Malignant neoplasms of urinary tract'
C69-C72 'Malignant neoplasms of eye, brain and other part
C73-C75 'Malignant neoplasms of thyroid and other endocri
C76-C80 'Malignant neoplasms of ill-defined, secondary and ur
C81-C96 'Malignant neoplasms, stated or presumed to be prim
C97-C97 'Malignant neoplasms of independent (primary) multip
D00-D09 'In situ neoplasms' (36)
D10-D36 'Benign neoplasms' (114)
D37-D48 'Neoplasms of uncertain or unknown behaviour' (17)
III 'Diseases of the blood and blood-forming organs and certain disord

Notes for C44 'Non-melanoma skin cancer'

Author1 (04/08/2010 12:45:35 PDT): [Reason for change]: Change in hierarchy
Author2 (04/08/10 04:39): Site suffix
Author1 (04/30/2010 03:22:45 PDT): Non-melanoma skin cancer
Author2 (04/30/2010 03:31:05 PDT): Cutaneous sarcomas
Author3 (08/05/2010 17:28:37 PDT): Re: Cutaneous sarcomas

Indicate the site of involvement with appropriate categories in the block on skin cancer as a 5th digit o o

0 Skin of the lip
1 Eyelid , including canthi
2 Ear and external auricular canal
3 Other and unspecified parts of face
4 Scalp and neck
5 Trunk
Anal: - margin
- skin
Perianal skin
Skin of breast
6 Upper limb, including shoulder
7 Lower limb, including hip

Figure 12.1 User interface of Collaborative Protégé. The left-hand-side panel shows the class hierarchy; the middle panel (hidden in this screenshot) shows the details of the selected class. The right-hand-side panel shows the collaborative panels. The entity notes panel allows users to attach notes and discussion threads to classes, properties, and individuals in the ontology.

in the left-hand panel, showing the class tree. The notes and discussions are stored in the context of the entity to which they refer. For example, the dialog about a specific class in an ontology, such as Non-melanona Skin Cancer (as shown in Fig. 12.1) is recorded directly in the context of this class, so that the users' comments can be easily retrieved at any time. The class tree shows a "note" icon next to the classes that have notes attached to them. For example, the Non-melanona Skin Cancer class has 5 notes attached to it and 13 notes attached to the children of this class (shown as a number in parentheses).

Displaying the number of children notes helps users to find notes that are attached somewhere deeper in the class tree. The number of notes in a branch is a good indication about the level of activity of that ontology branch and can help project managers focus their attention on "hot" areas in the ontology.

The notes and discussions are also threaded, so that a user may reply to an existing note in a manner similar to that of responding to a message in a discussion forum. The notes have different types, ranging from simple comments to structured proposals for ontology changes. The notes are stored as instances in the Protégé *Changes and Annotation Ontology* (ChAO) [14]. Figure 12.2 shows the class hierarchy of the note types with Annotation as the top-level class. All note types are represented as subclasses of the Annotation class. An instance of a subclass of Annotation, such as Comment, represents an actual comment that a user has added in the tool. Each note type has its own properties, such as the content of the note, author, and date. The properties associated with a given class of note allow Collaborative Protégé to define structured notes in a declarative way. For example, by editing the ChAO, we easily added a note type, NewEntityProposal, that enables users to add a proposal for new entities in the domain ontology under development.

A NewEntityProposal has additional properties, such as the preferred name of the new term that the user proposes to add to the domain ontology, a proposed identifier, a textual definition, and a list of synonyms. An example of such a new entity proposal is shown in Figure 12.3. Collaborative Protégé

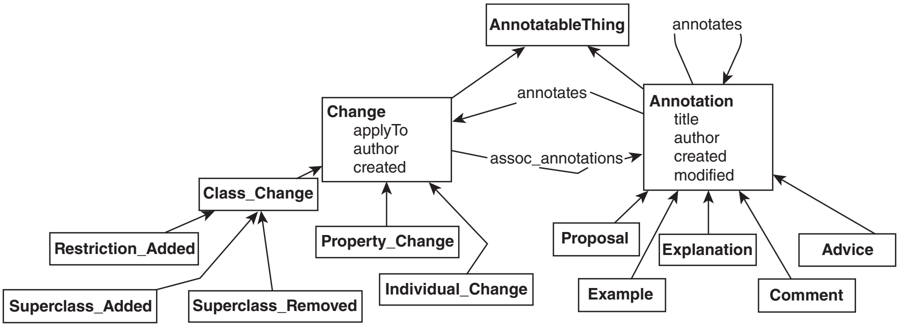
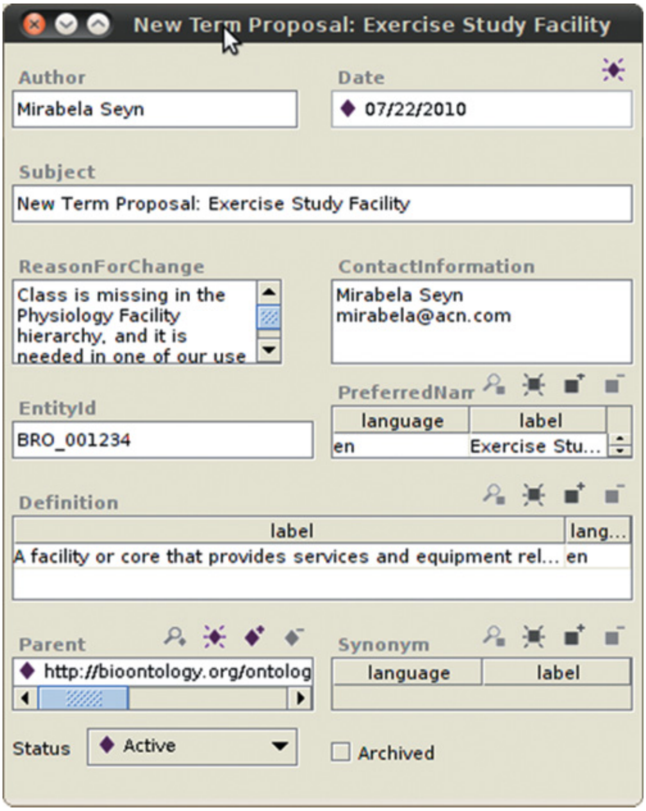


Figure 12.2 Changes and Annotation Ontology. Subclasses of the annotation class represent types of notes available to the user (e.g., proposal, advice). Changes are represented as instances of the change class.



New Term Proposal: Exercise Study Facility

Author: Mirabela Seyn

Date: ♦ 07/22/2010

Subject: New Term Proposal: Exercise Study Facility

ReasonForChange: Class is missing in the Physiology Facility hierarchy, and it is needed in one of our use

ContactInformation: Mirabela Seyn
mirabela@acn.com

EntityId: BRO_001234

PreferredNam

language	label
en	Exercise Stu...

Definition

label	lang...
A facility or core that provides services and equipment rel...	en

Parent: ♦ http://bioontology.org/ontolog

Synonym

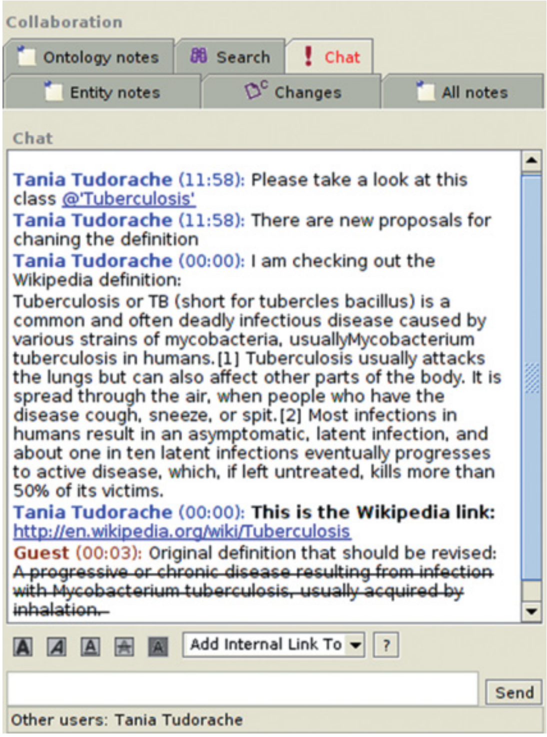
language	label

Status: ♦ Active

□ Archived

Figure 12.3 A new term proposal example showing the different fields of the proposal. The fields are defined as properties attached to the new term proposal class in ChAO.

offers a flexible and extensible mechanism for defining structured notes that allow users to define their own note types, which fit the specific needs of their project. Adding a new note type is as easy as adding a new subclass of the Annotation class in the ChAO, and defining the properties of the new class (the user guide for adding a new note type is available here: http://protegewiki.stanford.edu/wiki/Collaborative_Protege#Adding_your_own_note_type). There is no need to recompile the code or even to restart the application. Collaborative Protégé will automatically pick up the new note type and display it as one of the note-type choices in the user interface. Users may attach notes (Fig. 12.4) to any entity in an ontology, such as classes, properties, or individuals. It is also possible to add notes directly at the ontology level. Discussions that pertain to the entire ontology (e.g., naming conventions, modeling patterns) are usually attached at the ontology level rather than at the entity level. The ontology-level notes are displayed in the Ontology notes tab, seen in Figure 12.1.



Tania Tudorache (11:58): Please take a look at this class @'Tuberculosis'
Tania Tudorache (11:58): There are new proposals for chaning the definition
Tania Tudorache (00:00): I am checking out the Wikipedia definition:
Tuberculosis or TB (short for tubercles bacillus) is a common and often deadly infectious disease caused by various strains of mycobacteria, usuallyMycobacterium tuberculosis in humans.[1] Tuberculosis usually attacks the lungs but can also affect other parts of the body. It is spread through the air, when people who have the disease cough, sneeze, or spit.[2] Most infections in humans result in an asymptomatic, latent infection, and about one in ten latent infections eventually progresses to active disease, which, if left untreated, kills more than 50% of its victims.
Tania Tudorache (00:00): **This is the Wikipedia link:** http://en.wikipedia.org/wiki/Tuberculosis
Guest (00:03): Original definition that should be revised: ~~A progressive or chronic disease resulting from infection with Mycobacterium tuberculosis, usually acquired by inhalation.~~

Other users: Tania Tudorache

Figure 12.4 The chat panel allows users to send instant messages and direct links to ontology entities.

Collaborative Protégé also supports alternative workflows involving the notes. Notes may have a status property attached to them (visible in Fig. 12.3). For example, a proposal could have a status such as Under discussion, Accepted, or Rejected. Other status designations can be configured in the Changes and Annotations Ontology for different types of notes by simply creating instances of the Status class. Using this flexible approach, one can customize Collaborative Protégé to match a particular workflow for an application.

Our users have also requested a way to archive notes and discussions. A note can be archived by simply clicking on the Archive check box that is part of a note form. The user interface can be configured to show or hide archived notes using the Collaboration menu. This feature is particularly important in the workflow for ontology development: Once a proposal has been closed, either by implementing it or rejecting it, a user with enough privileges can archive the proposal and hide it, so that it does not clutter the display of active discussions related to an entity. For documentation and history purposes, the archived notes and proposals can be retrieved at any time.

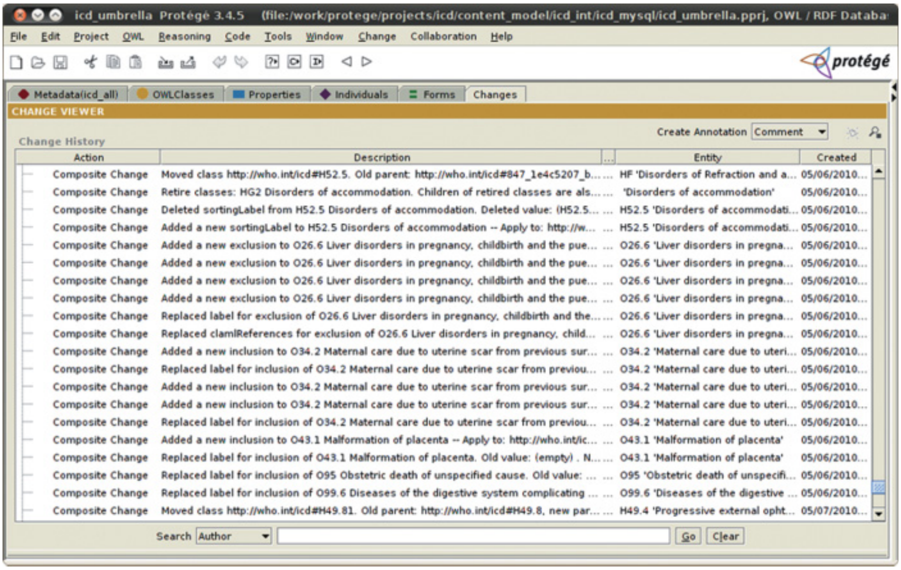
Users may also search the existing notes and discussions in the Search tab. The criteria for search include author name, note text, note type (e.g., Comment, Proposal), status, and begin and end date. These search criteria can be combined using the AND and OR logical operators.

12.2.2 Tracking Changes

Collaborative Protégé tracks all changes the authors make in the ontology. The Changes tab presents a list of changes in chronological order. Figure 12.5 shows the Changes tab with some of the changes performed during the ICD-11 development process. Each change in the table has metadata associated with it, including a human-readable description (e.g., "Added a new definition to A17 Tuberculosis of nervous system"), an author (not shown in Fig. 12.5), a date when the operation was performed, the entity on which the operation was performed (e.g., the class A17 Tuberculosis of nervous system), and the type of the change (e.g., Class created, Class deleted, Composite change, etc.).

The individual change records are stores as instances of the Change class in the ChAO ontology. Each subclass of the Change class represents a different type of change. Figure 12.2 shows some of the main subclasses of the Change class. We use the type of changes to compute statistics about these changes, such as the number of classes that have been moved from one branch of an ontology to another one.

A Composite change is an operation that the user perceives as atomic but which is composed of several other operations. For instance, retiring a class


icd_umbrella Protégé 3.4.5 (file:/work/protege/projects/icd/content_model/icd_int/icd_mysql/icd_umbrella.pprj, OWL / RDF Databa
File Edit Project OWL Reasoning Code Tools Window Change Collaboration Help
protégé
Metadata(icd_all) OWLClasses Properties Individuals Forms Changes
CHANGE VIEWER
Change History
Create Annotation Comment

Action	Description	...	Entity	Created
Composite Change	Moved class http://who.int/icd#H52.5. Old parent: http://who.int/icd#847_1e4c5207_b...	...	HF 'Disorders of Refraction and a...	05/06/2010...
Composite Change	Retire classes: HG2 Disorders of accommodation. Children of retired classes are als...	...	'Disorders of accommodation'	05/06/2010...
Composite Change	Deleted sortingLabel from H52.5 Disorders of accommodation. Deleted value: (H52.5...	...	H52.5 'Disorders of accommodati...	05/06/2010...
Composite Change	Added a new sortingLabel to H52.5 Disorders of accommodation -- Apply to: http://w...	...	H52.5 'Disorders of accommodati...	05/06/2010...
Composite Change	Added a new exclusion to O26.6 Liver disorders in pregnancy, childbirth and the pue...	...	O26.6 'Liver disorders in pregna...	05/06/2010...
Composite Change	Added a new exclusion to O26.6 Liver disorders in pregnancy, childbirth and the pue...	...	O26.6 'Liver disorders in pregna...	05/06/2010...
Composite Change	Added a new exclusion to O26.6 Liver disorders in pregnancy, childbirth and the pue...	...	O26.6 'Liver disorders in pregna...	05/06/2010...
Composite Change	Added a new exclusion to O26.6 Liver disorders in pregnancy, childbirth and the pue...	...	O26.6 'Liver disorders in pregna...	05/06/2010...
Composite Change	Replaced label for exclusion of O26.6 Liver disorders in pregnancy, childbirth and the...	...	O26.6 'Liver disorders in pregna...	05/06/2010...
Composite Change	Replaced clamlReferences for exclusion of O26.6 Liver disorders in pregnancy, child...	...	O26.6 'Liver disorders in pregna...	05/06/2010...
Composite Change	Added a new inclusion to O34.2 Maternal care due to uterine scar from previous sur...	...	O34.2 'Maternal care due to uteri...	05/06/2010...
Composite Change	Replaced label for inclusion of O34.2 Maternal care due to uterine scar from previou...	...	O34.2 'Maternal care due to uteri...	05/06/2010...
Composite Change	Added a new inclusion to O34.2 Maternal care due to uterine scar from previous sur...	...	O34.2 'Maternal care due to uteri...	05/06/2010...
Composite Change	Added a new inclusion to O34.2 Maternal care due to uterine scar from previous sur...	...	O34.2 'Maternal care due to uteri...	05/06/2010...
Composite Change	Replaced label for inclusion of O34.2 Maternal care due to uterine scar from previou...	...	O34.2 'Maternal care due to uteri...	05/06/2010...
Composite Change	Added a new inclusion to O43.1 Malformation of placenta -- Apply to: http://who.int/ic...	...	O43.1 'Malformation of placenta'	05/06/2010...
Composite Change	Replaced label for inclusion of O43.1 Malformation of placenta. Old value: (empty) . N...	...	O43.1 'Malformation of placenta'	05/06/2010...
Composite Change	Replaced label for inclusion of O95 Obstetric death of unspecified cause. Old value: ...	...	O95 'Obstetric death of unspecifi...	05/06/2010...
Composite Change	Replaced label for inclusion of O99.6 Diseases of the digestive system complicating ...	...	O99.6 'Diseases of the digestive ...	05/06/2010...
Composite Change	Moved class http://who.int/icd#H49.81. Old parent: http://who.int/icd#H49.8, new par...	...	H49.4 'Progressive external opht...	05/07/2010...

Figure 12.5 The changes tab plug-in shows a structured list of changes in the ontology. Each line in the table represents a change together with its metadata: type, author, description, entity on which the change occurred, and the date.

involves setting the type of the class to owl:DeprecatedClass and moving the class under the Retired parent class. Composite changes are similar to transactions in a database: If one of the subchanges fails, then the composite change will fail as well. We use composite changes to hide some of the internal operations from the user and to generate user-friendly descriptions of these operations.

The Changes tab also allows one to filter the list of changes by author, text in the description, type, and date. Rather than browsing the entire list of changes, a user may also select a class in the class tree and use the Changes tab in Figure 12.1 to see only the changes that are relevant to the selected entity. As we have already mentioned in Section 12.2.1, changes can also be annotated with different types of notes and discussions.

12.2.3 Instant-Message Exchange with Other Users

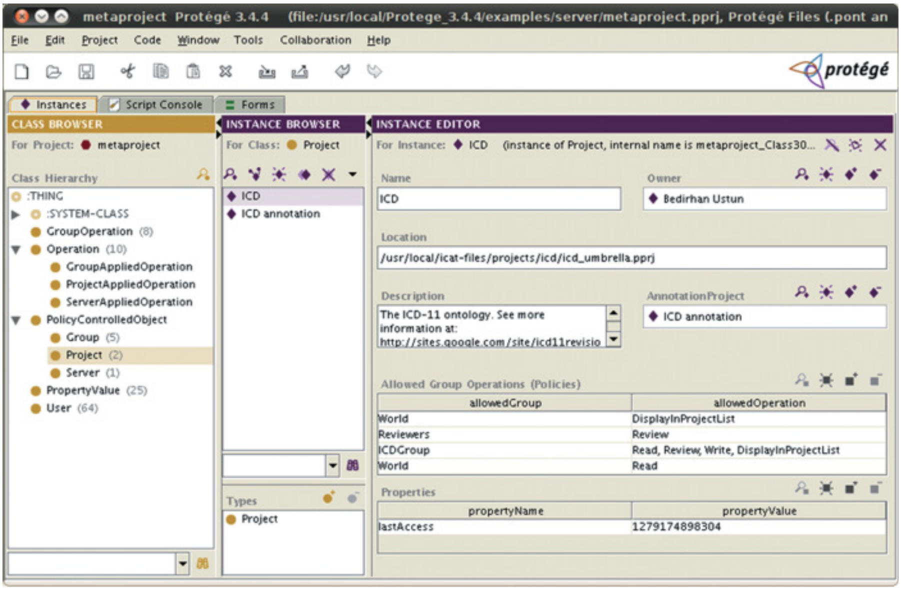
Collaborative Protégé allows users who are connected at the same time to a Protégé server to exchange text messages using an integrated chat client. The chat feature is available as one of the collaborative panels and is shown in Figure 12.1. The chat feature provides complementary functionality to notes and discussions during the ontology development process and allows for quick exchanges between users. The chat messages are also represented as instances in a Chat ontology.

One distinguishing feature of the chat is that it supports sending links to entities in an ontology using a custom syntax. For example, when discussing the Tuberculosis class in an ontology, a user may send a direct link to this class by using the syntax @tuberculosis, similar to a Twitter message. The chat panel will recognize this syntax as a link and will allow other users to click on it and directly browse the Tuberculosis class details. The direct link feature works with all the entity types in an ontology: classes, properties, and individuals. The chat feature is also available as a separate Protégé tab and can be detached from the main display of Protégé.

12.2.4 Access Policies

The support for different access policies is an important part of the collaboration infrastructure. In most community-driven projects, users take on different roles with different tasks and access rights. For example, in the development of the ICD-11 ontology, editors are allowed to edit the content of the ontology, whereas classification experts can change the class hierarchy but cannot edit the content. Everyone is allowed to add proposals or to make comments. Other projects may have other access policies set up.

Collaborative Protégé supports a flexible and extensible access policy mechanism. We represent the access policies and the Protégé server configuration using a simple ontology, called the metaproject. This ontology defines the projects, users, groups, and access policies related to one or more Protégé


metaproject Protégé 3.4.4 (file:/usr/local/Protege_3.4.4/examples/server/metaproject.pprj, Protégé Files (.pont an
File Edit Project Code Window Tools Collaboration Help
protégé
Instances
Script Console
Forms
CLASS BROWSER
INSTANCE BROWSER
INSTANCE EDITOR
Types

For Project: ● metaproject

Class Hierarchy

- :THING
- :SYSTEM-CLASS
- GroupOperation (8)
- Operation (10)
  - GroupAppliedOperation
  - ProjectAppliedOperation
  - ServerAppliedOperation
- PolicyControlledObject
  - Group (5)
  - Project (2)
  - Server (1)
- PropertyValue (25)
- User (64)

For Class: ● Project

- ◆ ICD
- ◆ ICD annotation

Types: ● Project

For Instance: ◆ ICD (instance of Project, internal name is metaproject_Class30...

Name: ICD

Owner: ◆ Bedirhan Ustun

Location: /usr/local/icat-files/projects/icd/icd_umbrella.pprj

Description: The ICD-11 ontology. See more information at: http://sites.google.com/site/icd11revisio

AnnotationProject: ◆ ICD annotation

Allowed Group Operations (Policies)

allowedGroup	allowedOperation
World	DisplayInProjectList
Reviewers	Review
ICDGroup	Read, Review, Write, DisplayInProjectList
World	Read

Properties

propertyName	propertyValue
lastAccess	1279174898304

Figure 12.6 Metaproject ontology used to configure the Protégé server and the access policies. The ICD project has several properties shown in the right-hand-side panel, such as name, owner, location, access policies, and so on.

servers. The metaproject can be edited directly in the Protégé user interface (Fig. 12.6) or can be edited while the server is running using the Server Admin application.

The Protégé server supports three types of access policies: (1) project policies, (2) group policies, and (3) server policies, modeled as subclasses of the PolicyControlledObject class in the metaproject. The *project policies* refer to permissions set for a particular project. For example, for the ICD project, a policy may say that only users from the group ICDGroup are allowed to write to this project. The *group policies* apply to groups. For example, one policy may state that only certain people are able to add users in this group. The *server policies* apply to the Protégé server itself and pertain to more "administrative" types of permission. For example, one server policy might say that only users from the Admin group are allowed to shut down the server.

The main classes in the metaproject ontology are shown in Figure 12.6. The central class is the PolicyControlledObject that has as subclasses all types of objects that can have policies attached to them (projects, groups, and servers). The Operation class represents the different types of operations that can be performed on the PolicyControlledObjects, such as Read and Write, Review, Shutdown server, and so on. The User class represents the users of the Protégé server, who are the performers of ontology operations. The screenshot in Figure 12.6 shows an example of the ICD Project instance, which has several properties attached to it: a name; a location on the server; an owner; a ChAO project that stores the users' notes, discussions, and change history; policies, shown as allowed group operations; and additional properties stored as key–value pairs. The ICD project instance is also associated with the ICD annotation project, which is a ChAO knowledge base containing instances representing the notes and changes for the ICD project.

The metaproject has some predefined operations and access policies that are enforced in Collaborative Protégé. For example, the Read and Write access policies are enforced: If a user does not have the Write privilege for an ontology, any attempt to edit the ontology using either a Protégé client or programmatic access will result in an error. The predefined operations and their documentation are available on the Protégé wiki (http://protegewiki.stanford.edu/wiki/Protege_Client_Server_Tutorial_Configuration).

It is also possible to define custom operations and access policies that meet the specific needs of a project. For example, the NCI has chosen to extend the metaproject with operations and access policies that prevent users from editing the properties in an ontology when using the Properties tab.

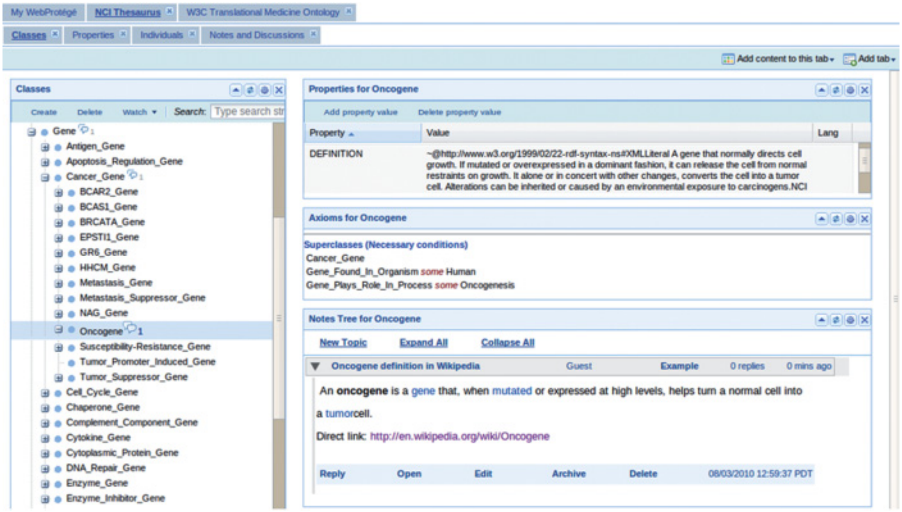
The access policies can be changed whenever a Protégé server is running and they take effect immediately. We have implemented a server-administration application, called Server Admin (a user guide for the administrative panel is available here: http://protegewiki.stanford.edu/wiki/Protege_Client_Server_Tutorial_Administration), which allows administrators to change policies and to add users or projects while the Protégé server is running. Administrators can also monitor the users who are currently logged into the server and browse other statistics related to the projects on the server.

12.3 WEBPROTÉGÉ

WebProtégé is a Web-based client for Collaborative Protégé [15]. WebProtégé provides functionality similar to that of Collaborative Protégé but has the advantage of running in a Web browser and of not requiring any installation. WebProtégé thus supports browsing and editing of ontologies in a collaborative setting on the Web.

12.3.1 WebProtégé User Interface

The user interface of WebProtégé is built as a portal, similar to iGoogle or myYahoo, and is shown in Figure 12.7. The main feature of a portal is that users may add components to the display and may lay them out in an optimal manner for performing their particular tasks. The interface components are called portlets and provide individual pieces of functionality. We have implemented a variety of portlets for WebProtégé that are commonly used in ontology development: a class-tree portlet, a properties-view portlet, a restrictions portlet, a properties-tree portlet, an individuals-list portlet, and so on.



Property	Value	Lang
DEFINITION	~@http://www.w3.org/1999/02/22-rdf-syntax-ns#XMLLiteral A gene that normally directs cell growth. If mutated or overexpressed in a dominant fashion, it can release the cell from normal restraints on growth. It alone or in concert with other changes, converts the cell into a tumor cell. Alterations can be inherited or caused by an environmental exposure to carcinogens.NCI	

Superclasses (Necessary conditions)
Cancer_Gene
Gene_Found_In_Organism some Human
Gene_Plays_Role_In_Process some Oncogenesis

Oncogene definition in Wikipedia Guest Example 0 replies 0 mins ago

An **oncogene** is a gene that, when mutated or expressed at high levels, helps turn a normal cell into a tumorcell.

Direct link: http://en.wikipedia.org/wiki/Oncogene

08/03/2010 12:59:37 PDT

Figure 12.7 WebProtégé user interface showing the NCI Thesaurus. The user interface is organized in portlets that provide independent pieces of functionality. The class tree portlet on the left-hand side shows the class hierarchy. The right-hand side is populated with three portlets for showing the properties of a class, the axioms, and the notes attached to a class.

Additionally, we have implemented portlets to support the collaborative aspects of the development, including a notes and discussions portlet, a review portlet, a change-history portlet, a watched entities and branches portlet, and so on.

The user interface of WebProtégé is organized as a series of tabs. We predefined some of the most common tabs that are available in other desktop ontology development tools, such as the Classes, Properties, and Individuals tabs. A user may change the default layout of these tabs by simply dragging and dropping the portlets in the display. New portlets can be added by making a selection in the toolbar. The user may select from a list of predefined portlets and then configure their layout. The user also has the possibility of adding new tabs to combine pieces of functionality that are not available in the predefined tabs by using the Add tab toolbar menu.

Portlets provide independent pieces of functionality (e.g., display the class tree, display the properties). Rather than hard coding the dependencies among the portlets, we implemented a generic selection model, and all portlets display the current selection in the selection model. For each tab, we define a controlling portlet that provides the selection for all other portlets. For example, the class-tree portlet is the controlling portlet in the Classes tab, which means that all other portlets in that tab will display the information related to the selected class in the class tree. The controlling portlet is specified in the configuration XML file and can be changed dynamically at runtime.

Once users have configured the user interface and saved the layout, the next time that they log into WebProtégé, the customized layout will be restored. WebProtégé stores user interface configurations per user and per project. This means that two users may have two completely different views of the same ontology. If no view is configured for a user and project, WebProtégé will display the default layout as shown in Figure 12.7.

The only way to support the flexible and customizable user interface was to make the user interface layout declarative. The layout configuration is stored as an XML file with a predefined schema (the user guide for the layout configuration can be found here: http://protegewiki.stanford.edu/wiki/WebProtegeLayoutConfig). By changing the XML configuration file, WebProtégé can be easily customized in Web applications tailored for a particular domain. We have used this feature to create a custom Web application used for the development of ICD-11 (The customized ICD-11 application is available at http://icatdemo.stanford.edu) [7, 8, 16] and for other collaboratively developed ontologies.

12.3.2 Browsing and Editing

Multiple users may browse and edit ontologies stored in WebProtégé. The default configuration of WebProtégé shows the Classes tab (Fig. 12.7), which presents in different portlets information related to a selected class in the class tree such as the annotation properties, the restrictions, and the notes and discussions attached to the class. Users may create new classes by using the *Create* toolbar button on top of the class tree. In a similar fashion, other portlets provide support for editing, either by means of toolbar buttons in the portlet or by simply double clicking on a value. Users may edit properties in the Properties tab and individuals in the Individuals tab.

In many cases, the editors of the ontology are domain specialists who are more familiar with Web forms than they are with editing the formal descriptions in the ontology. To accommodate these users and to make WebProtégé easier to use, we implemented a form-based input mechanism in which we can bind properties in the ontology to editing widgets, such as text fields, radio buttons, check boxes, tables, and so on. The form-based input mechanisms are implemented using the PropertyFormPortlet (the form-based input form documentation can be found here: http://protegewiki.stanford.edu/wiki/PropertyFormPortlet). The ICD-11 Web application uses this portlet to support domain experts in entering information related to particular diseases. The example can be browsed in the link that appears in http://icatdemo.stanford.edu.

12.3.3 Collaboration Support

WebProtégé is a client for Collaborative Protégé and therefore has access to all collaboration features presented in Section 12.2, including notes and discussions, change tracking, and access policies. WebProtégé has some additional collaboration features that are not supported in the rich client.

The notes and discussions feature is available in the Notes portlet that can be added to any of the tabs in the user interface, enabling those tabs to show the notes attached to the current selection in the tab (see Fig. 12.7). For example, if this portlet is used in the Classes tab, it will show the notes attached to the currently selected class; if it is used in the Properties tab, it will show the notes attached to the currently selected property, and so on. To help users identify discussion activity in the ontology, the class tree shows an indication that a class or its subclasses have notes attached to them by adding a comment icon next to the class name (see Fig. 12.7). The notes and discussions are shown in a threaded view. It is also possible to delete a note or to edit a note if the user is the creator of that note and no replies have been added to it. Once a note is not relevant anymore, users may archive a note with a simple mouse click.

In addition to attaching notes to entities in the ontology or to the ontology itself, WebProtégé supports adding notes at the level of a triple. For example, a user may add a note to a textual definition of a class that is represented as an annotation property on that class. This feature allows much finer control over the discussions and allows users to focus on a particular issue for an entity.

In a manner similar to that of the rich client, change tracking in WebProtégé is available as a complete change log or as the log of changes for a particular entity in the ontology. The Change history portlet shows the change history of the selected entity in the tab, whether it is a class, property, or individual. WebProtégé also provides a Change statistics tab that displays information about the changes in each branch of the tree. This feature provides a very useful tool for a project manager, who then can assess the level of group activity in each of the branches of the ontology. The change statistics can also be filtered to show only the changes for a specific time interval or for a specific set of users.

The watching and notifications features allow users to track more closely the changes and discussion activity in the ontology. A user may indicate his or her interest in an entity from an ontology (e.g., a given class) or an entire branch of the ontology by "watching" it. To watch an entity or a branch, the user can simply click on the Watch toolbar menu. Once the user logs into WebProtégé, the Watched entities portlet will display the changes that occurred in the watched entities and branches. The user may also configure the WebProtégé notification service to receive e-mails whenever changes to the watched entities occur.

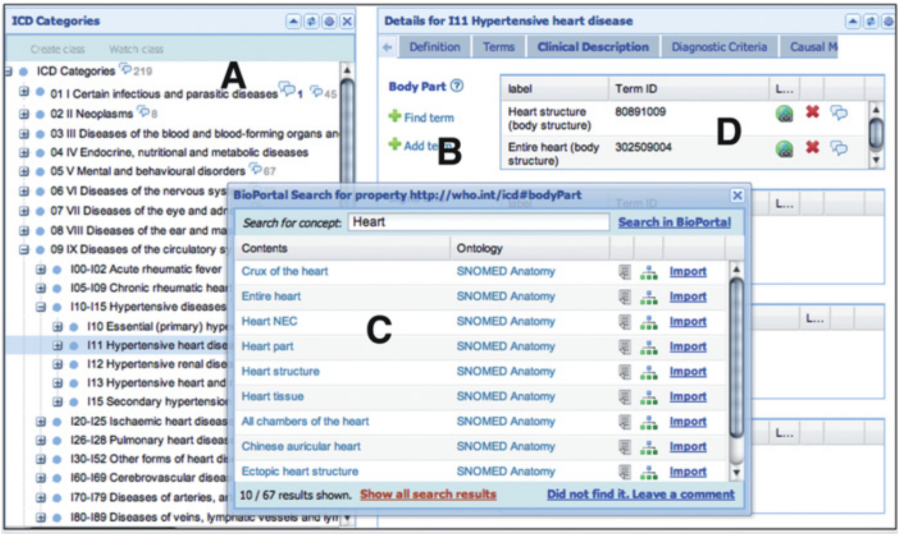
Many collaborative projects have as part of their workflows a reviewing phase: once the users have completed the content authoring in the ontology, reviewers will inspect the content and make recommendations. The Reviews tab in WebProtégé supports a simple reviewing process. A user with adequate privileges may request a review for an entity in the ontology. Once he or she has selected reviewers from a list of predefined reviewers, the request will be sent. The reviewers may enter their reviews directly in WebProtégé. Reviews are a specific note type and are represented as instances of a Review class in the ChAO.

12.3.4 Reusing Terms from Other Ontologies

It is a common case that large-scale biomedical ontologies need to import or reference terms in other biomedical ontologies and terminologies.

To support this task, we developed a generic Reference portlet that searches terms in ontologies stored in the BioPortal ontology repository. BioPortal offers an archive of over 200 ontologies and terminologies for the biomedical domain that can be accessed through a Web browser or through Web services. We describe the collaboration features of BioPortal in more detail in Section 12.5.

The Reference portlet uses RESTful Web services to search for terms in BioPortal. For example, one can imagine a scenario in which a property, say, bodyPart, for an entity Acute Myocardial Infarction, should be a reference to the term Heart in SNOMED-CT. The portlet allows the user to search for the string "heart" in the version of SNOMED-CT stored in BioPortal (Fig. 12.8). The search will return a list of matched terms for the search string "heart." To decide which SNOMED-CT term to import, the user may get more information about each search result either in textual form or as a graph visualization



ICD Categories 219
- 01 I Certain infectious and parasitic diseases 1 45
- 02 II Neoplasms 8
- 03 III Diseases of the blood and blood-forming organs an
- 04 IV Endocrine, nutritional and metabolic diseases
- 05 V Mental and behavioural disorders 67
- 06 VI Diseases of the nervous sys
- 07 VII Diseases of the eye and adr
- 08 VIII Diseases of the ear and ma
- 09 IX Diseases of the circulatory s
  - I00-I02 Acute rheumatic fever
  - I05-I09 Chronic rheumatic hea
  - I10-I15 Hypertensive diseases
    - I10 Essential (primary) hyp
    - I11 Hypertensive heart dise
    - I12 Hypertensive renal dise
    - I13 Hypertensive heart and
    - I15 Secondary hypertensio
  - I20-I25 Ischaemic heart diseas
  - I26-I28 Pulmonary heart diseas
  - I30-I52 Other forms of heart di
  - I60-I69 Cerebrovascular disea
  - I70-I79 Diseases of arteries, ar
  - I80-I89 Diseases of veins, lymphatic vessels and lym

label	Term ID
Heart structure (body structure)	80891009
Entire heart (body structure)	302509004

Search for concept: Heart

Contents	Ontology
Crux of the heart	SNOMED Anatomy
Entire heart	SNOMED Anatomy
Heart NEC	SNOMED Anatomy
Heart part	SNOMED Anatomy
Heart structure	SNOMED Anatomy
Heart tissue	SNOMED Anatomy
All chambers of the heart	SNOMED Anatomy
Chinese auricular heart	SNOMED Anatomy
Ectopic heart structure	SNOMED Anatomy

Figure 12.8 BioPortal reference portlet. The user selects a class in the class tree (A). In order to fill in the value for body part (B), she uses the widget to search BioPortal (C) and to select the appropriate body part from SNOMED-CT. The reference is then added as a link to SNOMED-CT.

that is also retrieved via Web service calls to BioPortal. The Reference portlet is also configurable. For example, a user can specify in what particular ontology from BioPortal the search should be performed.

A user may also restrict the search to a particular ontology branch when configuring the portlet (e.g., the Anatomy branch in SNOMED-CT).

Once the user clicks on the Import button, a reference to the term in the external ontology is created. The reference will have metadata associated with it: the name of the source ontology, the identifier of the referenced term, and a direct Web link to the term.

12.4 COLLABORATION ARCHITECTURE

A high-level overview of the collaboration infrastructure used in Collaborative Protégé and WebProtégé is shown in Figure 12.9. The collaboration framework is the core of the system and provides all collaboration services. Collaborative Protégé and WebProtégé are client applications that access the collaboration framework and display the relevant information.

The entire collaboration process is guided by a set of “meta”-ontologies that are part of the collaboration framework. The ChAO provides support for storing a structured log of ontology changes together with the metadata as well as notes and discussions represented as instances of predefined classes

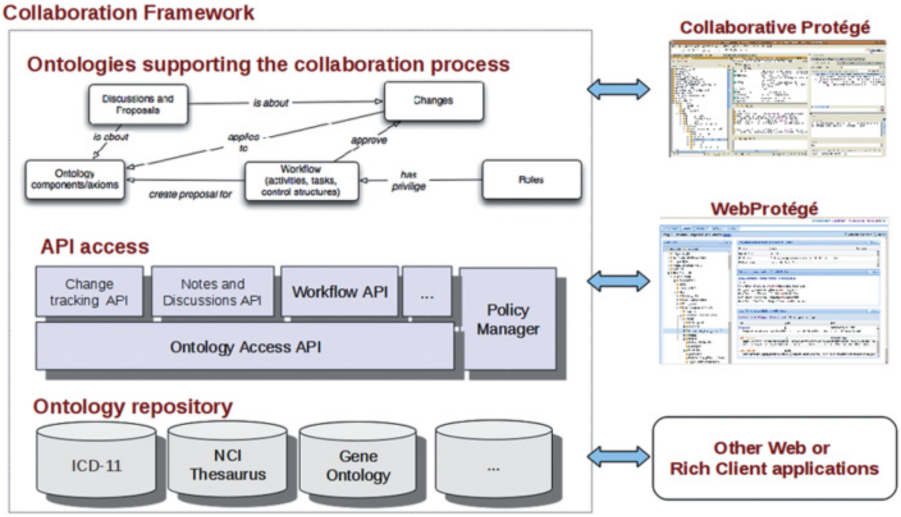


Figure 12.9 Collaboration architecture for Collaborative Protégé and WebProtégé. The collaboration framework (left side) provides all collaboration services used by the Collaborative Protégé rich client and WebProtégé. Ontologies guide the entire collaboration process. A layer of Java APIs provides access to the collaboration information. The ontology repository stores the ontologies available for collaboration.

(see Section 12.2.1). The Metaproject ontology is used to configure the ontology repository and the user access policies (see Section 12.2.4). The Workflow ontology (not discussed in this chapter) is used to configure the steps for a collaboration workflow [17, 18].

We developed a set of Java application programming interfaces (APIs) that provide independent pieces of functionality. Most of the Java APIs are used for accessing and modifying the information stored in one of the meta-ontologies used in the collaboration process. Several of these Java APIs have been generated automatically from the meta-ontology using a Java code generator. In this sense, we can say that the collaboration framework is an ontology-driven architecture. Some of the main Java APIs are shown in Figure 12.9. The Change tracking API and the Notes API retrieve the information stored as instances in the ChAO. The APIs provide convenient methods for the developers and hide the fact that the change and notes information is backed by an ontology. For example, a developer may call a method cls.getChanges() to retrieve all changes associated with a class; it is not important to know that the change information is stored as instances of the ChAO. The Ontology access API provides services to access the ontology content stored in the Protégé ontology repository. The Policy Manager library has services for accessing and modifying the user policies in a programmatic way.

The Protégé ontology repository stores all the ontologies that are available to the collaboration framework and hence also to the Collaborative Protégé and WebProtégé clients. Besides storing the domain ontologies, such as ICD-11 and the NCI Thesaurus, the repository also stores the associated ChAO instances for each of the domain ontologies. This association is stored in the Metaproject ontology (see Section 12.2.4). The collaboration framework is an essential part of the Protégé server (see Section 12.2). Because all clients (Collaborative Protégé, WebProtégé, and other applications) connect to a common collaboration framework, their users will see one another's changes to an ontology as they happen. For instance, if a user adds a class in a Collaborative Protégé–rich client, other users connected to the same server with WebProtégé will see the change in their clients right away. This feature allows users to edit a shared ontology in a collaborative environment with the tool of their choice.

12.5 PUBLISHING ONTOLOGIES WITH BIOPORTAL

Ontology authors and domain experts use Collaborative Protégé and WebProtégé to edit a shared ontology in a collaborative setting. It is a common scenario that once the ontology has reached a mature state, it will be published on the Web to allow the authors to get feedback from a broader user community [19, 20]. We present in this section a solution for publishing ontologies on the Web using BioPortal (http://bioportal.bioontology.org), and we describe how BioPortal services support collaboration and reuse.

BioPortal is an open library of biomedical ontologies [20] created by the NCBO. BioPortal adopts the philosophy of Web 2.0 applications to bring structure and order to the collection and dissemination of biomedical ontologies. The system enables users to provide and discuss a wide array of knowledge components, from submitting the ontologies themselves, to commenting on and discussing classes in the ontologies uploaded by other users, to reviewing ontologies in the context of their own ontology-based projects, to creating mappings between overlapping ontologies and discussing and critiquing the mappings.

BioPortal hosts more than 200 biomedical ontologies in OWL, Open Biomedical Ontologies (OBO), RDFS, and Protégé frame formats and contains more than 2.5 million ontology terms and over 4 million term-to-term mappings. All the information available in BioPortal is accessible via RESTful Web services, which encourages the mashing up of biomedical applications in a straightforward way.

BioPortal and WebProtégé provide complementary services to support collaborative ontology development and dissemination: WebProtégé supports the collaborative editing process, whereas BioPortal provides repository services that are crucial in later stages of the collaborative process, such as reviewing, versioning, creating mappings, and determining structural differences.

Once an author publishes an ontology in BioPortal, the larger community is able to browse and search for terms in the ontology, to review its content, and to add comments and proposals to classes in the ontology. Besides adding notes and proposals through the BioPortal Web interface, the system also supports the posting of notes via Web services. In this way, other applications can access the ontologies in BioPortal and can take advantage of all the services provided by the system programmatically.

BioPortal also supports *peer review* of ontologies. This is a very important feature that aids users who are looking for ontologies that they can reuse in their own projects and who want to know what other users think about alternative resources and problems they may have encountered. A key piece of information is the list of other projects that have used each ontology in BioPortal, and the suitability of the ontology for the tasks of each project. Thus, in addition to submitting ontologies to BioPortal, users may also submit descriptions of their ontology-based projects and link those descriptions to BioPortal ontologies. Registered users can provide comments on BioPortal ontologies along several different dimensions, such as degree of formality, documentation and support, usability, domain coverage, and quality of content.

BioPortal also supports *versioning* of ontologies. The development of large-scale biomedical ontologies usually takes place in iterative steps with several versions of the ontologies being published during the process. Users may store multiple versions of an ontology in BioPortal and can access their content through both the user interface and the Web services. Each version of the ontology has an associated *version identifier* that can be used in the Web Services for retrieving the content of the ontology. Besides the version identifier, each ontology also has an associated *virtual identifier* that points to the latest version of the ontology. The virtual identifier allows other users and applications to access the latest version of an ontology without needing to be aware of previous versions.

BioPortal can also compute the *structural difference* (diffs) between two versions of an ontology, and users can download the diff as an RDF file. The structural diffs provide crucial information for the users of the ontology who need to ensure that other ontologies or the software on which they rely for their applications evolve in concert with the primary changes.

12.6 DISCUSSION AND FUTURE WORK

We have described Collaborative Protégé, WebProtégé, and BioPortal—three tools that provide essential support for the collaborative development of large-scale biomedical ontologies. We have shown how Collaborative Protégé and WebProtégé help users in the collaborative authoring of ontologies, while BioPortal allows the publishing of ontologies and the gathering of feedback from the larger user community.

We envision that these tools will work increasingly seamlessly in the near future. For example, we are working to make it possible to upload an ontology directly from WebProtégé into BioPortal. In a similar way, a user should be able to start a WebProtégé editing session for an ontology in BioPortal with a simple click. We envision a model in which BioPortal will host not only the published versions of an ontology to which users may add comments and proposals but also a *working version* that is not made public and to which the ontology authors may apply the community's proposed changes using WebProtégé.

We are working to offer better support for managing the workflow required for proposals for ontology changes. Any registered user of BioPortal currently can create structured proposals for ontology changes. Ontology authors with the appropriate privileges (e.g., a curator) should be able to open an ontology in WebProtégé and to review the change proposals submitted through BioPortal. For appropriate proposals, curators should be able to apply the changes automatically. In the next step, the curator would set the status of the proposal to Implemented or Rejected, which should be reflected in the proposal status when viewed in BioPortal. Once all the changes have been addressed, the updated ontology should be published as a new version in BioPortal. To support this workflow, we need to implement several new features. First, we need to be able to browse the BioPortal proposals into WebProtégé. Second, curators need to be able to set the status of a proposal once a decision has been taken. For example, a new term proposal contains fields such as identifier, definition, synonym, and preferred name. We need to implement automatic importers for each of the common proposal types that would read the fields from the proposal and create the appropriate classes and properties in the ontology. Third, curators need to be able to set the status of a proposal once a decision has been take, and the status needs to be reflected back into BioPortal. We think that this type of proposal workflow is quite common, and we are working on designing and implementing it. Other biomedical projects may require different workflows, however. Our long-term goal is to provide support for generic collaboration workflows in our tools.

Both WebProtégé and BioPortal are in current production use for several biomedical projects, one of the most prominent ones being the collaborative development of ICD-11 [7, 8]. So far, the collaborative authoring has been performed mainly by WHO internal experts in WebProtégé. BioPortal has been used for importing and referencing terms from other biomedical ontologies and terminologies. In the next ICD-11 revision phase, WHO plans to open the process to a broader user community in which anyone can submit proposals that are reviewed and, if appropriate, integrated in ICD-11. In this new workflow, we envision that BioPortal will be the main platform on which users will add change proposals, and WebProtégé will be the tool in which the changes are applied. In order to have this workflow work, the two systems need to work seamlessly together.

Our experience with the ICD-11 development and with other projects convinces us that users will benefit from a closer integration of the editing and publishing processes as part of a streamlined workflow for collaborative ontology development.

ACKNOWLEDGMENTS

This work has been supported in part by research grant GM086587 from the National Institute of General Medical Sciences, by the Protégé Resource supported by grant P41LM007885 from the National Library of Medicine, and by the National Center for Biomedical Ontology, supported by grant U54HG004028 from the National Institutes of Health.

REFERENCES

1. Bodenreider O, Stevens R. Bio-ontologies: Current trends and future directions. *Brief Bioinform* 2006;7:256–274.
2. Rubin DL, Shah NH, Noy NF. Biomedical ontologies: A functional perspective. *Brief Bioinform* 2008;9(1):75–90.
3. Brickley D, Guha RV. (2004). RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation. W3C. Available: http://www.w3.org/TR/2004/REC-rdf-schema-20040210/.
4. Dean M, et al. Web Ontology Language (OWL) Reference Version 1.0. Available: http://www.w3.org/TR/2002/WD-owl-ref-20021112/, 2002.
5. Ruttenberg A, et al. Advancing translational research with the Semantic Web. *BMC Bioinform* 2007;8(Suppl 3):S2.

6. GOConsortium. Creating the gene ontology resource: Design and implementation. *Genome Res* 2001;11(8):1425–1433.
7. Tudorache T, Falconer S, Noy NF, Nyulas C, Ustun BT, Storey MA, Musen MA. Ontology development for the masses: Creating ICD-11 in WebProtege. In *Knowledge Engineering and Management by the Masses*, Proceedings of the 17th International Conference, EKAW 2010, Lisbon, Portugal, October 11–15, 2010: pp. 74–89. Berlin: Springer, 2010.
8. Tudorache T, Falconer S, Noy NF, Nyulas C, Musen MA. Will Semantic Web technologies work for the development of ICD-11? In *Proceedings of 9th International Semantic Web Conference (ISWC 2010), Shanghai, China*. 2010.
9. Fragoso G, de Coronado S, Haber M, Hartel F, Wright L. Overview and utilization of the NCI Thesaurus. *Compar Funct Genom* 2004;5(8):648–654.
10. Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu WL, Wright LW. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *J Biomed Inform* 2007;40(1):30–43.
11. Gennari J, et al. The evolution of Protege: An environment for knowledge-based systems development. *Int J Human-Computer Interact* 2003;58(1).
12. Tudorache T, Noy NF, Tu SW, Musen MA. Supporting collaborative ontology development in Protégé. Paper presented at the Seventh International Semantic Web Conference, Karlsruhe, Germany. Berlin: Springer, 2008.
13. Noy NF, Tudorache T, de Coronado S, Musen MA. Developing biomedical ontologies collaboratively. Paper presented at the AMIA 2008 Annual Symposium, Washington, DC. 2008.
14. Noy NF, Chugh A, Liu W, Musen MA. (2006). A framework for ontology evolution in collaborative environments. In *Proceedings of the Fifth International Semantic Web Conference, ISWC, Athens, GA*. Berlin: Springer.
15. Tudorache T, Vendetti J, Noy NF. Web-Protege: A lightweight OWL ontology editor for the Web. In *Proceedings of OWL-ED 2008,* CEUR-WS, Vol. 432. Karlsruhe, Germany.
16. Tudorache T, Falconer S, Nyulas C, Storey MA, Üstün TB, Musen MA. Supporting the collaborative authoring of ICD-11 with WebProtégé. Paper presented at the AMIA 2010 Annual Symposium, pp. 802–806. Washington, DC, 2010.
17. Sebastian A, Tudorache T, Noy NF, Musen MA. Customizable workflow support for collaborative ontology development. Paper presented at the 4th International Workshop on Semantic Web Enabled Software Engineering (SWESE) at ISWC 2008, Karlsruhe, Germany. 2008.
18. Sebastian A, Noy NF, Tudorache T, Musen MA. A generic ontology for collaborative ontology-development workflows. Paper presented at the 16th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2008), Catania, Italy. Berlin: Springer, 2008.
19. Noy NF, Tudorache T, Nyulas C, Musen MA. The ontology life cycle: Integrated tools for editing, publishing, peer review, and evolution of ontologies. Paper presented at the AMIA 2010 Annual Symposium, pp. 552–556. Washington, DC. 2010.
20. Noy NF, et al. BioPortal: Ontologies and integrated data resources at the click of a mouse. *Nucl Acids Res* 2009;37. Available: http://nar.oxfordjournals.org/content/early/2009/05/29/nar.gkp440.full.

13

STANDARDS FOR COLLABORATIVE COMPUTATIONAL TECHNOLOGIES FOR BIOMEDICAL RESEARCH

SEAN EKINS, ANTONY J. WILLIAMS, AND MAGGIE A. Z. HUPCEY

13.1 WHAT ARE STANDARDS?

Standards are a common way to ensure quality or provide a measure to which others should conform. Some everyday examples include weights and measures and railway track train gauges (http://en.wikipedia.org/wiki/Rail_gauge) and clearly standards are of value in these areas. There have long been standards for many specific domains and, for example, historical buildings document standards for units of measure such as bread loaf sizes: The exterior walls of Freiburg Minister in Germany displays different sizes for the years 1270 and 1317. Standards in engineering are important so that, for example, railway lines join correctly when started from opposite ends of a country, a nut fits on

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.

a bolt of a particular size, and tools match the dimensions of their associated fasteners. One benefit of standards is that if they are internationally adopted manufacturers of products can cater to a limited number of variants, thereby reducing overheads such as retooling. However, this is often not the case and standards are commonly national in nature. In Europe the metric standard of meters and kilograms is well established, while in the United States miles and pounds prevails. Our free-trade partner bordering our nation (United States), Canada, has adopted the metric system. Standards which are geographical in nature and represent such a "localized approach" will fail when data are expected to traverse organizational boundaries across the world and be mobile between multiple software applications.

As this chapter is written, World Standards Week, as organized by the American National Standards Institute (ANSI), has begun. This year national standards day was on September 23, 2010. All kinds of industries have adopted, are working on, or are renewing their standards on an ongoing basis following iterative gap analyses in capabilities and the need to support changes in technology. Many of these industries seek approval from the International Organization for Standardization (ISO) (http://webstore.ansi.org/SdoInfo.aspx?sdoid=39). Certification against such standards can be essential in industrial competitiveness. The United States also hosts the National Institute of Standards and Technology (NIST), which promotes measurement science, standards, technology, and industrial competitiveness.

In regards to the aims of this book there is a long history of standards development in the biomedical sciences. In terms of the domain of collaboration in informatics to benefit the life sciences we need to consider standards specifically for chemistry and biology. In chemistry specifically, efforts have been made to establish standards that allow for systematic nomenclature generation using the International Union of Pure and Applied Chemistry (IUPAC) rule standards [1], those that allow for the exchange of spectral data using the Joint Committee on Atomic and Molecular Physical Data (JCAMP) standard [2–4], and those that allow for chemical structure interchange and linking using the IUPAC International Chemical Identifier (InChI) [5]. Even chemical textbooks are labeled with the International Standard Serial Number (ISSN), which is used for many types of books. As for standards in biology there are many that have been developed around data sets (see Table 13.1). One could also focus on systems biology which has CellML as an open Extensible Markup Language (XML) standard, Systems Biology Markup Language (SBML) for machine-readable representations of networks, Systems Biology Graphical Notation (SBGN) for human-readable representations of biological networks, and finally BioPAX, a data exchange format for biological pathways [6].

The IUPAC naming conventions for chemicals mentioned previously are applicable to small molecules such as ligands and metabolites, and the Occupational Safety and Health Act (OSHA) standards for handling hazardous chemicals (http://www.osha.gov/Publications/osha3084.html) are essential

TABLE 13.1 Examples of Standards in Biomedical Sciences

Standard Name	Website
The Open Biological and Biomedical Ontologies (OBO)	http://www.obofoundry.org/
The Ontology for Biomedical Investigators (OBI)	http://obi-ontology.org/page/Main_Page
The Functional Genomics Data Society (MGED)	http://www.mged.org/index.html
Minimum Information About a Microarray Experiment (MIAME)	http://www.mged.org/Workgroups/MIAME/miame.html
The Minimum Information About a Bioactive Entity (MIABE)	http://www.psidev.info/index.php?q=node/394
Minimum Information for Biological and Biomedical Investigators (MIBBI)	http://www.mibbi.org/index.php/MIBBI_portal
Minimum Information for Publication of real time QT-PCR data (MIQE)	http://www.gene-quantification.de/miqe-press.html

in the workplace. Standards appear and soon proliferate with each new technology in the biosciences. Such standards are usually initiated by an organization, generally a nonprofit, that brings together key researchers to discuss the needs and approaches for harmonization [e.g., International Life Science Institute (ILSI), http://www.ilsi.org/Pages/AboutUs.aspx].

Many of the newer standards appear to be used as guidelines for publication of the various data types as well as to facilitate data exchange. In fact, many journals list an array of standards which articles must comply with, for example, deposition of Minimum Information About a Microarray Experiment (MIAME) compliant data (Table 13.1) in certain databases. Organizations are increasingly recognizing that adhering to standards is key to reinforcing quality.

In this book several chapters mention the term *standards*. For example, in Chapter 1 open standards are discussed briefly [7]. In Chapter 5 there is discussion of a lack of focus on creating standards or even setting requirements for collaborative technologies [8]. In Chapter 4 standards are discussed in the context of pharmaceutical companies for experiments, data storage, analysis, and recordkeeping (required for regulatory agencies) while these may not be so apparent in academia (especially in terms of the latter regulatory requirements) [9]. In Chapter 17 caBIG is described as providing a set of standards to enable sharing [10] while in Chapter 14 it is stated that users are helping define open standards for data collection and reading by software outside of that provided by the original equipment manufacturer as a lack of openness may become a competitive disadvantage for such companies [11]. In Chapter 28 we mention that the lack of formal standards to annotate publically available screening data limits their integration with other data sources [12]. Other common standards not described in these chapters are listed in Table 13.1.

Already there are many software products identified in this book that are used in the "biomedical collaboration" space though there has been little focus on integration to date. Despite the fact that there are many existing standards which may be embraced to allow integration, it seems appropriate at this juncture to propose standards for Collaborative Computational Technologies for Biomedical Research (CCTBR) as these are long overdue and in the near future will be essential.

13.2 WHY WE NEED STANDARDS FOR COLLABORATION

It is worth asking the question, why do we need such standards? In regards to collaboration in the life sciences we need standards to facilitate comparison of data, molecules, assays, experimental conditions, and so on. Simple issues such as how a company or research group draws a molecule structure in its corporate database can be quite heterogenous, and business rules can vary dramatically. The encoding of chemical structures in a consistent manner (e.g., representation of nitro groups as pentavalent nitrogens versus charge-separated, tautomer identification, chloride versus hydrochloride salts) can be an issue when companies merge and a single database needs to be compiled for the combined entity. Business database integration has a huge cost, and if companies adopted appropriate standards, then merger-related costs would be cheaper. Consideration of the freely available datasets on the Internet shows that in these collections there are also different rules (standards) for storing and displaying molecules.

Reading research papers in biology will quickly show that assay conditions can vary dramatically from one laboratory to another. A simple search in PubMed on "assay conditions and variability" retrieves over 5000 hits. Therefore it is key that if collaborating groups are going to share assay data then they should have identical protocols as minor changes can dramatically impact the resulting structure–activity relationships that could be extracted when assembling the data to model computationally. There are generally no accepted standards to our knowledge for how raw data should be presented or formatted for storage. Should numerical data be stored as text files, as comma-separated values, or as tab-delimited files? What should the orientation of data be in a table? Data output from different experimental hardware may not format the raw data in a manner that can then be readily imported into collaborative databases, something with which the authors of this text have considerable experience. There are also differences in approaches and standards for data cleanup before deposition in any database (missing data, outlier removal, etc.). These generally vary from one endpoint and database to another.

There are also human issues such as the ethical standards required for collaborative software, especially when applied to health and medical records. Obviously, involving human or animal data adds a layer of security and confidentiality compared to data derived from a simple enzyme assay. These human issues could extend to business agreements and intellectual property arrangements that may underpin any major collaboration, but when deposited later into public databases, it is unclear how these will be made transparent. For example, depositing data in the public domain does not mean that there are no constraints of licensing issues if the data are to be used for other purposes. For instance, compounds with associated public malaria data, such as the recently released Glaxo Smith Kline (GSK) data [13], may have been screened against other targets and the owner may have patents or prior art on other activities in which other scientists might be interested. There is thus a significant challenge as to how to make people aware of this. Should public depositors of data be required to reveal all associated constraints simultaneously?

Some collaborations may be very narrow, focusing on a specific target or molecule and requiring sharing of data on one project for only a defined time. There may be certain boundary conditions that could inhibit further collaborations. For example, if the collaboration was to be extended in new directions, then there may be challenges regarding whether any software would be available to integrate and share data with additional systems outside of the original collaboration. Software used for collaborations may not be integrated between two or more parties so the process of connecting data between all the key tools that may be used (e.g., chemistry and biology databases) can become an issue. Computer–computer interactions may simplify or complicate the process compared with human–computer interactions and hard-copy data sharing. The lowest common denominator between people with their different types of data before, during, and after a collaboration and their interactions with collaborative software become very important issues. Chemical structures, experimental data (both continuous and discrete), and computational models derived from such data may all need to be shared. There are, as yet, no agreed standards for quantitative structure–activity relationship (QSAR) model sharing while there are many standards for sharing molecular structures [simplified molecular input line entry specification (SMILES), InChI, SDF, Mol, etc]. Despite the fact that there are several sites that want to promote access to computational models [e.g., Chembench (http://chembench.mml.unc.edu/), Ochem (http://ochem.eu/, http://ochem.eu/static/home.do), and VCCLAB (http://www.vcclab.org/)], it is not yet clear whether any standards will emerge from these sites. QSAR-ML has been recently proposed as an XML exchange format for QSAR data, descriptors, software, and response data [14]. It will be interesting to see if it is accepted by users of software beyond the open-source Bioclipse workbench [15].

Even the manner in which data are uploaded into collaborative software platforms presently could be standardized, and such simplification combines a required format and data organization and should be a catalyst for increased collaboration by lowering the barrier to share data. Providing a simplified data upload standard would be a laudable first goal of any CCTBR standards development effort.

While there has been talk of how since its earliest days the Internet could be used for collaborative research [16], perhaps one of the reasons why biomedical research collaborations have been slow to take off is the lack of "standards." If we are to benefit from the Semantic Web [17] and advance translational research [18], this needs to happen soon.

13.3 HOW WILL WE GET THEM?

Standards can be the result of being mandated by a powerful organization (industry body, government, etc.), evolving into their final form by consensus over time, or becoming *de facto* by popular use (success or failure of other products) and evolution of a software product.

Many major universities have set up drug discovery screening centers primarily focused on high-throughput screening [19]. National Institutes of Health (NIH) funding of these may increasingly be tied to collaboration with demands for the provision of data in standardized formats, and we may see that the academic community or software vendors will drive standards development such as for CCTBR. This means that pharmaceutical companies themselves may not be necessarily "driving" such standards, and they may need to follow academia, unless they take the initiative based on pressure from their academic and public–private collaborations. However, some software standards are presently being driven through initiatives like the Pistoia Alliance [20] (see Chapter 1), which bring together pharmaceutical companies and software vendors as a way to reduce redundancy and repetition and increase cost effectiveness of informatics efforts.

If open-source research and development is really going to reinvigorate drug research [21] with computers at its heart, then collaborations will be integral and essential for validation of hypotheses and will require software to connect the disparate laboratories (whether real, virtual, or collaboratories via Web services; see Chapter 5) [22]. The recent development of SIMBioMS for information management across collaborations is an interesting example for the high-throughput space (OMICS) that supports various standards [23]. This suggests that standards for such collaborative software elsewhere (e.g., biology or chemistry data) are urgently needed and delay could hinder further progress.

Any company involved in defining or delivering such standards for collaborative software may have an advantage over their competitors. However, there could be more progress if such software were open source itself, with the community taking over its support, development, and extension such that a company does not need to pay for the platform development while at the same time it taps into a much bigger developer community. One problem with this approach is garnering the support of enough developers. While it is common for companies to not want to open up their collaborative software, this has changed in recent years as open-source approaches to product development have proliferated.

As a result of the recent recession there is a lot of drug discovery and development talent available now due to company lay-offs. If the software or other tools to enable this workforce to be productive and collaborate were available and they participated in the existing scientific collaboration networks, then there may be potential for enormous breakthroughs. Data are also becoming available from pharmaceutical companies at an increasing rate as part of precompetitive or other data-sharing initiatives [13, 24, 25]. The availability of collaborative computational technologies may help reengage unemployed pharmaceutical researchers as their own virtual medicinal chemistry or other departments. The timing may therefore be right for some of those at the forefront of collaborations and software development of collaborative computational technologies to develop the standards as progress in pharmaceutical research and development may depend on it.

ACKNOWLEDGMENTS

Dr. Barry Bunin is kindly acknowledged for discussions on this topic.

REFERENCES

1. IUPAC Commission on the Nomenclature of Organic Chemistry (CON) and IUPAC-IUB Commission on Biochemical Nomenclature (CBN). The nomenclature of cyclitols. Tentative rules. *Biochem J* 1969;112:17–28.
2. Bradley JC, Lancashire RJ, Lang AS, Williams AJ. The spectral game: Leveraging open data and crowdsourcing for education. *J Cheminform* 2009;1:9.
3. Lancashire RJ. The JSpecView Project: An open source Java viewer and converter for JCAMP-DX, and XML spectral data files. *Chem Central J* 2007;1:31.
4. Sharman GJ. Java applets for viewing one- and two-dimensional NMR spectra. *Magn Reson Chem* 2006;44:1008–1012.
5. Prasanna MD, Vondrasek J, Wlodawer A, Bhat TN. Application of InChI to curate, index, and query 3-D structures. *Proteins* 2005;60:1–4.
6. BioPAX—Biological Pathway Exchange. Available: http://www.biopax.org/.
7. Waller CL, Duravasula RV, Lynch N. The need for collaborative technologies in drug discovery. In Ekins S, Hupcey MAZ, Williams AJ, Eds. *Collaborative Computational Technologies for Biomedical Research*. Hoboken, NJ: Wiley, 2010.
8. Ekins S, Williams AJ, Pikas CK. Collaborations in chemistry. In Ekins S, Hupcey MAZ, Williams AJ, Eds. *Collaborative Computational Technologies for Biomedical Research*. Hoboken, NJ: Wiley, 2010.
9. Hunter J. Precompetitive collaboration in the pharmaceutical industry. In Ekins S, Hupcey MAZ, Williams AJ, Eds. *Collaborative Computational Technologies for Biomedical Research*. Hoboken, NJ: Wiley, 2010.

10. Komatsoulis GA. Collaboration in the cancer research community: The cancer biomedical informatics grid (caBIG). In Ekins S, Hupcey MAZ, Williams AJ, Eds. *Collaborative Computational Technologies for Biomedical Research*. Hoboken, NJ: Wiley, 2010.
11. Pratt B. Collaborative systems biology: Open source. open data, and cloud computing. In Ekins S, Hupcey MAZ, Williams AJ, Eds. *Collaborative Computational Technologies for Biomedical Research*. Hoboken, NJ: Wiley, 2010.
12. Williams AJ, Arnold RJ, Neylon C, Spencer RW, Schurer S. Current and future challenges for the collaborative computational technologies for the life sciences. In Ekins S, Hupcey MAZ, Williams AJ, Eds. *Collaborative Computational Technologies for Biomedical Research*. Hoboken, NJ: Wiley, 2010.
13. Gamo F-J, Sanz LM, Vidal J, de Cozar C, Alvarez E, Lavandera J-L. Thousands of chemical starting points for antimalarial lead identification. *Nature* 2010;465: 305–310.
14. Spjuth O, Willighagen EL, Guha R, Eklund M, Wikberg JE. Towards interoperable and reproducible QSAR analyses: Exchange of datasets. *J Cheminform* 2010;2:5.
15. Spjuth O, Alvarsson J, Berg A, Eklund M, Kuhn S, Masak C. Bioclipse 2: A scriptable integration platform for the life sciences. *BMC Bioinform* 2009;10:397.
16. Shortliffe EH, Barnett GO, Cimino JJ, Greenes RA, Huff SM, Patel VL. Collaborative medical informatics research using the Internet and the World Wide Web. *Proc AMIA Annu Fall Symp* 1996:125–129.
17. Neumann E, Prusak L. Knowledge networks in the age of the Semantic Web. *Brief Bioinform* 2007;8:141–149.
18. Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H. Advancing translational research with the Semantic Web. *BMC Bioinform* 2007;8 (Suppl 3):S2.
19. Roy A, McDonald PR, Sittampalam S, Chaguturu R. Open access high throughput drug discovery in the public domain: A Mount Everest in the making. *Curr Pharm Biotechnol* 2010;11(7):764–778.
20. Barnes MR, Harland L, Foord SM, Hall MD, Dix I, Thomas S, et al. Lowering industry firewalls: Pre-competitive informatics initiatives in drug discovery. *Nat Rev Drug Discov* 2009;8:701–708.
21. Munos B. Can open-source R&D reinvigorate drug research? *Nat Rev Drug Discov* 2006;5:723–729.
22. Ren J, Williams N, Clementi L, Krishnan S, Li WW. Opal web services for biomedical applications. *Nucleic Acids Res* 2010;38 (Suppl):W724–731.
23. Krestyaninova M, Zarins A, Viksna J, Kurbatova N, Rucevskis P, Neogi SG. A System for Information Management in BioMedical Studies—SIMBioMS. *Bioinformatics* 2009;25:2768–2769.
24. Ekins S, Williams AJ. Reaching out to collaborators: crowdsourcing for pharmaceutical research. *Pharm Res* 2010;27:393–395.
25. Bingham A, Ekins S. Competitive collaboration in the pharmaceutical and biotechnology industry. *Drug Disc Today* 2009;14:1079–1081.

14

COLLABORATIVE SYSTEMS BIOLOGY: OPEN SOURCE, OPEN DATA, AND CLOUD COMPUTING

Brian Pratt

14.1 INTRODUCTION

The term *systems biology* is a relatively recent coinage, and its meaning is still evolving. Here it is understood to encompass what some call "the omics"—genomics, proteomics, metabolomics, and so on, and the effort to synthesize knowledge gained about genes, proteins, metabolism, and so on, into an

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.

understanding of how organisms operate at the molecular level, and more importantly how they fail to operate properly and how that might be fixed. The key realization in systems biology is that none of these omics stand alone—it is not enough to determine "what this gene does" or "what that protein is for." All these mechanisms interact within the system that is an organism, and systems biology attempts to understand that interaction. Modeling the behavior of a living organism at the molecular level is an undertaking of breathtaking complexity made possible only by the advent of inexpensive and powerful computers. The simultaneous rise of systems biology and the Internet is no coincidence—the study of interconnectedness demands interconnectedness to meet the scale of the problem. The scope of systems biology is too great to be attacked without collaboration between researchers.

The author's company, Insilicos, has been very active in maintenance, integration, and improvement of various open-source software tools used by the proteomics research community, most notably the *Trans-Proteomic Pipeline* (TPP), originally developed at the Institute for Systems Biology (ISB), and LabKey Software's laboratory information management and data analysis system *LabKey Server* (Insilicos, ISB, and LabKey are all located in Seattle, WA). This is the keyhole through which the author views the systems biology world—as a scientific software toolmaker, as opposed to a scientific research practitioner.

14.2 TRADITION OF NOT VERY COLLABORATIVE SCIENCE

Since the early days of the Royal Society, science has largely advanced on a model of serial scientific collaboration. New scientific papers reference previously published papers, and the body of knowledge is built block by block. The intent of peer-reviewed publication is a marketplace of ideas which will result in the best being advanced and the worst being weeded out. However, the competitive nature of the scientific marketplace sometimes results in less than full disclosure of the ideas in question. Often the reader of a paper finds that promised data (and now, software) fail to appear or, when they do appear, their inner workings (raw data, source code) are not made available. A charitable view is that these habits of semicooperation are simply a result of the traditional inconvenience of handling physical media when sharing data and code. But in disciplines such as systems biology, which have largely arisen in the Internet age and in which code and data are the fundamental artifacts, there are new expectations as to what constitutes full disclosure.

14.3 IMPACT OF OPEN-SOURCE SOFTWARE ON TRULY COLLABORATIVE SCIENCE

The Internet has made collaboration in human endeavors possible in new ways and at unprecedented scale. The growing body of free and open-source general-purpose software projects such as Linux, GCC, Apache, and MySQL are the direct result of this new freedom to cooperate. Development is "crowdsourced"—there are often one or more core teams of developers with financial backing from commercial users of the software, and a large community of end users also contribute improvements that advance the projects much faster than would be possible with a single traditional developer team at work. For example, about 75% of new Linux kernel code is generated by teams inside normally competitive companies like IBM and Intel, and individual contributors account for at least 18% of ongoing efforts [1].

A grand tradition of academic thriftiness has led to the widespread adoption of these no-cost tools in the research community, and the "crowdsourcing" ethic has rubbed off onto scientists' thinking about their own work. In recent years there has been a flowering of open-source bioinformatics software and a move toward more open sharing of data. Indeed, many granting agencies now require an explicit plan for data sharing, although there is little agreement about what constitutes a reasonable plan [2].

14.4 OPEN DATA STANDARDS: ONTOLOGIES AND INTERCHANGE FORMATS

Sharing data requires mutual understanding of the content and format of the data, but achieving this understanding can be nontrivial. This is especially so when dealing with unprocessed, or "raw," data, which is typically written in some mysterious binary format closely held by each instrument manufacturer. The use of such closed formats is technically defensible as they are often the most efficient for rapidly storing data as it streams off an instrument, and they can be altered as needed by the manufacturer without worry of disrupting other software systems that read the data, since none exist. Of course, the fact that an ever-shifting and undocumented data format also binds the user to the data processing software sold by the instrument maker has long been seen as a happy side effect by the instrument makers, but not by instrument users. Increasingly, users are demanding and helping define open standards to allow the data they collect to be read and written by software agents other than those provided by the equipment manufacturer, and in many cases the manufacturers are now supporting these efforts lest a lack of openness become a competitive disadvantage. Developing open standards for describing processed data and results presents an even greater challenge as the very idea of "processing" and "results" is a rapidly moving target in the research world, and there is often little agreement in the terms of speech used in describing the domains themselves.

The first step in creating a data standard is to disambiguate the terminology used in the area of endeavor. This is most properly done by developing a structured, rigorous, and thorough description of the knowledge domain, or "ontology," while avoiding duplication of or conflicts with ontologies in related areas. This is a nontrivial and open-ended task requiring cooperation within and between research communities and the industries that serve them. In the life sciences, the Open Biological and Biomedical Ontologies "OBO Foundry" (www.obofoundry.org) [3] is the clearinghouse for free and open collaborations in developing interoperable ontologies. Ironically there is an ongoing debate over how best to represent ontologies themselves in a computer-readable manner, and the ontologies to be found on OBO Foundry are split between using the .obo and .owl formats.

Once the terminology has been nailed down, there is still the question of how data should be formatted. The use of a standardized data format (ideally accompanied by a reference implementation of code for readers and writers of the format) allows software developers to concentrate on novel algorithms instead of the drudgery of input/output issues, especially when implemented using standardized encoding rule sets such as XML (Extensible Markup Language) or JSON (JavaScript Object Notation). This author is most familiar with standards development in the area of proteomics-related mass spectrometry in the form of the mzXML and newer mzML [4] XML formats and can attest to the fact that these are long-term projects requiring serious effort from academia and industry alike and requiring constant extension as technologies evolve.

14.4.1 Some Notable Standards Efforts in Life Sciences

There are many complimentary and sometimes competing standards efforts ongoing at any given time and many that have been superseded or simply abandoned for lack of support or failure to keep up with rapidly evolving technologies. Some major active and ongoing efforts include:

- Systems Biology Markup Language (SBML), a computer-readable format for representing models of biological processes [5]. The project also includes a variety of software projects supporting the use of the format.
- The mzML mass spectrometry data format from HUPO-PSI [6] and its reference implementation in the ProteoWizard project [7], which contains many excellent support tools for converting from proprietary mass spectrometry vendor formats. Perhaps the greatest achievement of the ProteoWizard team is securing the cooperation of the mass spectrometry vendors in producing those converters, a political feat which previous proteomics mass spectrometry standards efforts did not achieve and which is a signal marker of the shifting attitudes toward openness in the sciences in academia and industry alike.
- Also in the proteomics field, HUPO-PSI is facilitating the development of mzIdentML as a standard for protein identification [8], along with the closely related mzQuantML standard for protein quantitation [9], and TraML [10] for the exchange and transmission of transition lists for selected reaction monitoring (SRM) experiments. These are just a few projects in this exceedingly active and successful group.

14.5 NOTE ON ASSESSING OPEN-SOURCE SOFTWARE

It is said that there is no such thing as a free lunch, but in the case of software it is probably more apt to say that the price of freedom is vigilance. The most important aspect of open-source software is not that it is free but that it is "open," which gives you the ability to make your own assessment of the correctness and stability of the tools you use in your research. Even if you can not really understand the code, you can read the comments in the code to understand the intent of the author. If there are no comments in the code, beware—this is indicative of a throwaway mindset and not a hopeful sign for using the code outside the environment in which it was developed.

Here are some things to think about when assessing open-source software:

1. *Software Generality—Are You Working the Same Problem as the Author?* Nobody sets out to write bad software, but in the research world authors do generally set out to write pragmatic software solutions which are just good enough to complete the task at hand. Beware of code which assumes certain truths about equipment or lab procedures which may not match your situation. How certain are you that there are no implicit assumptions that may apply to the author's experimental setup but not your own?

2. *Software Stability—Is the Software Rotting? Can Anyone Tell?* Even when software source code is untouched, the world around it changes and eventually the software does not work the way it needs to: "Software rots" goes the adage. Input formats are revised, feature requirements evolve with newer technologies, operating systems change, and eventually software maintenance is required. This is when things get dicey: it is easy to unknowingly break something while improving something else, especially in code written with the kind of ad hoc flair often found in research-grade software. Many, if not most, writers of open-source biological software are biologists and physicists trying to get a paper out, as opposed to engineers and computer scientists trained (and given time) to write software designed for maintainability and testing. Look at the source code directory tree for your software of choice—Are there any files in there with a name like "test"? If so, it is a hopeful sign. If not, how certain are you that the code you are trusting to analyze your data remains correct as it gets stretched in new directions? Exceptions to this generalization about lack of ongoing regression testing do exist: Projects such as LabKey Server [11], Skyline [12], and ProteoWizard [7] are very test focused and thus very stable, but in general due diligence is called for.

3. *License Terms—Can I Use This the Way I Want To?* Open-source software authors still have full copyright and usually specify the license terms under which others may use their work. Some licenses are quite generous: The Apache license, for example, does not place any restrictions on how you use the code beyond acknowledging the authors. Other licenses may prohibit or demand a fee for commercial use. Others may impose their terms on any software you wish to combine with—GPL is the prime example of a "viral" license—which can be an issue for would-be commercial adopters of things like libraries for reading and writing standard file formats who do not wish to open source their entire software offering.

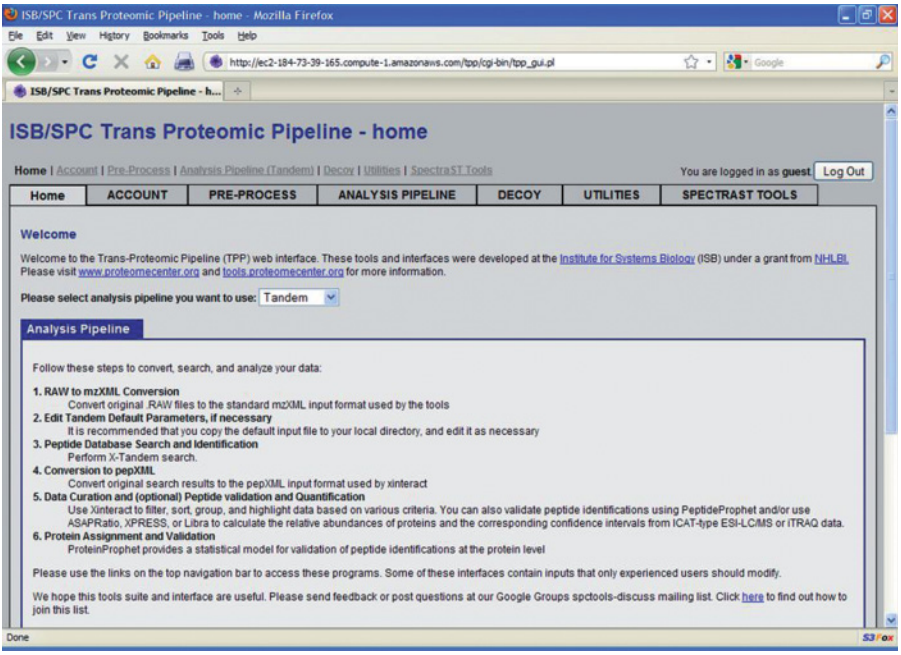
14.6 CONSTRAINTS ON OPEN-SOURCE SCIENCE

Open source and open data are making serious inroads into scientific research practice, but the cultural transformation is not complete:

- *Fear, Uncertainty, and Doubt* Researchers still have lingering suspicions that making source code publicly available before a paper on a project is published might somehow constitute "previous publication" and derail the paper in review. Many scientific academies and journals have taken steps to clarify their policies around this, but there is still confusion among authors, particularly when they have not yet chosen a journal for submission, and this can have a chilling effect on openness.
- *Data Privacy* There are also potential privacy issues that may make publishing some data more trouble than it is worth. A prime example of this is the health data security and privacy provisions of the Health Insurance Portability and Accountability Act of 1996 (HIPAA), which have been seen by many as having a chilling effect on research [13].
- *Data Value* And, of course, raw data are the ore from which researchers extract their nuggets of publishable work, and many are reluctant to share with potential claim jumpers.
- *Bureaucratic Barriers* Even researchers who see that more ore means more nuggets and agree to share and share alike may discover that granting would-be collaborators access to one's computing resources from outside the institutional firewall is an administrative nightmare. Happily (to paraphrase Internet pioneer John Gilmore), the Internet treats this as an error and routes around it in the form of cloud computing.

14.7 USING CLOUD COMPUTING TO ELIMINATE BARRIERS TO COLLABORATION

Say you wish to allow a colleague at another institution to analyze some of your data and share the results with you in an iterative, collaborative fashion. You have the data and you have chosen or developed the analysis software. You even have secured the blessing of your supervisors to work on this collaboration. All that remains is getting your institution's IT powers to allow your collaborators access to your network. This is probably going to be more bother than you would like, and may be even impossible, what with getting them visiting scholar status and other administrative details, negotiating your institution's IT policies just to open your firewall, and so forth. An attractive


ISB/SPC Trans Proteomic Pipeline - home - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://ec2-184-73-39-165.compute-1.amazonaws.com/tpp/cgi-bin/tpp_gui.pl
Google
ISB/SPC Trans Proteomic Pipeline - h...
Home | Account | Pre-Process | Analysis Pipeline (Tandem) | Decoy | Utilities | SpectraST Tools
Log Out
Home
ACCOUNT
PRE-PROCESS
ANALYSIS PIPELINE
DECOY
UTILITIES
SPECTRAST TOOLS
Done
S3Fox

ISB/SPC Trans Proteomic Pipeline - home

You are logged in as guest

Welcome

Welcome to the Trans-Proteomic Pipeline (TPP) web interface. These tools and interfaces were developed at the Institute for Systems Biology (ISB) under a grant from NHLBI. Please visit www.proteomecenter.org and tools.proteomecenter.org for more information.

Please select analysis pipeline you want to use: Tandem

Analysis Pipeline

Follow these steps to convert, search, and analyze your data:

1. RAW to mzXML Conversion
   Convert original .RAW files to the standard mzXML input format used by the tools
2. Edit Tandem Default Parameters, if necessary
   It is recommended that you copy the default input file to your local directory, and edit it as necessary
3. Peptide Database Search and Identification
   Perform X-Tandem search.
4. Conversion to pepXML
   Convert original search results to the pepXML input format used by xinteract
5. Data Curation and (optional) Peptide validation and Quantification
   Use Xinteract to filter, sort, group, and highlight data based on various criteria. You can also validate peptide identifications using PeptideProphet and/or use ASAPRatio, XPRESS, or Libra to calculate the relative abundances of proteins and the corresponding confidence intervals from ICAT-type ESI-LC/MS or iTRAQ data.
6. Protein Assignment and Validation
   ProteinProphet provides a statistical model for validation of peptide identifications at the protein level

Please use the links on the top navigation bar to access these programs. Some of these interfaces contain inputs that only experienced users should modify.

We hope this tools suite and interface are useful. Please send feedback or post questions at our Google Groups spctools-discuss mailing list. Click here to find out how to join this list.

Figure 14.1 TransProteomic Pipeline (TPP) running on Amazon Web Services EC2.

alternative is to use a cloud computing service such as Amazon EC2 (aws.amazon.com) to create a server for data storage and analysis which is physically outside of your respective institutions but still under your complete control for security and sharing. More and more software projects are adding cloud-friendly features to aid this scenario—TPP and LabKey Server both have EC2-aware configuration aids, for example (Fig. 14.1). Or, say you have developed software for a paper and would like to share it with the public as required by the terms of the funding grant but do not want to be responsible for distributing the software or adding features that people ask of you. This is a solved problem. There are many cloud-based services for collaborating on software development (SourceForge [14], GitHub [15], GoogleCode [16], etc.) as a direct result of the wider open-source software movement. These services are free of charge to open-source projects and allow you to place your code where users can get it without your intervention. You can also authorize certain trusted users to make modifications to the code for others to download. Enabling that kind of community-based software maintenance and support can give your code much more utility and lifespan than it would otherwise have. The TPP is an excellent example of this, having begun at the Institute for Systems Biology in Seattle but now having contributors in institutions and companies around the world [17].

14.8 ADDITIONAL BENEFITS OF CLOUD COMPUTING FOR SYSTEMS BIOLOGY

Cost and convenience are the other drivers in the move toward cloud computing in the sciences. Specifying, purchasing, and configuring a traditional compute cluster is an enormous undertaking of time and money. Operating and maintaining a cluster is no easier: Once installed, nodes begin to fail almost immediately, obsolescence begins its steady creep, and even under ideal operating conditions the cluster is almost never optimally used—it is either under capacity or maxed out. On the other hand, a cloud-based cluster requires no maintenance, requires no up-front money or planning, can be made just as large or small as is needed in the moment, and costs nothing at all when not actually in use. Compute power becomes more like a utility or reagent and frees up chunks of capital for more interesting lab equipment.

14.9 SOME EXAMPLES OF CLOUD-BASED SYSTEMS BIOLOGY TOOLS

Even when collaboration is not a goal, access to a system already configured with useful tools can save hours or days of setup efforts. Amazon EC2 provides online access to familiar Linux and Microsoft Windows setups (Amazon Machine Images, or AMIs) that can be readily customized, then saved for use by others, and has become a hotbed of helpful preconfigured images. Here is a small sample of useful Amazon EC2 machine images, all of which are publically available and cost about 10 cents (U.S.) per hour (paid to Amazon) to run:

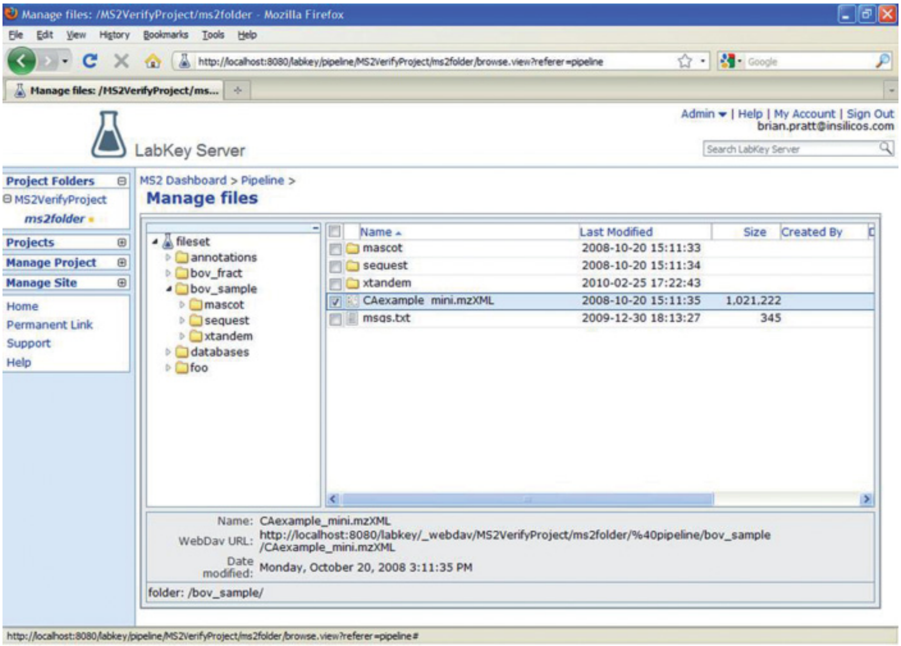
- ViPDAC (http://proteomics.mcw.edu/vipdac) provides a ready-to-run EC2 image of proteomics tools, including BLAST, OMSSA, and X!Tandem.
- CycleCloud for Life Sciences (http://my.cyclecloud.com/info/lifesciences/) provides EC2 implementations of many popular tools, including BLAST, Bowtie, X!Tandem, OMSSA, R, and many others. There is a paid option that includes support and extra features, but the standard offering costs you only what EC2 charges for the use of its compute nodes.
- BioConductor (http://www.bioconductor.org/) EC2 images maintained by Martin Aryee at Johns Hopkins University: http://www.biostat.jhsph.edu/~maryee/index.php/Cloud/BioconductorAMI.
- J. Craig Venter Institute's BioLinux (http://www.jcvi.org/cms/research/projects/jcvi-cloud-biolinux/overview) provides EC2 implementations of BLAST, glimmer, hmmer, phylip, rasmol, genespring, clustalw, the Celera Assembler, and the EMBOSS collection of utilities.

Others are readily found by filtering on "bio" in Amazon's list of public images.

14.10 SOME EXAMPLES OF OPEN-SOURCE SYSTEMS BIOLOGY TOOLS IN PROTEOMICS

The world of open-source systems biology software tools is dynamic and ever expanding. An excellent and well-maintained list of freely available tools in the area of mass spectrometry–based proteomics is available at http://www.ms-utils.org, and another is at http://www.expasy.ch/tools/. Some projects of note include:

- The *Trans-Proteomic Pipeline*, or *TPP* [18], is a suite of proteomics tools notable for the large and active user and developer community behind it.
- The *OpenMS Proteomics Pipeline*, or *TOPP* [19], is another proteomics suite with a smaller but still active community around it.
- *LabKey Server* is a Web-based suite of tools for organizing, processing, and sharing "all types of biomedical data including mass spectrometry, flow cytometry, microarray, microplate, ELISpot, ELISA, NAb and observational study information" [11] (Fig. 14.2). It integrates many components of the TPP for proteomics processing. Developed and maintained by a core group of software professionals largely funded by subcontracts to various prominent labs, this project sets the standard for software engineering discipline in the bioinformatics open-source world.


Manage files: /MS2VerifyProject/ms2folder - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/labkey/pipeline/MS2VerifyProject/ms2folder/browse.view?referer=pipeline
Google
Manage files: /MS2VerifyProject/ms...
Admin | Help | My Account | Sign Out
brian.pratt@insilicos.com
LabKey Server
Search LabKey Server
Project Folders
MS2VerifyProject
ms2folder
Projects
Manage Project
Manage Site
Home
Permanent Link
Support
Help
MS2 Dashboard > Pipeline >
Manage files
fileset
annotations
bov_fract
bov_sample
mascot
sequest
xtandem
databases
foo
http://localhost:8080/labkey/pipeline/MS2VerifyProject/ms2folder/browse.view?referer=pipeline#

	Name	Last Modified	Size	Created By
□	mascot	2008-10-20 15:11:33		
□	sequest	2008-10-20 15:11:34		
□	xtandem	2010-02-25 17:22:43		
☑	CAexample mini.mzXML	2008-10-20 15:11:35	1,021,222	
□	msqs.txt	2009-12-30 18:13:27	345	

Name: CAexample_mini.mzXML
WebDav URL: http://localhost:8080/labkey/_webdav/MS2VerifyProject/ms2folder/%40pipeline/bov_sample/CAexample_mini.mzXML
Date modified: Monday, October 20, 2008 3:11:35 PM
folder: /bov_sample/

Figure 14.2 Labkey Server.

- *Skyline* is “a Windows client application for building Selected Reaction Monitoring (SRM)/Multiple Reaction Monitoring (MRM) methods and analyzing the resulting mass spectrometer data” [12]. This project is also notable for its very high standard for rigorous software engineering practices, which is not surprising as its lead developer is a LabKey alumnus.
- *Cytoscape* is “an open source bioinformatics software platform for visualizing molecular interaction networks and integrating these interactions with gene expression profiles and other state data” [20]. The project is notable for its consistent progress and support from academia and industry alike.
- *ProteoWizard*, already discussed above, hosts several proteomics-related projects in addition to the mzML reference implementation and converters, including the aforementioned Skyline and various tools for data visualization and working with metabolic labeling and protein metabolism experiments.

14.11 PUBLIC DATA REPOSITORIES

The ability for different research teams to assess data sets that other teams have used is critical in evaluating the merit of new systems biology techniques. The following list, while by no means complete, illustrates the rich and still developing culture of public life sciences data repositories (much of which is made possible by the emergence of open-data standards):

- *PeptideAtlas* is “a multi-organism, publicly accessible compendium of peptides identified in a large set of tandem mass spectrometry proteomics experiments” [21] with many contributors. Many of the data sets are associated with publications.
- *Amazon EC2* has a growing number of data sets ready for use in its cloud environment; the biology data sets are mostly for genomics such as Unigene and GenBank [22].
- *Uniprot* provides the Swiss-Prot and TrEMBL protein sequence databases used by many protein search tools [23].
- *Protein Data Bank* from Research Collaboratory for Structural Bioinformatics (RCSB) is “the single worldwide repository of information about the 3D structures of large biological molecules, including proteins and nucleic acids” [24].
- The MGED Society is “an international organization of biologists, computer scientists, and data analysts that aims to facilitate biological and biomedical discovery through data integration” [25]. While not actually a repository, it is notable for creating the *Minimum Information About a Microarray Experiment* (*MIAME*) standard and particularly for running a service to facilitate data deposition for published papers and a watch-

dog service calling out authors who fail to provide data as required by its funders or publishing journals.

- *BioTorrents* [26] and *ProteoCommons* [27] are two different projects looking to apply peer-to-peer file-sharing technologies to sharing biological data sets.
- *PRoteomics IDEntifications database* (*PRIDE*) is a public data repository for proteomics data "developed to provide the proteomics community with a public repository for protein and peptide identifications together with the evidence supporting these identifications." [28].
- Sage BioNetworks is developing *Sage Commons*, an ambitious open repository for systems biology tools and datasets: "The Commons will be a novel computational environment for shared research and development of biological network models and their application to human disease and biology. It will consist of very large network datasets, tools and models organized within conventions governing user participation" [29]. While in its infancy, this should be an interesting project to watch as its backers are well connected and well funded.

14.12 CONCLUSION

Perhaps it is no surprise that many scientists working in systems biology are comfortable operating in and contributing to its open-source and open-data ecosystems. While it is difficult to say precisely what one will get back when one puts one's energy into extending or simply supporting others in the use of the available open offerings, the understanding that it *is a system* which requires feeding seems to motivate many to be active in the community. As a young discipline attacking problems of breathtaking complexity and scope and with many young researchers who have never experienced a world without an Internet and open source, in systems biology collaboration may be the instinctive norm.

REFERENCES

1. Asay M. Intel claims No. 2 Linux contributor spot as hedge against Microsoft. Available: http://news.cnet.com/8301-13505_3-10288910-16.html.
2. Pryor G. Multi-scale data sharing in the life sciences: Some lessons for policy makers. *Int J Digital Curation* 2009;4:71–82.
3. Smith B, et al. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007;25:1251–1255.
4. HUPO Proteomics Standards Initiative mzML 1.1.0 Specification. Available: http://www.psidev.info/index.php?q=node/257.
5. SBML.org The Systems Biology Markup Language. Available: http://www.sbml.org.

6. The HUPO Proteomics Standards Initiative. Available: http://www.psidev.info/.
7. Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: Open source software for rapid proteomics tools development. *Bioinformatics* 2008;24: 2534–2536.
8. mzIdentML: Exchange format for peptides and proteins identified from mass spectra. Available: http://www.psidev.info/index.php?q=node/403.
9. Orchard S, et al. Annual spring meeting of the Proteomics Standards Initiative. *Proteomics* 2009;9:4429–4432.
10. Draft TraML Specification. Available: http://www.psidev.info/index.php?q=node/405.
11. LabKey Software. Available: http://www.labkey.com/.
12. Skyline Targeted Proteomics Environment. Available: http://brendanx-uw1.gs.washington.edu/labkey/project/home/software/Skyline/begin.view.
13. Nass SJ, Levit LA, Gostin LO, Institute of Medicine (U.S.). Committee on Health Research and the Privacy of Health Information the HIPAA Privacy Rule. *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health through Research*. Washington, DC: National Academies Press, 2009.
14. SourceForge: Find and develop open source software. Available: http://sourceforge.net/.
15. github: Social coding. Available: http://github.com/.
16. GoogleCode. Available: http://code.google.com/.
17. TPP: Authors and contributors. Available: http://tools.proteomecenter.org/wiki/index.php?title=TPP:Authors_and_Contributors.
18. Seattle Proteome Center (SPC)—Proteomics Tools. Available: http://tools.proteomecenter.org/software.php.
19. OpenMS Welcome. Available: http://open-ms.sourceforge.net/news.php.
20. About Cytoscape. Available: http://www.cytoscape.org/.
21. Peptide Atlas. Available: http://www.peptideatlas.org/.
22. Amazon Web services public data sets. Available: http://developer.amazonwebservices.com/connect/kbcategory.jspa?categoryID=246.
23. About UniProt. Available: http://www.uniprot.org/help/about.
24. About the PDB archive and the RCSB PDB. Available: http://www.pdb.org/pdb/static.do?p=general_information/about_pdb/index.html.
25. MGED Society mission. Available: http://www.mged.org/Mission/.
26. BioTorrents. Available: http://www.biotorrents.net/.
27. About ProteomeCommons.org. Available: https://proteomecommons.org/.
28. Vizcaino JA, et al. A guide to the Proteomics Identifications Database proteomics data repository. *Proteomics* 2009;9:4276–4783.
29. Sage BioNetworks Repository: A growing catalogue of globally-coherent datasets. Available: http://sagebase.org/commons/repository.html.

15

EIGHT YEARS USING GRIDS FOR LIFE SCIENCES

Vincent Breton, Lydia Maigne, David Sarramia, and David Hill

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.

15.1 INTRODUCTION

According to Wikipedia (http://en.wikipedia.org/wiki/Grid_computing), "distributed" or "grid" computing in general is a special type of parallel computing that relies on complete computers [with onboard central processing units (CPUs), storage, power supplies, network interfaces, etc.] connected to a network (private, public, or the Internet) by a conventional network interface, such as Ethernet. This is in contrast to the traditional notion of a supercomputer, which has many processors connected by a local high-speed computer bus.

Grid technology promises to revolutionize many services already offered by the Internet because it offers rapid computation, large-scale data storage, and flexible collaboration by harnessing together the power of a large number of commodity computers or clusters of other basic machines. The grid was devised for use in scientific fields, such as particle physics and bioinformatics, in which large volumes of data or very rapid processing or both are necessary [1, 2]. Since 2002, our group has been exploring how to use grids for the life sciences. We have seen the emergence of the technology in the DataGrid and Enabling Grids for E-science (EGEE) projects [3], and we are now witnessing the emergence of production grids, called e-infrastructures, at national and European scales.* It is still not easy to use grids for someone who is not familiar with Linux; however, the quality of service is considerably better than eight years ago. One can compare the present situation to a glass which is half full: Looking at the filled half, one can appreciate the capacity to access tens of thousands of cores on demand and now have robust data management tools. Looking at the empty half, one can still complain about the lack of advanced data management service and the limited user friendliness. But everybody will agree that (at the time of writing in August 2010) the existing grid services are better than they have ever been. In other words, grids offer more opportunities than ever to their users to do science differently.

The goal of this chapter is to explain how grids can be used to address the grand scientific challenges of the twenty-first century in a more innovative and collaborative way. We will provide examples from our experience over the last eight years to show how grids can be used today by researchers. We will also discuss how new technologies are emerging and are going to enrich the services offered to the grid users.

15.2 GRIDS FOR E-SCIENCE

Two trends are irreversible in science: First, more and more scientific data are produced and must be analyzed. Examples are easy to find in almost every field of science. In high-energy physics, chasing for new particles at the large hadron collider requires selecting a few events per year out of billions taking

*See the website of the e-infrastructure Reflection Group and references therein at http://www.e-irg.eu/.

place every second [1]. In health care, evidence-based or personalized medicine requires collecting genomics or other omics data for each individual [4]. In earth sciences, the study of climate change requires collecting data all around the world to adjust all the parameters of the climate models. Second, more and more science is done in silico, through simulation and modeling. Dismantling a detector on the large hadron collider takes months; testing all the existing druglike compounds on a biological target costs, many millions of euros (or other currency). Simulating all elements of the high-energy physics detector and their response to the highly radioactive environment of the large hadron collider allows avoidance of breakdowns, sparing billions of euros and gaining precious time [1]. Using docking software to compute the binding energy of tens of millions of druglike compounds to the active site of a biological target allows the timely selection of the most promising ones for in vitro testing and verification [5].

The concept of e-science has emerged to describe the creation involved with the design of in silico experiments. It is about inventing and exploiting new advanced computational methods:

- To generate, curate, and analyze data coming from experiments, observations, and simulations
- To develop and explore models and simulations combining computation and data at an unprecedented scale to achieve quick, reliable, and relevant results
- To help the setup of distributed virtual organizations to ease collaboration and sharing of resources and information with guaranteed conditions of security, reliability, responsibility, and flexibility

Grids open new avenues to e-science:

- They allow the researcher to think much bigger (in terms of calculations or processes) than in the past, because they give access to extended computing resources on demand.
- They allow sharing data where it is produced, because they federate data sources.
- They allow the creation of virtual research communities that share services and tools across frontiers and administrative borders.

We are now exploring these avenues with present-day technologies in different fields of biomedical research.

15.3 GRIDS TO THINK BIGGER

15.3.1 Introduction

The first advantage of grids that comes to mind is their capacity to aggregate and therefore make available to scientists a much larger number of processors than an individual cluster. Compared to supercomputers, the main difference is that these resources are available on demand without evaluation or previous agreement by a scientific committee. This opens new perspectives to scale up existing strategies by a factor of 10, 100, or even 1000 or to explore new approaches without guaranteed success. The change of scale is a major driver for scientific progress. The history of physics shows that scaling up in energy and in luminosity was the key to discover new phenomena.

Our group was involved in two biomedical projects where the capacity to think bigger was exploited and led to very significant scientific results.

15.3.1.1 Example of WISDOM Drug Discovery Platform The pharmaceutical research and development (R&D) enterprise presents unique challenges for information technologists and computer scientists. The diversity and complexity of the information required to arrive at well-founded decisions based on both scientific and business criteria are remarkable and well recognized in the industry. Drug discovery is the process by which drugs are discovered and/or designed. Drug candidates are inputs to the drug development process. Current efforts within the pharmaceutical industry are directed at reducing the time and costs for drug development. Recent progress in genomics, transcriptomics, proteomics, high-throughput screening, combinatorial chemistry, molecular biology, and pharmacogenomics has radically changed the traditional physiology-based approach to drug discovery where the organism is seen as a black box.

An important step in the drug discovery process is virtual screening, which is about selecting druglike molecules active on a specific biological target by computing the binding energy of the molecule to the target active site [6]. The prerequisite for the use of virtual screening is to know the three-dimensional (3D) structure of both the druglike molecules and the target active site. The 3D structures of more than 3 million chemical compounds are now available in public databases like ChemBridge and ZINC while the Protein Data Base provides the structure of more than 50,000 proteins of biological interest [7]. Since 2004, the WISDOM initiative [5] has successfully deployed large-scale virtual screening computations on grid infrastructures in order to find new drugs against malaria, avian flu, and diabetes. Meanwhile, it has also grown into a multidisciplinary collaboration of biologists, biochemists, bioinformaticians, and e-scientists from Africa, Asia, and Europe. More than a thousand CPU years have been used since 2004 on e-infrastructures in France, Africa, America, Asia, Open Science Grid (OSG), and of course on EGEE [3], which has provided the majority of the resources. About 20% of the druglike molecules selected in silico have been confirmed by in vitro tests to be active inhibitors and most promising molecules have been patented [8]. Today, WISDOM is a success story from a grid deployment point of view because it has demonstrated the potential impact of e-infrastructures for virtual screening.

15.3.1.2 Example of Protein Database Refinement During the spring of 2007, a large-scale application [9] was deployed on grid resources in order to refine the Protein Data Bank (PDB) [10]. The goal of this application was to recalculate 19,000 X-ray structures in the PDB. Indeed, structural biology, homology modeling, and rational drug design require accurate 3D macromolecular coordinates. However, the coordinates in the PDB have not all been obtained using the latest computational methods. The study showed that they can be improved in terms of fit to the deposited experimental X-ray data as well as in terms of geometric quality.

The re-refinements of the structure models were performed on a hybrid computing environment consisting of two virtual organizations of the EGEE grid infrastructure and several clusters provided by bioinformatics institutes in Europe within the framework of the EMBRACE project [11]. On a single CPU, the entire calculations would have taken about 17 years. With our grid-and-cluster computing approach, more than 90% of the total calculation was finished in only two months—this shows the clear time advantage arising from the usage of modern computing technology. All 16,807 successful re-refinements were complete after four months; the vast majority were done after three weeks.

By employing methods such as translation/libration/screw (TLS) motion refinement that represents the displacement of groups of atoms that behave as (quasi) rigid bodies, 10,046 out of 15,034 structure models (67%) were improved [9]. These results showed that re-refinement of existing PDB entries was worthwhile and, because the method is fully automated, little time investment was needed to re-refine a single structure model. PDB entries are now routinely re-refined before they are used for molecular dynamics, homology modeling, or drug design.

15.4 GRIDS TO SHARE DATA WHERE IT IS PRODUCED

15.4.1 Introduction

Everybody understands the importance of sharing data in modern-day science, but one may wonder why it is important to share data where it is produced. The important concept is to allow a data owner to keep control of who accesses his or her own data. Indeed, in order to publish in peer-reviewed journals, a scientist must demonstrate creativity and present ideas and results that are beyond the present state of the art. One way is to work with experimental data that have not been previously produced or analyzed. Making these data publicly available represents a big risk of losing a competitive advantage and therefore many researchers are reluctant to share their data out of fear they are overtaken by competitors and exploited without due credit or attribution in general. Not all scientific communities show this behavior. For instance, this fear does not exist with high-energy physics experimental data because they are unusable without an in-depth knowledge of the detector used to produce them [1]. In the field of molecular biology or astronomy, data are made available after one or two years so that the scientists have enough time for publishing a few papers before releasing them for a wide use by the community.

Researches in health care are severely hindered by this culture of secrecy as well as patient confidentiality and the difficulty for researchers to access innovative medical data. Data from patient cohorts and clinical trials are like a treasure that is carefully protected because it holds the promise for publication.

The grid technology allows data owners to share their data while keeping them. In other words, they keep full control of who can access their data, which subset of the data is accessed, which parameters can be extracted, and for how long. The data owner can at any time decide not to share his or her data any more or to change the policy for accessing them without the boundaries of a level agreement.

The following example of a cancer surveillance network under deployment in Auvergne illustrates this concept.

15.4.2 Example of Cancer Surveillance Network in Auvergne

15.4.2.1 Introduction In this example, we propose a very innovative approach to facilitate both cancer screening and epidemiology using the grid technology in the Auvergne region. How do we federate medical data in a secure and reliable way without adding any complexity to the existing dial operations performed in the different medical structures?

In France, cancer screening structures are in charge of providing a second diagnosis on the mammograms and have to follow up medical data sheets describing the tumor characteristics from cytopathology laboratories. Presently, medical data sheets are faxed or posted by the patient to the associations where information is registered again. This process is human costly and promotes errors as data have to be registered and reinterpreted twice.

The solution proposed, fully grid compliant, provides each medical staff member the ability to query pathology databases located directly in the laboratories. The grid architecture, federating the laboratories (see Fig. 15.1), provides a secured framework and easy usage in order that no added actions are required from the daily practice of the physicians.

15.4.2.2 Network Architecture As shown in Figure 15.1, the cytopathology laboratories own the first information concerning diagnosis on cancer; this information is the basis of the full follow-up of the patient and is the key to preparing health care delivery to the patient (surgery, radiotherapy, or chemotherapy). These different laboratories host different software systems and local databases for medical data management.

On the other side, cancer screening associations need to register and follow up the information concerning the cancer diagnosis if it happened. Those associations therefore have to link the patient's identity to the medical data sheets.

If a sentinel network is able to federate anatomical pathology databases, it can be used also by the epidemiological services of the National Institute for Sanitary Watch (Institut National de Veille Sanitaire) and the regional epide-

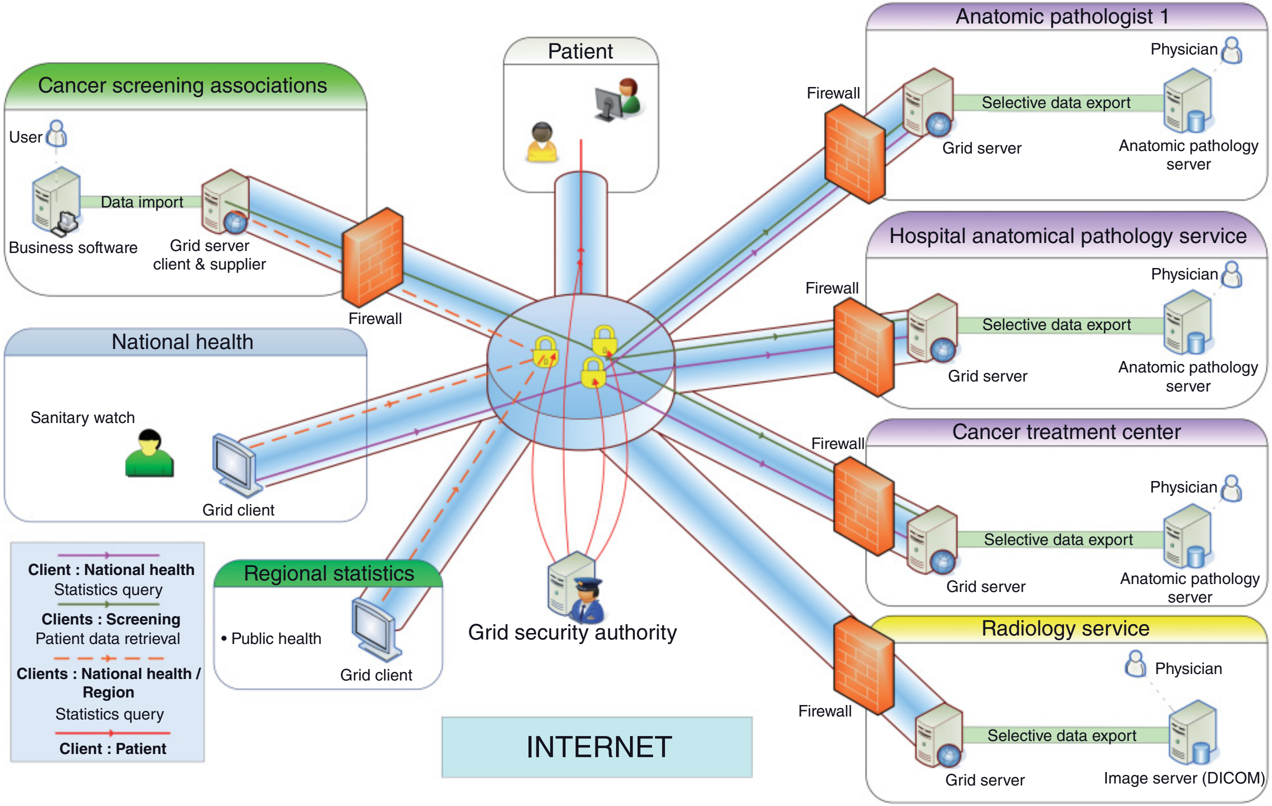


Figure 15.1 Sentinel network architecture.

miological observatory to build epidemiological studies. Contrary to cancer screening associations, personal information on the patient (especially name, surname, and social security number) is not relevant for epidemiological structures to produce statistics; on the other hand, exhaustive and disambiguation on the cancer information for each patient are required in order to produce reliable statistics on cancer incidence in a region.

The proposed grid architecture is built upon a central server hosting security features and core grid services:

- The AMGA (ARDA Metadata Grid Application) [12] server, which provides a way to access and store metadata. Especially when dealing with medical images, the use of metadata is mandatory therefore AMGA is able to glue the DICOM servers with the grid middleware [13]. AMGA is a very attractive software to fulfill the strong security requirements and access right management of medical data in a grid infrastructure.
- Pandora GateWay is a set of software designed as a service-oriented architecture (SOA) developed by the maatG company. Pandora GateWay is used to address medical data accessibility, exchange, and processing while guaranteeing a high level of security for sensitive data. The main added value of GateWay, compared to a classic SOA platform, is the high-level security. The GateWay authentication service is based on several security checkpoints required for login. The access point is a two-factor authentication based on user certificate and pin code followed by an authentication process using a Virtual Organization Membership Service (VOMS) grid proxy [14]. VOMS is an authorization manager which implements a public key infrastructure (PKI)-based authentication with certificates delivered by trusted certification authorities (CAs):

  GridFtp server for data transfer [15]

  Logical file catalog (LFC) server, for data management, included in the gLite middleware

Then, in each medical structure which takes part in the sentinel network, a grid node is deployed with two interfaces:

- The first one is linked to the medical database inside the private medical structure network. This connection is an automatic export system [Structured Query Language (SQL) query + standardized output]. It is less intrusive, offers enlarged customization for integration and standardization of data, and does not overload the medical database during working hours.
- The second interface is linked to the sentinel network (Internet), enabling external users to query the grid data server, according to the local security and authentication policy fixed by VOMS.

For a better readability of cytopathological data, data sheet standardization is used to simplify the integration of medical sheets in the database without interfering with other data.

15.4.2.3 Network Security The security and privacy requirements for a distributed medical data querying system are critically important, and data protection is essential. Within the cancer Biomedical Information Grid (caBIG) and ACGT projects, different studies about security, privacy, ethical, and legal requirements for distributed architecture have been published [16, 17]. The European Union (EU) released a document [18] relative to personal data process, treatment, and movement issues.

To answer security issues, we have made compliant the security infrastructure of the network with French regulation on medical data transfers and exchanges. Users of the sentinel network are authenticated using recognized accreditation tools like the Carte de Professionnel de Santé (CPS) health care professional smartcard (http://gip-cps.fr) [19] released by the French health ministry. These cards will be available throughout the EU (http://www.hprocard.eu). The chip contains an X509 grid-compatible certificate issued by a trusted CA. The authentication process and the data encryption are then ensured by these cards.

15.4.2.4 Patient Identification Throughout the health care systems, cases of false patient identification are numerous and could be responsible for mistakes in drug delivery to the patient. Due to lack of a global identification system, there is no solution to address a distributed patient identification. Most countries in the EU already have a robust identification system. In France, the usage of the social security number (SSN) is strictly prohibited for data linkage as it contains privacy data about gender and date and place of birth. Moreover, the accuracy and reliability of these numbers are reconsidered: the SSN in the United States presents a high risk of identity. Aware of this issue, the EU has launched the European Patients Smart Open Services (EPSOS) program (http://www.epsos.eu/) in order to build a European Electronic Health Record while the French government released guidelines to build a national health identifier (http://www.asipsante.fr/). Despite this, there is no suitable solution today; therefore a dedicated solution has been designed for this project.

The patient can be identified using different medical folder numbers regarding the different laboratories he or she visited. In order to link all the information stored in the multiple medical databases all over the world, an additional identifier has been created for the sentinel network. This identifier consists of a random number generated as defined in RFC 4122 [20]) for each patient. This identifier is created only for data linkage and is always encrypted using different keys in each database to protect patient privacy. When a data provider downloads some new data from his or her local data server to the local grid server, the Pandora Gateway is in charge of searching all the local databases with respect to information on the patient. It will produce a unique identification number corresponding to the medical data if two identifiers are correlated to the same patient (Fig. 15.2).

The distributed identity management requires specific ability to compare records and link identities. The entire reliability of the sentinel network depends on a good record linkage.

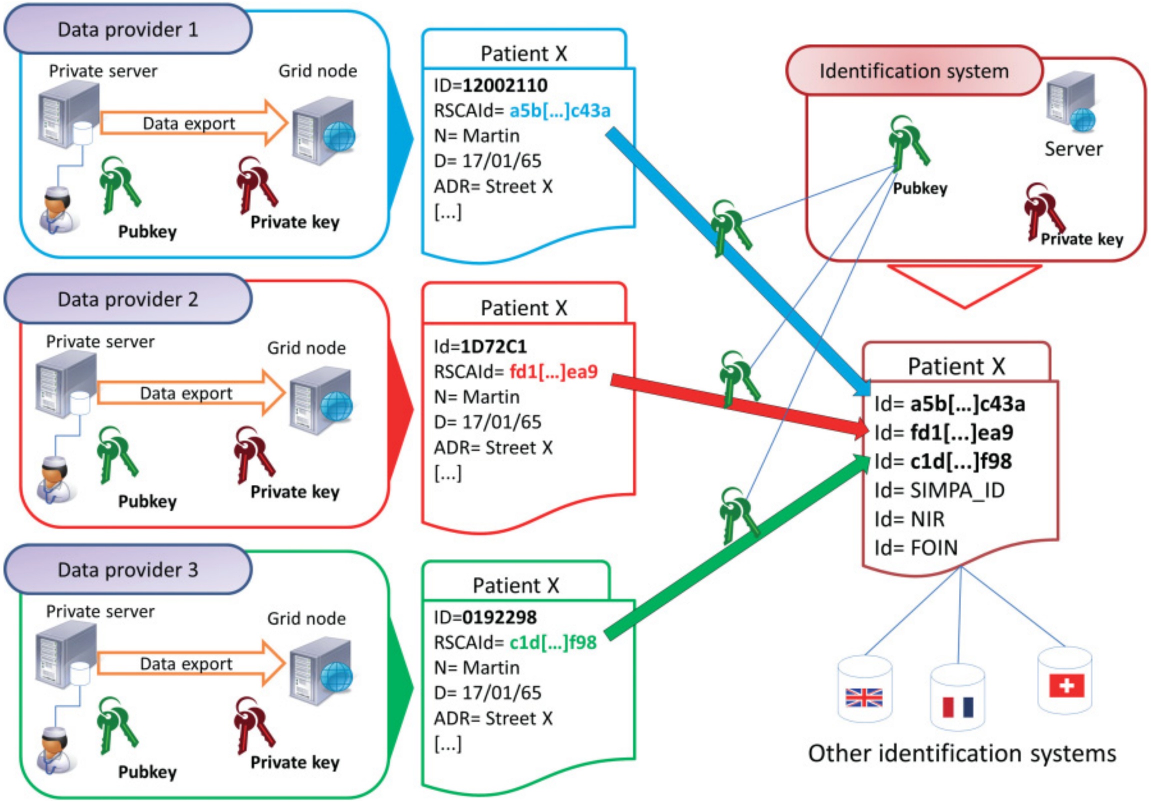


Figure 15.2 Identification system.

15.4.2.5 Patient Data Linkage Data linkage does not consist of a simple string comparison; the two main problems are related to looking through a patient's information (homonyms, same address, equivalent birthdates) and overall errors in names. Three levels of errors appear:

- Typographical errors (despite known spelling)
- Cognitive errors (comprehension problem)
- Phonetic errors (similar spelling)

The errors and variations are mainly related to the typing of handwritten data, keyboard neighbors (k–i, e–r, etc), data input during a telephone conversation, and software or database limitation of input fields (length limitation) that force the use of abbreviations or initials. Several matching techniques aim to measure similarity between strings. Two different approaches can be adopted:

- Pattern matching for flexible matching between two strings
- A combination of phonetic encoding and exact matching

The similarity measurement is generally normalized: two strings are equivalent with score = 1 and if totally different score = 0.

The efficiency of the solution will impact the percentage of automatic matching. This ratio must be as high as possible while guaranteeing a lower level of false positive. For this linkage process the usage of a combination of Jaro-Winkler [21] and Phonex [22] (French) algorithms are used. According to the relevance and accuracy of information in the data set, different weights are attributed.

For each field, four different criteria define how to interpret matching scores according to field types:

- Accuracy, which defines the relevance of information
- Blocking, in case of false matching (under threshold), where the correspondence would be automatically rejected
- Weight (similar), which represents a factor attributed in case of similarity (over threshold)
- Weight (different), in case of false matching, a divide factor attributed to global similarity

Weight distinction between similar and different matching is necessary. As in the following example: The probability of having a last name different for only one patient in distributed databases is small so it considerably reduces the matching chance. However, having two entries with the same address does not mean that the patient is identical for these two entries. Table 15.1 summarizes the proposition of criteria adjustment for automatic record linkage. A weight factor is attributed for each field and is submitted as input for the linkage algorithm.

TABLE 15.1 Relevance of Information for Selected Fields

	Last Name	First Name	Sex	Maiden Name	Birth	Address	Region	Postal Code	City	Physician
Type	String	String	Digit	String	Date	String	Digit	Digit	String	String
Accuracy	•••	•••	•••	•	•••	•	•	••	•	•
Blocking	X		X		X					
Weight	•••	•••[a]	••[a]	•	•••	••	•	••	••	••
Weight	•••	•[a]	••	••	••	•	••	•	•	•

[a] Only if previous fields match.

A global score is attributed for each *n*-to-*n* comparison and is submitted as input for the matching process

By using a distributed identification mechanism in combination with data linkage techniques, patient matching is quite easily fixed.

15.5 GRIDS TO CREATE VIRTUAL RESEARCH COMMUNITIES

15.5.1 Introduction

More and more discoveries are going to come out of collective scientific efforts. Building collaboration between research groups that are remote from a geographical point of view requires sharing common scientific tools. Grid infrastructures are designed for hosting virtual organizations which gather scientists across national and administrative borders. We are going to illustrate the collaborative power of the grid on the example of a surveillance network for emerging diseases.

15.5.2 Surveillance Network for Emerging Diseases

We live in a small world. Air travel and the Internet have considerably reduced perceived distances and increased communication. There is also a growing awareness that the whole of humanity is confronted by challenges that it has to address together in order to achieve success. Viruses are not stopped by frontiers and recent pandemics like H1N1 or AIDS have highlighted the need for good practices in the sharing of data for improved health care monitoring [23]. There is also a growing need to provide services to the scientists who are on the front line of emerging diseases to enable the public health authorities to take the most accurate decisions at the earliest stages.

A concrete example of a present-day concern is avian flu. While the world was anxiously going through the H1N1 pandemic, interest in the media for H5N1 disappeared. However, the capacity of the H5N1 to mutate into a strain with human-to-human transmission remains a very significant threat to public health [24]. The H5N1 keeps mutating, as can be observed in the regular outbreaks taking place in Southeast Asia [25].

Molecular epidemiology of influenza virus strains provides scientists with clues about the temporal and geographic evolution of the virus. Researchers from France and Vietnam are developing a global surveillance network based on grid technology: The goal is to federate influenza data servers and automatically deploy molecular epidemiology studies [26]. A first prototype based on AMGA [12] and the WISDOM production environment [26] extracts daily from the National Center for Biotechnology Information (NCBI) influenza H1N1 sequence data which are processed through a phylogenetic analysis pipeline deployed on EGEE [3] and AuverGrid (http://www.auvergrid.fr) e-infrastructures. The analysis results are displayed on a Web portal (http://g-info.healthgrid.org) for epidemiologists to monitor H1N1 pandemics.

15.6 PERSPECTIVES

15.6.1 Introduction

IT technology is constantly evolving and new concepts have been emerging in recent years. The new popular concept heavily promoted by the largest IT companies is cloud computing. Integrating private or public clouds on e-infrastructures would enrich the services offered to their customers. However, a number of questions are still open related to the interoperability of the grid middleware and cloud services, to the business model of the private clouds, and to the security framework required for their integration.

Another very promising approach to increase the computing resources available to the scientific community is to use graphical processors.

In this chapter, we will present activities we are currently developing on graphical processors applied to life sciences. We will also discuss how all the developments we have made for eight years are now converging toward multiscale modeling for system radiobiology.

15.6.2 Graphical Processors

General-purpose graphical processing units (GP-GPUs) were designed to process more than the regular computer graphics, but while the classical CPU computation performance evolution recently began to slow down, the GP-GPU has continued to provide very significant speedup. Ten years ago, developers of high-performance computing applications started to port scientific software from CPU to GP-GPU to make the most of it [27]. While a CPU possesses few cores, each of them allowing the execution of one thread at a time, a GP-GPU possesses a small number of streaming multiprocessors, each of them allowing the parallel execution of numerous threads, supporting vector computing in a SIMD (single instruction multiple data) approach. After the initial success, GPU manufacturers started to work on friendlier application programming interfaces (APIs) for general-purpose computation, and one is now able to develop directly in languages which are close variants of the C language [28]. In the same way, other hardware accelerators like field programmable gate arrays (FPGA) have been considered to speed up various parallel applications [29], including multiscale simulations [30].

Current GP-GPUs [31] now allow the execution of hundreds of threads with a regular PC hosting a device card. This capability can be exploited in the case of life science applications when we have to compute the same algorithm many times. Since the introduction of Tesla boards by Nvidia, the single-precision performances show very interesting improvement even when compared to the latest CPU processors. The interconnection of GP-GPU boards and servers is used to build clusters [32], and nowadays grids of hybrid machines are even on track and used for computer-intensive bioinformatics application [33]. Different brands of GP-GPU exist and ATI is also proposing very interesting cards. In the case of the widely spread Nvidia Tesla 10, the board proposes 240 vector cores split in 30 streaming multiprocessors (SM) with eight thread processors (SP thread processors) each. Each streaming multiprocessors can run a set of 32 threads (warp) with the same control flow for different data leading in one GPU cycle (each SP computes four identical operations per GPU cycle); and since an SM can schedule up to 32 warps at a time, it leads to potentially 1024 threads running concurrently on each of the 30 SMs. The GP-GPU programming environment is proposed by the manufacturing company: for instance, CUDA (Compute Unified Device Architecture*) in the case of Nvidia or a portable programming standard usable for different manufacturing brands (OpenCL[†]).

However, programming GP-GPUs can be tricky. The main difficulty lies in the memory manipulation since GP-GPU have various levels of memory with different performances. The GP-GPU global memory which can be accessed by any thread at any time has very important access latency. The shared memory available for each streaming multiprocessor inside a GP-GPU does not have such latency, but this memory is only shared by threads running on the same multiprocessor. In addition, the number of concurrent threads in a streaming multiprocessor is limited. Moreover, memory transfer between the host computer and the GP-GPU device can severely damage the global speedup if the computation time is not significant enough in comparison with the data transfer time. With CUDA, all the threads needed to execute a kernel must be grouped in blocks, and all these blocks must have the same, limited number of threads. All the threads of a block are executed on the same multiprocessor and therefore can make use of its shared memory. To get the best results from GP-GPUs, we have to place data in shared memory. This is the fastest memory managed by a streaming multiprocessor. The latency for accessing global memory is very high, and we have to limit its access. Thus,

*What is CUDA? See the CUDA website: http://www.nvidia.co.uk/object/cuda_what_is_uk.html. Accessed January 20, 2010.

[†] OpenCL Overview. See the OpenCL website from Khronos: http://www.khronos.org/opencl/. Accessed January 20, 2010.

dependent threads needing fast communications have to use the shared memory within the same SM.

The latest generation of GP-GPU proposed by Nvidia, known as the Fermi architecture, significantly improves the memory bandwidth with a much larger and reconfigurable cache, ECC (error correction code) memory, and impressive peak performances in standardized Institute of Electrical and Electronics Engineers (IEEE) double precision with 512 CUDA cores compared to what was obtained a few years before [34]. In order to deploy a large number of GPUs in data centers, ECC memory was needed to detect and correct errors introduced during storage or data transmission. In addition, the size of addressable global memory has been considerably augmented, up to 1 terabyte for a single GP-GPU board, to meet the needs of some supercomputing applications.

Dealing with parallel stochastic simulations, there is a need of rigor in the parallelization of random streams [35]. We have proposed a survey of the current pseudo-random-number generators (PRNGs) available on GPU and we have given a particular focus to the recent Mersenne Twister for Graphics Processors (MTGP) that has just been released by Saito on Matsumoto's homepage. We have empirically checked thousands of PRNGs with the most stringent testing suite, TestU01 "Big crush" from [36]. The dedicated GP-GPU generators have been created with the Dynamic Creator software designed to propose independent MTGPs, 30% of them found to be weak according to the current level of statistical tests. A current challenge is to prevent potential bias introduced by the parallelization pseudo-random-number streams in grid computing and particularly when using the latest Mersenne Twister generator dedicated to GP-GPU.

15.6.3 New Challenge: System Radiobiology

Understanding the impact of radiation on living organisms is of crucial importance to both biology and health care. Living organisms, including humans, are constantly exposed to ionizing particles through sunlight and radioactive materials in the ground. Radiation exposure results in damage to the cellular genome. This damage can kill the cell or result in mutations. Radiation exposure is therefore an initiator for evolution and also one of the earliest identified causes of cancers. Radiation is also used to kill cancerous cells through brachytherapy and radiotherapy treatment.

Radiobiology aims at reaching a deeper understanding of the interaction of radiation with living organisms. Progress in acceleration and imaging techniques as well as high-throughput sequencing opens new avenues for a quantitative assessment of the damage one ionizing particle can produce on the genome of a model organism or a human cell.

These data are needed for the modeling of living organisms under radiation exposure. Such modeling requires simulating the living organism, the interaction of ionizing particles inside it, in particular the damage to its DNA, and the repair of DNA. Having a complete model would open very interesting avenues for health care. For instance, it would allow personalized treatment planning taking into account the patient's genomic data. But it involves difficult challenges:

- Interaction of ionizing particles with DNA takes place through direct and indirect processes which are too complex to describe and simulate.
- An accurate multiscale model of the living organism is needed. The multiscale approach is needed because damage to DNA takes place at a molecular level on an extremely short time scale while DNA repair is a global response of the organism.
- DNA repair mechanisms are still poorly understood.

Finally, accurate modeling is very computationally intensive because multiscale approaches involve handling large and complex images describing the geometry of the medium in which the interaction of ionizing particles has to be tracked down to very low energies to correctly describe indirect damage to DNA.

Since 2002, our research groups have been building the software pieces relevant to this global task of modeling living organisms under radiation exposure. Our modeling efforts have been mostly deployed on grid infrastructures and driven by experimental evidence. The choice of deploying our computations on grids was driven by opportunity.

It is well known that building a model involving all the biological levels cannot be achieved, but the current challenge is to provide a fully integrative framework on the various levels of a whole organism to understand the connections between them. So, as a first step, we have decided to split a radiobiology study in three parts: the irradiation of the real organism, the simulation of the radiation process and its impact on biological structures, and finally the simulation of organism development after radiation exposure.

The Geant4/GATE [37, 38] toolkit will be used to simulate physical interactions between particles and organisms down to the cellular and DNA levels. The specific goal of the Geant4-DNA project started in 2010 is to provide probabilistic damage onto the DNA structure after radiation exposure of several particle types [39].

Simulating the development of an organism is a great challenge. One modeling approach is to use the most common models for the various biological processes. However, this will lead to a set of unlinked models. The complexity can be tackled by several means: using a few common modeling concepts that are valid through scales, using a biologylike process such as morphogenesis principles to simulate growth and evolving, and using coupling schemes to make the various models work together [40, 41].

Our approach is to use cellular automata [42] as a common framework, but by relaxing it from the classic definition to a more suitable one for biological applications and by proposing a multiscale model. In order to follow as much

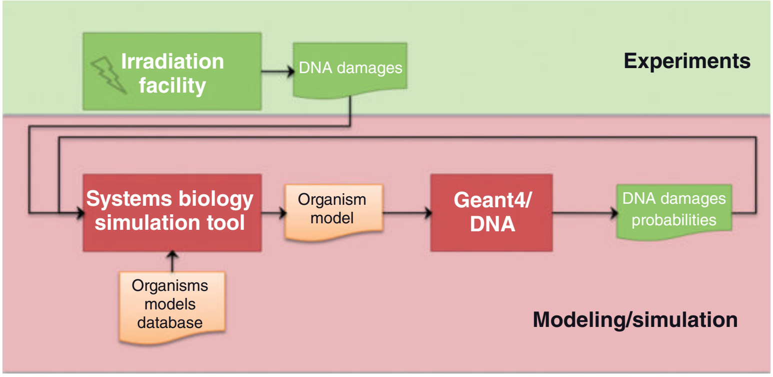


Figure 15.3 Proposed radiobiology framework.

as possible morphogenesis principles, we will introduce bioinspired metaheuristics (such as genetic algorithm, swarm intelligence, simulated annealing) to mimic the biological selection of active genes. They are used to constrain/drive cellular automata to follow a particular development cycle (which can be the normal one or a pathological one). One interesting point with this approach is that it will enable us to build patient dedicated organs using data from imaging devices as some of the quantitative values (such as volume, surface, number of cells) needed by the model are patient dedicated. The great challenge will be to include the good modeling scales with the right sharpness to learn about the impact of radiation exposure. Scales such as genotype and phenotype seem to be necessary to understand and to study the impact of radiation from a quantitative point of view (which is rather common) and from a qualitative point of view (which is less common).

Thus, an iteration of the framework we propose (Fig. 15.3) is a sequential process with three steps. First, the organism model is built from cells to the full organism. Next the organism model is irradiated in Geant4-DNA to produce probabilities for DNA damages. Finally, some simulation scenarios are designed according to those probabilities and the organism is simulated with those input parameters.

15.7 CONCLUSION

Through this chapter, we have shared our experience with using grids as an infrastructure to investigate life sciences in an innovative way. After eight years of working together with biologists, chemists, computer scientists, physicians, physicists, and many others, we are deeply convinced that grids provide a unique framework to build multidisciplinary collaborations in the field of simulation and modeling because they are about sharing resources and therefore ideas, We have provided a few examples of successful scientific initiatives and proposed the very challenging perspective of system radiobiology at the interface of biology, computer science, and physics. Technology will certainly continue to evolve at an incredible speed, but the real treasure that has to be protected is the human network of researchers from all around the world sharing a common scientific tool.

ACKNOWLEDGMENTS

The work described in this chapter was partly supported by grants from the European Commission (EGEE, EMBRACE, EUAsiaGrid, EUMedGrid-Support), the French Ministry of Research (GWENDIA, GINSENG), and the regional authorities (Conseil Régional d'Auvergne, Conseil Général du Puy-de-Dôme, Conseil Général de l'Allier).

The Enabling Grids for E-sciencE (EGEE) project is cofunded by the European Commission under contract INFSO-RI-031688. The EMBRACE project is cofunded by the European Commission under the thematic area "Life sciences, genomics and biotechnology for health," contract number LHSG-CT-2004-512092. Auvergrid is a project cofunded by CNRS, Conseil Regional d'Auvergne, and the European Commission. The GWENDIA and GINSENG projects are supported by the French ministry of Research. EUAsiaGrid ("Towards a common e-Science infrastructure for the European and Asian Grids") is a project cofunded by the European Commission as a Coordinated and Support Action within the 7th Framework Programme (FP7-INFRA-223791). EUMedGrid-Support is a project cofunded by the European Commission as a Coordinated and Support Action within the 7th Framework Programme.

REFERENCES

1. LHC computing grid, Technical Design Report. CERN-LHCC-2005-024. See also http://lcg.web.cern.ch/LCG/tdr/.
2. Breton V, Jacq N, Kasam V, Salzeman J. Deployment of grid life sciences applications. In Talbi E-G, Zomaya A, Eds. *Grids for Bioinformatics and Computational Biology*. Hoboken, NJ: Wiley, 2007.
3. Gagliardi F, Jones B, Grey F. Building an infrastructure for scientific Grid computing: Status and goals of the EGEE project. *Philos Trans R Soc A Math Phys Eng Sci* 2005;363(1833):1729–1742.
4. Olive M, Rahmouni H, Solomonides T, Breton V, Legré Y, Blanquer I, Hernandez V. SHARE roadmap for HealthGrids: Methodology. *Int J Med Inform*, 2009;78(Suppl 1):S3–S12.
5. Jacq N, Salzemann J, Legré Y, Reichstadt M, Jacq F, Medernach E, Zimmermann M, Maaß A, Sridhar M, Vinod-Kusam K, Montagnat J, Schwichtenberg H, Hofmann M, Breton V. Grid enabled virtual screening against malaria. *J Grid Comput* 2008;6(1):29–43.
6. Lyne PD. Structure-based virtual screening: An overview. *Drug Discov Today* 2002;7:1047–1055.
7. Congreve, M, et al. Structural biology and drug discovery. *Drug Discov Today* 2005;10:895–907.

8. Degliesposti G, Vinod Kasam, Ana Da Costa, Hee-Kyoung Kang, Nahyun Kim, Do-Won Kim, Vincent Breton, Doman Kim, Giulio Rastelli. Design and discovery of novel plasmepsin II inhibitors using an automated workflow on large scale grids. *Chem Enabling Drug Discov* 2009;4(7):1164–1173.
9. Joosten RP, Salzemann J, Bloch V, Stockinger H, Berglund A-C, Blanchet C, Bongcam-Rudloff E, Combet C, Da Costa A, Deleage G, Diarena M, Fabbretti R, Fettahi G, Flegel V, Gisel A, Kasam V, Kervinen T, Korpelainen E, Mattila K, Pagni M, Reichstadt M, Breton V, Tickle I, Vriend G. PDB_REDO: Automated rerefinement of X-ray structure models in the PDB. *J Appl Cristallogr* 2009;42:1–9.
10. Berman HM, Henrick K, Nakamura H. *Nat Struct Biol* 2003;10:980.
11. Pettifer Steve, Ison J, Kalas M. The EMBRACE web service collection. *Nucl Acids Res* 2010;38(Suppl):W683–688.
12. Koblitz B, et al. the AMGA Metadata service. *J Grid Comput* 2008;6:61–76.
13. Erberich SG, et al. Globus MEDICUS—Federation of DICOM medical imaging devices into healthcare Grids. *Stud Health Technol Inform* 2007;126:269–278.
14. Alfieri R, et al. From grid-map file to VOMS: Managing authorization in a grid environment. *Future Generation Comput Syst* 2005;21(4):549–558.
15. Allock V, et al. The Globus striped gridFTP framework and server. Paper presented at the ACM/IEEE Conference on Supercomputing. 2005:54–64.
16. ACGT, legal and ethical requirements. Available: http://www.eu-acgt.org/documents/public-deliverables.html.
17. Manion FJ, et al. Security and privacy requirements for a multi-institutional cancer research data grid: An interview-based study. *BMC Medi Inform Decision Making* 2009;9:31.
18. Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Official Journal of the European Communities, L281/31 23.11. 1995.
19. Fortuit P. Professional health cards (CPS): Informatic health care system in France. *Ann Pharm* 2005;Fr63(5):350–355.
20. Leach P, Mealling M, Salz R. A Universally Unique IDentifier (UUID) URN Namespace. IETF RFC 4122 (2005). Available: http://www.ietf.org/rfc/rfc4122.txt.
21. Winkler WE. Overview of record linkage and current research directions. Technical Report RR2006/02. Washington, DC: U.S. Bureau of the Census, 2004.
22. Russell RC. U.S. Patent 1435663, 1922.
23. Breton V, et al. Innovative in silico approaches to address avian flu using grid technology. *Infect Disord Drug Targets* 2009;9(3):358–365.
24. Juckett G. Avian influenza: Preparing for a pandemic. *Ameri Family Phys* 2006; 74(5):783–790.
25. World Health Organization (WHO). Cumulative number of confirmed human cases of avian influenza A/(H5N1) reported to WHO. Available: http://www.who.int/csr/disease/avian_influenza/country/cases_table_2010_08_31/en/index.html. Accessed September 2, 2010.
26. Doan T-T, Bernard A, Da-Costa A-L, Bloch V, LE T-H, Legré Y, Maigne L, Salzemann J, Sarramia D, Nguyen H-Q, Breton V. Grid-based International

Network for Flu Observation (g-INFO). Proceedings of Healthgrid Conference. *Studies Health Technol Inform* 2010;159:215–226.

27. Trendall C, Stewart AJ. General calculations using graphics hardware with application to interactive caustics, In *Rendering Techniques '00 (Proc. Eurographics Rendering Workshop)*, Springer, June 2000, pp. 287–298.
28. Owens JD. A survey of general-purpose computation on graphics hardware. *Computer Graphics Forum* 2007;26(1):80–113.
29. Woundeberg M. Using FPGAs to speed up cellular automata computations, Master's thesis. University of Amsterdam, 2006.
30. Caux J, Siregar P, Hill D. Accelerating 3D cellular automata computation with GP-GPU in the context of multiscale integrative biology. In *Cellular Automata*. Intech, accepted for publication 2010.
31. Luebke DL, Harris M, Krüger J, Purcell T, Govindaraju N, Buck I, Wooley C, Lefohn A. GPGPU: General purpose computation on graphics hardware. In *SIGGRAPH'04, Proceedings of the SIGRAPH 2004 Conference, Course Notes*. New York: ACM Press, 2004.
32. Fan Z, Qiu F, Kaufman A, Yoakum-Stover S. GPU cluster for high performance computing. In *SC'04: Proceedings of the 2004 ACM/IEEE Conference on Supercomputing*, Washington, DC: IEEE Computer Society, 2004, p. 47.
33. Ritchie DW, Venkatraman V, Mavridis L. Using Graphics processors to accelerate protein docking calculations. Proceedings of Healthgrid Conference. *Studies in Health Technol Inform* 2010;159:146–155.
34. Goddeke D, Strzodka S, Turek S. Accelerating double precision (FEM) simulations with (GPUs). *In Proceedings of ASIM 2005—18th Symposium on Simulation Technique* 2005. SCS European Publishing House, Erlangen, Germany.
35. Hill D. Practical distribution of random streams for stochastic high performance computing. In *Proceedings of the 2010 International Conference on High Performance Computing and Simulation*, Smari WW, McIntire JP (Eds.), HPCS 2010, June 28–July 2, 2010, Caen, France. IEEE 2010, pp. 1–8.
36. L'Ecuyer P, Simard R. TestU01: A C Library for empirical testing of random number generators. *ACM Trans on Math Software* 2007;33(4):22:1–40.
37. Agostinelli S, et al. GEANT4: A simulation toolkit. *Nucl Instrum Meth A* 2003;506:250–303.
38. Allison J, et al. GEANT4 developments and applications. *IEEE Trans Nucl Sci* 2006;53:270.
39. Incerti S. Comparison of GEANT4 very low energy cross section models with experimental data in water. *Med Phys* 2010;37(9):4692–4708.
40. Engquist WEB, Li X, Ren W, Vanden-Eijnden E. Heterogeneous multiscale methods, a review. *Commun Comput Phys* 2007;2:367–450.
41. Ingram GD, Cameron IT, Hangos KM. Classification and analysis of integrating frameworks in multiscale modelling. *Chem Eng Sci* 2004;59(11):2171–2187.
42. Von Neumann J. *Theory of Self-Reproducting Automata*. Chicago, IL: University of Illinois Press, 1966. Edited and completed by A. W. Burks.

16

ENABLING PRECOMPETITIVE TRANSLATIONAL RESEARCH: A CASE STUDY

Sándor Szalma

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.

16.1 INTRODUCTION

The difficulty of reconciling animal model data with clinical outcomes has been leading to a growing consensus that the most valuable data source for biomedical discoveries is derived from human samples. This recognition is clearly reflected in the increasing number of translational medicine and translational sciences departments across pharmaceutical companies as well as academic and government-supported initiatives such as Clinical and Translational Science Awards (CTSA) in the United States (http://www.ctsaweb.org/) and the Seventh Framework Programme (FP7) of the European Union (EU) [1], which puts strong emphasis on translating research for human health.

The recent advancement of the idea of precompetitive sharing [2–5] has been quickly gaining ground. Bioinformaticians from the pharmaceutical industry are proposing improved collaboration in computational biology and chemistry between the public domain and the industry [5] by virtualizing informatics tools, services, and infrastructure. Some notable successes have already been achieved.

One such example is Merck's partnership with the Moffit Cancer Center and Research Institute [6]. They have developed a system which enables sharing of human subject data in oncology trials. This system is built from proprietary and commercial components such as Microsoft BizTalk business process server and Tibco and Biofortis LabMatrix applications but does not address any data-sharing issues outside of the two institutions.

Some preliminary pilot studies in other pharmaceutical organizations have been reported [7], but to date no solid evidence for production-level systems being deployed has been found.

Another example of shared infrastructure is the case of CTSA awardees. These institutes have recognized the need for more efficient data sharing (see, e.g., the proposal for the CTSA Human Studies Database (HSDB) Project [8]; Fig. 16.1). The proposed system concentrates on the study results emanating from the CTSA awardees and does not capture the wealth of information generated by institutes which are not part of this grant, and also the proposed system has not yet been put into wide use according to our best knowledge.

The pharmaceutical companies of Johnson and Johnson have established translational and biomarker departments and implemented translational informatics approaches, including building a data warehouse and data-mining application. The solution is heavily reliant on open-source components, and thus the implemented resource and the standardized framework it was built on can form the basis of precompetitive sharing of studies involving samples from human subjects. This in turn can lead to better understanding of human biology and pathophysiology, ultimately leading to more effective management and treatment of diseases in a collaborative setting. This infrastructure is a combination of dedicated people, robust processes, and informatics solution and is called tranSMART [9, 10]. In this chapter the process of building the system, the technical solution, and application examples are described.

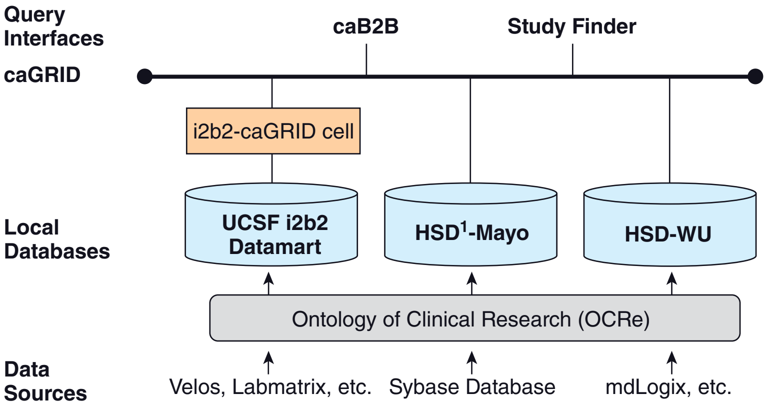


Figure 16.1 Proposed Human Studies Database (HSDB) to enable sharing data between CTSA awardee institutes.

16.2 ESTABLISHING TRANSLATIONAL RESEARCH INFRASTRUCTURE

First, we have established a strong cooperative team across the research and development (R&D) organization of the pharmaceutical companies of Johnson & Johnson. We extended the team with open innovation partnerships with Rutgers University and the Cancer Institute of New Jersey (CINJ). The informatics and information technology group worked in close collaboration with business partner stakeholders from discovery biology, translational medicine, biomarker, and clinical organizations with a goal to develop a system which enables democratic access to all the data generated during target validation, biomarker discovery, preclinical and translational studies, and clinical development.

Change management is an important aspect of successfully introducing a paradigm shift within a large pharmaceutical organization. From the beginning of the project we relied on the collective wisdom of biologists, pharmacologists, and physicians from the therapeutic areas to guide us through the development and eventually to help champion the adaptation of the newly developed translational infrastructure.

16.3 WHY DATA WAREHOUSING

There are multiple approaches one can take to implement a translational data repository. If an organization has established a set of strong primary data source repositories such as databases for omics data from preclinical and clinical studies and clinical databases, one can choose to develop a federated system which will integrate all these sources into a translational system. Special consideration needs to be given to ensure that the source systems have clean data and that there are appropriate identifiers present so that federation can be implemented using cross-references between the data elements in the different source systems.

We chose another route—data warehousing so that a consistent database can be built up by extracting, curating, and transforming the data from multiple primary data sources and finally loaded in a consistent manner into a warehouse schema. The primary reason for this decision was that many experimental modalities lack appropriate primary data repositories and many primary sources lacked standardized dictionaries. Thus a road through cleaning and curating data was deemed to be the most effective way forward.

16.4 BUILDING DATA WAREHOUSE

The translational medicine data warehouse—tranSMART—was developed in partnership with Recombinant Data Corporation (Fig. 16.2). We built the system for a set of translational use cases, such as:

- What is the correlation between animal models and human data?
- What is the best biomarker strategy for a given compound?

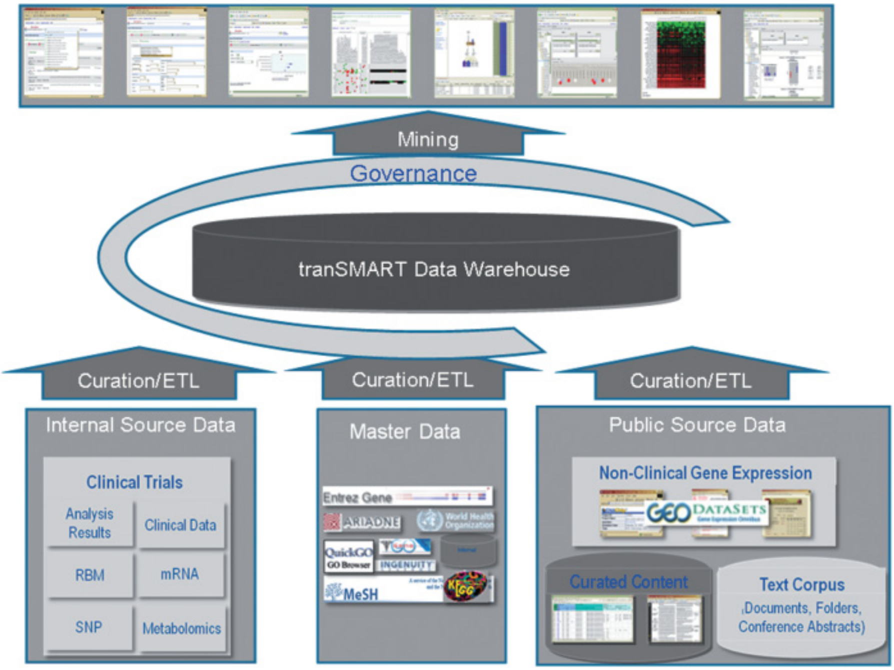


Figure 16.2 Schema of tranSMART data warehouse.

- What is the best indication for a given compound?
- How should a researcher design a trial based on the experience from previous internal and public trials?
- How should a researcher stratify a disease based on clinical data?
- Is there support for a target of interest based on clinical data?

Collecting use cases early in the project and periodically revalidating and refining with the stakeholders is important for a project with longer timelines.

After collecting the use cases we had the first prototype deployed with some basic data in three months. After the successful first demonstration we used the agile software development methodology to build iterations and demonstrate to business partners for feedback and defining the next iteration. The typical cycle time was about 3–5 weeks. The first full deployment of the system was 12 months after the first demonstration of the prototype. By this time we had developed a detailed data governance model in collaboration with data owners, developed publication strategy and training materials, and loaded 10 trials; basic data mining and analysis workflows were available for biologists and physicians.

16.5 CONTENT

At the time of writing the system is one year old and it has more than 30 internal trials with access to deidentified clinical, laboratory chemistry, genomics, protein profiling, metabolomics, proteomics, flow cytometry, protein assay, and single-nucleotide polymorphism (SNP) data at the subject level and 34 public studies with phenotype and genomics data aligned. Furthermore, subset analyses (A versus B comparisons or contrasts) of gene expression or protein profiling data is available for 10 internal studies and more than 9000 public sets. We have curated more than 100,000 biomarker assertions and we also loaded almost 100 studies from the Dana Farber Cancer Institute curated collection.

The data warehouse also provides integrated access to internally developed tools such as an integrative pathway and gene set enrichment analysis tool called Pictor and a gene index and gene information integration resource called Hydra and several third-party tools such as GeneGo's MetaCore and Ariadne Genomics Pathway Studio.

A set of standard dictionaries, ontologies, and curated metadata provides the master data backbone of the data warehouse, including gene and protein names and synonyms from Entrez, gene name mapping vocabulary for Affymetrix, Illumina and Agilent gene expression probeset ID, SNP identifiers, pathways from the Gene Ontology Consortium (GO) [11], Kyoto Encyclopedia of Genes and Genomics (KEGG) [12], GeneGo, Ingenuity, Ariadne, and MSigDB [13], diseases from MeSH [14] and the International Classification of Diseases, 10th revision (ICD-10), clinical trial observations from CDISC SDTM [15], and a dictionary of curated inhibitors. Our curators created an internal J&J drugs dictionary, clinical trials metadata dictionary, and a cell line dictionary.

16.6 DEVELOPMENT METHODOLOGY

The development took about 12 months after the first demonstration prototype using the agile software development methodology. A team of scientists, bioinformatics professionals, and software engineers designed and built the system. The implementation required several distinct efforts. First, we needed to develop appropriate rules and regulations for adding data to the system and granting access to users. Second, we needed to design a system for efficiently storing and querying data. Finally, we needed to develop a system for securely accessing stored data.

A set of appropriate policies for adding data to the system and granting access to data is necessary to ensure cooperation from the data owners and to avoid compliance issues. We modeled the system on processes that were used in academic medical centers in the United Sates. Academic medical centers usually restrict data access to qualified researchers. Researchers are granted access only to data for specific studies and are only allowed to access the data after approval by an institutional review board. Additionally, clinical research data are protected by law through HIPAA (Health Insurance Portability and Accountability Act) and HITECH (Health Information Technology for Economic and Clinical Health Act) regulations. In addition to these restrictions, pharmaceutical companies need to be also bound by U.S. Food and Drug Administration (FDA) regulations and health care compliance.

To develop the appropriate policies for tranSMART, we formed a data governance working group and created a new data steward role in the organization. The data steward is responsible for the data acquisition process, including negotiating with the source data owners, ensuring safe transfer of the data from source systems to the staging area of tranSMART, and the security and integrity of the data in transit.

Publication rules govern the publication of findings resulting from mining the data warehouse. Finally, the data governance working group led an effort to develop guidelines for user training and ethics training which are prerequisite to access. All users are required to attend the official training before being given access to the system.

Two parallel efforts were undertaken for software implementation. First, a team of application developers designed and built a Web-based application to provide graphical user access to the data. The application developers worked closely with research scientists to design the system.

Second, a team consisting of data curators and ETL (extract, transform, and load) engineers worked with research scientists, biostatisticians, and informat-

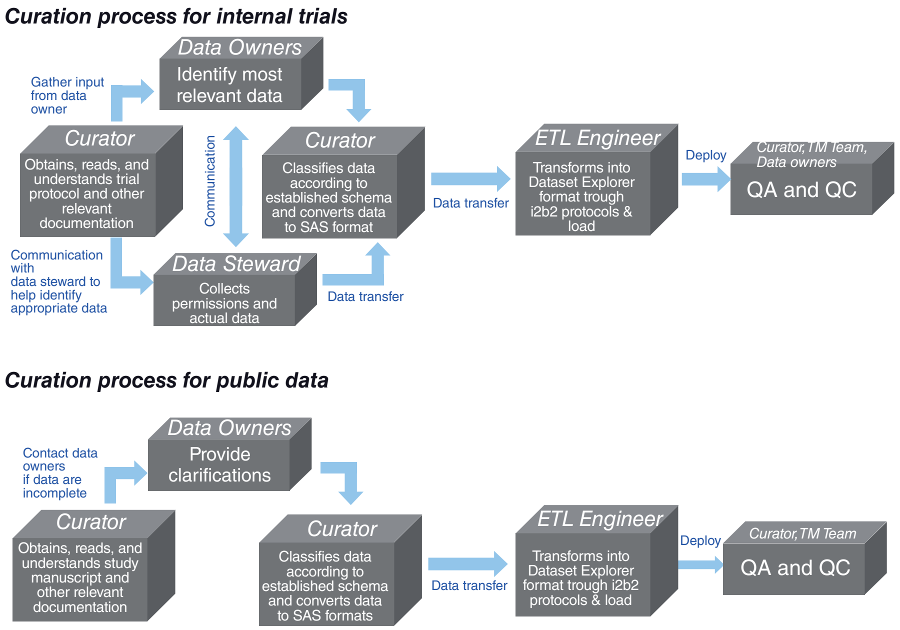

Curation process for internal trials
Data Owners
Identify most relevant data
Gather input from data owner
Curator
Obtains, reads, and understands trial protocol and other relevant documentation
Communication
Curator
Classifies data according to established schema and converts data to SAS format
Data transfer
ETL Engineer
Transforms into Dataset Explorer format trough i2b2 protocols & load
Deploy
Curator, TM Team, Data owners
QA and QC
Communication with data steward to help identify appropriate data
Data Steward
Collects permissions and actual data
Data transfer
Curation process for public data
Data Owners
Provide clarifications
Contact data owners if data are incomplete
Curator
Obtains, reads, and understands study manuscript and other relevant documentation
Curator
Classifies data according to established schema and converts data to SAS formats
Data transfer
ETL Engineer
Transforms into Dataset Explorer format trough i2b2 protocols & load
Deploy
Curator, TM Team
QA and QC

Figure 16.3 Data acquisition, curation, and data-loading process for internal trial data and public studies.

ics professionals to clean and add data to the system. Curators were responsible for understanding each data set added to the system, researching and reconciling inconsistencies in the data, and tagging documents with metadata to facilitate search. Additionally, product manager and curators were responsible for validating data after it was loaded into the system for scientific utility. ETL developers were responsible for loading the data sets and metadata into the database, normalizing data, and making the data available through the application (Fig. 16.3).

16.7 tranSMART DESCRIPTION

All users of the system are authenticated using the same enterprise security processes that the pharmaceutical companies of Johnson and Johnson require for other internal systems. Because tranSMART contains sensitive data, we developed a fine-grained security model for managing information in the system. Each clinical trial data set has a specific owner who can control access to information about and data in the trial.

The tranSMART system was built by maximally optimizing the reuse of open-source and open-data projects, including Lucene [16], i2b2 (http://www.i2b2.org) [17], GenePattern [18], Gene Expression Omnibus (GEO) [19], MeSH [13], and Entrez [20]. The graphical user interface enables data access through two paradigms: a search interface and a hypothesis-testing interface. The search engine allows users to create queries through combinations of elements from a biological subject dictionary, including pathways, genes, diseases, trials, and compounds. The analysis engine (called Dataset Explorer) provided utilities for analyzing phenotypes and genomic data at a cohort level, including trial observations and endpoints. In the following sections we describe in greater depth the functionality of these two engines.

16.7.1 Dataset Explorer

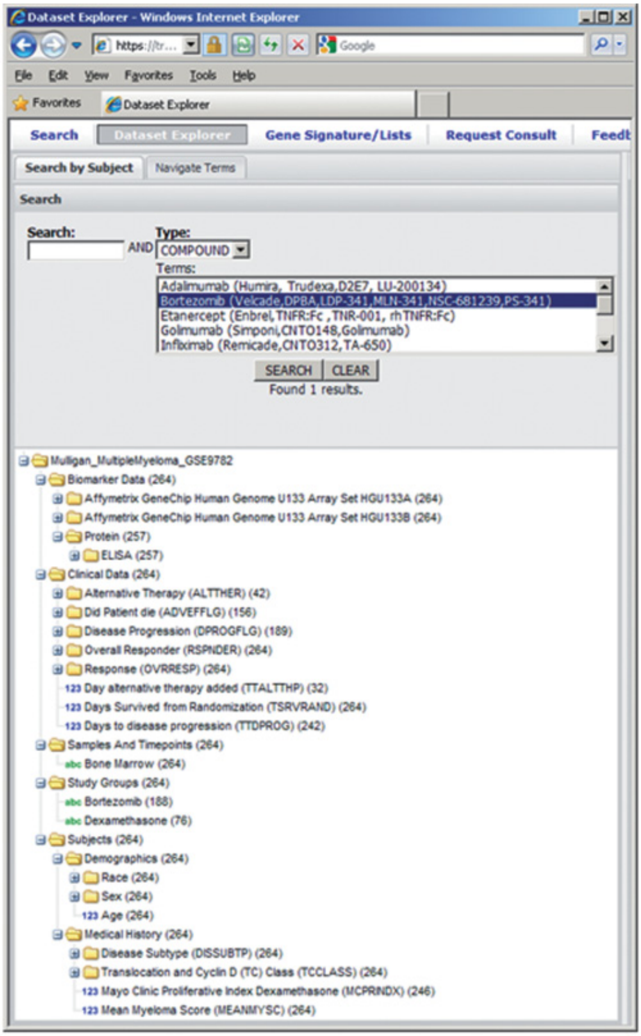
Dataset Explorer provides the scientists and physicians with a unique in silico hypothesis-testing facility. The i2b2 software [17] was used as the basis for this component, but key modifications and additions were made. The users can create virtual cohorts using characteristics from a predefined proprietary ontology. Not all data fields were comparable across studies, so the ontology was organized by study. Within each study, information was categorized using a common structure, including demographic information, clinical data, sample data, and biomarker information (Fig. 16.4). For unique data elements which are comparable across studies we developed a separate ontology and user paradigm to enable cross-comparing multiple studies for a given phenotype.

We created a selection tool for defining multiple cohorts and rapidly establishing statistical comparisons between those cohorts using a *t*-test or χ^2-statistics and visualizing the results using pie charts, histograms, and box plots (Figs. 16.5 and 16.6). Additionally, we implemented components for querying, viewing, and comparing the associated biomarker information such as gene expression, proteomics, rules-based medicine protein panels, metabolomics, and SNP data. Simple (Fig. 16.7) and clustered heatmaps are provided for viewing expression data through the GenePattern tool [18], and linkage disequilibrium maps for viewing SNP data are implemented through the Haploview application [21].

A comparative marker selection [22] module of GenePattern is deployed to develop biomarker hypotheses, and a meta-analysis across multiple gene expression experiments is also enabled for data sets which were measured on comparable platforms (Fig. 16.8).

The system is designed to be able to meet the needs of the casual user so it is envisioned that there are cases where the analysis capabilities implemented are not sufficient for the scientific task at hand. Therefore the aligned and cleaned data can easily be exported for more in-depth analysis.

Finally, we added fine-grained access control to fit our environment as described above. Thus each user has a different view of the study tree according to their access rights: the ones they are granted access by the study owners can be opened in the tree view for exploration but the ones that have no access rights can only see the metadata. Everyone has access rights to the public studies.



Search: AND Type: COMPOUND

Terms:
Adalimumab (Humira, Trudexa,D2E7, LU-200134)
Bortezomib (Velcade,DPBA,LDP-341,MLN-341,NSC-681239,PS-341)
Etanercept (Enbrel,TNFR:Fc ,TNR-001, rhTNFR:Fc)
Golimumab (Simponi,CNTO148,Golimumab)
Infliximab (Remicade,CNTO312,TA-650)

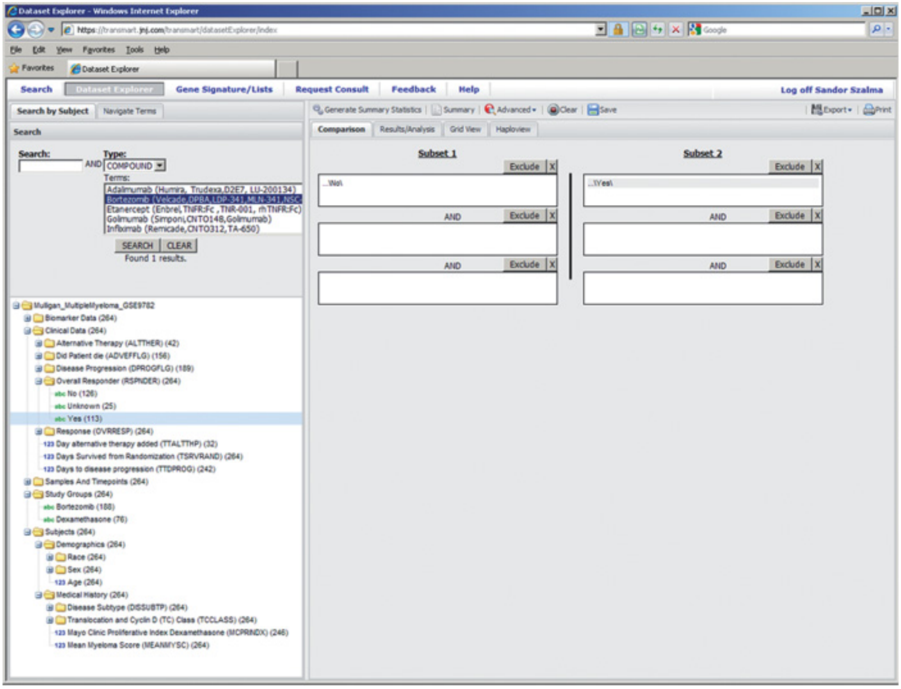
Found 1 results.

```
Mulligan_MultipleMyeloma_GSE9782
  Biomarker Data (264)
    Affymetrix GeneChip Human Genome U133 Array Set HGU133A (264)
    Affymetrix GeneChip Human Genome U133 Array Set HGU133B (264)
    Protein (257)
      ELISA (257)
  Clinical Data (264)
    Alternative Therapy (ALTTHER) (42)
    Did Patient die (ADVEFFLG) (156)
    Disease Progression (DPROGFLG) (189)
    Overall Responder (RSPNDER) (264)
    Response (OVRRESP) (264)
    123 Day alternative therapy added (TTALTTHP) (32)
    123 Days Survived from Randomization (TSRVRAND) (264)
    123 Days to disease progression (TTDPROG) (242)
  Samples And Timepoints (264)
    abc Bone Marrow (264)
  Study Groups (264)
    abc Bortezomib (188)
    abc Dexamethasone (76)
  Subjects (264)
    Demographics (264)
      Race (264)
      Sex (264)
      123 Age (264)
    Medical History (264)
      Disease Subtype (DISSUBTP) (264)
      Translocation and Cyclin D (TC) Class (TCCLASS) (264)
      123 Mayo Clinic Proliferative Index Dexamethasone (MCPRINDX) (246)
      123 Mean Myeloma Score (MEANMYSC) (264)
```

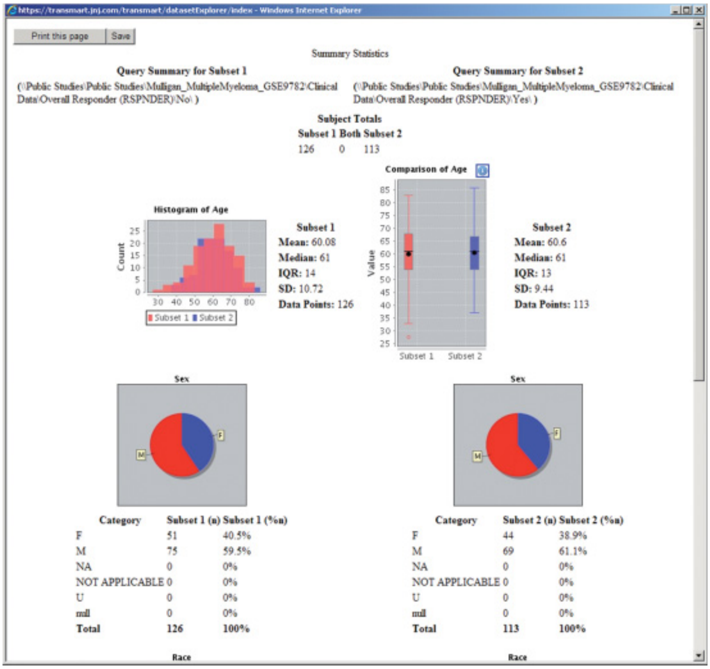
Figure 16.4 Study ontology of a particular public study is shown as implemented in the Dataset Explorer interface of tranSMART. The hierarchy is standardized at minimum at the main-node level with terms such as *biomarker data*, *clinical data*, *samples and timepoints*, *study group*, and *subjects*. Some subnodes such as demographics and medical history are also used across all studies and, where applicable, CDISC SDTM terminology is implemented.

16.7.2 Search Interface

When designing the search interface, the request from scientists was to model the system on familiar user paradigms such as the most frequently used search engine Google. This goal was met by providing scientists with a simple and intuitive widget which allows the scientists and physicians to quickly retrieve


Dataset Explorer - Windows Internet Explorer
https://transmart.jnj.com/transmart/datasetExplorer/index
File Edit View Favorites Tools Help
Favorites
Dataset Explorer
Search
Dataset Explorer
Gene Signature/Lists
Request Consult
Feedback
Help
Log off Sandor Szalma
Search by Subject
Navigate Terms
Search
Generate Summary Statistics
Summary
Advanced
Clear
Save
Export
Print
Comparison
Results/Analysis
Grid View
Haploview
Subset 1
Subset 2
Exclude
AND
..\No\
..\Yes\
Search:
AND
Type:
COMPOUND
Terms:
Adalimumab (Humira, Trudexa,D2E7, LU-200134)
Bortezomib (Velcade,DPBA,LDP-341,MLN-341,NSC
Etanercept (Enbrel,TNFR:Fc ,TNR-001, rhTNFR:Fc)
Golimumab (Simponi,CNTO148,Golimumab)
Infliximab (Remicade,CNTO312,TA-650)
SEARCH
CLEAR
Found 1 results.
Mulligan_MultipleMyeloma_GSE9782
Biomarker Data (264)
Clinical Data (264)
Alternative Therapy (ALTTHER) (42)
Did Patient die (ADVEFFLG) (156)
Disease Progression (DPROGFLG) (189)
Overall Responder (RSPNDER) (264)
abc No (126)
abc Unknown (25)
abc Yes (113)
Response (OVRRESP) (264)
123 Day alternative therapy added (TTALTTHP) (32)
123 Days Survived from Randomization (TSRVRAND) (264)
123 Days to disease progression (TTDPROG) (242)
Samples And Timepoints (264)
Study Groups (264)
abc Bortezomib (188)
abc Dexamethasone (76)
Subjects (264)
Demographics (264)
Race (264)
Sex (264)
123 Age (264)
Medical History (264)
Disease Subtype (DISSUBTP) (264)
Translocation and Cyclin D (TC) Class (TCCLASS) (264)
123 Mayo Clinic Proliferative Index Dexamethasone (MCPRINDX) (246)
123 Mean Myeloma Score (MEANMYSC) (264)

(*a*)



Summary Statistics

Query Summary for Subset 1

(\\Public Studies\Public Studies\Mulligan_MultipleMyeloma_GSE9782\Clinical Data\Overall Responder (RSPNDER)\No\ )

Query Summary for Subset 2

(\\Public Studies\Public Studies\Mulligan_MultipleMyeloma_GSE9782\Clinical Data\Overall Responder (RSPNDER)\Yes\ )

Subject Totals

Subset 1	Both	Subset 2
126	0	113

Subset 1
Mean: 60.08
Median: 61
IQR: 14
SD: 10.72
Data Points: 126

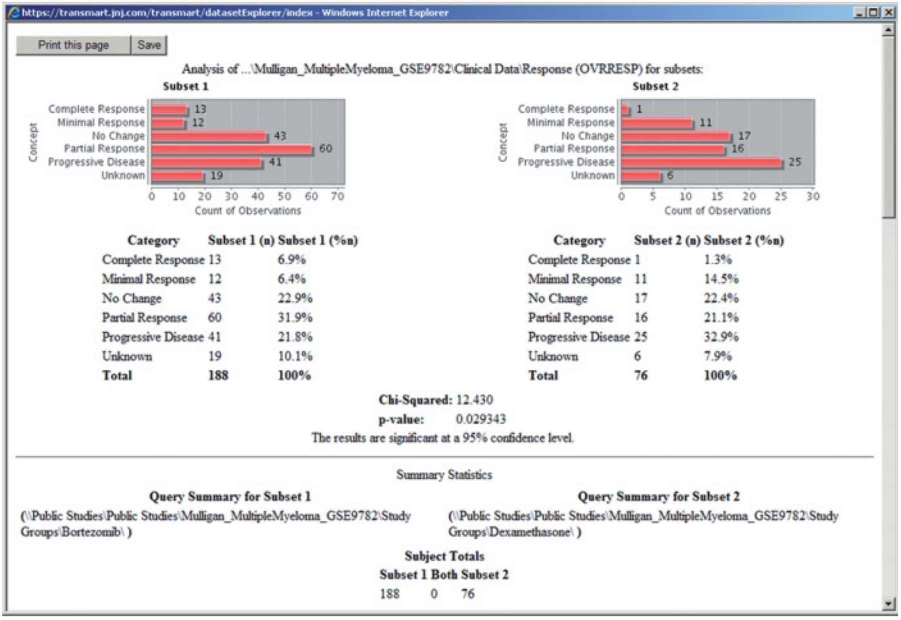
Subset 2
Mean: 60.6
Median: 61
IQR: 13
SD: 9.44
Data Points: 113

Category	Subset 1 (n)	Subset 1 (%n)
F	51	40.5%
M	75	59.5%
NA	0	0%
NOT APPLICABLE	0	0%
U	0	0%
null	0	0%
Total	**126**	**100%**

Category	Subset 2 (n)	Subset 2 (%n)
F	44	38.9%
M	69	61.1%
NA	0	0%
NOT APPLICABLE	0	0%
U	0	0%
null	0	0%
Total	**113**	**100%**

(*b*)

Figure 16.5 (*a*) Basic statistics can be generated for two cohorts when two concepts are selected from the study ontology using drag-and-drop paradigm. (*b*) These selected concepts will define the subjects, which then form the basis of a simple statistical calculation, which then is presented as a set of plots showing distribution of age, gender, and race for the two cohorts.



Analysis of ...\Mulligan_MultipleMyeloma_GSE9782\Clinical Data\Response (OVRRESP) for subsets:

Category	Subset 1 (n)	Subset 1 (%n)
Complete Response	13	6.9%
Minimal Response	12	6.4%
No Change	43	22.9%
Partial Response	60	31.9%
Progressive Disease	41	21.8%
Unknown	19	10.1%
Total	**188**	**100%**

Category	Subset 2 (n)	Subset 2 (%n)
Complete Response	1	1.3%
Minimal Response	11	14.5%
No Change	17	22.4%
Partial Response	16	21.1%
Progressive Disease	25	32.9%
Unknown	6	7.9%
Total	**76**	**100%**

Chi-Squared: 12.430
p-value: 0.029343
The results are significant at a 95% confidence level.

Summary Statistics

Query Summary for Subset 1
(\\Public Studies\Public Studies\Mulligan_MultipleMyeloma_GSE9782\Study Groups\Bortezomib\ )

Query Summary for Subset 2
(\\Public Studies\Public Studies\Mulligan_MultipleMyeloma_GSE9782\Study Groups\Dexamethasone\ )

Subject Totals

Subset 1	Both	Subset 2
188	0	76

Figure 16.6 Phenotype distribution for two cohorts can be calculated once the two cohorts are defined. Through a simple drag-and-drop action, the scientist can identify the concept for which the distribution is calculated and plotted across the two cohorts as well as calculate and plot the appropriate statistics. For continuous variables *t*-statistics and for categorical variables χ^2-statistics are calculated.

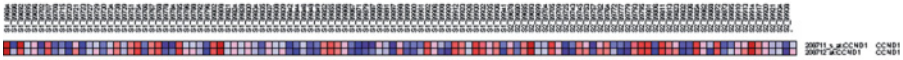
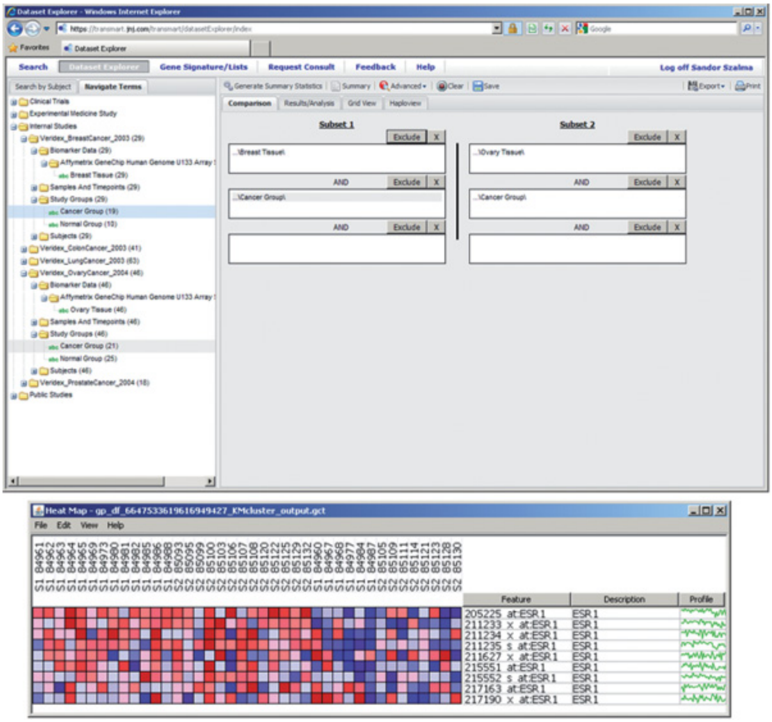


Figure 16.7 If the concept of interest is a biomarker such as gene expression or protein expression, heatmaps can be generated to show the distribution of the levels across the samples of the two cohorts.

information about genes, pathways, compounds, diseases, drug trials, or combinations of therein. The underlying data consist of analyzed proprietary nonclinical and clinical biomarker studies, including gene expression and proteomics data. Additionally, data from public sources were added to the warehouse, including data sets from the National Center for Biotechnology Information (NCBI) GEO [19], EBI Array Express [23], Dana Farber Cancer



Feature	Description
205225 at:ESR1	ESR1
211233 x at:ESR1	ESR1
211234 x at:ESR1	ESR1
211235 s at:ESR1	ESR1
211627 x at:ESR1	ESR1
215551 at:ESR1	ESR1
215552 s at:ESR1	ESR1
217163 at:ESR1	ESR1
217190 x at:ESR1	ESR1

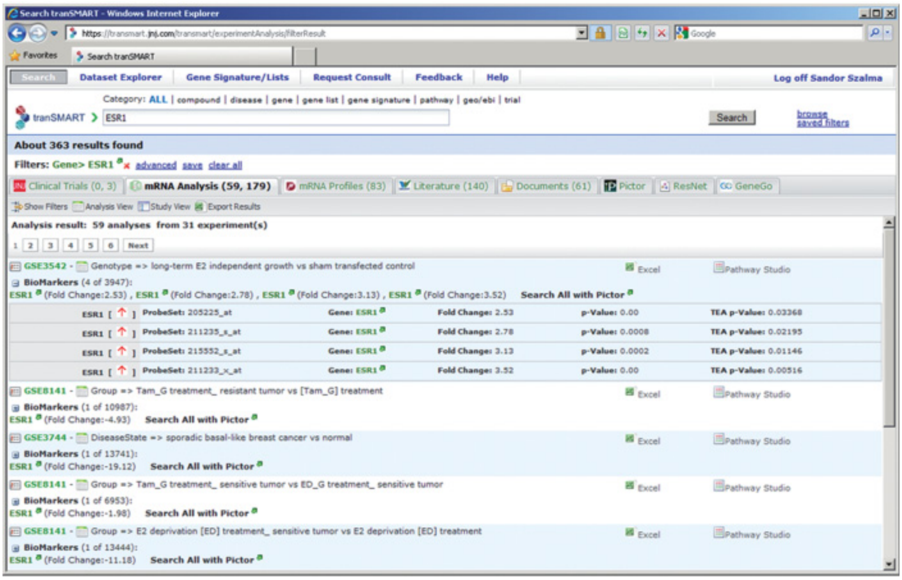
Figure 16.8 Meta-analysis can be carried, as shown. Two cohorts are selected from two studies which used the same platform to measure biomarkers. In this case gene expression was measured in a set of breast and ovary tissues (a) and the *ESR1* gene distribution is shown after *k*-means clustering is applied to the data set using $k = 2$ stratifying some breast tumor tissue samples (denoted as S1 …) as high expressors (red) and as low expressors (blue) and similarly for the ovarian tumor tissue samples (denoted as S2 …).

Institute GeneChip Oncology Database [24], and other sources of gene expression data.

Search results are presented on a series of tabs. Each tab shows a different type of information in an appropriate format. Documents are presented with short summaries, tables of characteristics, and links to the documents. Data are presented in data tables (Fig. 16.9) and heatmaps. Filters are provided to enable users to refine queries.

16.7.3 Signatures

It is a very complex process to select the appropriate indication for an asset in the development stage or even during life-cycle management. Many aspects should be considered, such as a regulatory path for filing, potential market size, differentiability of the therapeutic and experience with and difficulty to



ESR1

About 363 results found

Filters: Gene> ESR1 advanced save clear all

Analysis result: 59 analyses from 31 experiment(s)

GSE3542 - Genotype => long-term E2 independent growth vs sham transfected control
BioMarkers (4 of 3947):
ESR1 (Fold Change:2.53) , ESR1 (Fold Change:2.78) , ESR1 (Fold Change:3.13) , ESR1 (Fold Change:3.52) Search All with Pictor

ESR1 [↑]	ProbeSet: 205225_at	Gene: ESR1	Fold Change: 2.53	p-Value: 0.00	TEA p-Value: 0.03368
ESR1 [↑]	ProbeSet: 211235_s_at	Gene: ESR1	Fold Change: 2.78	p-Value: 0.0008	TEA p-Value: 0.02195
ESR1 [↑]	ProbeSet: 215552_s_at	Gene: ESR1	Fold Change: 3.13	p-Value: 0.0002	TEA p-Value: 0.01146
ESR1 [↑]	ProbeSet: 211233_x_at	Gene: ESR1	Fold Change: 3.52	p-Value: 0.00	TEA p-Value: 0.00516

GSE8141 - Group => Tam_G treatment_ resistant tumor vs [Tam_G] treatment
BioMarkers (1 of 10987):
ESR1 (Fold Change:-4.93) Search All with Pictor

GSE3744 - DiseaseState => sporadic basal-like breast cancer vs normal
BioMarkers (1 of 13741):
ESR1 (Fold Change:-19.12) Search All with Pictor

GSE8141 - Group => Tam_G treatment_ sensitive tumor vs ED_G treatment_ sensitive tumor
BioMarkers (1 of 6953):
ESR1 (Fold Change:-1.98) Search All with Pictor

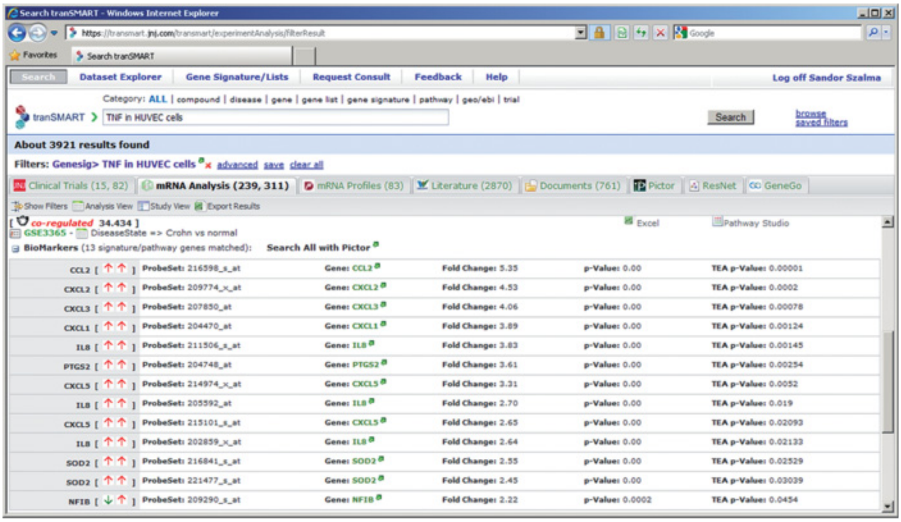
GSE8141 - Group => E2 deprivation [ED] treatment_ sensitive tumor vs E2 deprivation [ED] treatment
BioMarkers (1 of 13444):
ESR1 (Fold Change:-11.18) Search All with Pictor

Figure 16.9 Simple search is illustrated by using a gene name (*ESR1*) as a search term. The set of hits from different data sources are presented in a tabbed window. The public gene expression hit list is shown.

carry out clinical trials in the disease of interest. Besides these dimensions one should also consider if there is strong supporting scientific evidence for the involvement of the targeted pathway in the pathophysiology of the disease. We chose to establish this evidence statistically by mining a large corpus of biological data such as gene expression experiments associated with diseases. In order to accomplish that, we introduced the notion of signatures and contrasts in the data warehouse.

First, drug signatures can be uploaded—these are genes which are differentially expressed when a drug candidate is added to an in vitro system which is believed to mimic some aspects of the biology of a disease. Here the gene expression is measured using a gene chip before applying the drug and after and the difference is evaluated using, for example, a simple *t*-test. Genes which pass quality control (QC) criteria and some predefined threshold then comprise the signature of the drug candidate. These genes can be either upregulated or downregulated and they are recorded as such in the signature.

Second, we stored in the data warehouse a corpus of gene expression contrasts, that is, A versus B comparisons with associated diseases or drugs from public databases described above. Each contrast contains a set of disregulated genes (fold change with directionality and associated *p*-values) which pass predefined quality criteria and metadata, including disease, drug, or phenotype information. These contrasts are processed by Omicsoft Corporation using



TNF in HUVEC cells

About 3921 results found

Filters: Genesig> TNF in HUVEC cells advanced save clear all

[co-regulated 34.434]
GSE3365 - DiseaseState => Crohn vs normal
BioMarkers (13 signature/pathway genes matched):

CCL2 [↑ ↑]	ProbeSet: 216598_s_at	Gene: CCL2	Fold Change: 5.35	p-Value: 0.00	TEA p-Value: 0.00001
CXCL2 [↑ ↑]	ProbeSet: 209774_x_at	Gene: CXCL2	Fold Change: 4.53	p-Value: 0.00	TEA p-Value: 0.0002
CXCL3 [↑ ↑]	ProbeSet: 207850_at	Gene: CXCL3	Fold Change: 4.06	p-Value: 0.00	TEA p-Value: 0.00078
CXCL1 [↑ ↑]	ProbeSet: 204470_at	Gene: CXCL1	Fold Change: 3.89	p-Value: 0.00	TEA p-Value: 0.00124
IL8 [↑ ↑]	ProbeSet: 211506_s_at	Gene: IL8	Fold Change: 3.83	p-Value: 0.00	TEA p-Value: 0.00145
PTGS2 [↑ ↑]	ProbeSet: 204748_at	Gene: PTGS2	Fold Change: 3.61	p-Value: 0.00	TEA p-Value: 0.00254
CXCL5 [↑ ↑]	ProbeSet: 214974_x_at	Gene: CXCL5	Fold Change: 3.31	p-Value: 0.00	TEA p-Value: 0.0052
IL8 [↑ ↑]	ProbeSet: 205592_at	Gene: IL8	Fold Change: 2.70	p-Value: 0.00	TEA p-Value: 0.019
CXCL5 [↑ ↑]	ProbeSet: 215101_s_at	Gene: CXCL5	Fold Change: 2.65	p-Value: 0.00	TEA p-Value: 0.02093
IL8 [↑ ↑]	ProbeSet: 202859_x_at	Gene: IL8	Fold Change: 2.64	p-Value: 0.00	TEA p-Value: 0.02133
SOD2 [↑ ↑]	ProbeSet: 216841_s_at	Gene: SOD2	Fold Change: 2.55	p-Value: 0.00	TEA p-Value: 0.02529
SOD2 [↑ ↑]	ProbeSet: 221477_s_at	Gene: SOD2	Fold Change: 2.45	p-Value: 0.00	TEA p-Value: 0.03039
NFIB [↓ ↑]	ProbeSet: 209290_s_at	Gene: NFIB	Fold Change: 2.22	p-Value: 0.0002	TEA p-Value: 0.0454

Figure 16.10 The tumor necrosis factor (TNF) signature comprising significantly differentially expressed genes from two sets of samples from human umbilical vein endothelial cells (HUVECs) treated with TNF vs. nontreated samples was generated. This signature then was uploaded into tranSMART and used in an enrichment search using the target enrichment analysis (TEA) algorithm. Here a statistically significant hit is presented—GSE3365—comparing Crohn's disease patients' peripheral blood mononuclear cells (PBMCs) with normals. All the genes from the signature which have corresponding genes in this contrast are shown in the table with the *p*-values and fold changes calculated from GSE3365 as well as the so-called TEA scores. The total score is shown on the top.

state-of-the art methodology and strict QC based on the requirements of Johnson and Johnson informatics scientists.

Finally, we deployed a statistical method to measure the enrichment of a signature across disease comparisons. The resulting statistically significant hits then comprise the disease indication hypothesis for the drug candidate. Such an enrichment analysis is shown in Figure 16.10.

16.7.4 Federation

We ultimately chose data warehousing as a solution for slowly changing or internal data. For more dynamic content we followed the data federation approach. Thus a set of external data sources are federated when searching for genes, drugs, and pathways and the results are rendered in the application. Primarily, gene-centric information is linked from public sources using Entrez gene id—direct link to Entrez Gene, Entrez Global search, and Google Scholar; licensed sources—GeneCards (http://www.genecards.org); and internal sources—a gene index application called Hydra and a pathway integration application called Pictor. This latter application integrates multiple-pathway databases and does an enrichment analysis over GO and KEGG and licensed applications such as MetaCore from GeneGo, IPA from Ingenuity, and Pathway Studio from Ariadne Genomics. For drugs the application federates the National Library of Medicine Drug Information Portal.

16.7.5 Text Indexing and NLP

Access to text sources is provided via four paradigms. First, as described previously for accessing textual reference information for a gene of interest, it is provided by federating the PubMed interface of MEDLINE abstracts and Google Scholar.

Second, simple text indexing is provided through Lucene [16], an open-source text indexer. Currently, the system indexes text sources such as group folders and repositories of abstracts of scientific conferences.

Third, we decided to curate a set of scholarly articles for biomedical assertions. At the start of the project a set of important biomedical concepts were identified by the users—such as a set of particular diseases and targets. Published scholarly articles contain a wealth of information about these topics of interest, and it is a challenge to use computational approaches such as text indexing or even natural language processing (NLP) algorithms [25] to extract quantitative facts with high accuracy. Quantitative facts such as the number of subjects or percent of observations within the study used to establish an assertion are critical for decision making in areas such as biomarker and disease indication selection. We used the services of a team of biologists who extracted biomedical facts from a selected set of journal articles. The extraction was done using a predefined structured template and the data were subsequently tagged, stored, and made searchable within tranSMART.

While the manual process provides high-accuracy extraction of biomedical assertions it is very resource expensive and conversely its coverage is limited. Therefore, fourth, tranSMART also provides access to assertions about biomarkers generated from MEDLINE by the Ariadne Genomics MedScan Reader engine. For this application we have been devoting considerable resources to improve the accuracy of the NLP engine and the fidelity of assertion extraction for specific scientific subdomains such as immunology, oncology, and clinical trials by developing specific text-mining cartridges.

16.7.6 Workflows

In silico analyses rarely consist of only one step—they are typically a result of multistep, complex workflows. The system addresses some of the workflows by extending the data export capability and deploying specific interfaces to other applications such as Microsoft Excel and Ariadne Genomics Pathway Studio for further mining and modeling of data. For example, gene expression data can be exported directly to Ariadne Genomics Pathway Studio by a single click and further analyzed in the external application. Here pathway and network enrichment or network building can be carried out for in-depth exploration of the data.

An even more interesting workflow was enabled when we connected the manually curated biomedical assertions described previously to Ariadne Genomics Pathway Studio. In this module the quantitative biomedical assertions are translated into semantic triples by a software algorithm, exported into the appropriate XML format, and passed on for network visualization to Pathway Studio (Fig. 16.11).

16.8 STRATEGIC CONSIDERATIONS

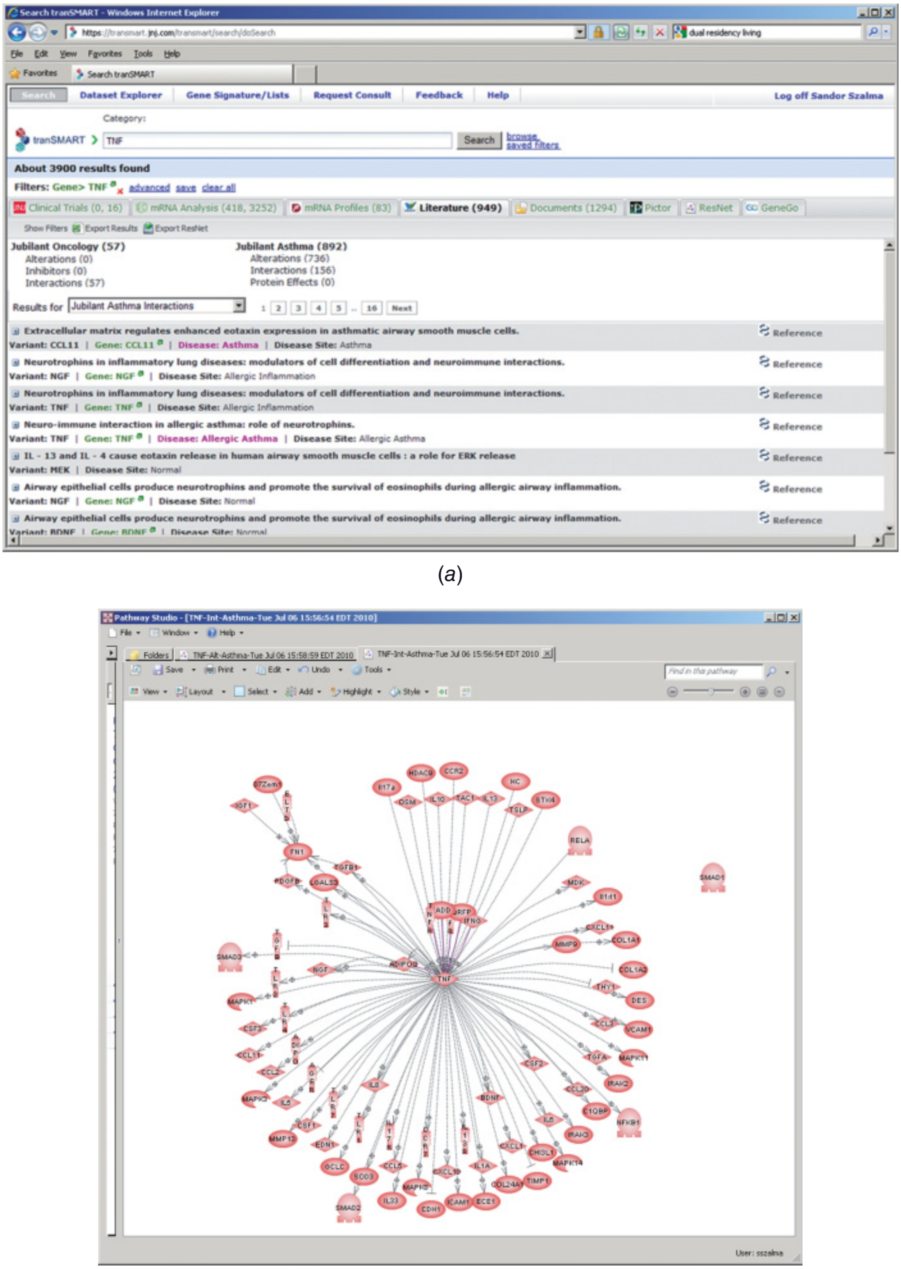
In building tranSMART, we made several strategic decisions that have proven very beneficial. First, we have selected the core technologies such that we can build on previous efforts in a most efficient way. After rigorous evaluation, we decided to build tranSMART on open-source frameworks such as i2b2, GenePattern, and Haploview to enable data sharing and partnering. Alternatives were considered such as proprietary coding or several commercial offerings, but we were intrigued by the possibilities of taking the open-source approach. The result has been an excellent and robust system, access to an outstanding community of academic partners, and the ability to share and propagate our success.

Second, a preferred academic partner was selected as core partner for the program. In this case, we chose the Cancer Institute of New Jersey (CINJ) for its informatics and medical expertise which we utilized throughout the project. In return, the institute will receive a full version of the software for internal implementation. This provides an obvious economic "win–win" for both sides and serves as a model for further academic and nonprofit collaborations that is being established.

Finally, tranSMART is the first Johnson and Johnson application to be hosted externally on the Amazon Elastic Compute Cloud (EC2). Cloud computing offers the advantages of low entry cost, a pay-per-use system, high speed, ready accessibility, easier access for external partners, and a potential to efficiently integrate with other public available resources hosted on the same cloud. At Johnson and Johnson, we achieved a 12-fold reduction in hosting costs, as well as increased "up-time," without incurring any additional risks associated with protection of information assets.

16.9 DISCUSSION

The system has been deployed for almost a year at the time of this writing. During this time we have trained approximately 200 internal scientists and physicians and the system was queried close to 4000 times. The data warehouse contains more than 60 internal and public human studies with the access to



Search tranSMART - Windows Internet Explorer

https://transmart.jnj.com/transmart/search/doSearch

dual residency living

File Edit View Favorites Tools Help

Favorites | Search tranSMART

Search | Dataset Explorer | Gene Signature/Lists | Request Consult | Feedback | Help | Log off Sandor Szalma

Category:

tranSMART > TNF | Search | browse, saved filters

About 3900 results found

Filters: Gene> TNF advanced save clear all

Clinical Trials (0, 16) | mRNA Analysis (418, 3252) | mRNA Profiles (83) | Literature (949) | Documents (1294) | Pictor | ResNet | GeneGo

Show Filters | Export Results | Export ResNet

Jubilant Oncology (57)
Alterations (0)
Inhibitors (0)
Interactions (57)

Jubilant Asthma (892)
Alterations (736)
Interactions (156)
Protein Effects (0)

Results for Jubilant Asthma Interactions | 1 2 3 4 5 .. 16 Next

Extracellular matrix regulates enhanced eotaxin expression in asthmatic airway smooth muscle cells. Reference
Variant: CCL11 | **Gene:** CCL11 | **Disease:** Asthma | **Disease Site:** Asthma

Neurotrophins in inflammatory lung diseases: modulators of cell differentiation and neuroimmune interactions. Reference
Variant: NGF | **Gene:** NGF | **Disease Site:** Allergic Inflammation

Neurotrophins in inflammatory lung diseases: modulators of cell differentiation and neuroimmune interactions. Reference
Variant: TNF | **Gene:** TNF | **Disease Site:** Allergic Inflammation

Neuro-immune interaction in allergic asthma: role of neurotrophins. Reference
Variant: TNF | **Gene:** TNF | **Disease:** Allergic Asthma | **Disease Site:** Allergic Asthma

IL - 13 and IL - 4 cause eotaxin release in human airway smooth muscle cells : a role for ERK release Reference
Variant: MEK | **Disease Site:** Normal

Airway epithelial cells produce neurotrophins and promote the survival of eosinophils during allergic airway inflammation. Reference
Variant: NGF | **Gene:** NGF | **Disease Site:** Normal

Airway epithelial cells produce neurotrophins and promote the survival of eosinophils during allergic airway inflammation. Reference
Variant: BDNF | **Gene:** BDNF | **Disease Site:** Normal

(*a*)

(*b*)

Figure 16.11 (*a*) Simple search results from the curated biomedical assertions can be exported to (*b*) Ariadne Genomics Pathway Studio application for further analysis and mining. Assertions with respect to TNF interactions with other genes in the context of asthma and other related pulmonary diseases are turned into semantic triples and visualized in a network diagram.

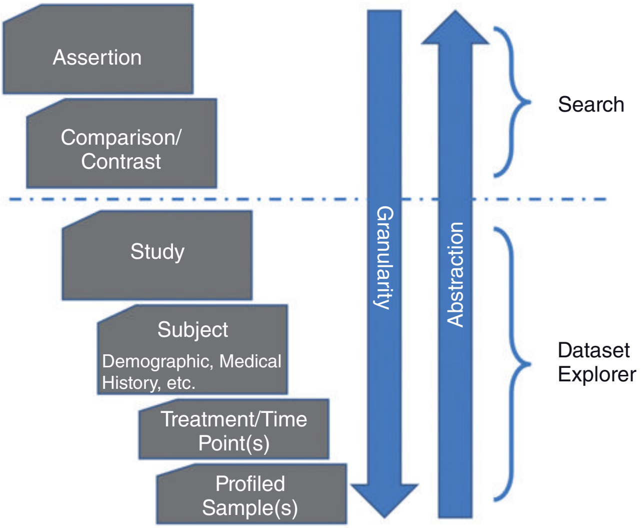


Figure 16.12 Different levels of granularity and abstraction are captured from a study in the tranSMART data warehouse. The two graphical user interfaces then enable access of these levels of data through different query and presentation paradigms.

subject-level data, more than 9000 curated in vitro, in vivo, and clinical contrasts, and 100,000 curated assertions in a hierarchical structure, as depicted in Figure 16.12. The primary users of the systems are scientists from the immunology and oncology therapeutic areas, but there are data available from psychiatry and cardiovascular studies.

As described above, the data warehouse infrastructure consists of open-source derived software as well as content in the form of well-curated data sets derived from public studies. Both "public-derived" aspects of the data warehouse make this resource well positioned for precompetitive sharing of education, training, and best practices which are considered lacking for many collaborations to be truly innovative and less siloed [26]. Indeed, we are currently working with CINJ and several other academic, nonprofit organizations and consortia to deploy and utilize the data warehouse in this type of setting. Each of these partners may bring in new perspectives such as oncology clinical practice or connection to medical records, and through these collaborations the understanding of biology and translational medicine may increase in a much more optimal way. Moreover, the systems supporting these endeavors will improve in an innovative, rapid, and robust manner.

We have succeeded in developing a system to bridge source systems and enable translational research with breaking down most of the data silos (Fig. 16.13). Future extensions are being worked on, such as connection to the internal biobank, connecting more internal source systems, and extending the capabilities to cover more types of data modalities and deploy advanced analysis methodologies such as network analysis.

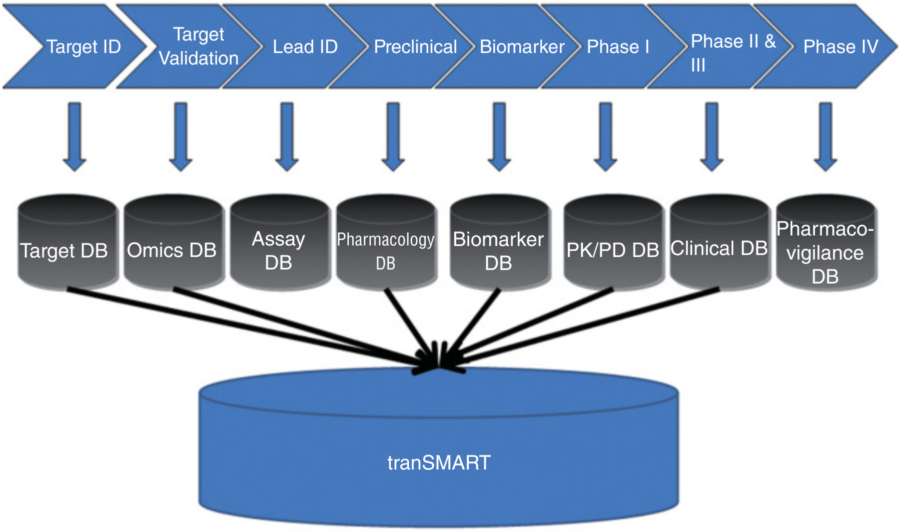


Figure 16.13 tranSMART integrates data from multiple data silos across the drug discovery and development process.

ACKNOWLEDGMENTS

We wish to thank the following individuals for their outstanding effort and dedication during the tranSMART project: Eric Perakslis, Jason Smart, Greg Leibfreid, Robert Gruninger, John Boles, George Kuebrich, Dorian Tisdale, Venkat Koka, Eudean Shaw, Hongtao Fan, Xiang Yao, Jean-Marc Neefs, Hans Winkler, Elliot Barnathan, Anuk Das, and Chris Ward (employees of the pharmaceutical companies of Johnson and Johnson); Dan Housman, Bruce Johnson, Jinlei Liu, Joe Adler, Anthony Ioven, Tarang Jatkar, Jean Avitable, Haiping Xia, and Aaron Mandel (employees of Recombinant Data Corporation); Tatiana Khasanova and Svetlana Bureeva (employees of GeneGo); and Gyan Bhanot and Joe Irgon (Rutgers University).

REFERENCES

1. Seventh Framework Programme. http://ec.europa.eu/research/fp7/index_en.cfm?pg=health.
2. Stoffels P. Collaborative innovation for the post-crisis world. The Boston Globe 2009;February 2, A13.
3. Friend SH. The need for precompetitive integrative bionetwork disease model building. *Clin Pharmacol Ther* 2010;87:536–539.
4. Hunter J, Stephens S. Is open innovation the way forward for big pharma? *Nature Rev Drug Disc* 2009;9:87–88.
5. Barnes MR, et al. Lowering industry firewalls: Pre-competitive informatics initiatives in drug discovery. *Nature Rev Drug Disc* 2009;8:701–708.

6. Cancer center builds gene biobank. http://www.bio-itworld.com/BioIT_Article.aspx?id=49382.
7. Human studies database. http://rctbank.ucsf.edu/home/hsdb.html.
8. Shon J, Varma R, Mahuvakar V, Vig J, Gonzaludo N, Chiu S. Integrating pre-clinical and clinical data sets to enable translational science queries and comparisons of multiple biological result types. Paper presented at the 2010 Summit on Translational Bioinformatics, San Francisco, March 10–12, 2010.
9. Perakslis ED, Van Dam J, Szalma S. How informatics can potentiate pre-competitive open source collaboration to jump-start drug discovery and development. *Clin Pharma Ther* 2010;87:614–616.
10. Szalma S, Koka V, Khasanova T, Perakslis ED. Effective knowledge management in translational medicine. *J Transl Med* 2010;8:68.
11. The Gene Ontology Consortium. Gene ontology: Tool for the unification of biology. *Nat Genet* 2000;25:25–29.
12. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* 2010;38:D355–360.
13. Subramanian A, Tamayo P, Mootha V, Mukherjee S, Ebert BL, Gillette MA, Pomeroy S, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005;102(43):15545–1550.
14. MeSH. http://www.nlm.nih.gov/mesh/.
15. CDISC SDTM. http://www.cdisc.org/sdtm.
16. Lucene. http://lucene.apache.org/java/docs/index.html.
17. Murphy S, et al. Instrumenting the health care enterprise for discovery research in the genomic era. *Genome Res* 2009;19:1675–1681.
18. Reich M, et al. GenePattern 2.0. *Nature Genet* 2006;38:500–501.
19. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002;30:207–210.
20. Sayers EW, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2010;38:D5–16.
21. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005;21:263–265.
22. Gould J, Getz G, Monti S, Reich M, Mesirov JP. Comparative gene marker selection suite. *Bioinformatics* 2006;22:1924–1925.
23. Parkinson H, et al. ArrayExpress update—From an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res* 2009;37:D868–872.
24. GCOD. http://compbio.dfci.harvard.edu/tgi/cgi-bin/tucan/tucan.pl.
25. Novichkova S, Egorov S, Daraselia N. MedScan, a natural language processing engine for MEDLINE abstracts. *Bioinformatics* 2003;19:1699–1706.
26. Wagner JA. Open-minded to open innovation and precompetitive collaboration. *Clin Pharma Therap* 2010;87:511–515.

17

COLLABORATION IN CANCER RESEARCH COMMUNITY: CANCER BIOMEDICAL INFORMATICS GRID (caBIG)

George A. Komatsoulis

17.1 INTRODUCTION

Cancer in its myriad forms is the second leading cause of death in the United States, exceeded only by heart disease. In 2006 alone (the latest date for which complete information is available) 1,479,350 Americans received a cancer diagnosis and 562,340 persons died of cancer [1]. Even within a particular disease site, cancer is actually a collection of diseases that have different

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.

underlying causes. This complexity, which often does not manifest at the phenotypic level, makes it particularly difficult to classify disease variants and select the most effective treatment for a specific patient.

The only obvious mechanism to resolve this problem is to assess the particular genetic basis of an individual patient's disease and then select a treatment known to be effective against that particular abnormality. This "personalized medicine" paradigm has been a dream of many in the biomedical community but interestingly has been successfully used since the 1970s in the treatment of childhood acute lymphoblastic leukemia (ALL). Five-year survival for ALL patients aged 10–14 at diagnosis have increased from 58.8% for children diagnosed between 1975 and 1977 to 79.7% for those diagnosed between 1996 and 2002. This increase can be attributed to several factors, notably the fact that children are treated in an environment that blends care and research and the use of (for the time) modern molecular biology technology, in this case karyotypes (a view of the number and gross physical structure of the chromosomes in a cell) [2]. By combining genetic technology with outcomes research, it was possible to determine a patient's risk profile and provide a treatment that was most likely to achieve a desirable outcome. This model (the blending of care and research coupled with the use of genomics technology to provide a molecular characterization of the patient's specific disease) is the goal of those that wish to bring personalized medicine to all. However, its success is dependent on removing the barriers between care and research and ensuring that data from a wide range of sources (clinical outcomes, genomics, pathology, images) can be successfully integrated.

This need to provide data liquidity (that is, the free flow of information among those authorized to use it) required by the molecular medicine paradigm was the genesis of the cancer Biomedical Informatics Grid (caBIG®) program (http://cabig.cancer.gov). The program was initiated in 2004 (at the request of the NCI's National Cancer Advisory Board) to be overseen by the Center for Biomedical Information and Information Technology (CBIIT) of the U.S. National Cancer Institute (NCI), a part of the National Institutes of Health (NIH), U.S. Department of Health and Human Services. It was tasked with creating a virtual network of interconnected data (including clinical, pathology, genomics, and imaging data), individuals, and organizations whose goal is to redefine how research is conducted, care is provided, and patients/participants interact with the biomedical research enterprise—in effect, to create a "world wide web" for cancer among a highly diverse collection of stakeholders (cancer centers, cooperative groups, individual researchers, etc.) that were widely distributed across the United States. Further, these researchers tended to exist within professional silos (pathology, clinical research, genomics, information technology, etc.) To accomplish this goal, therefore, caBIG needed to resolve two fundamental issues: First, it had to create a technology platform that would allow for interoperability between and among various biomedical information systems, and it had to address the social issues associated with the large-scale data sharing required by the personalized medicine paradigm and modern biomedical research. Ultimately, caBIG is a model that is extensible to other research and care communities to enable the large-scale collaborations that are needed in the fight against complex disease.

17.2 caBIG COLLABORATION STRATEGY: OVERVIEW

As described in the introduction, the collaboration strategy taken by caBIG® involved two major components: providing the technical infrastructure required in a large-scale collaborative effort and working with the community to remove the social barriers associated with such collaborations. In order to accomplish these two goals, caBIG adopted a set of core principles [3] to guide its activities:

Open Access The caBIG program is open to all, enabling access to tools, data, and other infrastructure by the biomedical research community.

Open Development All caBIG software development is driven by the needs of the cancer research community as defined by caBIG participants. Although these projects are built by designated teams, all artifacts, from use cases to software to test and bug logs are publically available to all.

Open Source All caBIG products that are created using National Cancer Institute (NCI) funding are released under a nonviral open-source license.* This license is designed to enable the creation of both open- and closed-source derivative works and explicitly allows for commercialization of products that are based on caBIG technology.

Federation Ultimately, by providing an interoperable infrastructure, caBIG allows information to remain under the control of those that are responsible for its integrity and security while still enabling the sharing that is required for personalized medicine. The analogy is to the federal system of government in the United States, with the central government providing a set of capstone policies and regulating interstate trade while leaving other powers to the states. Similarly, caBIG provides a set of standards to enable sharing while ensuring that data providers have ultimate control over their information.

Ultimately, caBIG is about meeting the needs of the oncology and biomedical research community, and while actively managed by the NCI, its priorities and

* In the context of open source, a viral license is one that places restrictions on "derivative works," that is, software that uses or extends the open-source software. The GNU Public License (GPL) is a viral license in the sense that all derivative works are made open source by virtue of the terms in the license of the original work. The terms of a nonviral license do not pass through to derivative works, so that the derivative work can be open source or closed source at the discretion of the developer of the derivative product.

activities are those of oncology researchers and caregivers as defined by their representatives. For this reason, caBIG has devoted substantial effort toward community-based collaboration strategies.

17.3 caBIG COLLABORATION STRATEGY: COMMUNITY

As indicated, convening the community has been a major focus of caBIG from the beginning of the program. Doing so provides two major benefits: First, it provides a forum for community members to express their needs in the realm of biomedical informatics and to have those requirements translated into software systems. Second, it provides a means to address the legal, ethical, and social barriers to the data liquidity required by the new molecular medicine paradigm.

The basic unit of organization for convening the community in caBIG is called a "workspace," and it consists of individuals that have a set of common interests. A workspace has a workspace lead (generally a subject matter expert from the caBIG prime contractor), an NCI facilitator (a CBIIT staff member), and a technical lead (a subject matter expert from the entity that manages caBIG technology development). In the first several years of the program the workspace and technical lead were often the same individual, but this was changed to reflect the divergent duties of those two roles. A workspace meets regularly by Web and teleconference (the frequency varies from every other week to monthly, depending on the workspace) and has two or three face-to-face meetings each year. Face-to-face meetings are typically held at cancer centers that participate in the workspace and last from two to three days. The agenda at workspace teleconferences varies from workspace to workspace, but common activities include requirements gathering, reports from working groups, updates on software development efforts, and presentations on activities in other workspaces or other non-caBIG programs. Although the NCI provides support for setting and communicating agendas, the actual content of the agendas is driven by the needs and desires of the cancer research community. Face-to-face meetings are meant to provide an opportunity for more extensive discussion and debate than can be accommodated on the shorter (1- to 2-hr) workspace calls and to provide a venue for formal decision taking, training, and social interaction that are essential to collaboration.

Membership in workspaces is drawn from the community, broadly defined. Any person, whether from academia, industry, government, the advocacy community, or elsewhere, is welcome to join any workspace, and there are no restrictions on the number of workspaces that an individual may join or the number of representatives from a particular organization in a workspace. The membership of workspaces tends to be highly diverse, ranging from informatics staff to bench scientists and clinical researchers and the biomedical informaticians that bridge the gaps between researchers and information technology (IT) personnel. The level of time commitment is similarly varied; some par-

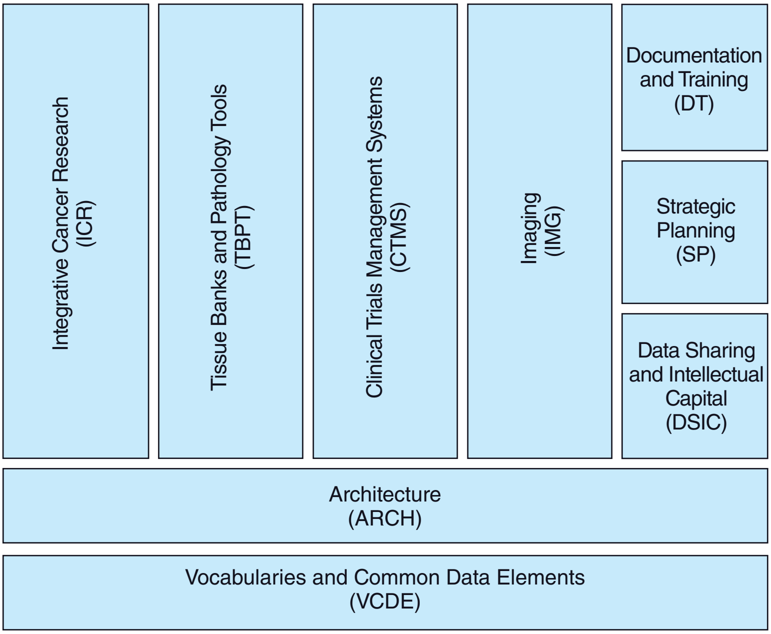


Figure 17.1 Organization of caBIG workspaces.

ticipants engage multiple workspaces and are present at all meetings, others only attend when items that are of interest will be discussed. Where there are needs for specific skills, the NCI provides a minimal amount of support for subject matter experts to be part of caBIG workspaces, but the vast majority of participants are unpaid volunteers. All members, whether supported or volunteer, have the same standing in caBIG workspaces and the same level of input into that workspace's activities. In that sense, caBIG workspaces function as virtual town hall meetings, with the workspace lead and facilitator working to identify a consensus among a group of peers.

There are three classes of workspaces within the caBIG program (Fig. 17.1). The first type is known as a "domain" workspace and is organized by scientific discipline. The current domain workspaces are clinical trials management systems (CTMS), tissue banks and pathology tools (TBPT), imaging (IMG), and integrative cancer research (ICR), the latter covering basic biological research. There are two "cross-cutting" workspaces that provide services across caBIG, the architecture workspace (ARCH) that is charged with providing the technical underpinnings for caBIG and the vocabularies and common data elements workspace (VCDE) that is responsible for semantics and data standards. Finally there are three "strategic" workspaces: documentation and training (D&T), strategic planning, and data sharing and intellectual capital (DSIC), the latter workspace charged with easing the legal, regulatory, and social roadblocks to data sharing.

Once the caBIG community (via the workspaces) and the NCI have defined their needs and priorities, the caBIG program endeavors to implement those requirements in nonviral open-source software or policies. Generally speaking, such activities are funded by the NCI using the standard federal contracting system described by the Federal Acquisition Regulations, or FAR. In alignment with the core principle of open development, all artifacts and activities of caBIG software and policy development activities are open and made available to the community as they are created. This includes requirements documents, software deliverables, quality assurance (QA) test results, and bug lists. Groups that are funded to develop these tools regularly update interested workspaces on their progress so that they can report back to their institutional users and help develop the next generation of requirements for these tools.

17.4 caBIG COLLABORATION STRATEGY: TECHNOLOGY

While convening the community is a necessary precondition to enabling large-scale collaboration in the cancer community, it also requires a technical infrastructure that can support those collaborations. Further, the NCI's stated principle of federation ruled out the creation of a small number of highly centralized repositories; instead caBIG would depend on creating the infrastructure to enable interoperability among biomedical information systems. The Institute of Electrical and Electronics Engineers (IEEE) *Standard Computer Dictionary* defines interoperability as the "ability of two or more systems or components to exchange information and to use the information that has been exchanged" [4]. Interoperability actually requires successful completion of two processes: First, the system (or its data) must be made accessible to another system and second the data must then be usable, that is, it must exist in a form that can be understood by the receiving system. These two elements are generally referred to as "syntactic interoperability" and "semantic interoperability," respectively, and both are required to enable interoperability. Two individuals who speak different languages are an example of a syntactically but not semantically interoperable system; information is exchanged but it cannot be used. Two systems that both encode the same questions using the same controlled terminology but exist behind completely closed firewalls are semantically but not syntactically interoperable; the information could be understood and used if only if it could be transported from one system to the other.

The infrastructure implemented by caBIG must therefore resolve the intertwined problems of syntactic and semantic interoperability. The strategy taken by caBIG was a semantic services-oriented architecture (sSOA). In a services-oriented architecture, individual systems support well-defined services that provide access to data and/or resources; a semantic SOA provides a mechanism to communicate semantic information in addition to the data itself.

An essential element to the construction of a semantic SOA is the use of data standards whenever appropriate. These can include standard controlled biomedical terminology (such as the NCI Thesaurus [5], LOINC [6] for laboratory tests, the Common Terminology Criteria—Adverse Events, or CTCAE [7] for adverse-event reporting, Gene Ontology [8]), data types [such as the National Organization for Standardization (ISO) 21090 [9]], common data element standards [such as the Clinical Data Acquisition Standards Harmonization, or CDASH, standards [10] promulgated by the Clinical Data Interchange Standards Consortium (CDISC)], standardized clinical case report form modules [see, e.g., the NCI's standard Case Report Form (CRF) modules], or common information models such as the HL7 reference information model (RIM) [11] or the BRIDG (Biomedical Research Integrated Domain Group) model [12]. The use of these standards allows for clear and easy aggregation of information collected by disparate groups, easier interoperability of information systems (see below), and easier design of clinical trials and other research activities. The preference within the caBIG program is to adopt existing standards rather than creating new ones. This preference is entirely pragmatic; the use of existing standards, where adequate, allows caBIG participants to reap the benefits of the substantial effort already undertaken by other groups and ensures that new data can be compared to existing data.

The initial technical solution was to leverage and extend two existing technologies, the cancer common ontologic representation environment (caCORE) [13] and emerging grid technologies developed as part of the Globus project [14]. The caCORE is a software toolkit that provides a number of the capabilities required to enable interoperability. It includes:

1. *Enterprise Vocabulary Services (EVS)* A series of tools for creating, managing, and delivering controlled biomedical terminology (and other terminological content) for use by electronic information systems [5]. EVS provides the underlying semantic content needed to understand information when exchanged.
2. *Cancer Data Standards Repository (caDSR)* A repository of metadata (often called data about data) based on an extension of the ISO 11179 metamodel. The caDSR and its tools are used to create a description of the information being transferred based on controlled biomedical terminology provided by EVS. The basic unit of metadata in the caDSR is called a Common Data Element (CDE) and it describes a single atom of information collected during a trial or recorded by an information system. A CDE has a formal semantic definition, a human-readable description, and a "value domain" that describes what constitutes a "valid" response in that field.
3. *caCORE Software Development Kit (caCORE SDK)* A set of tools for creating information systems that include application programming interfaces (APIs) that facilitate access to a system and the tools necessary to describe the information system in the form of caDSR CDEs based on controlled biomedical terminology.

The caCORE technology provided for basic levels of interoperability but was somewhat technology dependent (e.g., all APIs* were implemented in Java) and it lacked both a means to identify which systems had data that were relevant to a particular biomedical question and a robust, federated security infrastructure.

The solution to these problems was provided by an extension of globus [14] based grid technology. Grid technology provided three major benefits to caBIG. First, it allowed for the creation of a technology-neutral adapter layer that could be used to provide access to APIs regardless of the technical implementation of the underlying system. Second, it provided for discovery and advertising of services. Finally, grid technology contained the seeds of a powerful, federated security environment that could be extended to meet the needs of caBIG. The specific implementation of grid technology used by caBIG is called caGrid [15], and it is shown schematically in Figure 17.2. Conceptually, caGrid can be subdivided into five functional areas: data and analytical services, metadata services, higher order or workflow services, security services, and client applications. Data and analytical services contain the information and tools that are provided by and made available to the cancer research community. Higher order/workflow services such as the federated query processor (that mediates queries into multiple systems) and workflow engines [caBIG utilizes both Business Process Execution Language (BPEL) and the Taverna [16] workflow engine] enable service marshaling to perform complex tasks beyond the scope of any single system. Metadata services provide access to information that describes the systems or the data contained in those systems.

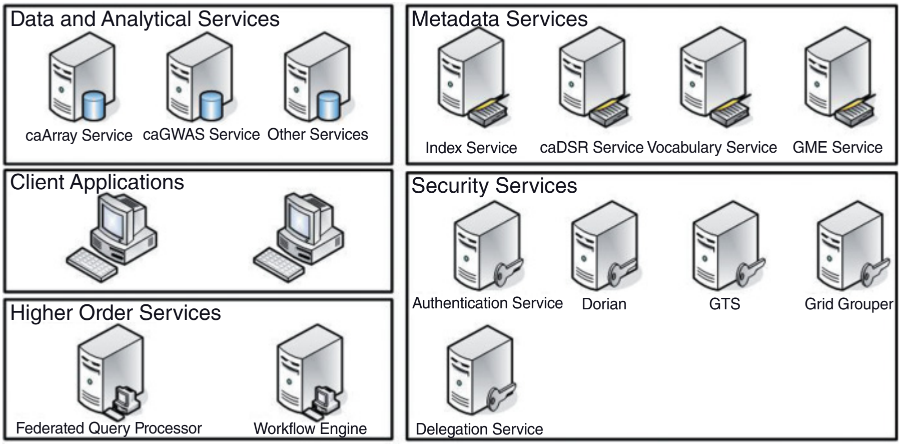


Figure 17.2 Schematic representation of services in caGrid.

*An API is software that enables other programmers to access data or capabilities that exist within a computer system. The presence of public (i.e., described and accessible) APIs greatly facilitates integration of software systems and prevents "vendor lock."

The Index Service provides a listing of available services, including representations of the information contained in that system, as well as status updates on that system. The Global Model Exchange (GME) supports the index service by providing access to the XML schemas of the services described in the index. The caDSR and vocabulary or EVS services provide access to CDEs and terminology that are used to code information in caGrid connected information systems as well as describe those systems themselves. Security services provide the necessary capabilities to support federated authentication and authorization as well as establish the chain of trust necessary to support collaboration among the cancer research community. The details of the caGrid security framework will be described in more detail in a later section. Finally, there are the client applications that most users will interact with to access information or resources on the grid. It is important to note that caGrid is a specialized set of services based on open standards running across the commodity internet; it is not a dedicated network.

In addition to the underlying technology, caBIG has created a series of applications that meet these caBIG interoperability requirements. All caBIG applications have several common characteristics. First, they provide mechanisms for an external entity to access the data or resources of the system, including at least one native, object-oriented API (generally Java for caBIG funded projects) that is used to support a Globus-based grid service (WSRF) and a WS-I-compliant Web service. REST* and other APIs as well as user interfaces are often available as well. Second, a formal information model [in the Unified Modeling Language (UML)] of the interface is provided to enable easier use and integration. Finally (and perhaps most importantly) the information model (and hence the interface) is fully described by a set of CDEs defined by the caDSR extended ISO 11179 metamodel, providing the semantics that are required to understand the information that is contained within the system. As described earlier, all of these applications are released under a nonviral open-source license, available for both commercial and noncommercial use as well as open- or closed-source derivative works. These applications have been deployed at a number of sites across the United States and beyond (see later); in addition, the NCI usually provides a hosted instance for use by community members that do not have the resources or the desire to maintain their own servers.

In order to facilitate adoption and use, caBIG has created two sets of application bundles, the Life Sciences Distribution (LSD) and the caBIG Clinical Trials Suite (CCTS), to support the needs of large communities. The LSD [17] is composed of eight applications [16] that support basic biomedical research:

1. *caArray* Microarray repository [18].
2. *caTissue* Biospecimen management system [19].

* REST (short for representational state transfer) is a lightweight interface protocol based on the Hypertext Transfer Protocol (HTTP) used by Web browsers.

3. *caGWAS* Data mart for Genome Wide Association Studies [20].
4. *CTODS* Clinical Trials Object Database, a system for providing deidentified data from clinical trials [21].
5. *National Biomedical Imaging Archive (NBIA)* Repository for Digital Imaging and Communication in Medicine (DICOM) images [22].
6. *caIntegrator 2* A tool for setting up custom caBIG-compatible Web portals for integrative research [23].
7. *geWorkbench* caGrid-enabled tools for performing genomic analysis [24].
8. *caB2B* a caGrid-enabled tool for performing in silico experiments by leveraging caBIG-compatible data repositories and analytical tools [25].

The CCTS [26] provides interoperable software to support clinical trials and consists of the following applications:

1. *C3PR* caBIG Clinical Participant Repository, a patient registration tool [27].
2. *PSC* Patient Study Calendar, a tool for scheduling patients on clinical trials [28].
3. *caBIG Clinical Connector* Software to assist in the integration with a clinical data management system. Currently C3D and OpenClinica are supported [29].
4. *caBIG Adverse Event Reporting System (caAERS)* A tool for managing and reporting adverse events during clinical trials [30].
5. *caBIG Lab Viewer* Laboratory results viewer that integrates with the rest of the clinical trials suite [31].
6. *caBIG Integration Hub (formerly caXchange)* An open-source, enterprise service bus to enable seamless integration of CCTS components with each other and existing systems at hospitals and research centers [32].

As with all caBIG software, the LSD and CCTS provide semantically annotated interfaces to enable interoperability with other systems.

In the first generation of caBIG systems, the services provided to the SOA tended to be large and complicated. Indeed, it was not uncommon to have a single API specification (implemented as described above in multiple technologies such as Java, Web Service, and Grid Service) that provided access to all of the data and resources that were made available by that system. While this greatly facilitated integration compared to a more traditional system, it became clear that a different level of granularity would make it simpler to support service marshaling and other integration activities. In addition, the model of one system producing one service tended to introduce undesirable variation in the way that particular classes of data were represented. For example, both a microarray repository and a biospecimen repository need to represent a source tissue, but since the two service specifications covered the entirety of the systems (and hence would not be the same), there were not appropriate incentives to standardize on the representation of a part of the model.

To resolve this problem, the next generation of caBIG systems are being designed to utilize services that are defined at a much different granularity to improve reuse and promote working interoperability between systems. This is accomplished by requiring that systems demonstrate compliance with specific conformance statements that exist within the service specification itself, rather than simply demonstrating that the system provides an arbitrary semantically defined service. The services themselves are defined through a formal enterprise architecture process based on the reference model for open distributed processing (RM-ODP) and the specifications are defined by the NCI's enterprise conformance and compliance framework (ECCF). The ECCF is an implementation of the HL7 services aware interoperability framework (SAIF) and provides three levels of specification: a conceptual model that describes the function of the service and the types of data that it will utilize/provide, a platform-independent model that includes the domain analysis model, and an implementable platform-specific model that is defined for a particular technology binding. By having three levels of specification, it is possible for groups to interact at different levels (using the same technology binding, domain model, or conceptual model) with a clear understanding of the level of effort required to enable those systems to interoperate.

Based on the requirements of the community, caBIG has devised an initial catalog of needed services which will be implemented for use in the next generation of systems. This "periodic table of services" is shown in Figure 17.3.

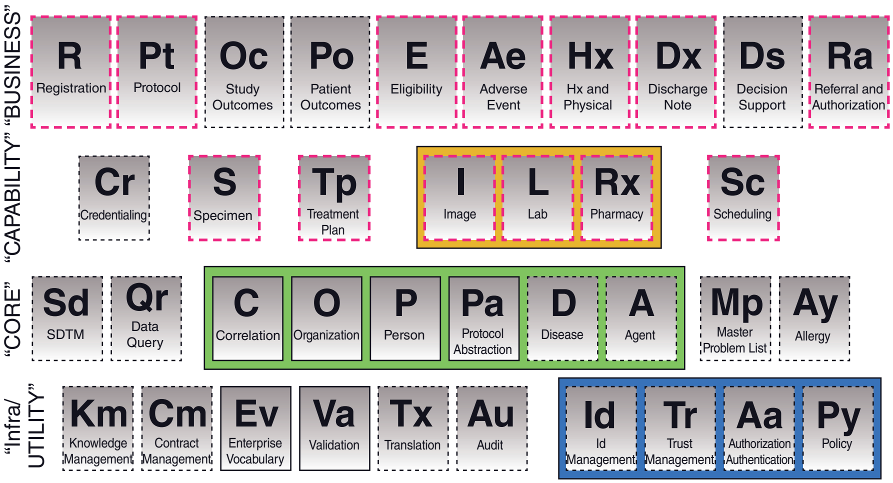


Figure 17.3 Periodic table of services to support oncology research and care.

caBIG services are classified into four primary types: Infrastructure/utility services are those that are required or utilized by virtually all other services and include semantic services, identity management, security services, and audit. Core services provide key information components that are utilized in higher level business capability and business/process services. Examples are the "COPPA" services that support protocol abstractions (PA), persons (P), organizations (O), and correlations (C) and services such as diseases (D) and agents (A). Business capability services provide "business atoms," data obtained from core and other services that are utilized by business processes; examples of these business atoms are models for specimens (S), treatment plans (Tp), and schedules (Sc). Finally business/process services provide arbitrarily complex capabilities that utilize the other three service types to carry out business functions such as registration (R), outcomes (Po), eligibility (E), and adverse events (Ae). This list of services is currently under development and always evolving; the current set of candidate services and their specifications as well as API specifications for those services that already have reference implementations (eventually created for all caBIG services) are available from the NCI Services Wiki at https://wiki.nci.nih.gov/display/EAWiki/Candidate+NCI+Enterprise+Services.

As alluded to above, there are a number of usage patterns that can be utilized with caBIG services. At one end of the spectrum an entity may choose to implement a service specification using their own technology bindings, interoperating with the caBIG implementation at either the platform-independent or the platform-dependent level (depending on the technology chosen). Alternatively, the entity could utilize the reference implementations that are provided for all caBIG services, integrating that software into their own system and allowing for interoperability at the platform-specific level. Most reference implementations of caBIG services provide Java, Web services and grid services APIs; some also provide support for REST and/or PERL. Finally, the NCI hosts instances of these services that can be used by groups that do not wish to deploy their own infrastructure.

17.5 caBIG COLLABORATION STRATEGY: SECURITY

Ensuring the security of information (particularly health information) is an essential part of facilitating collaborations. As a result, security and privacy have been major concerns of the caBIG program since its inception. As with all other activities within caBIG, a two-pronged strategy that involves both technology and sociology was selected to address these concerns. The social aspects of security and privacy are handled by the DSIC workspace while the ARCH workspace actually creates the security technology.

The technical implementation of caGrid security is the grid authentication and authorization with reliably distributed services (GAARDS) framework. It is shown schematically in Figure 17.2. GAARDS is composed of five primary services. Authentication is handled by an authentication service that provides a consistent front end to a variety of technologies such as Lightweight Directory Access Protocol (LDAP) or Shibboleth (http://shibboleth.internet2.edu/) and Dorian, a caBIG service that accepts signed Security Access Markup Language (SAML) assertions from the authentication service and provides X.509 proxies (a type of digital certificate) for invoking secure services. The grid trust service (GTS) mediates trust by verifying that SAML assertions or X.509 certificates are from trusted sources, while the credential delegation service (CDS) provides a means to delegate credentials during workflows. Authorization is delegated to the receiving service, which can utilize the caBIG developed grid grouper to assign roles and attributes to credentials that are requesting access. The general flow of accessing a secured grid service is shown in Figure 17.4.

On the policy side, the DSIC workspace and its associated knowledge center (see below) provide leadership to the caBIG community. The primary product of the DSIC workspace is the caBIG data sharing and security framework (DSSF), a collection of policies, procedures, and model agreements that can be used to support data sharing [33]. The central element of the DSSF is the DSSF decision support tree that is used to help classify the level of sensitivity of data. Supporting the primary DSSF components are a series of decision support tools for human research, privacy, contract terms, and intellectual property issues—a model informed consent document and other associated

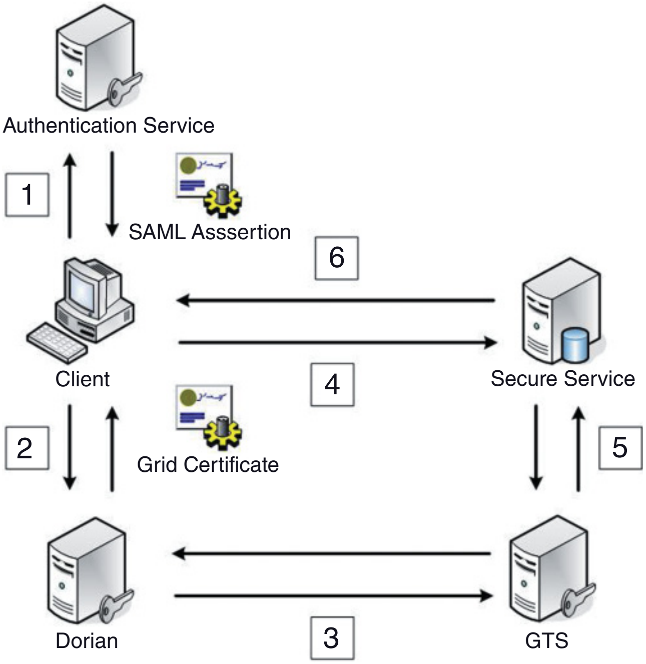


Figure 17.4 Invoking a secure service in caGrid.

materials. On the security front, the DSIC workspace has developed the security policies for caGrid. These are described in two works, a policy manual ("caBIG Security Program Policy," commonly called the "thin book") and an implementation guide based on the policy document ("caBIG caGrid Toolkit," or the "thick book"). The DSIC workspace works closely with the architecture workspace and NCI security staff to ensure that caBIG policy and technology remain aligned and that both remain aligned with federal law and best practice in security and privacy.

17.6 caBIG COLLABORATION STRATEGY: SUPPORT

The NCI recognized early in the program that successful use of caBIG technology and processes to enable collaboration would require a robust support infrastructure that could work effectively with the cancer research community. Part of that support mechanism is provided by caBIG program activities (teleconferences, face-to-face meetings, etc.) but other types of support were clearly required. Even prior to the launch of the caBIG program, NCI CBIIT has maintained an application support helpdesk to assist users.* However, to meet the much greater needs of the caBIG community, the NCI created the caBIG Enterprise Support Network. Comprised of a group of knowledge centers and support service providers, it is designed to provide enterprise-level support to a diverse collection of caBIG stakeholders.

Knowledge centers (KCs) are the primary mechanism for obtaining information about caBIG program activities, tools, and processes. Funded by the NCI to provide this support, there are six KCs that cover specific parts of caBIG program activities (Table 17.1). Each KC is responsible for providing phone and e-mail support for products in its area of capability and for performing routine bug fixes on that software. They are also sources of information to the community in their areas of expertise, maintaining listservs and wikis for those tools. The NCI funds the KCs so that some level of support is available to anyone in the community. However, KCs generally operate during normal business hours, and while they provide information as quickly as possible, they do not have formal service-level agreements with various elements of the community. Further, there are support activities that it would be inappropriate to use federal tax dollars to support, such as highly specialized integration activities that would only benefit the specific site where the integration work is taking place. To address these issues, caBIG created the support service provider (SSP) program. The SSPs perform work for clients under business arrangements (generally fee for service) made between the SSP and the client that requires the service. To become an SSP, an organization must demonstrate that it meets a minimum set of standards in one or more of four

* CBIIT support can be reached at 1-301-451-4383 or toll free at 1-888-478-4423 (U.S. only) or by email at ncicb@pop.nci.nih.gov.

TABLE 17.1 caBIG Supported Knowledge Centers and Their Host Institutions

Knowledge Center	Host Institution
Clinical Trials Management Systems	Duke University Comprehensive Cancer Center (prime) Robert H. Lurie Comprehensive Cancer Center (Northwestern) Cancer and Leukemia Group B Semantic Bits, Inc.
Molecular Analysis Tools	Columbia Herbert Irving Comprehensive Cancer Center (prime) Broad Institute of Harvard and MIT
Tissue Bank and Pathology Tools	Siteman Cancer Center, Washington University at St. Louis
Vocabulary	Mayo Clinic
caGrid	Ohio State University and OSU Comprehensive Cancer Center (prime) University of Chicago Argonne National Laboratory
Data Sharing and Intellectual Capital	University of Michigan

capability areas: help desk support, adaptation and enhancement of caBIG-compatible software applications, deployment support for caBIG software applications, and documentation and training materials and services. Organizations that meet these standards are eligible to receive a license that allows them to use the caBIG trademark to indicate their status as an SSP in marketing and other communications. The NCI does not support SSPs; all work performed as caBIG SSPs is paid for by other organizations that have decided to utilize the SSP's offerings. New applications for caBIG SSP are accepted on a regular basis. More information on the SSP program is available at the caBIG website: https://cabig.nci.nih.gov/esn/service_providers.

17.7 CANCER CENTER DEPLOYMENT

Ultimately, caBIG technology is most useful when it is deployed broadly across the cancer research enterprise. Although caBIG technology is open source (and hence free from software licensing costs) and while many caBIG applications can be deployed on laptops, it is designed to be enterprise software. When deployed to support an enterprise, it requires standard types of IT infrastructure (servers, connectivity, administrative staff, etc.) to function. Further, a decision to deploy and support enterprise systems should be coupled with an enterprisewide review of biomedical informatics needs to ensure that scarce resources are appropriately utilized.

To support this goal, the NCI Cancer Centers Branch, the National Community Cancer Centers Program (NCCCP), and caBIG partnered to create the caBIG cancer center deployment program. NCI-designated cancer centers that were interested could apply to receive a supplement to their primary cancer center grant equivalent to half of a full-time equivalent to support biomedical informatics deployments at their sites. To qualify, the sites would need to provide a matching level of support, designate a full-time staff member to lead the deployment [this person was generally referred to as a center deployment lead (CDL)], work with caBIG staff to create a strategy for biomedical informatics, deploy at least one application to caGrid, and begin the process of implementing that strategy by the conclusion of the first year of funding. Funding was renewable twice, for a total of three years. An important point about the deployment is that this activity was meant to support biomedical informatics, not caBIG per se. Obviously, caBIG software was available to the deployment sites, but sites did not have to use that caBIG software, even for the system that was deployed to caGrid. Sites could choose to utilize caBIG tools if it met their needs (adopt a tool) or they could choose to use non-caBIG technology to connect to caGrid by implementing APIs with semantics (adapt a tool) either by hand or using existing toolkits such as the caCORE SDK (see earlier). All told, 68 NCI-designated cancer centers (59 in year 1, 6 in year two, and 3 in year 3) and all 10 NCCCP organizations (that did not receive a separate supplement) decided to participate in the deployment.

Deploying sites went through a three-part process to assist with the creation of their strategic plan for biomedical informatics. In the first phase, the sites completed an IT readiness self-assessment. This assessment was designed to help assess their capabilities to deploy enterprise systems, develop custom extensions to software, and support the tools that they did deploy as well as assess their existing capabilities to support biomedical informatics. This document was evaluated by caBIG staff members who would then be able to provide advice to centers during the later steps of the deployment process. Next, each site produced a goals document that covered the types of capabilities desired, including a set of goals for data sharing. This document (which needed to be signed by the cancer center director) was used as the basis for the final component, an implementation plan for biomedical informatics (also signed by the cancer center director).

To ensure that the deployment proceeded smoothly, caBIG provided a variety of resources to assist sites that were carrying out the deployment. First, a group of staff members was provided to manage the deployment, providing a clear point of access to caBIG senior leadership, program staff, software developers, and deployment experts. During the first year (when the need was considered the greatest), a team of deployment specialists was maintained to provide technical assistance to cancer centers and to transfer knowledge to those sites. The caBIG knowledge centers (and other parts of the enterprise support network) were also marshaled to support the deployment. Finally, a series of center deployment lead-centric activities were scheduled to help nucleate a community of practice. These included a monthly teleconference with caBIG program staff, software developers, data sharing, and intellectual capital experts, and so on, gatherings at the caBIG annual meeting and ultimately a separate CDL face-to-face meeting. In addition, a series of CDL initiated and led working groups began meeting on topics of importance to the cancer centers. Ultimately, the purpose of these activities was to foster interaction among the CDLs, ensuring that sites could learn from each other's successes and problems and to drive the next generation of caBIG activities and software.

17.8 INTERNATIONAL EFFORTS AND BIG HEALTH

While the bulk of caBIG activity occurs within the United States, caBIG is reaching out to a variety of organizations with the goal of creating a global collaborative infrastructure. The longest running of these collaborations is with the National Cancer Research Institute (NCRI) in the United Kingdom. The NCRI and caBIG collaborate on technical infrastructure (ONIX, the NCRI infrastructure platform is technically compatible with caGrid) and joint interoperability projects as well as support a joint annual meeting that alternates between Washington, D.C., and London. In addition to the NCRI, caBIG has launched collaborative activities in Jordan (the new King Hussein Cancer Center in Amman will be utilizing caBIG technology), Pakistan (a new clinical trial unit at the Aga Khan University has adopted the caBIG Clinical Trials Suite), India (where a caBIG SSP already exists), Latin America, and China. The caBIG program plans to continue to develop these international collaborations with the goal of supporting a globalized research infrastructure.

Although the mission of the NCI (which sponsored caBIG) is to reduce the burden of cancer, nothing about caBIG (save perhaps the data in its repositories) is fundamentally specific to cancer. Indeed, virtually every piece of software, every underlying capability (terminology, common data elements), can be utilized to support biomedical research for other diseases. The Cardiovascular Research Grid (CVRG) supported by the National Heart Lung and Blood Institute uses the same infrastructure as caBIG, and caGrid (and some of its supporting tools) are in use by Clinical and Translational Science Award (CTSA) sites. With this in mind, NCI CBIIT chose to launch a new activity, the BIG Health Consortium, dedicated to implementing caBIG collaboration beyond the realm of cancer.

The BIG Health Consortium is a coalition of organizations dedicated to using next-generation information technology (so-called Web 2.0 capabilities) to implement personalized medicine. In contrast to caBIG, which is actively managed by the NCI, the BIG Health Consortium is convened by the NCI, where it is one equal partner among many. The NCI's contribution to BIG Health is caBIG and its associated technology as well as the BIG Health Enterprise Architecture Specification (BIG HEAS), which defines interoperability requirements for the consortium. Within BIG Health, projects are initiated, funded, and managed by champions belonging to one or more partners. By working through the BIG Health Consortium (and accepting BIG HEAS), a partner has access to consortium members and specialized capabilities provided by those members. In particular, they have access to various project action groups that provide a mechanism for other partners to self-identify their willingness to support or participate in a project and they have access to the IT leadership group that provides assistance in developing the IT architecture for the project.

Several BIG Health projects are already underway, bringing novel capabilities to biomedical research. These include the Health of Women (HOW) project, an online cohort study of unprecedented scope, and project Athena, another long-term study that has the potential to revolutionize the treatment of breast cancer in the United States. The HOW study is being spearheaded by the Dr. Susan Love Research Foundation (DSLRF) in cooperation with the City of Hope, HealthCare IT, and NCI CBIIT. Leveraging the DSLRF Army of Women and caBIG developed capabilities, the HOW study was able to obtain the first report on almost 30,000 women in a single month. Over the next several years, these women will continue to receive questionnaires about their health and lifestyle choices, while additional capabilities to capture images and genetic data will be integrated from the existing caBIG toolkit. Project Athena will similarly follow a cohort of women, in this case women who wish to participate that receive regular breast care within the University of California system. Women who chose to join the cohort will have additional biospecimens collected for the purpose of genetic testing, and these results will be aggregated with long-term disease status, images, and so on, to provide improved risk and outcomes assessments for physicians.

17.9 CONCLUSION

The cancer Biomedical Informatics Grid (caBIG) has demonstrated that it is possible to bring together a large, diverse community of biomedical researchers using information technology. This achievement required several key elements: a shared value proposition (in this case translational research and a learning health care system), a neutral party that could act as an organizer (the NCI), a policy of openness, and support from leaders and researchers throughout the community. The caBIG experiment was considered quite radical when it was initiated in 2004, advancing the cutting edge in both technology and community building. The program and the cancer research community still seek to remain on the cutting edge: continuously improving its infrastructure and tools and constantly expanding the boundaries of the cancer research community. To support this, caBIG is working with colleagues at the American Society of Clinical Oncology, the NCCCP program, and vendors to define a specification and open-source reference implementation for an oncology extended electronic health record (caEHR) that meets caBIG interoperability specifications. With a broadly deployed caEHR, it should be possible to create a virtuous cycle in which research defines the next generation of care and care drives the next generation of research, using collaborations enabled by information technology to achieve the NCI's mission of reducing the burden of cancer in our society.

REFERENCES

1. Altekruse SF, Kosary CL, Krapcho M, Neyman N, Aminou R, Waldron W, et al. *SEER Cancer Statistics Review, 1975–2007*. Bethesda, MD: National Cancer Institute, 2010.
2. Ankathil R, Stephen J, Vasudevan DM, Kusumakumary P, Pillai GR, Nair MK. Prognostic significance of karyotype analysis in children with acute lymphoblastic leukemia. *Hematol Oncol* 1992;10:339–344.
3. Center for Biomedical Informatics and Information Technology (CBIIT) NCI. caBIG Guiding Principles. Available: http://cabig.cancer.gov/caBIGstory/principles/, 2010.
4. Staff, I.o.E.a.E.E. IEEE Computer Dictionary—Compilation of. 1990. p. 610.
5. Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu WL, Wright LW. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *J Biomed Inform* 2007;40:30–43.
6. McDonald CJ, et al. LOINC, a universal standard for identifying laboratory observations: A 5-year update. *Clin Chem* 2003;49:624–633.
7. National Cancer Institute NIoH, U.S. Department of Health and Human Services. *Common Terminology Criteria for Adverse Events (CTCAE)*, version 4.0. Washington, DC: National Cancer Institute, 2009.
8. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;25:25–29.
9. Standardization IOf. ISO/DIS 21090—Health Informatics—Harmonized data types for information interchange. Available: http://www.iso.org/iso/catalogue_detail.htm?csnumber=35646, 2010.
10. CDASH CDISC. http://www.cdisc.org/cdash, 2010.
11. 7 HL. HL7 Reference information model page. Available: http://www.hl7.org/Library/data-model/RIM/modelpage_mem.htm, 2010.
12. Fridsma DB, Evans J, Hastak S, Mead CN. The BRIDG project: A technical report. *J Am Med Inform Assoc* 2008;15:130–137.
13. Komatsoulis GA, Warzel DB, Hartel FW, Shanbhag K, Chilukuri R, Fragoso G, et al. caCORE version 3: Implementation of a model driven, service-oriented architecture for semantic interoperability. *J Biomed Inform* 2008;41:106–123.
14. Foster I. The Globus Project. Available: http://www.globus.org/, 2010.
15. Oster S, Langella S, Hastings S, Ervin D, Madduri R, Phillips J, et al. caGrid 1.0: An enterprise Grid infrastructure for biomedical research. *J Am Med Inform Assoc* 2008;15:138–149.

16. Hull D, Wolstencroft K, Stevens R, Goble C, Pocock MR, Li P, Oinn T. Taverna: A tool for building and running workflows of services. *Nucleic Acids Res* 2006;34: W729–732.

17. Klemm J, Basu A, Fore I, Floratos A, Komatsoulis G. The caBIG® life sciences distribution. In Ochs MF, Casagrande JT, Davuluri RV, Eds. *Biomedical Informatics for Cancer Research*. New York: Springer, 2010, pp. 253–266.

18. National Cancer Institute NIoH, U.S. Department of Health and Human Services. caArray. Available: https://cabig.nci.nih.gov/tools/caArray, 2010.

19. National Cancer Institute NIoH, U.S. Department of Health and Human Services. caTissue. Available: https://cabig.nci.nih.gov/tools/catissuesuite, 2010.

20. National Cancer Institute NIoH, U.S. Department of Health and Human Services. caGWAS. caBIG. Available: https://cabig.nci.nih.gov/tools/caGWAS, 2010.

21. National Cancer Institute NIoH, U.S. Department of Health and Human Services. CTODS. Available: https://cabig.nci.nih.gov/tools/CTODS, 2010.

22. National Cancer Institute NIoH, U.S. Department of Health and Human Services. NBIA. caBIG. Available: https://cabig.nci.nih.gov/tools/NCIA, 2010.

23. National Cancer Institute NIoH, U.S. Department of Health and Human Services. caIntegrator2. Available: https://cabig.nci.nih.gov/tools/caIntegrator2, 2010.

24. National Cancer Institute NIoH, U.S. Department of Health and Human Services. geWorkbench. Available: https://cabig.nci.nih.gov/tools/geWorkbench, 2010.

25. National Cancer Institute NIoH, U.S. Department of Health and Human Services. caB2B. Available: https://cabig.nci.nih.gov/tools/caB2B, 2010.

26. Speakman J. The caBIG® clinical trials suite. In Ochs MF, Casagrande JT, Davuluri RV, Eds. *Biomedical Informatics for Cancer Research*. New York: Springer, 2010, pp. 203–214.

27. National Cancer Institute NIoH, U.S. Department of Health and Human Services. C3PR. Available: https://cabig.nci.nih.gov/tools/c3pr, 2010.

28. National Cancer Institute NIoH, U.S. Department of Health and Human Services. PSC. Available: https://cabig.nci.nih.gov/tools/PatientStudyCalendar, 2010.

29. National Cancer Institute NIoH, U.S. Department of Health and Human Services. caBIG Clinical Connector. Available: https://cabig-kc.nci.nih.gov/CTMS/KC/index.php/Clinical_Connector_Documentation, 2010.

30. National Cancer Institute NIoH, U.S. Department of Health and Human Services. caAERS. Available: https://cabig.nci.nih.gov/tools/caAERS, 2010.

31. National Cancer Institute NIoH, U.S. Department of Health and Human Services. Lab Viewer. Available: https://cabig.nci.nih.gov/tools/LabViewer, 2010.

32. National Cancer Institute NIoH, U.S. Department of Health and Human Services. caXchange. Available: https://cabig.nci.nih.gov/tools/caBIGIntegrationHub, 2010.

33. National Cancer Institute NIoH, U.S. Department of Health and Human Services. Data sharing and security framework. DSIC Wiki. Available: https://cabig-kc.nci.nih.gov/DSIC/KC/index.php/Data_Sharing_and_Security_Framework, 2010.

18

LEVERAGING INFORMATION TECHNOLOGY FOR COLLABORATION IN CLINICAL TRIALS

O. K. Baek

18.1 INTRODUCTION

In this chpater the technologies that enable clinical collaboration will be discussed. In the connected health future, sophisticated technologies will be available to all clinicians through connected collaboration and use of electronic clinical information, such as health records. Meanwhile, today, in order for information technology (IT) to drive health care transformation, adoption

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.

of IT by office-based physicians is essential for the delivery of care in the long term.

The topics covered in this chapter are clinical trials in general, challenges and areas for improvement for clinical trials, and technologies to be leveraged to address the challenges.

The chapter starts with an overview of clinical trials and key challenges for successful trials based on the publically available information through the Internet, specifically the public websites for the U.S. Food and Drug Administration (http://www.fda.gov/), the U.S. National Institutes of Health (http://health.nih.gov/), and the UK National Health Services (http://www.nhs.uk/), among others.

Then, the concept of online collaboration approaches and mechanisms will be discussed, such as social computing and virtual workplace and benefits with leverage of IT and the global network infrastructure, the World Wide Web, and the Internet.

Finally, there is a discussion of how some of the selected technologies, such as weblogs, virtual workplace, and secure e-mail, can help address the challenges for higher efficacy and efficiency through optimization of the processes associated with clinical trials.

18.2 WHAT IS A CLINICAL TRIAL?

Clinical trials are research studies that involve patients or healthy people and are designed to test new treatments. The new treatments cover a wide range of health care approaches, including drugs, vaccines, other approaches to disease prevention, surgery, radiotherapy, physical and psychological therapies, educational programs, and methods of diagnosing disease.

The U.S. Food and Drug Administration defines clinical trials as a research study to answer specific questions about vaccines or new therapies or new ways of using known treatments. The U.S. National Institutes of Health defines clinical trials as a research study in human volunteers to answer specific health questions. The UK National Health Services [1] defines clinical trials as research studies that involve patients or healthy people and are designed to test new treatments.

In this context, treatments refer to a wide range of health care approaches that can be tested in a clinical trial, including drugs, vaccines, other approaches to disease prevention, surgery, radiotherapy, physical and psychological therapies, educational programs, and methods of diagnosing disease. Interventional trials determine whether experimental treatments or new ways of using known therapies are safe and effective under controlled environments. Observational trials address health issues in large groups of people or populations in natural settings.

Each institution defines clinical trials in a slightly different way, but the main objective of a clinical trial is the same in the sense that it is to test if a new or different treatment is safe and to evaluate how well the new or different treatment works. Clinical trials cover a broad range of different types of research. For example, they are used to test new drugs, including vaccines, but also to look at new combinations of existing medicines. In addition, they are used to test if administering a treatment in a different way would improve its efficacy or reduce any side effects. And, in a nutshell, clinical trials, which are also referred to as medical research studies, are used to determine whether new drugs or treatments are both safe and effective and to find the best way to prevent disease and reduce the number of people who become ill; treat illness to improve survival or increase the number of people cured; improve the quality of life for people living with illness, including reducing symptoms of disease or the side effects of other treatments; and diagnose diseases and health problems.

Carefully conducted clinical trials are the fastest and safest way to find treatments that work in people. Trials are in four phases [2], where “phases” of clinical trials refer to the different stages of clinical trials:

- Phase I tests a new drug or treatment in a small group and is mainly aimed at finding out how safe a drug is.
- Phase II expands the study to a larger group of people and is aimed at measuring the safety and side effects and also to see if the drug has a positive effect in patients.
- Phase III expands the study to an even larger group of people and is mainly aimed at comparing the effects of a new drug with the existing treatment, finding out how well the drug works and how long the effects last, and finding out how common and serious any side effects or risks are and about any potential long-term effects.
- Phase IV takes place after the drug or treatment has been licensed and marketed and is mainly aimed at finding out how well the drug works when it is used widely as well as the long-term benefits and risks and potential rare side effects.

18.3 KEY CHALLENGES OF CLINICAL TRIALS

Key challenges of clinical trials are patient recruitment and retention because clinical trials involve a large number of people for a long time to get the results, timely access to accurate data with assured confidentiality, and statutory or regulatory compliance, especially for protection of personal privacy.

One of the critical success factors (CSFs) of clinical trials is recruitment of a large number (generally thousands) of people to take part, because the difference between the effects of different treatments is quite often small and also the effects are heavily dependent upon the patients’ characteristics, such as hereditary traits, dietary habits, environmental parameters, life styles (e.g., smoking, drinking, exercise), and socioeconomic status.

In addition, it takes a long time to get the results of particular trials due to the lead time to recruit the trial participants, it often involves treatment over a long period of time, and the lead time to monitor and gather reliable data to analyze the long-term effects of a treatment. Patient recruitment and retention in clinical trials are widely recognized as the bottleneck in the new drug development pipeline and is essential to conducting successful trials because adequate enrollment provides a base for projected participation retention, resulting in evaluative patient data.

At the same time, monitoring daily activities and collecting reliable and accurate information and maintaining confidentiality are important. Much of the information collected is sensitive and personal data, and therefore the trial participants need an assurance that the data collected will neither be accessed by unauthorized people nor be misused outside the original intended purposes.

More importantly, the trial process needs to adhere to the statutory regulations for protection of personal privacy such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States or the European Union (EU) Privacy Directives in the EU member countries. Although each country has different regulations in the details for protection of personal privacy, the main concept is the same: Protect by law any information that can be used to identify a particular individual. Therefore, all personal information, such as name, address, telephone number, photo, e-mail ID, social security or insurance number, driver's license number, and bank account number, is legally protected by the federal statute.

Sharon Jameson [3] identified the following challenges in clinical trials:

1. Improper management of informed consent to comply with the statutory requirement for protection of personal privacy—protection of the human research subject is the first obligation by obtaining informed consent and thoroughly documenting the associated process.
2. Failure to prevent or filter out inaccurate or falsified data.
3. Failure to maintain adequate source documentation—source documentation tells the story of the patient's experience while on trial and provides proof of oversight.
4. Failure to comply with protocols—failure to follow the protocol can compromise patient safety and does compromise the integrity of the study.
5. Delinquent or inaccurate data submission—multiple queries and late data lead to no payment or delayed payment for the staff.

One of the most critical success factors of clinical trials can be summarized, as Genevieve Frank pointed out in her paper "Current Challenges in Clinical Trial Patient Recruitment and Enrollment" [1], as well-maintained open communication among the health care providers, clinical trial sponsors, the media, and the public to overcome real and perceived barriers to clinical research participation.

Advancement of technologies and wide adoption of the World Wide Web and the Internet provide many possibilities and opportunities for online education of prospective participants, online patient enrollment, remote visits, and data capture: for example, medical devices connected to personal computers (PCs) or mobile phones that can be used to transmit data to a central control point in near-real time and across time zones and geographical boundaries, Web-based self-assessment and reporting instruments, text messaging to remind patients of things they need to do and collecting data from patients, remote patient visits using PCs and Web-based collaboration tools that enable visual inspections and video communications.

18.4 TRANSLATIONAL RESEARCH

The term "translational" quite often refers to high-speed utilization of the latest high-throughput research, such as discovering new biomarkers in a rather narrow definition, but one would argue that it is in effect an extension of the principles of evidence-based medicine and involves passing from advanced medical research to clinical applications. The primary concept is to bridge the patient care domain and the clinical research domain to allow the medical researchers to leverage the clinical outcomes observed by the health care practitioners. It will also allow the health care practitioners to leverage the research outcomes derived by the medical researchers.

Through full leverage of IT for facilitating translational research, we can realize and enable the concept of the proactive, patient-centered, outcome-driven, wellness-oriented continuum of healthcare service.

Translational research can be enabled by leveraging IT and other related technologies such as artificial intelligence, robotics, cybernetics (e.g., machine learning and autonomous computing), and nanoscience (e.g., nanobots). Translational research is anticipated to help us realize a personalized health care in consideration of an individual's genotypic, phenotypic, environmental, and lifestyle variances and to help the health care professionals improve the accuracy of diagnosis, the efficacy of treatment, and patient safety (i.e., minimal adverse drug responses). Translational research may also help us realize a preventative or predictive health care through in silico disease modeling and organ simulation to complement experiments based on animal models. Other industries, such as the automobile and aircraft industries, as well as the U.S. National Aeronautics and Space Administration (NASA) have long been exploiting IT for modeling and simulation for design of automobiles, aircraft, and space shuttles, while the health care industry has relied on experiments based on animals and observation of only external symptoms.

The key issues in the health care industry include:

- Health care providers heavily rely on their personal experience and opinion as the main reference of diagnosis, treatment, and prognosis.
- The information or knowledge gained by the health care providers through years of practice and experience is not shared among the health care professionals who can benefit from it for more accurate diagnosis or for better treatment for higher efficacy (e.g., a survey conducted by the American Medical Association suggested that physicians in the United States would rather share their toothbrushes than share their patient information [4]).
- The health care services are rather fragmented with lack of continuum of services from diagnosis and treatment through prognosis and from primary care to tertiary care.
- The healthcare services are provided reactively as episodic events.
- All the patients are treated equally for diagnosis and prognosis with little consideration of their genotypic or phenotypic variances.
- Consequently patient safety is significantly compromised.

Translational research can be enabled by an integrated IT solution that enables interdisciplinary collaboration across multiple institutions, provides transparent access to multidimensional data and tools to analyze unstructured data as well as structured data, manages vast amount of multidimensional data, provides high-performance computing, and incorporates robotic instruments as part of the workflow management. This integrated IT solution refers to a set of integrated systems that draws upon relevant information and context to enhance the activity and performance of people, robotic instruments, systems, and organizations through an online collaboration across the boundaries of various disciplines, organizations, countries, and geographical locations.

The goal of the IT solution is to assist, while leveraging IT and related technologies, a group of multidisciplinary researchers in navigating the information spaces in "real" and "virtual" environments, orienting and guiding them based on the research themes, interacting with and leveraging others to find their way in the information spaces, and sharing discoveries and knowledge among them. The main focus is to foster trust and community in research teams, whether colocated or geographically distributed, and to support long-running, contextual interactions rather than short-term, task-focused activities, for a holistic management of collaboration among a group of multidisciplinary researchers.

The key characteristics of the IT solution to enable interdisciplinary collaboration across multiple institutions are as follows:

1. The data to be analyzed are mainly unstructured, vast, and yet scarce and noisy.
2. The users of the system, also referred to as "actors," include autonomous robotic systems as well as human users.

3. The internal users and the external collaborators need to cooperate as an integrated cohesive team, but the intellectual properties are yet to be protected.
4. The laboratory information systems for experiments and assays need to be integrated with the system.
5. The external reference systems and databanks need to be integrated with the system through a dynamic data filtering, ingesting, and transformation for transparent access to the relevant data.
6. A virtual workplace needs to be provided for knowledge sharing and online collaboration across the organizational boundaries.
7. Semantic search and undirected data mining against multidimensional data are essential analysis tools.
8. Very high performance computing resources, typically MPP (massively parallel processing) clusters, in the range of trillions of floating point operations per second (teraflops) need to be provided as shareable services.
9. A very high throughput content/storage management system needs to be provided for policy-driven migration and recall of vast amounts of data in the range of quadrillion bytes (petabytes).

The key enabling technologies include the following:

- Electronic health records to be shared among the health care professionals and the medical researchers
- Electronic data capture (EDC) system to be used by the health care professionals at the sources of data
- Standard ontology and taxonomy for data representation and exchange
- Global patient identifier to link fragmented patient data generated and collected at various sources (e.g., general physician's office, diagnostic laboratories, private clinical offices, hospitals, clinical trial laboratories, research laboratories)
- Cross-enterprise vocabulary services to be used among multidisciplinary professionals
- Integrated security services and privacy framework to ensure data confidentiality and integrity and to comply with the statutory requirements for protection of personal privacy
- Undirected data-mining facility for discovery of associations and covariance among multidimensional heterogeneous data as well as directed data-mining facilities for validation of hypothesis
- Knowledge sharing and online collaboration facility among multidisciplinary professionals

Computer-assisted workflow management helps the researchers navigate through tools and data based on research scenarios, use cases, personal preferences, and usage patterns of individual researchers for usability and high productivity. Online collaboration and a knowledge sharing framework allow the researchers to interact with coresearchers through integrated e-mail, calendaring, instant messaging, and Web conferencing. The researchers will be able to insert annotations, research outcomes, and discover knowledge and share them with coresearchers even in a different scientific discipline for interdisciplinary research.

18.5 SOCIAL COMPUTING

An emerging computing model for society on the Web is referred to as "social computing" and can address the most critical success factors for clinical trials: well-maintained open communication among the health care providers, clinical trial sponsors, the media, and the public to overcome real and perceived barriers to clinical research participation.

According to Wikipedia (http://en.wikipedia.org), social computing is defined as "systems that support the gathering, representation, processing, use, and dissemination of information that is distributed across social collectivities such as teams, communities, organizations, and markets." The key objectives are to find like-minded people, develop trusted relationships, and share information (raw data, preprocessed data) and knowledge (result data, discoveries).

The current younger generation are believed to fully leverage IT for personal networking and communications. In fact, they depend heavily on technology (e.g., cell phone, iPod, digital camera, PC, calculator) for their daily life, such as text messaging with friends, listening to music, and writing reports for school assignments. From a technology perspective, as commodity hardware and software technologies reach the masses, computing power migrates to the edge of the network and the computing power of the hand-held devices becomes as powerful as that of desktop computers.

With the high-speed wireless network and powerful hand-held devices such as cell phones and personal digital assistants (PDAs), social computing expanded its playing field to the pervasive mobile infrastructure. Real-time text messaging became the primary means of communication among the younger generation.

If the health care providers, clinical trial sponsors, the media, and the public use real-time text messaging to exchange ideas and up-to-date information, some of the issues in clinical trials, such as overcoming real and perceived barriers to clinical research participation and collecting current and accurate data in a timely manner, can be fully addressed. Of course, security issues and compliance with privacy laws must be addressed in order to use text messaging as a means for communication and data collection in clinical trials.

Key functional building blocks of social computing framework include:

- Federated identity management for uniquely identifying people across organizational boundaries and geographical boundaries
- Location-based presence detection for discovering who is online or available and where and to determine primary means to contact
- Interactive multiway communication for real-time interactions for exchange of ideas and information in a global cross-institutional virtual space
- Online communities for forming and managing communities of interest in a global cross-institutional virtual space
- Security and privacy for assurance of confidentiality and protection of personal privacy so that a fully trusted virtual space may be established for the participants
- Context-based definition of relationships for describing how the individuals are related within a certain context and their roles
- Context-based semantic search and information sharing for finding people and information of interests and for sharing things that are meaningful and relevant to the participants

The first four capabilities have been, more or less, implemented and various solutions are currently available in the market.

The functional capability to establish a fully trusted virtual space across the public network infrastructure, the Internet or the cellular network, is a tall order and has many challenges. This topic is discussed in detail later in this chapter.

The last point on the above list may be one of the essential building blocks for the social computing framework and is often referred to as "social search" capability. Social search is not the traditional search based on key words to locate information but a context-based semantic search to find people of interests and relevant information based on semantics such as ontology. Here, the term "context" spans diverse domains, from geographic locations to social connections to behavior trajectories.

Google.com became one of the most successful (if not *the* most successful) Internet service providers in a very short time period due to its innovative search algorithm and information indexing mechanism. "Google" is used interchangeably with the word "search" among the younger generation. They use phrases like "Google it" to find information or "Google shows" for a map or direction to a certain place. Matt Cutts (www.mattcutts.com/blog/), head of Google's Webspam Team, asserts that new search tools will provide personalization, a completely new user interface, semantic understanding of queries and documents, social search for unlocking the power of people, and universal search for multidimensional and heterogeneous data.

Google.com is very widely and extensively used for finding things on the Internet. There is a saying that if Google cannot find something most likely it does not exist on the Internet. In addition to Google for search, there is another social computing tool very widely adopted and extensively used on the World Wide Web.

Weblogs, commonly referred to as "blogs," are burgeoning across the Internet as a means to improve social interactions, according to META Group (www.metagroup.com) articles published in recent years. "Blogging" became a pervasive means for improved information/expertise sharing, collaboration, and community building. Focusing on the connections of people to teams, communities, process, and information in evolving workplaces became a vital discipline for adaptive organizations trying to improve operational effectiveness and efficiency not only to reduce the cost for higher profit but also to drive competitive advantage.

Organizations have launched various corporate initiatives under the name "business transformation" to fully assess the implications of the Web and to capitalize on the value of the World Wide Web and the Internet. Many organizations are investing heavily on development of Web portals to provide a consolidated access point for data and tools and the backend content management for improved quality and integrity of corporate data as an attempt to establish a technology foundation for "social computing," a frequently cited buzz word these days in addition to SOA, or service-oriented architecture.

For years, organizations have deployed other tools such as e-mail, instant messaging, and discussion forums under various names such as blogs and wikis. Those tools are being integrated along with advanced search and cognitive user interfaces as the enablers of social computing. Blogs in particular have become incredibly popular in areas of social discourse and community building as well as many Internet portal, news, sports, entertainment, and technology sites as a means of disseminating information, generating commentary, and engaging a self-selecting audience. Blogs are used to post a diary or journal metaphor on the Web to convey information in conversational dialog. A simpler form of blogs, called twitter, is gaining its penetration. In twitter short messages are used for the same purpose, while narrative sentences are used in blogs.

As Google has become a new word for finding information on the Internet, blog has become a new word for exchanging information over the Internet. The act of posting ideas, opinions, or news on the Web is referred to as "blogging." While people can read blogs just as they do pages found on other websites, one capability that makes them different is that they can be subscribed to via a syndication technology called RSS, which stands for really simple syndication or rich site summary [5].

Blogs enable us to improve information sharing, community building, and collaboration and thus is helping us to socialize in a "virtual space", that is, over the network, although the quality of information in terms of currency, accuracy, and reliability is in serious question because anyone can blog anything as there is no control mechanism or governance process for checks and balances. Nevertheless, blogs are helping people to refine and adapt their cognitive model on various activities by observation and by making people aware of credible resources and consequently becoming the most efficient and widely adopted tool for social computing.

Allowing the participants of a clinical trial to blog among themselves to share information about how a new treatment works for them and what kind of side effects they experience may be a very effective and efficient means to collect the data in near-real time. The categorized blogs can be used for clinical trials: blogs used internally among the trial participants and sponsors; blogs open to the general public; blogs focused on a particular event; blogs used exclusively among trial sponsors; and blogs used by an individual.

Blogs can be offered as an ad hoc channel for trial participants to publish and be listened to only by those that care to subscribe to a specific topic or clinical trials of a treatment in this case. The trial sponsors and researchers can monitor how their trial participants share information, communicate, and collaborate. For example, blogs can be used for sharing laboratory notes, peer review commentary, and analysis of test cases for clinical trials.

18.6 VIRTUAL WORKPLACE

A virtual workplace can be used as a controlled and managed blog to enable multidisciplinary collaboration for effective and efficient clinical trials with a large group of participants in various geographies.

A virtual workplace is adaptable to many scientific disciplines and environments for collaboration over the network, obviating the need for face-to-face meetings requiring not only the extensive travel time that elongates the duration of the study but also the travel expenses that increase the cost of the study.

The requirements for a virtual workplace to support all activities relevant to a scientific research and development project, including clinical trials, can be summarized as:

- Real-time collaboration across scientific disciplines and institutional boundaries
- Knowledge sharing across scientific disciplines and institutional boundaries
- Intellectual property and rights management across institutional boundaries
- Workforce management through semantic search for locating resources based on expertise, projects, focus areas, organizations, geographical locations, and so on
- Integrated activity management to manage the end-to-end research and development process from raw data collection to knowledge discovery

A virtual workplace needs to be open, adaptable, and extensible. The solution components should be able to communicate via open standards to facilitate integration and collaboration with new entities such as for a rich set of collaboration and workflow management tools, scalability to handle a very large number of collaborators from many institutions, and interoperability with tools from various sources and various platforms.

The following functional attributes are essential for an effective virtual workplace:

- Integrated communication with various teams and subteams and support for asynchronous and synchronous communication methods such as secure e-mail, instant messaging with presence awareness, video and audio conferencing [e.g., Voice over Internet Protocol (VoIP)], calendaring, blogging, podcasting, RSS, wikis, and other communal postings, and online conference forums where many different groups participate and topics may be brainstormed, discussed, and expanded
- Teaming to support multiple but distinct collaboration communities and multiple teams within each community, address book and bios of all team members that are searchable for skills as well as phone numbers, integrated calendar with team activities and important conferences and meetings, bulletin board with notices and announcements, and informal virtual place for people to hang out and socialize
- Education for courses on use of the virtual workplace and collaboration tools and technical topics relevant to the collaborators to facilitate interdisciplinary interactions (e.g., distance learning offered by colleges and universities)
- Ease of use for interaction and access with intuitive look and feel, ideally requiring zero learning curve and data entry as easy as writing a paper notebook entry

18.7 SECURITY AND PRIVACY

Assurance of confidentiality and compliance with the statutory requirements for protection of personal privacy to establish a "trusted" virtual workplace is a critical success factor for leverage of collaboration in clinical trials

Security and privacy are main concerns of individuals and organizational entities whether they are commercial institutions, nonprofit organizations, or charity organizations. These issues become much more sensitive and critical, especially when personal health information is involved, as in clinical trials. The general public is very much concerned about potential identity theft and misuse of personal information by their employers and government agencies. In addition, there is a growing public mistrust of government.

People have a pretty good understanding of the meaning of *security* and the implications associated with compromised security but still are trying to comprehend the true meaning of *privacy* and the potential risks associated with noncompliance to the statutory requirements for protection of privacy. It is critical to distinguish privacy from security: Privacy is concerned with personal control of the collection, use, and disclosure of personal information, while security is concerned with control of access to assets (information and resources) that are used in a business context. Security is an important part of privacy as security is an essential building block for the implementation of privacy policies. If privacy of health-related information (e.g., the protected health information defined by the HIPAA of the United States) is compromised, for instance, we are talking about bankruptcy of an organization that is very healthy in all other aspects of business and about criminal prosecution of the C-level executives of that organization.

The biggest challenge with privacy compliance is limiting the use of personal information to the intended purposes stated originally at the time of collection. In addition, protection of individual identity becomes a bigger challenge in this information age when our daily life depends on digital information.

The traditional IT security model focuses on physically securing computers and protecting users from outsiders attempting to access computers and data in an organization. The premises behind the traditional IT security model are that selected people within an organization can be fully trusted, security threats are outside the organization, and masquerading computer systems is practically impossible. However, traditional assumptions are no longer valid due to technological changes, such as powerful portable computers and high-speed global computer networks. As a result, the traditional security model cannot ensure confidentiality and personal privacy or conform to the statutory requirements.

With the ubiquitous adoption of the Internet and the globalization of the marketplace, corporations around the world face a new challenge for protecting their assets. Corporations convert and maintain data in electronic format to be shared among the business units and business partners around the world to conduct business in the global marketplace with leverage of the Internet. Consequently, the majority of intellectual properties exist in digital format. Protection of corporate intellectual property is becoming a real challenge, especially because the people who are trusted and charged with safeguarding the corporate assets, such as IT managers, CIOs, and CTOs, are engaging in acts of digital espionage.

There are three aspects of security:

- Physical security (e.g., locks for buildings, badge access to secure rooms)
- Logical security (e.g., passwords for computers or networks, smart cards)
- Operational policies and procedures (e.g., oath of office, management approval)

Adequately protecting assets and assuring personal privacy require attention. Technology protection is just one aspect of security. Logical security plays a major role in ensuring that proper access and security policies are enforced. Physical security is as important as (if not more important than) logical security for biotech companies for protection of tangible assets such as invaluable samples and expensive instruments such as mass spectrometers.

Logical security must assure authentication, authorization, confidentiality, integrity, protection of privacy, nonrepudiation, and availability. One aspect of security alone does not assure the needed protection. All three aspects of security, that is, physical, logical, and operational security provisions, must be enforced. The e-business solutions need to provide integrated security infrastructure to address all the logical security services in conjunction with the applications that provide authorization services and privacy.

Identity management is a critical functionality to establish a trusted environment for collaboration in a virtual workplace. The IT term for this is *authentication for positive identification and validation* of an entity, an individual in this case. The traditional method of authentication based on what you know (e.g., password) is vulnerable because the shared secret can be exposed to unauthorized users. The new authentication mechanism based on what you have (e.g., seal, smartcard) has a challenge in distribution and revocation of the entities. This mechanism has been used for centuries. For example, personal seals are used in many countries for business transactions, and government and academia are using the institutional seals to issue certificates.

The method of identification and validation of an entity based on the unique physical attributes of individuals has been around for a long time. For example, thumbprints have been used in many countries for many years. What is new is that a similar method is being adopted for electronic transactions. The IT industry is moving toward a new mechanism of authentication based on what you are, or what you are born with, which is referred to as biometrics.

We use face recognition for identification of a person or voice recognition in our daily lives. We use signatures for business transactions. The U.S. INS (Immigration and Naturalization Services) uses palm prints for the border crossing at airports.

There are three approaches for validation of identity, which is referred to as authentication: (1) based on what they know (e.g., user id and password), (2) based on what they have (e.g., seal, badge, smartcard), and (3) based on what they are, (e.g., facial geometry, voice pattern, fingerprint). Signatures are commonly used for our business transactions for nonrepudiation of commercial or legal transactions. Seals are commonly used to certify official documents.

The new industry trend is to use a combined approach of smartcards activated with biometrics for leverage of the advantages of both technologies. For example, smartcards are used in some countries in place of health insurance cards, where the smartcard needs to be activated by the owner's fingerprint along with detection of body temperature.

It is critical for an institution to conform to the statutory requirements associated with protection of privacy, especially for customer/consumer profiling, which may deal with personal information. The potential for exposure of private information to unauthorized entities has increased dramatically. The potential liability a company faces by not taking appropriate measures to protect this information could bankrupt a successful business.

Several governments throughout the world and various industries are addressing the issue of privacy through legislation, standards efforts, and products. Most of the countries around the world, such as Australia, Belgium, Canada, France, Germany, Japan, New Zealand, South Africa, Sweden, United Kingdom, and The Netherlands, have federal statutes for protection of personal privacy and have established federal government agencies responsible for governance of compliance to protection of privacy.

Here is a short list of privacy laws around the world:

- Freedom of Information and Protection of Privacy Act (FIPPA) and Personal Information Protection and Electronic Documents Act (PIPEDA), also known as the Electronic Privacy Act, Canada
- Directive on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data, also known as the EU Privacy Directive, The European Union
- The U.S. HIPAA

The statutory requirements for protection of privacy differ country by country. However, there are common themes and principles. The statutory requirements, as common themes and principles, impose the following requirements for protection of personal privacy and these are relevant to data generated from clinical trials:

- Organizational entities are required to obtain an individual's informed consent before collecting or disclosing personal information in any medium.
- Individuals have legally guaranteed access to the personal information collected about them.
- Organizational entities are prohibited from using personal information for any purposes other than for which it was originally collected.

The U.S. Federal Trade Commission recently presented a report to Congress on "Privacy Online: Fair Information Practices in the Electronic Marketplace" [6]. The report states that consumer-oriented businesses would be required to comply with four widely accepted information practices—notice, choice, access, and security:

- Notice requires a disclosure of what information they collect, how they collect it, and how they use it.
- Choice requires consumers' ability to choose how information will be used beyond the purpose for which it was collected.

- Access requires that consumers have reasonable access to the information collected, including the opportunity to review, correct, or delete that information.
- Security requires performing reasonable steps to protect the security of information they collect.

The privacy laws require a data collection audit. Data collection, such as the names and addresses, is simply a part of doing business. Other information, such as demographic or personal financial data, is frequently collected to find out as much as possible about consumers and retailers. The law will force that consent be obtained before collecting, using, and/or disclosing such personal data.

The central obligation under the legislation is the need for data collectors to provide transparent privacy policies so the individuals are accurately informed about who is collecting their data, why it is being collected, and how it will be used. As such, organizations need to define a security policy and procedure to accurately inform individuals as to what data are being collected and how they will be used by providing mechanisms for:

- Obtaining an individual's informed consent before collecting or disclosing personal information
- Allowing individuals to access to information collected about them

In addition, the policies and procedures should be in place to prohibit the employees from using personal information for any purposes other than for which it was originally collected.

Notice that the consent is to be "specific and informed." If applied literally, for some secondary research this would require solicitation of more-focused consent than is now sought.

18.7.1 HIPAA Compliance

Compliance to the U.S. federal privacy law (HIPAA) means a legal assurance for protection of personal privacy in addition to traditional security requirements such as authentication, authorization, confidentiality, data privacy, non-repudiation, and auditability. The HIPAA prohibits anyone from collecting or disclosing an individual's personal information without explicitly stating the purpose(s) of collecting or disclosing the individual's personal information and acquiring an explicit informed consent from the individual.

Therefore, any documents containing protected personal information, specifically personal health information (PHI) in clinical trials, must be de-identified before they are shared with any unauthorized persons or when the documents are used other than the original authorized purposes. The term *de-identification* is a legal term for provision of an assurance of removal of any personal information and also other information unique to an individual that can be used to identify a particular person. As such, the medical information needs to be de-identified before leaving the hospital premises, so that it may not be reassociated with a particular individual. The challenge of de-identification arises when the personal information is passed outside the organizational premises. For example, a personal health record transferred from a medical center to a research center needs to be de-identified since the medical information contains personal information which is protected by the privacy laws. Currently, there are quite a few algorithms to solve certain parts of the de-identification problem, although most of the algorithms are proprietary and address only certain parts of the de-identification task on a given health standard with given de-identification rules.

18.8 CLINICAL E-MAIL SYSTEM

Good old e-mail will still play a role in the world of the new medicine, though privacy and security become even more important for protection of personal privacy and also nonrepudiation—an assured mechanism to prevent an individual from denying one's involvement in an activity, whether it is a commercial or legal transaction, and security audit become critical functions. In addition, the network infrastructure needs to provide sufficient bandwidth for an exchange of large volumes of data such as medical images.

The electronic messaging system for electronic mail exchange among health care professionals is referred to as a "clinical electronic mail" system. This implies that management of "good old e-mail" has to be enhanced for clinical use. When the electronic messaging or mail system is used among the health care professionals, there is a set of unique additional requirements. This is because health care professionals most likely exchange sensitive personal information and so necessarily "disclose" (hopefully to an accredited recipient) personal information. Disclosure of such information is of course protected by the statutory requirements for protection of privacy, such as the U.S. HIPAA, PIPEDA and FIPPA in Canada, and the EU Privacy Directive in Europe. That does not mean that you can not send e-mail about a patient, but it does mean that the appropriateness to do so is a matter which IT can manage.

For clinical trials involving participants across multiple organizations and multiple geographies, incompatible electronic mail systems may require a significant amount of time and effort to provide *gateway functions* for validation of credentials and transformation of messages among the disparate e-mail systems. These are literally the systems that allow the e-mail in and out and in theory at least let it in or out with censorship of parts of the e-mail. The primary function is to assess the originator of a message and its validity and integrity of the message.

For assurance of validity and integrity of messages, that is, to ensure that the messages were sent from the legitimate sources and that the messages are delivered to the intended destination without being tampered with, the following capabilities need to be provided for clinical e-mail systems:

- Confidence that a message really is coming from the purported source
- Confidence that it has not been altered in transit
- Ability to identify which population the message comes from (different population group may have different assumptions about security)
- Ability to identify the exact *individual* from whom the message comes
- Ability to identify each individual or institution by a unique identifier
- Point-and-click features so that a reference to an abnormal laboratory result, for example, can be linked to the relevant laboratory reporting system and record
- Support of digital signatures for nonrepudiation
- Cryptography that can prevent any attempts at intervention and eavesdropping
- Ability to mark messages with an indication of clinical importance or "level of emergency"
- Delivery confirmation for audit
- Automatic expiration after the date when they cease to be relevant
- Long-term archiving for future references

Special capabilities are required for provision of confidentiality and protection of privacy for the clinical e-mail messages or attachments. They include the ability to control access based on roles and identity; the ability to limit copy, print, and transfer, leaving an audit trail of everyone that has received and read sensitive information; and delegation and escalation. That last item is the ability to authorize a third party (say an assistant) to review a practitioner's mail. There should be the ability to link a clinical message to a specific patient, provider, and setting. When patients are referred to in messages, they must accompany some kind of a link to unique identifier or some other strategy for ensuring that the communication gets filed correctly. Various types of images, video clips, audio clips, and other attachments must be supported. There must be automatic sorting by patients to see all communications about an individual, by a practitioner to see all consults from a particular practitioner, by a provider (e.g., hospital) to see all communications pertaining to a particular practice environment, by priority, and by expiration date. There should be an auto-archiving feature. There should be automatic filing of messages by patient, or some other attribute, including automatic forwarding of messages to an electronic medical record system (after de-identification, of course!). There should also be clipboard support, that is, the ability to quickly copy materials from a message into a medical record, for example.

In addition to the special properties of the clinical e-mail system required for secure exchanges of health-related messages and documents with confidentiality, integrity, privacy (e.g., informed consent, de-identification of personal information), nonrepudiation (e.g., using digital signature and delivery confirmation), and auditability, the clinical e-mail system also needs to provide mechanisms for message priorities and message expiration. The sender should be able to mark the message urgent or high priority not only for a prompt attention from the recipient but also to expedite delivery of a message by the network transport and routing protocols. On the flip side, undelivered messages must expire and need to be discarded after a certain period. We do not want a diagnostic report for a person to be queued in a mail server in the network for several years and then delivered to their children. E-mail messages cannot be stuck for several years in a mail server but could be queued for delivery when there is a message routing error, as in the case of the domain server omega.univ.edu, which was briefly described by Greenberg et al. [7]. According to the story, a computer assigned to the domain name omega.univ.edu was decommissioned and was turned off. After a few years, upon purchase of a new computer, the network domain name was reassigned to the new computer and activated to receive the e-mails, many of which were over three years old. The e-mail messages had been stored "pending delivery" in mail relays on the Internet.

The key properties of a clinical e-mail system, such as message validity and integrity, are provided by encrypting the e-mail messages and attachments and digitally signing them to protect personal privacy and to ensure data confidentiality and integrity. Maintaining e-mail logs will provide nonrepudiation and security audit trails.

In order to comply with the statutory requirements for protection of privacy, an individual must be notified about the purpose of collecting the personal information as well as the legal authority for doing so. The HIPAA defines personal information as "any recorded information about an identifiable individual. Institutions must protect that personal information by appropriate security. All the demographic and other information associated with registration must be consolidated and securely managed in one logical location for confidentiality and integrity of personal information and for accurate eligibility assessment. Access to a personal health record may be granted at the following levels: role based (access privileges will be assigned to a set of users based on the role they perform), group based (access privileges will be assigned to a set of users that are members of a defined group), and individual based (access privileges will be assigned to individual health care providers). Access privileges to a personal health record should be constrained for delegation, referral, or escalation based on the role, purpose, and logical location, and all transactions should be logged. The transactions in this context are any activities involving the patient data, such as creation, viewing, sharing with others, updating, correcting, archiving, or deleting a health record or related information.

18.9 GREEN HEALTH CARE

Green health care is an emerging discipline of sustainable health care to keep people healthy and to protect the environmental and medical resources (www.greenhealthcare.ca/). The first step toward realizing the concept of green health care is to develop much more effective and much less harmful treatments. The essential step toward developing safe, effective, and preventative medicine that keeps people healthy and also keeps the environment healthy will be effective clinical trials tested over large groups of the population globally and also over a long period of time. By doing so, the long-term benefits and risks of a treatment will be fully assessed. Green health care spans from building environmentally friendly homes and offices (health care centers), offering affordable, sustainable, and renewable patient care, to promoting community and environmental health.

Effective clinical trials to fully assess the long-term benefits and risks of a treatment require multidisciplinary and cross-institutional collaboration to exchange strategies and pioneer ideas by enabling interactive dialogue in real time among the stakeholders of the trial, including the trial participants. A multidisciplinary and cross-institutional collaboration can be enabled with the leverage of various technologies which are essential building blocks for social computing where people can exchange ideas and information in real time in a trusted virtual space, as discussed earlier in this chapter.

In the near future, as asserted by Baek and Robson [4], "your doctor could screen you for known diseases, simply by taking a few drops of your blood, and prescribe the best medication for your condition based on your personal genotypic and phenotypic profile." Doctors, having a variety of safe and effective treatments readily available, may be able to choose the best treatment for each patient in consideration of the individual patient's hereditary traits and environmental and lifestyle variances. Doctors can now provide personalized holistic patient care by screening treatment options in microfluidic chips, predictively simulating the effects and risks of treatment options in virtual reality in computers, and even synthesizing a combination of treatments in microlaboratories.

REFERENCES

1. Genevieve F. Current challenges in clinical trial patient recruitment and enrollment. *SoCRA SOURCE* 2004;Feb:30–38.
2. http://www.nlm.nih.gov/services/ctphases.html.
3. Jameson S. The benefits and challenges of conducting clinical trials. *Community Oncol* 2006;Mar.
4. Baek O, Robson B, Eds. *The Engines of Hippocrates: From the Dawn of Medicine to Medical and Pharmaceutical Informatics*. Hoboken, NJ: Wiley Interscience, 2009: pp. 349–350.
5. RSS. Available: http://www.rssboard.org/rss-specification.
6. FTC. Available: www.ftc.gov/reports/privacy2000/privacy2000.pdf.
7. Greenberg M, Byington JC, Harper DG. Mobile agents and security. *IEEE Commun* 1998;36:76–85.

PART III

TOOLS FOR COLLABORATIONS

19

EVOLUTION OF ELECTRONIC LABORATORY NOTEBOOKS

KEITH T. TAYLOR

19.1 INTRODUCTION

Paper-based laboratory notebooks have a long history (see Fig. 19.1). Leonardo da Vinci recorded his observations in 13,000 loose-leaf pages [1]. Subsequently they were collected into notebooks and survive in public collections; the

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.

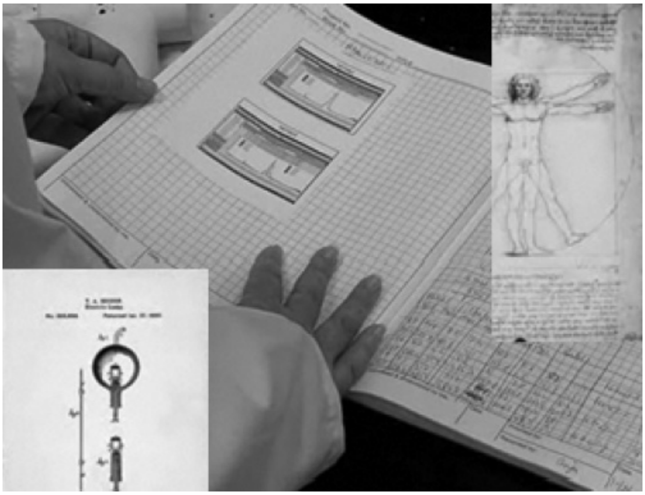


Figure 19.1 Paper laboratory notebooks: past and present.

largest collection made up from 12 volumes and over 1100 pages is the Codex Atlanticus [2]. The Royal Institution (London) also has notebooks from Sir Humphrey Davy and Michael Faraday in its archives [3]. These notebooks were essentially personal productivity tools that recorded experimental procedures, the results, and related observations. Subsequent experiments were then designed based on the scientist's knowledge and intuition. Their use expanded to include intellectual property (IP) protection in countries, like the United States, that issued patents on a first-to-invent basis. In such areas rules evolved around laboratory notebook records to substantiate the first to invent claim. It is not sufficient to record an idea to claim an invention. Due diligence must be followed in reduction to practice; that is, the idea must be converted to an actual work product in a reasonable time. A combination of conservatism and a lack of case law concerning the admissibility of electronic records in U.S. courts retarded the adoption of electronic laboratory notebooks (ELNs) even as research became increasingly electronic.

The U.S. Patent and Trademark Office (USPTO) issued an Official Gazette notice on March 10, 1998, relating to the use of electronic records in patent interferences that asserted that, pursuant to 37 Code of Federal Regulations (CFR) 1.671, electronic records are admissible as evidence in interferences before the Board of Patent Appeals and Interferences to the same extent that electronic records are admissible under the Federal Rules of Evidence (FRE) [4]. This paved the way for the acceptance of electronic records, but it was not tested until 2002 in RE: JOLLEY, 01-1646 (U.S. Federal Circuit 2002) [5]. In this case an e-mail sent from one coinventor to another was deemed to contain the conception of an invention and the priority date for the invention was set as the date of the e-mail

In countries where patents are issued on a first-to-file basis notebooks served the original purpose of personal record archive, but in countries such as the United States they serve a secondary purpose, that is, to claim the priority date for an invention to support the first-to-invent principle on which a patent is granted.

Prior to 1998 a U.S. patent could only be awarded to inventions that occurred within the United States, but in that year the United States adopted legislation that removed this discrimination [6]. In order to be awarded a patent it was still necessary to prove first invention, and this required the keeping of U.S.-style laboratory notebooks that met USPTO standards. Scientists outside the United States were not used to keeping paper laboratory notebooks. This provided the drive to ELNs. Over the next decade ELN applications evolved and the technical and legal issues that impeded their acceptance have been overcome.

19.2 EARLY ELNS

A pioneer in the field of ELNs was Richard Lysakowski, the founder of CENSA [7]. CENSA was ahead of its time and was unable eventually to contribute to the growth of ELNs, but it did provide the first definition of an ELN: "An electronic notebook is a system to create, store, retrieve and share fully electronic records in ways that meet all legal, regulatory, technical and scientific requirements." Subsequently this definition was adopted by Bristol Myers Squibb (BMS), a founding member of the organization.

Initially, the focus of ELN development centered on IP protection. Progress was slow but eventually electronic technology was employed to produce robust and secure electronic notebook records. The crucial components for gaining a U.S. patent are:

- *Conception*—record the idea behind the invention.
- *Reduction to practice*—carry out the invention.
- *Diligence*—execute the invention in a timely manner.
- *Corroboration*—prove that the work was done and when it was done.

Critical to the process is time stamping of the records and all changes to a record. The timestamp must be verifiable, for example, by making it a standard operating practice to reset the clock against a public resource such as the National Institute for Standards and Technology (NIST) [8]. An additional, and safe, practice is to back up the electronic records frequently and retain the backups. A combination of auditing, time stamping, calibration of the timestamps, and availability of extensive backups provides a chain of custody that supports the authenticity of the ELN record.

Important to this process is the fact that electronic records are never original records and are always copies. It is, therefore, important to demonstrate a chain of custody for the record that shows that it has not been tampered with. Furthermore, it is likely that the person who made the original record may not be available to defend its authenticity. Legally this makes the record hearsay, and normally hearsay is not admissible as evidence. FRE 802 [9] covers this situation. It states that "hearsay is not admissible except as provided by these rules," and hearsay is defined as "a statement other than one made by a declarant while testifying in court, offered in evidence to prove the truth of a matter asserted" (FRE 801) [10].

Companies must put in place a foundation that supports the electronic records in an ELN. ELN vendors can assist in the process by including auditing functionality, but the final responsibility lies with the company implementing the ELN. The foundation must include the following policies known to the relevant employees:

- Record activities as part of their normal course of duties.
- Make the record at or near the time of the activity.
- Create and maintain the records according to company policies.

The consequences of not having adequate procedures can be dire. In *Chen (BMS) v. Bouchard (RPR) Interference No. 103,675* [11] the following was decided: "Thus, we [the court] simply do not have adequate information on which to find that Ms. Wei's laboratory notebooks were 'kept in the course of regularly conducted business activity, and if it was the regular practice of [BMS] to make the [record].' " Consequently the patent was declared invalid.

Companies must also establish a custodian for the electronic records; this person will normally present the evidence in court and establish the business exception to the hearsay rule described above.

The records must be in human-readable form and be maintained using industry standards for records management. It is also advisable to test the security of the system using ethical hacking procedures.

Once IP protection was no longer an issue, ELNs began to evolve into a wider solution (see Fig. 19.2).

19.3 CENTERPIECE OF SCIENTIST'S DESKTOP

An ELN can be workflow specific, for example, supporting medicinal chemists who produce new chemical entities, or more generic, where a framework provides for the secure, audited capture of information. Many organizations use Microsoft Word or Excel to capture the information and tie them to a document management system, such as Documentum, to provide security. If the information to be captured is extremely unstructured, this can be sufficient, but if there are structured elements, for example, repetitive calculations, then structure around the repetitive tasks is desired by the users. Adoption by academic scientists has been low, but academic institutes are now very aware of the value of IP and seek to capture it securely so that it can be commercialized.

If a dedicated ELN solution is adopted by the organization, then it becomes an agent for change. The ELN then becomes the central application that the

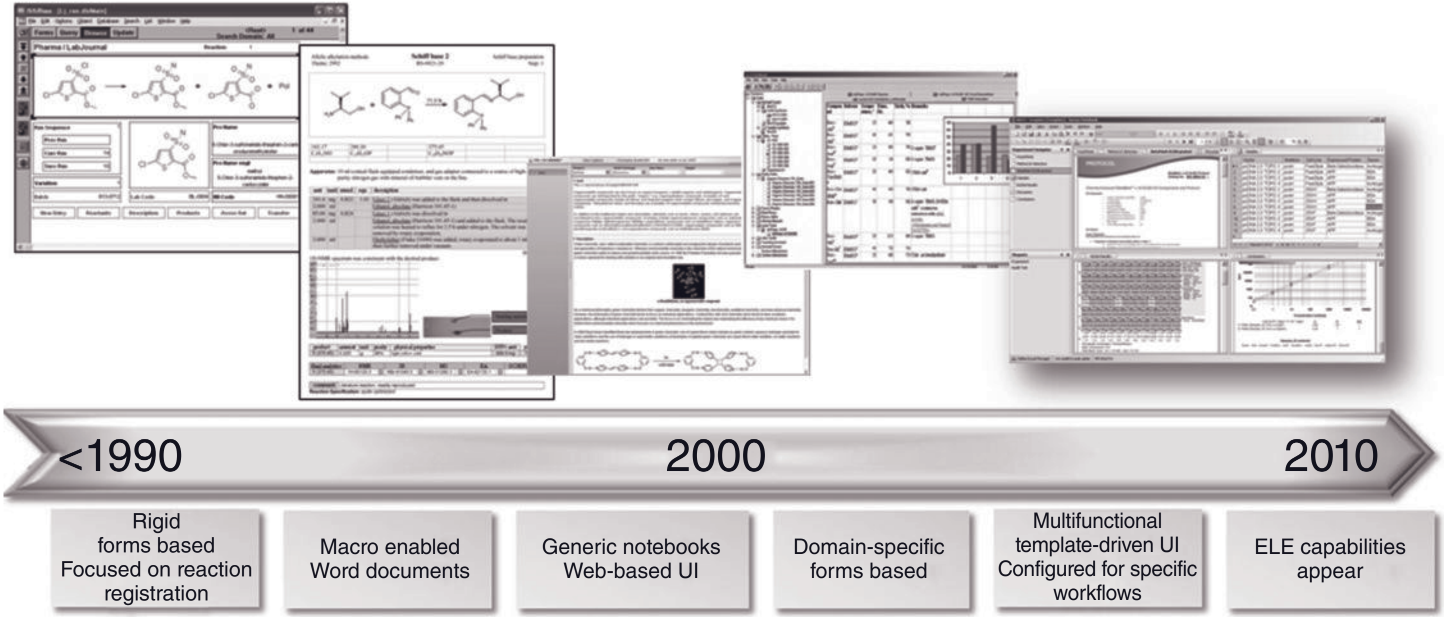

<1990
2000
2010
Rigid forms based Focused on reaction registration
Macro enabled Word documents
Generic notebooks Web-based UI
Domain-specific forms based
Multifunctional template-driven UI Configured for specific workflows
ELE capabilities appear

Figure 19.2 Timeline: evolution of ELNs.

scientist uses many times a day, every day. The scientist expects, and needs, all other applications either to feed information into it or extract information from it. This drove the evolution of the ELN into a hub application that communicates in both subscribe and publish modes with all the other applications that the scientist needs to use:

- Databases are searched to help define the experimental protocol.
- Materials are resourced and ordered.
- Equipment is allocated.
- If equipment needs to be certified, then the certificate information is imported.
- The experiment is performed.
- Data that support the invention are imported.

There are of course many other benefits of an ELN, including the ability to perform general searching across historical content, integration to structure drawing tools, generic algorithms for systematic nomenclature generation, and physicochemical property prediction as well as handling spectral data (see Fig. 19.3). The scientists have a general interface for accessing the majority of data manipulation and handling they will require in terms of developing and documenting a reaction.

19.4 A CORPORATE RESOURCE

Information captured in paper notebooks is essentially personal and with a limited lifetime. It is very expensive to index the data to enable reuse. Information could only be retrieved if you knew someone who might know when and who did similar work. This led notebook records often to be sketchy and illegible—the chance that my experimental details would be available to other scientists was minimal. This happened even though the USPTO requires that the information must be sufficiently detailed to enable someone skilled in the art to repeat an experiment.

Once electronic records were being captured and indexed in databases, there was a good chance that someone else would read the information; the more scientists who benefited, the better it was for the author. A side benefit of electronic notebook records is that the quality of the records has improved dramatically, and because they are not handwritten, their legibility has greatly improved. ELNs now provide a corporate resource of high-quality information, not just details of experiments that worked but also what failed. Now the organization can learn and benefit from the body of work in the ELNs. Companies now report that a significant proportion of new experiments are cloned (essentially copied and edited) from a previous record; Millennium reports that 50% of the experiments are now cloned.

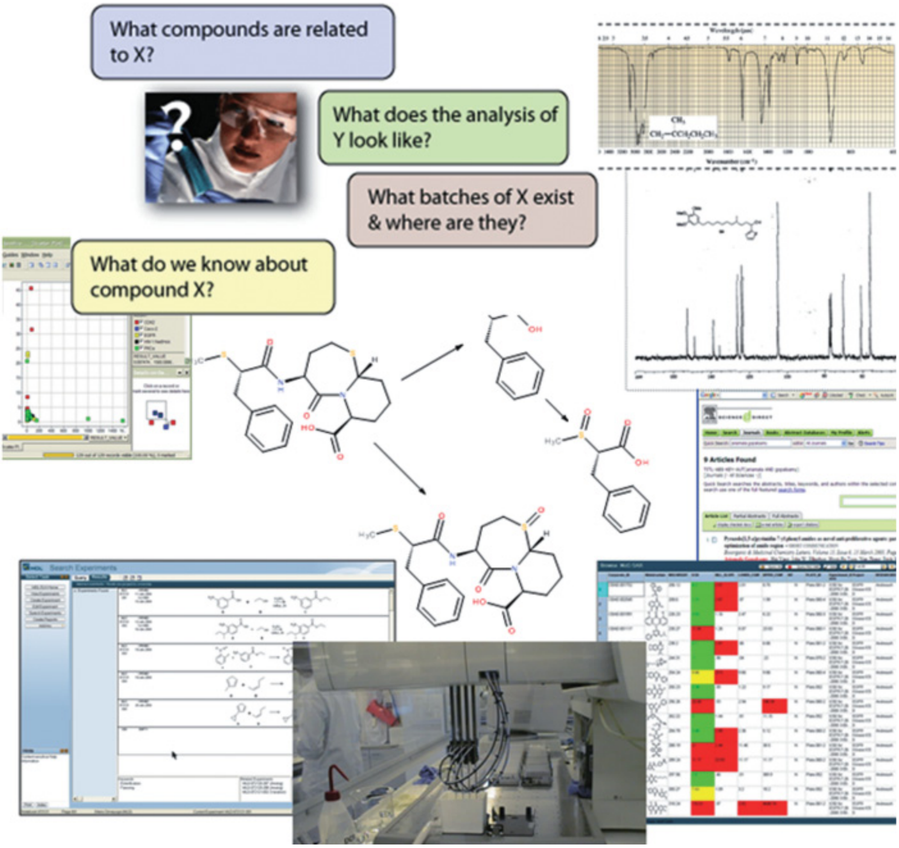

What compounds are related to X?
What does the analysis of Y look like?
What batches of X exist & where are they?
What do we know about compound X?

Figure 19.3 Wide range of data types need to be accommodated in an ELN.

It is common to repeat experiments with minor variations, for example, scale up in a drug development laboratory. In the paper world a new record had to be entered in its entirety. This is both tedious and susceptible to errors. In the electronic environment a simple search discovers previous work; this work is then copied (cloned) to a new record and adapted to the new requirement—larger equipment and material quantities are entered, the equipment is allocated, and the experiment is run. Companies report large improvements in productivity due to the time saved by cloning. Now that a positive value can be assigned to the use of ELNs, companies are keen to leverage their use to find other areas where costs can be saved.

19.5 COLLABORATION

Collaboration means different things to different scientists. Traditionally discovery chemists have worked in semi-isolation. Collaboration to them means finding out what has been done before from the literature and in discussion with their peers. Discovery chemists like to keep their work to themselves until it is completed, so collaborative aspects of the ELN were not highly developed. In principle, once all the information is electronic, it can be searched by anyone, anywhere in the organization. Biologists, however, frequently work in teams where collaboration is the norm. Once an invention moves from research to development, the need for collaboration grows.

Multinational companies particularly could benefit from the sharing of information. Here a scientist in one continent can benefit from the knowledge acquired by a colleague in another continent even though they may be working on different projects. More recently companies have gone further, and they have adopted an aggressive policy of outsourcing research and development to third parties, often a contract, or clinical research organization (CRO) in China or India. The outsourcing of clinical research was a normal part of research in the 1990s; companies such as Quintiles, Huntingdon Life Sciences, and Covance benefited greatly from this approach. Extension to chemical and biological research and development was then natural.

Outsourcing is driving the collaborative aspects of the ELN and placing strict constraints on who collaborates with whom. The ELN also provides a common language and format for the exchange of information; here describing procedures in terms of predefined terms improves clarity. A procedural step can be described in a number of languages and translated into the local form based on the user's preference. This improves clarity of understanding and reliability of information transfer.

Company employees need to be able to view results from the CRO, and the CRO needs to be able to see information relevant to their work that is available within the company, but they must be restricted from viewing other work. Cloud computing is currently in vogue and is seen as a way to deliver selective collaboration because cloud computing environments have to be set up to guarantee the isolation of groups of users.

The cloud is not the only way to share data. More traditional Web-based hosted systems, such as that offered by Collaborative Drug Discovery (CDD), also allow users to share data selectively. This company also offers a public site where data can be made openly visible. Cloud computing vendors and companies like CDD do not yet offer ELNs, but their technology can provide the foundation for a collaborative ELN environment.

19.6 PISTOIA ALLIANCE

The Pistoia Alliance [12] (Chapter 1) was set up following an ad hoc meeting of scientists from GlaxoSmithKline (GSK), AZ, Pfizer, and Novartis who all identified similar challenges and frustrations in discovery informatics. The Alliance gets its name from the town of Pistoia where the group had congregated for an Accelrys user group meeting. The group is founded on the following:

- *Mission* Standardize and streamline data interchange in life science research and development (R&D).
- *Method* Precompetitive collaboration between life science, academia, and commercial partners.
- *Result* Standardization will drive down the cost of data exchange, cloud computing, and process outsourcing.
- *Benefit* Informatics organizations can streamline commodity services and focus on innovation in R&D.

As of April 2010, the Alliance had over 30 members from life science companies and vendors.

A number of projects were initiated. The ELN query service is the most relevant here. The purpose of this project is to derive a set of standards that ELN vendors would implement to enable facile data extraction and interchange between ELNs. This is a concrete example of the need to collaborate driving system design. If successful, it will change the market, reducing the feeling that an organization is locked into a vendor, and CROs who have to collaborate with multiple companies will only need to support one ELN that can share information with all the ELNs that the contracting organizations use.

19.7 QUALITY BY DESIGN

Capture of all the data and information associated with an experiment gives the potential to incorporate knowledge into experiments as they are executed. Repetitions of a process allow weaknesses to be detected, evaluated, and improved. As the body of knowledge increases, users can evaluate a process and predict where the weaknesses are and correct them before execution. The process is then designed to be of high quality from the beginning rather than improve it in a reactive manner over time [13].

19.8 ACADEMIC PROJECTS

ELN development has been largely a commercial initiative, with many companies competing to provide a solution. This means that there is little public information on the design and use of ELN systems. There are, however, two notable projects: open-notebook science and SmartTea. These are discussed below.

19.9 OPEN-NOTEBOOK SCIENCE

The term open-notebook science was coined by Bradley [14] (and discussed in more detail in Chapter 25). In essence it is a real-time notebook that is accessible on the Web as it is created and indexed by standard search engines. This means that anyone can see the experiment as it is being executed and share in the results. Predominately open-notebook science has been adopted by academics but may become more popular inside commercial organizations in the future.

The concept, while attractive from a collaborative standpoint, has its detractors. The most significant issue is that publication in this manner constitutes prior publication for patent purposes and will prevent the issuance of a patent derived from the work. Clearly this limits the use of the approach in commercial enterprises who would normally seek to protect their discoveries with a patent. Many scientists, however, do not like to expose their work until it is completed. This group of users will not adopt the open-notebook science approach willingly.

A technical disadvantage of many Web-based notebooks is that the included data are made available as dead images. Links can be provided to a chemically aware object, but this is undesirable because it introduces a point of weakness, as the integrity of a link has to be maintained. A significant feature of ELNs is that they make it possible to base the next experiment on previous work. That work need not belong to you. The availability of the data electronically means that queries can be posed that will retrieve work similar or identical to that planned for the next experiment. Rather than laboriously copying the work into the notebook, a process that is both time consuming and error prone, the scientist can copy or clone the previous experiment, modify the details, and start executing the experiment. In such circumstances it is useful to retain the link back to the original work so that the system can develop metrics on most cloned work. Such experimental procedures can then acquire a quality standard. Apart from the improved likelihood that the experiment will be successful, much time is saved and that reduces costs significantly.

Open-notebook science does, however, demonstrate that collaboration is practical and, truly interactive ELNs being possible, a company could exploit the approach within its firewall or installed in a cloud computing environment.

19.10 SMART TEA

The SmartTea project [15] was predicated on a simple (very British) idea: how to describe the preparation of the perfect cup of tea in a manner that allowed a second user to replicate the work. While the task seems simple, capturing all the requirements in a concise and unambiguous manner proved a challenge. Variability in the write-up and variability in the interpretation of the written procedure lead to much irreproducibility in the output. With a variety of chemists and computer scientists the process led to a shared understanding of the difficulty of the task. An ontology was developed that enabled the information to be translated in a machine-understandable format [16].

Although there appears to be no activity at present, the work is still available on the Web. The project was funded by the EPSRC at the University of Southampton (UK) and led by Jeremy Frey. This project evolved into a broader investigation of the possibility of remote control of experiments and is a pointer to the electronic laboratory environment (ELE).

SmartTea now appears to have evolved into a more general e-science project [17]. Current research is focusing on the challenges of the control, monitoring, analysis, and dissemination of laboratory physical chemistry experiments using Semantic Web and broker technologies, including environmental factors associated with experiments.

19.11 THE OTHER ELN

Currently one size is expected to fit all. All work is entered into the ELN, including experiment design, some of which may be merely conjecture. Such work has no place in a legal notebook—those that support patent applications—as it can weaken the claims that are made based on the work. This type of information should be contained in a personal notebook. If scientists still have to carry around paper notebooks to capture this type of information, then much of the value of the electronic environment is lost. Of course this information is truly personal, and it will not be published to the organization but the owner can promote it to become the basis of a full notebook entry.

A major barrier to the adoption of an ELN will be ease of data entry. Most people can still write faster, albeit with limited legibility, than they can type. Voice input is attractive in the right circumstances; it is very fast but is not practical in meetings. Tablet devices are the appropriate vehicle for containing the information, but data entry will need to be simplified. Tablet devices rely on the relatively clumsy, human fingers to input data; they do not have the precision of a mouse. This is inconvenient in, for example, chemical structure drawing and the entry of data into cells. Typically, only one hand is available for the operation. This suggests that they will be most useful when much of the data can be captured electronically, in process development and execution, or where voice input is acceptable such as in field trials.

Voice input still has its place and is a largely unserved technology; scanners will have to become part of the tablet, as will information retrieval based on scanned barcodes, and identifiers such as digital object identifiers (DOIs) will also need to be supported as standard.

19.12 STRUCTURED AND UNSTRUCTURED DATA

Experimental records consist of both structured and unstructured data. An experiment that strives to generate a new chemical or biological entity will contain mostly unstructured data, that is, practices and observation, whereas an analytical or clinical experiment will contain mostly structured data, in this case analytical results. Laboratory information systems (LIMSs) are designed to handle large quantities of structured data, and ELNs have evolved to handle a mix of unstructured and structured data with unstructured data predominating.

Laboratory information systems have been in use for over 30 years and predate ELNs. Superficially they do the same things as an ELN. They record the details of an experiment, record details of the reagents used and their source, allocate necessary equipment, and then record the data for the experiment. Finally a report is generated. LIMSs are most commonly found in high-throughput analytical laboratories, for example, a QA (quality assurance) laboratory, or a laboratory servicing clinical trials. This means that they need to operate in environments regulated by the U.S. Food and Drug Administration (FDA) under GLPs (good laboratory practices) or GMPs (good manufacturing practices).

The major difference between an ELN and a LIMS can be characterized on the basis that a LIMS handles structured data, whereas an ELN handles mostly unstructured data. Although there has been a move to simplify the vocabulary used to describe the experiment through the use of predefined terms, the record is much more variable than a LIMS record. A LIMS is focused on capturing a limited number of data types from a large collection of similar samples, whereas an ELN is designed to capture data from a diverse range of experiments with a limited number of samples but a diverse range of tests for each sample. In the extremes ELNs and LIMSs can be differentiated. A discovery chemistry laboratory running three of four syntheses a week with the need to run confirmatory analyses from a range of techniques is best served by an ELN. An analytical chemist receiving 10 samples a day from a 1000 patients and running one or two tests on each sample is best served by a LIMS. But ELNs and LIMSs meet in a gray area where both are equally as appropriate. The automation experience that has been developed by LIMS vendors means that they can make inroads into the evolving ELN market where the ELN is the hub application, but they lack the domain expertise of vendors that have developed ELNs based on their knowledge of the scientific research area.

LIMS vendors recognize the overlap between ELNs and LIMSs. Most have responded by providing ELN functionality in their LIMS offerings. Waters [18] has released a stand-alone LIMS product, Agilent acquired an ELN [19], and Thermo [20] chose to provide ELN capabilities by partnering with Symyx Software (now Accelrys) [21].

19.13 ELECTRONIC LABORATORY ENVIRONMENT

The future of ELNs is as the central component of the electronic laboratory environment (ELE). This concept envisages all information flows being electronic: to seek first information to assist with its design, then resourcing of

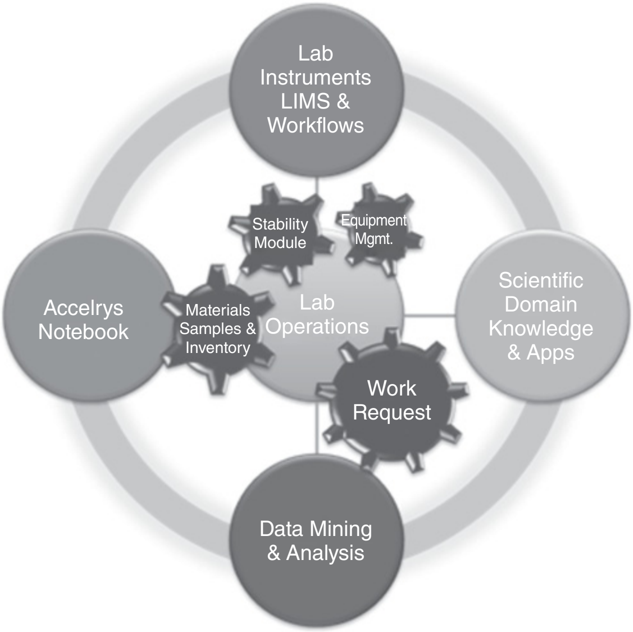


Figure 19.4 Electronic laboratory environment (ELE).

equipment and materials, and finally a collection of experimental results. The concept is simple, and current technology is up to the task, but the variety of instruments and data formats that need to be supported is daunting, and the workflows are highly variable. The task is similar to a standard manufacturing operation, but in this case a holistic approach would probably be taken to the selection of equipment and information protocols. Research scientists will currently not accept this restriction—they always want the freedom to select the best—but scientists in development will be more likely to accept design restrictions as it has the potential to make their life easier.

The focus has moved from the ELN to its place in the overall scientific workflow. The new goal is the ELE (see Fig. 19.4). The ELE is a natural extension of the hub aspects of an ELN. It merges the strengths of an ELN with those of a LIMS:

Strengths of ELN

- Capture free form data from a range of experiments
- Authenticate and invention
- Present a flexible and easy-to-use interface to the scientist

Strengths of LIMS

- Interface to a wide range of equipment
- Submit requests for analysis

- Retrieve results
- Record results automatically
- Support chain of custody for the invention

The scope of this work should not be underestimated. The ELE requires that all equipment be online. This will mean that companies will standardize on models, and perhaps they will need to replace much of their current inventory, particularly for the more mundane pieces such as balances.

19.14 DARK LABORATORY

The ELE facilitates the concept of the dark laboratory, where a scientist identifies structures for testing. The substances are acquired through purchase or synthesis and submitted to the biological assay queue. Results are posted back to the requestor. This will drive the need for global standards for acquisition and storage of data.

The work being undertaken at Southampton University is driving toward the dark laboratory. The experiments being conducted involve mostly physicochemical measurements with robots controlling the equipment. It is not a major extension to use this approach to handle biological screening, but progressing to new chemical entities is more difficult due to the unpredictability of chemical reactions. The move to biological entities as drugs is amenable to remote control. Much of the work is already done by automated synthesizers and could be adapted to be fully dark.

19.15 FUTURE OF ELN

At a high-level laboratory workflows share many components. First comes the concept, then design. Once there is a design, then equipment, laboratory space, and materials must be sourced. The experiment is executed and monitored. The work product can be a physical entity such as a chemical or biological substance, a cell line, a plasmid, a set of numbers, or a graphical object such as a spectrum or chromatogram. Finally the experiment has to be dismantled and the components cleaned or sent for disposal.

The ELN sits at the center of the workflow. It captures the concept; this is perhaps the simplest step. A design evolves that taps the body of knowledge that ELN usage has delivered. External sources of information are queried and relevant information is brought back into the design notebook. The concept notebook is a legal record; it provides evidence of first to invent. The design notebook may not be essential to support a patent, but it does provide evidence of due diligence in the reduction to practice. Now the experiment is set up, the design is transferred to the experiment record, and from here equipment and materials are sourced; links to internal inventories and external suppliers are needed. During execution samples may be taken and analyzed;

ideally the sampling and analysis are handled automatically or with minimal intervention from the experimenter. The results are compared with records held in internal database systems such as OpenLab (Agilent) or Nugenesis (Waters), and the results are used to direct the execution of the experiment. Finally the work product is obtained and processed and registered into the company's knowledge store. A physical entity requires more care, as it needs to be housed in a safe and accessible place and associated with property information.

Although the workflows are similar between research disciplines, the details differ considerably, and this is where the challenges lie. In addition, there is little standardization in input and output formats of the various pieces of equipment. There have been repeated attempts to standardize the format of analytical information, initially through the Joint Committee on Atomic and Molecular Physical Data (JCAMP) [22] format and more recently through the Analytical Markup Language (AniML) format [23]. But there has been little pressure on vendors to deliver standards as users are more driven by best of the breed as the selection criterion and do not adhere to standards. This must change as the ELE evolves. This will be less of an issue with scientists in development environments, especially those regulated by the FDA. Repetition, accuracy, and precision are vital to their work, and automation is the way to deliver it. For these reasons the ELE will first become real in development laboratories and the transition is already in progress.

The ELE presents enormous challenges to niche suppliers of ELN systems. These companies do not have the resources needed to develop such a diverse environment. As the ELE evolves, companies will look to one system to administer. This reflects the current and inevitable trend toward cost containment.

The only companies with the resources to deliver such a system are Accelrys following its merger with Symyx and perhaps the larger instrument makers, Agilent and Waters. In principle an instrument company is in a good position to deliver the complete ELE, but history shows that a software mindset cannot accommodate instrumentation, and conversely an instrument mindset does not accommodate software. Closer collaborations between instrument and software companies appear to be the more likely route to deliver a fully integrated ELE.

19.16 ACCELRYS' EXPERIENCES

Users of Accelrys (formerly Symyx) notebooks are active at the company's user group meetings and frequently contribute to the company's in-house magazine, *Molecular Connections*. A number of detailed case studies are also available on the company's website [24].

Return on investment (ROI) is a critical part of the purchase justification process, and for the early adopters there had to be a leap of faith in assuming that the significant financial and personnel investments required to implement an ELN system would be matched by an attractive ROI. Today it is easier because companies are comfortable with publishing the savings and other benefits that have followed their ELN deployment.

Better access to information smoothes the transition of information between development and manufacturing, and AstraZeneca reports [24] a 50% saving in time using an ELN; Kalexyn reports a 25% saving in report creation and patent preparations. Lilly, an early adopter of the Intellichem product, uses the system across discovery chemistry and development and projects a $75,000 cumulative net saving per ELN user over the period since deployment. Cloning of experiments is widely reported as a benefit and cost saving; Millennium reports that 50% of new experiments are cloned, and Johnson and Johnson reports an astounding 90%. BMS was an early adopter of the Intellichem product in chemistry development and estimates 10% time saving in formulation, analytical, and process pharmaceutical development together with a 20–40% saving in method execution times.

Evidence of the ELN driving equipment standardization is beginning to appear. Pfizer determined [25] that it had more than 100 electronic balances across two large sites in two countries. Previously the purchase decision was made locally leading to no consistency in the makes and models in use. This complicated interfacing to the Accelrys notebook. Rather than spend time providing custom interfacing, it decided on a phased replacement of the balances, focusing on one vendor and selecting models that were compatible with the Accelrys notebook. Several other speakers at the 2010 Symyx User Group Meeting in Barcelona discussed the need to standardize, and Stephen Taylor [26] introduced the concept of the highly automated lab (HAL). In a not-unsurprising parody, he illustrated a verbal dialog with HAL, where the system prevented the scientist from executing a badly designed experiment.

ELN systems are complex to set up and manage. This is a major barrier to implementation at smaller companies. A solution to this problem is remote hosting of the application. Accelrys has put in place a hosting environment and has a number of companies using the approach.

Hosting presents challenges. Security is an issue, but there are proven technologies that address this. More challenging is communication between local services, such as a balance, spectrometer, or even the corporate registry, and the remote application. Once technology can address this communication barrier, we can imagine a completely hosted informatics system that links the ELN with registration, inventory, decision support, and data repositories all remotely hosted.

What goes around comes around! Thirty years ago most computer systems were based on local, dumb terminals communicating with a remote host (normally within the company); systems then adopted a client–server architecture followed by a three-tier architecture. The future seems to be with an externally hosted server system with local terminals that are used in a relatively dumb way so that users can connect to their ELNs.

REFERENCES

1. Science and inventions of Leonardo da Vinci. Available: http://en.wikipedia.org/wiki/Science_and_inventions_of_Leonardo_da_Vinci.
2. Codex Atlanticus. Available: http://en.wikipedia.org/wiki/Codex_Atlanticus.
3. Sir Humphrey Davy at Royal Institution of Great Britain. Available: http://www.aim25.ac.uk/cats/17/2882.htm and http://www.aim25.ac.uk/cats/17/2785.htm.
4. Federal Court Rules and Policies. Available: http://www.uscourts.gov/RulesAndPolicies.aspx.
5. United States Court of Appeals,Federal Circuit: In Re: Scott T. Jolley. Available: http://caselaw.findlaw.com/us-federal-circuit/1343257.html.
6. http://www.foley.com/publications/pub_detail.aspx?pubid=608.
7. Rubacha M, Rattan AK. Selecting the Right ELN. *Scientific Computing*. Available: http://www.scientificcomputing.com/Articles-IN-Selecting-the-Right-ELN-062210.aspx.
8. NIST Internet Time Service (ITS). Available: http://www.nist.gov/pml/div688/grp40/its.cfm.
9. Rule 803. Hearsay Exceptions; Availability of Declarant Immaterial. Available: http://www.law.cornell.edu/rules/fre/rules.htm#Rule802.
10. Hearsay Exceptions; Availability of Declarant Immaterial: (6) Records of regularly conducted activity. Available: http://www.law.cornell.edu/rules/fre/rules.htm#Rule802.
11. United States Court of Appeals, Federal Circuit. Shu-Hui Chen and Vittorio Farina, Appellants, v. Herve Bouchard, Jean-Dominique Bourzat, and Alain Commercon, Appellees. No. 03-1037. October 22, 2003. Available: http://caselaw.findlaw.com/us-federal-circuit/1330376.html.
12. Pistoia Alliance: Open standards for data and technology interfaces in the life science research industry. Available: http://www.pistoiaalliance.org/overview.
13. Bronfield J. Strategy-driven informatics. Paper presented at Symyx International User Conference, Barcelona, May 2010.
14. UsefulChem at Open Notebook Science. Available: http://usefulchem.wikispaces.com/All+Reactions.
15. The Smart Tea Project. Available: http://www.smarttea.org/.
16. Smarttea.or: Making Tea. Available: http://www.smarttea.org/tea-experiment.pdf.
17. Frey JG, Wilson S. Control, monitoring, analysis and dissemination of laboratory physical chemistry experiments using semantic web and broker technologies. Boston: ACS, August 23, 2010.
18. SDMS Vision Publisher an analytical electronic laboratory notebook (ELN). Available: http://www.waters.com/waters/nav.htm?cid=10067209.
19. Kalabie Electronic Lab Notebook (ELN) from Agilent. Available: https://www.chem.agilent.com/en-US/Products/software/labinformatics/openlab/eln/Pages/gp66396.aspx.
20. Electronic Laboratory Notebook (ELN) partnerships. Available: http://www.thermoscientific.com/ecomm/servlet/productscatalog_11152_L11005_81853_-1_4.

21. Symyx Notebook by Accelrys. Available: http://accelrys.com/products/eln/.
22. JCAMP-DX Protocols. Available: http://www.jcamp-dx.org/protocols.html.
23. Analytical Information Markup Language. Available: http://animl.sourceforge.net/.
24. Symyx Notebook by Accelrys Case Studies. Available: http://accelrys.com/micro/notebook/#second.
25. Quinn R, Chabot A. Overcoming development and implementation challenges in an enterprise deployment of Symyx Notebook at Pfizer. Paper presented at Symyx International User Conference, Barcelona, May 2010.
26. Taylor SPB. How we will do science. Paper presented at Symyx International User Conference, Barcelona, May 2010.

20

COLLABORATIVE TOOLS TO ACCELERATE NEGLECTED DISEASE RESEARCH: OPEN-SOURCE DRUG DISCOVERY MODEL

ANSHU BHARDWAJ, VINOD SCARIA, ZAKIR THOMAS, SANTHOSH ADAYIKKOTH, OPEN SOURCE DRUG DISCOVERY (OSDD) CONSORTIUM, AND SAMIR K. BRAHMACHARI

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.

20.1 INTRODUCTION

Neglected diseases, also called neglected tropical diseases (NTDs), are a group of preventable illnesses largely affecting tropical countries that lead to severe disability or even death. More than a billion people, referred to as the "bottom billion", suffer from one or more such diseases [1]. Particularly in the developing countries in Africa and Asia, infectious diseases, including lower respiratory infections, HIV/AIDS, diarrheal diseases, tuberculosis, and malaria, are the major cause of death [2]. Furthermore, approximately 2.7 billion people are at risk of contracting NTDs [1]. Hence, there is a clear need for research into NTD to improve the lives of the afflicted individuals. However, NTDs are overlooked in terms of research and funding given difficult targets and low commercial incentives. Thus, more innovative and cost-effective ways need to be developed for their treatment. The information technology (IT) industry has been an early adopter of open-source models, for example, the Linux operating system, which was created with community participation. Of late, life scientists have also been able to harness the power of collaborative problem solving, for example, the Human Genome Project and more recently an online game to solve protein structures, Foldit [3]. These efforts indicate a movement toward collaborative efforts in solving challenging scientific problems.

An open innovative approach has been espoused through the Open Source Drug Discovery (OSDD) project, a team India initiative led by the Council of Scientific and Industrial Research, India, with global participation to discover affordable drugs for infectious diseases. The OSDD project is focusing on tuberculosis as the first target disease. Tuberculosis (TB) is next only to HIV as the leading global cause of death from infection. Nine million people around the world develop active TB every year. India has the largest incidence of TB in the world with over a quarter million deaths in 2008. This disease has stalked humanity far longer than any other disease, as evident from the fossils, indicating that TB has haunted humans for more than half a million years. Even though existing drugs can actually cure most cases of the disease, the lengthy treatment regimen is a major obstacle. The current frontline drugs were developed in the middle of the twentieth century and require six to nine months treatment [4].

One of the major impediments in discovering new drugs for infectious diseases is the inadequate understanding of the biology of the bacteria causing the disease. The lack of market-based incentives also deters investments into such research by major pharmaceutical companies. Many organizations working in this field have considerable resource and manpower constraints. An open and collaborative approach is an alternative model of innovation for discovering affordable drugs at a faster pace. There is a need to integrate knowledge and human resources for finding solutions to challenges which elude conventional models. Connecting experts and their knowledge across various domains is a technological challenge that calls for a framework for interaction and collective data sharing.

As biology is fast turning to information-driven and quantitative science, data are generated faster than can be analyzed and understood. Thus, there is a pressing need to develop technologies not only to handle this massive amount of data but also for performing comprehensive analyses. The realization of these technologies depends on developing standard ontologies for defining data and its properties (see Chapter 5). The Semantic Web technologies offer powerful new ways to integrate data from disparate sources. This also provides us with meaningful new ways to query data and is a promising approach for integration of biological information.

Facilitating collaboration involving interdisciplinary areas encompassing a wide spectrum of capabilities and skills requires a robust IT infrastructure. Unlike the earlier collaborations that spanned a couple of laboratories, the present collaborations span across the world with individual expertise drawn from diverse fields. In a conventional collaborative model the partners are predecided and known. In the open-source model the crowdsourcing of unknown varying expertise makes the collaboration seamless and dynamic. This implies that the collaborative infrastructure should also scale to keep pace. In this chapter we discuss how the OSDD's Semantic Web-based portal (http://sysborg2.osdd.net) enables collaborative networking using the Internet.

20.2 SEMANTIC WEB-BASED PORTAL TO LINK MIND AND MACHINES

This OSDD Web portal (developed pro bono by Infosys Technologies) is the interface for presentation and exchange of information over the Internet. It is used by the OSDD community of over 4000 participants from more than 120 countries. A Semantic Web-based portal has been developed for seamless information search, access, extraction, interpretation, processing, and sharing between the community members. At the core of this portal is a resource description framework (RDF) data store. Every data point on the portal is linked to another with a relationship and this subject–predicate–object triple forms the unit for RDF store.

The need for interoperable data has led to development of a large number of biological databases with the World Wide Web Consortium (W3C) RDF triples. Gene Ontology (GO) [5], CHEBI [6], and SNOMED (http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html) are examples of the most widely used ontologies in the biomedical domain (see Chapter 5). Efforts are being made to define standard ontologies. The W3C Health Care and Life Sciences Interest Group (http://www.w3.org/blog/hcls) and Open Biomedical Ontologies (OBO) [7] are actively defining common controlled vocabularies and making them available as an RDF. The standards adopted by OSDD conform to the standards that are being developed worldwide.

Web 2.0 tools are changing the way we communicate on the Web. Web 1.0 was about displaying information. Web 2.0 is about conversations and about participation in the flow of information. Web 2.0 uses many new approaches for dealing with information, including wikis, blogs, syndication, forums, and mash-ups. These require and facilitate the active participation of users and have been used to create hugely popular social media sites, such as Facebook and YouTube. The tools of Web 2.0 are chipping away at the tightly guarded boundaries of scientific communication. The tools used by social networks enable researchers to connect with similar projects and locate persons with similar or complementary skills. Researchers are sharing and discussing their work on blogs, wikis, and social networks of Web 2.0. The OSDD portal couples these Web 2.0 tools for scientific collaboration to usher in science 2.0 for TB research.

20.3 DESCRIPTION OF THE PORTAL: COLLABORATIVE WORKSPACES

The main aim of the portal is to provide a virtual framework for discussions and data sharing along with other essential functionalities for managing and organizing the research projects. Thus, in the OSDD portal all the components ranging from social networking to open electronic laboratory notebooks and project management utilities are coupled to enable seamless searching through projects and their results. The best-of-breed open-source applications have been customized and interfaced to achieve the desired functions of a collaborative portal. The portal is also amenable for data and tools integration from third-party resources in addition to accepting and collating data and tools generated by the community.

The OSDD portal design has incorporated all the components of the drug discovery pipeline. The OSDD process divides the entire drug discovery process into 10 work packages under which individual projects are formulated and worked upon by researchers. This involves a chain from ideation to devel-

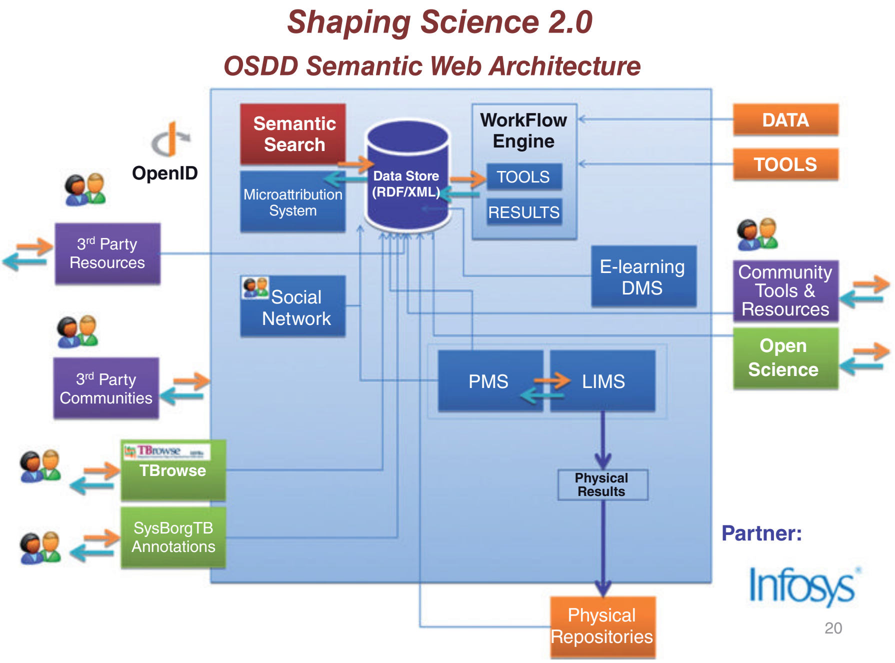


Figure 20.1 OSDD portal components. All the components of the portal are open-source applications. At the core of the portal is the RDF data store which integrates data from all the applications integrated into the portal. The portal has been developed in collaboration with Infosys Technologies.

opment of projects to posting of the results on an open laboratory notebook to facilitating comments on these. In addition, a number of researchers offer their own informatics tools, with their own servers, for facilitating research. The portal framework is designed to link these resources in a federated manner. Components of the portal (Fig. 20.1) are discussed below.

20.3.1 Single Sign On

The OSDD portal is an integrated system using applications with their own authentication mechanism which has been merged under a single sign-on (SSO). The SSO is responsible for authentication of OSDD portal users and establishment of SSO across all systems. This federated system ensures more collaboration giving more freedom to the developer and also helps to integrate newer applications easily.

20.3.2 OpenID Server

The OSDD portal hosts an openID server and helps the members establish an open id account with OSDD and use it to log on to the portal. It is based on the Lightweight Directory Access Protocol (LDAP), which is an application protocol for querying and modifying data of directory services implemented in Internet Protocol (IP) networks (http://en.wikipedia.org/wiki/LDAP) which are configured for authentication and user management across all systems integrated into the OSDD portal.

20.3.3 Portal Backbone

Liferay, a leading open-source portal, has been used as the backbone of the OSDD portal. It has been customized and interfaced with other open-source applications to accomplish the desired functions. The primary modules in Liferay provides out-of-the-box and customized features like social networking, forum discussions, blogs, portlets post, and so on. It also provides flexibility to change ideas, projects, laboratory notebooks, and so on. Liferay has been seamlessly interfaced with a project management system (PMS), lab information management system (LIMS), workflow system, and learning management system (LMS). It also provides portlets to display summary-level information from the PMS, LIMS, LMS, and workflow system. The portal also provides access interfaces for a full-text semantic search mechanism and an interface to view credit points due to users "contributions". The portlets invoke services to access data in the portal database and content management system (CMS) and invoke RDF application programming interfaces (APIs) to aggregate information in the RDF format.

20.3.4 Data Store

At the core of the OSDD portal is the data store, which is customized to store data generated via the portal into the RDF format. It also includes APIs to convert different content generated in each system into the RDF and also for third-party Web services to be aggregated into the data store. The format of the data in RDF makes it amenable for semantic search and hence provides a scalable system for integrating data and concepts. As opposed to integrating individual databases of the different open-source applications, the current system is designed to capture data from all the federated resources conforming to specific ontologies and APIs. Portlet services are developed for data transfer between portlets and portal database/CMS. As each system is represented as a portlet in Liferay, it is imperative to display the information from all the systems on the homepage of the user. Hence each system's Web services are customized to expose information that is required by corresponding portlets to display summary-level information. APIs for each system function to aggregate information in an RDF format.

20.3.5 Project Management System

The function of a PMS is to allow creation and monitoring of projects and various project parameters. DotProject, a Web-based project management application, has been selected as it is designed to provide project layout and control functions. It also captures basic project information from the group to dates to budgets and owners. The projects and tasks can be color coded. Also, files can be uploaded against specific tasks and projects. An interesting module in dotProject is the provision to create subprojects as a "dynamic" task. It allows one to create tasks with a hierarchy including dependencies, priorities, resource assignments, and dates. Tasks can be created and tagged as a milestone. A Gantt chart illustrates the project schedule and shows all tasks and milestones.

20.3.6 Search

Search engines capture the content from the Web, mostly searching the contents of the Web pages in the form of documents. These do not search across databases with different standards. In the OSDD portal, a semantic search engine has been incorporated which performs full-text semantic search on the RDF data store and other databases associated with each system. A semantic search helps improve accuracy of the search by contextual meaning of the terms in structured data and generates more relevant results delivering the information rather than a list of search results based on keywords.

20.3.7 Microattribution

Appropriate attribution is at the core of scientific communication. A microattribution system is put in place to ensure that each contribution on the OSDD portal is attributed to the author. Microattribution is a concept to allocate credit points to all users based on their inputs and contributions to the OSDD community. This ensures that every contribution is tracked and that credit points are attached to each contribution. To keep track of the contributions made by each user and assign credit points to them, a microattribution system is interfaced with all the other systems to capture contributions. This system helps to track discussions on a forum, blog entries, contribution to ideas, projects, and laboratory notebooks and awards points to the associated users. It also facilitates users to redeem points accumulated by them on the portal. The users are to be assigned blue/silver/gold/platinum membership and are part of corresponding user groups in the portal based on points earned. It is anticipated that assignment of credits will enable it to incentivize contributions.

20.3.8 Workflow System

With the advent of high-throughput technologies, there has been an unprecedented increase in the amount of data generated. It is critical to analyze these data comprehensively in a collaborative and reproducible manner. For enabling the OSDD community to perform sharable high-throughput computational analysis, the Galaxy workflow engine [8] has been customized and interfaced with the data store and laboratory notebook. This allows members to invoke services for performing computational analysis of the data present in the data store as well as data available from third-party resources. The computational workflow and the results generated through this system are linked to open laboratory notebooks. In order to provide a comprehensive resource, the OSDD community has integrated more than 300 modules from various sources into Galaxy. There are standard APIs available for integrating any application that the user would like to incorporate in the workflow. This system also lists the tools contributed from the community as well as the usage of each module. There is a provision to create and execute nested workflows and all the data generated through this computational analysis get automatically linked to an open laboratory notebook entry and are stored in the RDF data store, which may be used for further analyses.

20.3.9 Lab Information Management System

LIMS is an application to manage project resources, allocate resources, and store laboratory experiment results. This system includes a database of its own to hold its data. The LIMS interface displays all the resources available. It helps users place an order on resources available for any projects. Only the members of the project are allowed to place the order. Once the order is placed, the workflow for resource approval process kicks in. The order is sent to the review committee group, financial committee group, and project director for approval. During the process, committee members, with whom the order is pending, can approve or reject it and provide comments. The comments are visible to the OSDD community. The LIMS interface allows users to contribute external resources for OSDD which are exported into the portal as XML. For higher value projects, the approval is obtained offline from a high-power science committee and eventually by the chief mentor.

20.3.10 Learning Management System (LMS/eLearning System)

Learning management is a space for managing, training, and education-related activities for the OSDD community members. It mainly concentrates on providing online learning, online assessment, and training materials which are managed and stored in a CMS. This system's interface helps OSDD members to author courses and tests, and upload documents, take-up courses, tests, and so on, and has been customized from Moodle, an open-source community-based e-learning management tool (http://moodle.org/).

20.4 SOCIAL NETWORKING FOR RESEARCH

The OSDD community works together on a virtual distributed laboratory seamlessly interacting and sharing their ideas and results over the Internet through the portal. Hence it was a challenge to interface the state-of-the-art social networking open-source applications with the core components of the OSDD cycle starting from idea to project, results, and more ideas. It is important to create a Facebook or Orkut-like social networking application linking it to scientific projects and experiments and the discussions related to them. Each of the networking modules exists as portlets in the OSDD portal and provides support for Internet-based collaboration. The OSDD portal is comprised of customized Liferay out-of-the-box portlets such as summary portlet, friends and communities portlet, activities portlet, wall portlet, blog portlet, chat portlet, message board portlet, alerts and announcements portlet, really simple syndication (RSS) feed portlet, and document library portlets to meet the requirements of the OSDD community.

20.4.1 Summary

This portlet shows the image of the user along with the number of forum and blog contributions made along with the skills.

20.4.2 Create Virtual Teams

My communities portlet is customized to display members as groups (i.e., within communities) of the virtual team. This portlet is customized to create a community of participating members of an idea/project/laboratory notebook on the portal. These members are added to the friends list as well. The communities, once created, exist even when the project is closed or when the idea or laboratory notebook is no longer in use. Users have the option to join or leave the community.

20.4.3 Friends Requests and Friends

This portlet helps manage friend requests that are placed from the user's page (summary portlet). This allows adding friends who are already not part of any ideas or projects or laboratory notebooks or communities.

20.4.4 Activities

The activities portlet in Liferay is used out of the box to show the recent activities of the user. It shows the friends requests that have been accepted, forum messages, and blogs posted by the user. It is customized to show contributions made by the user. This includes creating idea/project/laboratory notebook entry, running workflows, and so on.

20.4.5 Scrapbook

The wall portlet serves as scrapbook that can be viewed by all users. Only friends have access to write on the wall.

20.4.6 Blog

The blogs portlet in Liferay is used to implement the blog requirement for OSDD. The blogs are configured to be accessed by authenticated users only and are customized to get projects/ideas/laboratory notebook entries so that users can link blogs to them. All blogs linked to projects, ideas, and laboratory notebook entries are credited with points under microattribution.

20.4.7 Chat

The chat portlet is used out of the box and is part of all the pages by default. This is configured to allow community members and other friends to be able to chat with each other.

20.4.8 Forums

Liferay's message board portlet is used to implement the forum requirement for OSDD. For every project, idea, and laboratory notebook entry created in the corresponding system, a subcategory is created in Forum under the corresponding category—ideas/projects/laboratory notebooks. Contributions in those forums that are under projects, ideas, and laboratory notebooks categories are considered for credit under microattribution. The idea/project/laboratory notebook entry and its forum created are linked at the database end. The skills portlet effectively captures the skills and expertise of each and every member of the sysborg2 family. A project manager initiating a project in sysborg2 can search for team members with ease by just browsing through the required skill sets in this portlet. Tracking people with a particular skill and getting the entire skill set of a particular member are some of its salient features.

The RDF API aggregates forum category, threads in the category, and messages in threads as individual objects and provides the relationship between them.

20.4.9 Alerts/Announcements

The alerts and announcements portlets are used out of the box and are part of a user's homepage. Only administrators are allowed to publish site-wide alerts and announcements to all users. Community-wide alerts can be issued by managers of that community.

20.4.10 Document Library

All the documents posted on any idea/project/open laboratory notebook become a part of a document library, an out-of-the-box portlet in Liferay. It has been customized to interface with other portlets and applications. For every project, idea, and lab notebook entry created in the corresponding system, a subfolder is created in the document library under the corresponding folder for the three core components of the portal—ideas/projects/laboratory notebooks. The idea/project/laboratory notebook entry and its document folder created are linked at the database end. The portal API service creates a folder in the document library for each component. The documents are stored in Liferay in the file system. RDF aggregation is done for the title and description of each document and is stored in the RDF database. Their relationship with ideas, projects, and laboratory notebooks are also stored in the RDF.

20.4.11 Web Content

Web content display is Liferay's out-of-the-box portlet which is used to publish content on the Web page. This portlet displays news and updates and is a part of the organization page. The administrators or users with writer role are allowed to create Web content.

20.4.12 RSS Feeds

The RSS feed portlet is a Liferay out-of-the-box portlet. It is used to store feeds and display them in the portlet. It has been customized to dynamically add feeds to the portlet.

20.4.13 Friends Activities

The friends activities portlet is a Liferay out-of-the-box portlet. It is used to display the activities that friends perform in the portal, such as adding blogs, forum entry, and so on.

20.4.14 Calendar

The calendar portlet is a Liferay out-of-the-box portlet and is part of the user page. This portlet allows the user to create events, appointments, and so on.

20.5 MOVING FORWARD: FUTURE OF VIRTUAL COLLABORATIVE RESEARCH

Virtual collaborations have far-reaching benefits in research as they leverage the existing intellectual strength, increase productivity, create a self-organized working group, and decreases travel time, costs, and so on. A successful virtual platform should have ease of access and search and use of the data and other information related to collaborative projects. In addition, alerts to important updates, team dynamics, facilitated discussions, scheduling, and communication are important for successful collaborations. Tools for document sharing, project repositories, tools for process tracking and instant notification, instant messaging, free video and voice calling, and so on, exist in open source or otherwise (http://en.wikipedia.org/wiki/List_of_collaborative_software), but there exists no platform that can provide all of these functionalities on a single platform for drug discovery. This means that researchers are bound to use these applications independently. Moreover, there are no existing technologies to track intellectual contributions of researchers across different applications. Hence it was felt necessary to develop applications and integrate them in a manner that allows for seamless online collaborations and track contributions. The OSDD portal congregates the best-of-breed open-source applications to achieve the functionalities of a virtual collaboration platform for drug discovery. It makes full use of existing open-source tools that promote effective collaborations.

A major challenge while designing the portal was to implement a SSO across applications, which required customizing their authentication system as well as interfacing their database with the portal data store. This architecture, although challenging to implement, ensures scalability for integrating data and tools from any resource.

The OSDD portal provides a robust scalable platform for collaborations and managing drug discovery projects. In OSDD, large complex problems are broken down into simpler tasks that can be implemented with the help of community members. An excellent example to highlight this approach is the development of TBrowse with OSDD community members. TBrowse is an integrative genomics map of *Mycobacterium tuberculosis* and attempts to connect nearly a million data points from various databases, including computational predictions and data from the literature [9]. Also, it is well known that the process of drug discovery is highly complex and challenging. Integrating projects, resources, and intellect becomes the key to facilitating collaborations for solving challenging problems. Such collaborations need compute infrastructure for their speedy implementation.

SysBorg (Systems Biology of the Organism) is the core component of the OSDD portal (http://sysborg2.osdd.net). This component is comprised of subcomponents, namely, open ideas, open project space, open laboratory notebook, resources, and document repository. Any project starts with an idea which is translated into a project with well-defined deliverables with specific timelines. The OSDD portal provides space for posting ideas and developing projects on these ideas. There is also provision to develop subprojects on existing projects. The project space has a predefined template and the participating members of the project automatically become part of the community for that subproject. In addition to community members, other members are also welcome to comment and provide their inputs on the project, thus improving the work design and providing crucial inputs.

Sharing negative results is a big challenge in research as most often these are not published and remain with individual researchers or groups. In OSDD, the community is welcome to share their negative results, which can save time and resource and also provide cues for designing better experiments. For computational projects, the Galaxy workflow engine is customized to provide an easy user interface for creating workflows and sharing them with other community members. The workflow itself and the results generated by running the workflow are automatically extracted to open laboratory notebooks. A version track mechanism exists for storing different versions of the same document for ease of understanding the project plan and results at different time points. A project management system is linked to each project and holds information on the deliverables (tasks) on a scheduler. This system helps in closely monitoring the projects. The overdue tasks are differently colored to highlight bottlenecks in the system. Such tracking and monitoring systems are crucial for online collaborations as all the participating members can access and review the project status, discuss the issues, and get help from the community in resolving bottlenecks.

A drug discovery pipeline is full of failures and challenging bottlenecks. The OSDD portal may help in bypassing at least some of these issues by providing a common platform for communication among experts from different domains. A cross-domain interaction may further the pace of scientific research. The portal also serves as a means for faster communication as the documents or data exist on a common sharable platform and hence are advantageous over normal e-mail traffic as the data exist on an online archive which is searchable. It also helps in evading management overhead by providing PMS applications closely interfaced with the projects and a versioning mechanism to keep track of the document versions. In addition, it is also important to share results in real time to make decisions on future experiments. The open laboratory notebooks may be used for sharing results and discussing the experiments. Moreover, each project is also linked to the resources (chemicals, funds, staffing, etc.), and this information is updated as when the resources are allocated and utilized in the project.

The OSDD project has created open-access repositories for clones, proteins, and small-molecule libraries. These resources are acquired, generated, or contributed by the members of the OSDD community. These resources are also listed for the community and can be readily accessed (http://oar.osdd.net/).

The key component of the OSDD portal is the semantic search, which is based on Semantic Web technologies and uses the RDF data store to find relevant information. In this era of data deluge, it becomes a challenge to generate knowledge from data. In order to make best use of available information, it is imperative to organize and manage data in such a way that it may be searched to answer questions than merely finding data for specific keywords. This is the challenge of the Semantic Web and demands data in standard ontology so that a search may encompass every possible resource and find answers to relevant questions. In the OSDD portal, users can create their own query and find relevant information on the data that exist on the portal as well as data incorporated from third-party resources.

In a nutshell, the OSDD portal enables end-to-end process integration from ideation to drug discovery. It uses a Web 2.0 technology framework to integrate people, process, and information to facilitate the drug discovery process. It features a suite of open-source tools for project management or workflow management, laboratory information management, e-learning, and SSO. The portal features a semantic search tool on the RDF data store, the heart of the system, and also a microattribution system and algorithm to assign credits to contributors. The portal is accessible by scientists and researchers from across the world and the information generated is also available for all registered members of OSDD.

The changing mode of drug discovery in the open-source model will hopefully allow researchers to develop affordable drugs faster in collaboration rather than institutes or companies working in isolation. Thus, open-source collaborative platforms such as those described in this chapter provide a new way to solve challenging problems by community participation.

REFERENCES

1. Payne L, Fitchett JR. Bringing neglected tropical diseases into the spotlight. *Trends Parasitol* 2010;26:421–423.
2. Lopez AD, Mathers CD. Measuring the global burden of disease and epidemiological transitions: 2002–2030. *Ann Tropical Med Parasitol* 2006;100:481–499.
3. Cooper S, Khatib F, Treuille A, Barbero J, Lee J, Beenen M. Predicting protein structures with a multiplayer online game. *Nature* 2010;466:756–760.
4. Yew WW, Cynamon M, Zhang Y. Emerging drugs for the treatment of tuberculosis. *Expert Opin Emerg Drugs* 2010;16(1):1–21.
5. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;25:25–29.
6. Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A. ChEBI: A database and ontology for chemical entities of biological interest. *Nucleic Acids Res* 2008;36:D344–350.
7. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007;25:1251–1255.
8. Goecks J, Nekrutenko A, Taylor J. Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 2010;11:R86.
9. Bhardwaj A, Bhartiya D, Kumar N, Scaria V. TBrowse: An integrative genomics map of *Mycobacterium tuberculosis*. *Tuberculosis (Edinburgh, Scotland)* 2009;89: 386–387.

21

PIONEERING USE OF THE CLOUD FOR DEVELOPMENT OF COLLABORATIVE DRUG DISCOVERY (CDD) DATABASE

Sean Ekins, Moses M. Hohman, and Barry A. Bunin

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.

21.1 INTRODUCTION

Biomedical scientists in academia and industry are rarely collaborative and open with their data until publication or patenting, resulting in considerable redundancy, wasted expenditures, and unnecessary delays. While collaboration may be difficult, the pathway can be negotiated [1] and some have even devised simple rules to make them successful [2]. While chemists and biologists may not always see eye to eye, they bring different perspectives and their collaboration is essential if progress is to be made in biomedical research [3]. It takes a tight collaboration between biologists and chemists in order to effectively translate molecules into potential drug candidates. Until recently there has been limited discussion of how such collaborations could be initiated [4], negotiated [1], or successfully enabled [2]. We are also seeing a number of initiatives such as precompetitive collaboration [5–8], competitive collaboration [9], crowdsourcing [10], and open innovation [11–13], which strongly suggests collaborative drug discovery will be the future paradigm of biomedical research [14–16]. There is also a growing list of publicly accessible databases and Internet-based collaborative tools for chemistry [6, 9, 16–19] that make data more accessible and may increase scientific research efficiency and collaboration. These databases can be used for computational modeling such as quantitative structure–activity relationships [17] and relatively rapid lead identification [18].

It is clear to us and others that such collaborations will be facilitated by computational tools and databases such that data could be shared and stored securely and when desired published. Currently available computational database tools for drug discovery and chemistry in particular are not collaborative and are of limited application for drug development [19]. Recent studies have suggested that there are productivity benefits of collaboration [20, 21] and the formation of collaboration networks [22]. We have previously described novel Web-based tools that combine chemistry informatics, biology, and social networks for drug discovery [19]. Building networks of researchers is important as the impact of these tools would be expected to increase in an exponential manner as a function of the number of interconnected users, for example, like telephones and the Internet. Several examples of network-based technologies exist for business and social environments such as LinkedIn and Facebook. Only recently have these types of technologies begun to impact drug discovery as it becomes more fragmented, with large drug companies relying more on outsourcing and collaborations. Open-access chemistry databases and Internet-based collaborative tools such as LabMeeting, myExperiment, DIYbio, Open wetware, Open Notebook Science, Laboratree, and Science Commons are now available for the science community, but they have limited or no capability to mine the data based on chemical structure and they do not have collaboration features or enable data sharing (open-source public data exchange). In addition, these commercially available tools do not foster community-based models for drug discovery and are relatively costly to maintain and support [9, 10, 19, 23]. On the other side, open, public chemistry and biology data repositories (PubChem, ZINC, eMolecules, ChemSpider, etc. [24, 25]) focus on publicly available data and are not designed for comprehensive data archiving. For example, recent articles have assessed the expanding public and commercial databases containing bioactive compounds [8, 9, 18, 19, 23, 27]. Such open repositories clearly lack the ability to specify private data or limit sharing to selected groups so users have been forced to make a choice between sharing all or none of their data. What is perhaps needed is a selective, secure, collaborative software. For community-based drug discovery to work within the larger biopharmaceutical industry, a platform must have:

1. Strong privacy, security, and collaborative software features
2. Ability to handle both free text and complex, heterogeneous drug discovery data and molecular structures, capturing not just small molecules but larger products seen in the biotechnology industry
3. Data presentation and organization that allow both humans and computers to easily draw conclusions and prioritize experiments, leading to new insights

In general, tools that enable the *selective* sharing of diverse data would be a valuable asset, especially within the area of neglected disease drug development for which the need for collaborative efforts has been well documented [26, 27]. Due to relaxed commercial considerations, neglected disease researchers are certainly more open to selective, appropriate sharing and they are therefore an ideal, forward-thinking community to evaluate innovative data-sharing concepts, with potentially transformational implications for all drug discovery efforts [28].

Exploiting more than six years of experience of using cloud computing and Web 2.0 [24], Collaborative Drug Discovery (CDD) has developed a unique Web-based software currently helping scientists optimally identify and advance novel drug candidates. The software allows scientists to not only manage and analyze their data more effectively but also optionally share their data effortlessly and securely to the degree they want, with whomever they wish, at the time of their choosing [19]. It allows them to easily toggle between and simultaneously mine across private, shared, and public data sets. The CDD software and existing user network are uniquely positioned to improve collaborations in the neglected disease space, thereby increasing the efficiency of drug discovery and development [4].

As a test case for collaborative drug discovery, we will describe our plans to further develop CDD using intelligent informatics. This will facilitate collaboration among researchers that are working on similar projects or diseases and bridge across private data networks to retrieve more meaningful pubic data. Because of the secure and configurable sharing capabilities, researchers from pharma, academia, government, and foundations are open to sharing selected data and algorithms via CDD. We will also describe how this database:

1. Enables researchers to securely collaborate both within and across their research communities
2. Integrates with other public databases
3. Will require an ontology
4. Can be extended to enable CDD to store, mine, and share data on larger molecules such as antibodies, RNAi [29], proteins, or even eventually generic objects like reagents, tissues, cells, and patient subpopulations

21.2 BRIEF HISTORY OF THE CLOUD

CDD was the first collaborative drug discovery data-hosting platform in the cloud [using the application service provider (ASP) or software as a service (SaaS) model] to the best of our knowledge. Today this is a model being emulated by every major drug discovery informatics company, because once one is comfortable with the security requirements, it is a fundamentally more economical model for the customer e.g., (no software support or other related outlay).

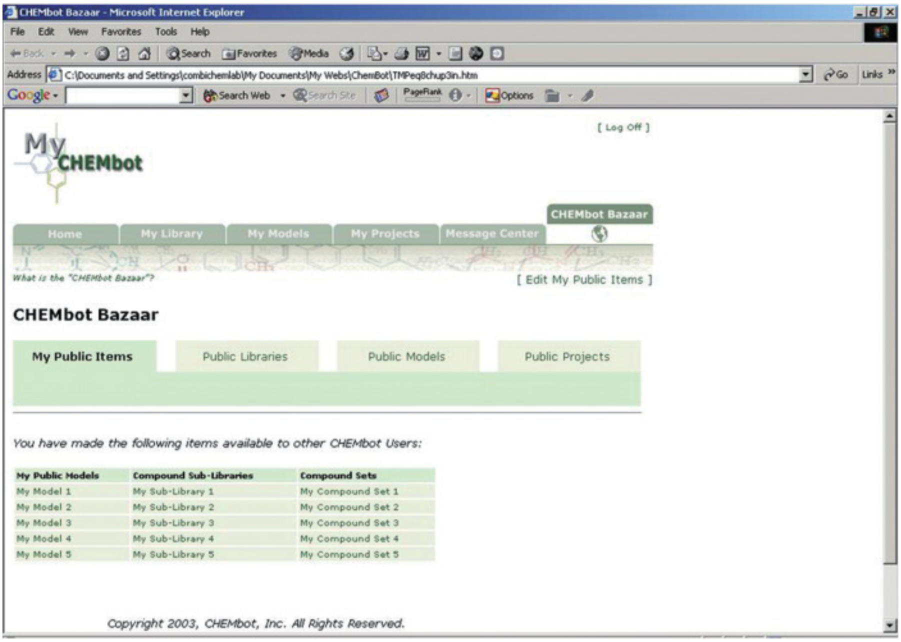
When first envisioned (2003), CDD was a pet project within the Eli Lilly e.Lilly think tank and was then called ChemBot. Initially the business model for ChemBot was unclear, but it was obvious there was a disruptive potential bringing together different researchers, molecules, models, and assays. A very early prototype is show in Figure 21.1, from which CDD has evolved.

As it was becoming clear that the Internet was the primary cultural and economic revolution of our times, the e.Lilly team began to envision what the future pharmaceutical drug discovery model could look like in the Internet era—and how one might have a competitive advantage by moving onto the cloud.

The basic tenet was that in a diversified world the winning pharma company would be the one with the best software "glue" to hold projects together and facilitate them efficiently.

In 2003 and 2004, working with consultants Dr. Will Welch and Cignex using the Zope Plone Content Management System as a backend, a working prototype was created. Furthermore it was demonstrated that there was a market need with multiple customers signing up. Most notable was the laboratory of Professor Jim McKerrow at the University of California at San Francisco (UCSF), which acted as that all-important partner with a real need at the earliest stages of the project. From day one CDD was a collaborative, Web-based platform with requirements both to respect intellectual property (IP) while empowering real collaborations.

In 2004, the Chembot project was officially spun out of Eli Lilly as an independent company called Collaborative Drug Discovery (CDD). Eli Lilly maintained a minor position and coinvested in a syndicate with Omidyar Network and Founders Fund a year later. Today CDD is a profitable platform with researchers from six continents logging into their CDD Vaults™ using


CHEMbot Bazaar - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address C:\Documents and Settings\combichemlab\My Documents\My Webs\ChemBot\TMPeq8chup3in.htm
Go Links
Google Search Web Search Site PageRank Options
[Log Off]
My CHEMbot
CHEMbot Bazaar
Home
My Library
My Models
My Projects
Message Center
What is the "CHEMbot Bazaar"?
[Edit My Public Items]
My Public Items
Public Libraries
Public Models
Public Projects

CHEMbot Bazaar

You have made the following items available to other CHEMbot Users:

My Public Models	Compound Sub-Libraries	Compound Sets
My Model 1	My Sub-Library 1	My Compound Set 1
My Model 2	My Sub-Library 2	My Compound Set 2
My Model 3	My Sub-Library 3	My Compound Set 3
My Model 4	My Sub-Library 4	My Compound Set 4
My Model 5	My Sub-Library 5	My Compound Set 5

Copyright 2003, CHEMbot, Inc. All Rights Reserved.

Figure 21.1 Chembot and Chembot Bazaar prototypes would later evolve into the Collaborative Drug Discovery (CDD) platform in 2007.

CDD Collaborate™ for private collaborations together with data in CDD Public™.

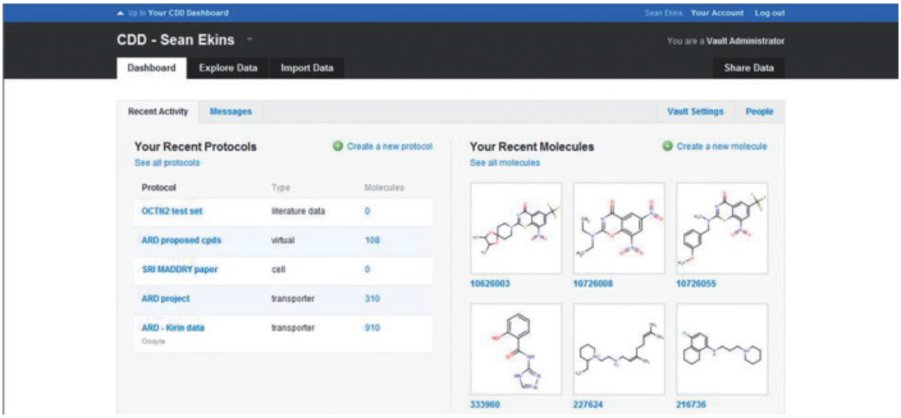
21.3 CDD DATABASE TECHNICAL DETAILS

The development of the CDD database (Burlingame, CA) has been described previously with applications for collaborative malaria research [19]. While the following may be of lesser interest to general readers, it provides useful insight into how the technology is implemented in comparison to alternative approaches. The CDD database brings the power of cloud computing to drug discovery, enabling collaborators to share research data securely within and across organizations without the need to install and maintain complex software. CDD runs on a fault-tolerant infrastructure providing redundant storage, compute nodes, power, heating, ventilation, and air conditioning (HVAC), and backbone connections. The infrastructure is also redundantly secure, protected by multiple layers of host-based, network, and physical security measures.

CDD software runs on a MySQL database and was developed using the Ruby and Java programming languages, leveraging particularly the power of the Ruby on Rails Web application framework. Ruby on Rails is a modern framework noted for enabling productive, "quick and clean," well-factored object-oriented software development, strong Web standards adherence, thorough automated software testing, and horizontal scaling via a shared-nothing architecture. CDD's unusually disciplined approach to automated software testing and a well-factored code base translate into a nearly bug-free user experience for its growing user community which interacts with millions of compounds and tens of millions of data points on a daily basis.

The CDD database can archive and mine a broad range of diverse objects that can later be selectively and securely shared with other researchers (or permanently kept private, which is the default behavior). The CDD database is a hosted collaborative system with an important advantage over traditional PC-based database systems since it can enable secure login into the database from any computer using any common browser (e.g., Firefox, Internet Explorer, Chrome, or Safari). This unique capability for a database system provides flexibility for the users. The CDD Web-based database architecture handles a broad array of data types and is arranged as three specific modules. First, the CDD Vault securely stores and enables mining of private data which are hosted and managed by CDD. Second, CDD Collaborate enables confidential exchange of data between vaults as selected by users. Third, CDD Public hosts public data sets that can be mined. The CDD platform incorporates Marvin, calculated plug-ins for physical chemical calculations, and the JChem Cartridge for structure searching from ChemAxon (Budapest, Hungary) within the application as the chemistry engine. This allows one to do sophisticated structure–activity relationship (SAR) analysis, including chemical pattern recognition (e.g., similarity and substructure searching), physical chemical property calculations, Boolean search and save capabilities for potency, selectivity, toxicity, and other experimentally derived properties. The database can handle heterogeneous data files as well as standardized csv and sdf file convertible formats that represent the chemical and biological data. In particular, CDD is tailored for common data formats used by biologists such as Microsoft Excel (.xls) and text (.txt) files. The technology can mine against a variety of values, including concentration, time, percent, real, integer, textline, cpm, rlu, Z/Z' plate statistics, and IC_{50} (log IC_{50}, R^2 values, Hillslope, etc.). The outputs of such mining can be saved, exported, shared inside CDD, or plotted with an integrated plug-in.

The researcher can control which data to keep 100% private, share with groups of individual researchers, or share more generally with the public. A further unique capability of CDD is the ability to compare all or subsets of public access data with private data simultaneously in a single container as well as analyze multiple vaults to which the user has access. The power of this collaborative approach to drug discovery can be seen in different types of community-based research projects. These range from traditional completely private collaborations, to temporally private collaborations which may become more open following a privacy escrow period, to completely open collaborations where researchers can blog about the experiments as they occur [19].



Protocol	Type	Molecules
OCTN2 test set	literature data	0
ARD proposed cpds	virtual	108
SRI MADDRY paper	cell	0
ARD project	transporter	310
ARD - Kirin data Oocyte	transporter	910

Figure 21.2 Example of CDD dashboard illustrating recent protocols and molecules.

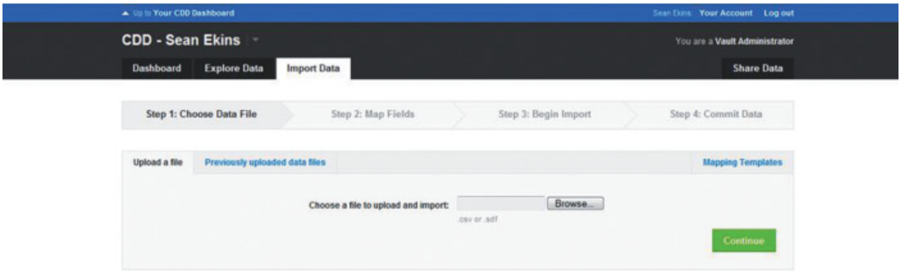
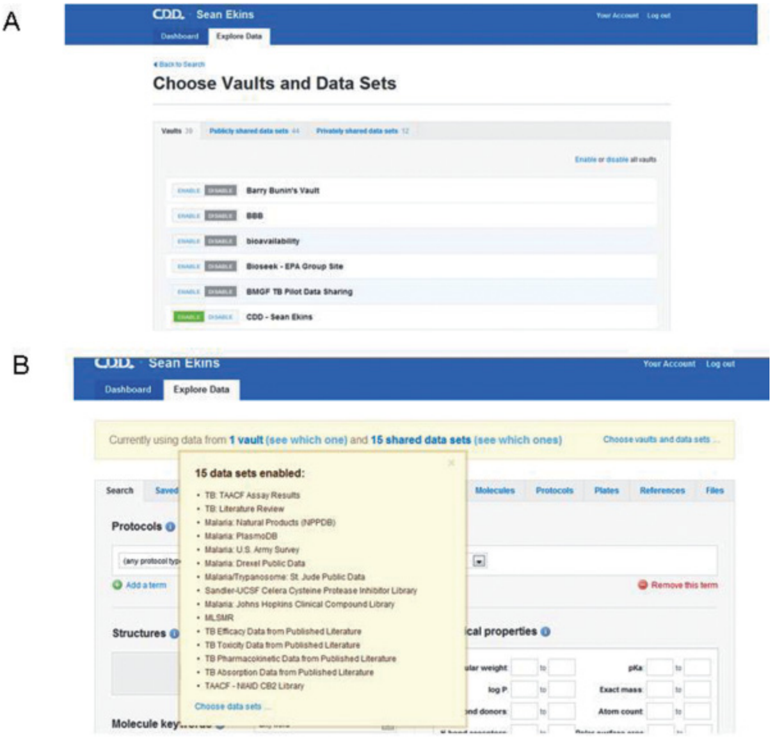


Figure 21.3 CDD 4 step data import process.

When users log in to CDD, they have the option to select the vault which they want to look at; once selected, they will then see a dashboard (Fig. 21.2) which summarizes recent protocols and molecules which they have accessed as well as a listing of recent activity and messages. Data import into CDD is currently a simple four-step process from a .csv or .sdf and mapping a data set to a user-defined protocol if required (Fig. 21.3). Data can be readily mined in CDD, and in addition the user can specify which private vaults and public data sets to use (Fig. 21.4). A full Boolean search is possible specifying protocol, run, readout, chemical properties, keywords, and so on (Fig. 21.5). If molecules are selected, CDD also provides a link to find more information in external databases like ChemSpider (Fig. 21.6). Data in CDD can also be plotted graphically using an interactive visualization which also provides a snapshot of the molecule and data upon mousing over *X*, *Y* coordinated (Fig. 21.7). This may allow a simple SAR analysis, as in the case of the example shown relating to the relationship between calculated log *P* and the human ether-à-go-go



Choose Vaults and Data Sets

Enable or disable all vaults

- Barry Bunin's Vault
- BBB
- bioavailability
- Bioseek - EPA Group Site
- BMGF TB Pilot Data Sharing
- CDD - Sean Ekins

Currently using data from 1 vault (see which one) and 15 shared data sets (see which ones)

15 data sets enabled:

- TB: TAACF Assay Results
- TB: Literature Review
- Malaria: Natural Products (NPPDB)
- Malaria: PlasmoDB
- Malaria: U.S. Army Survey
- Malaria: Drexel Public Data
- Malaria/Trypanosome: St. Jude Public Data
- Sandler-UCSF Celera Cysteine Protease Inhibitor Library
- Malaria: Johns Hopkins Clinical Compound Library
- MLSMR
- TB Efficacy Data from Published Literature
- TB Toxicity Data from Published Literature
- TB Pharmacokinetic Data from Published Literature
- TB Absorption Data from Published Literature
- TAACF - NIAID CB2 Library

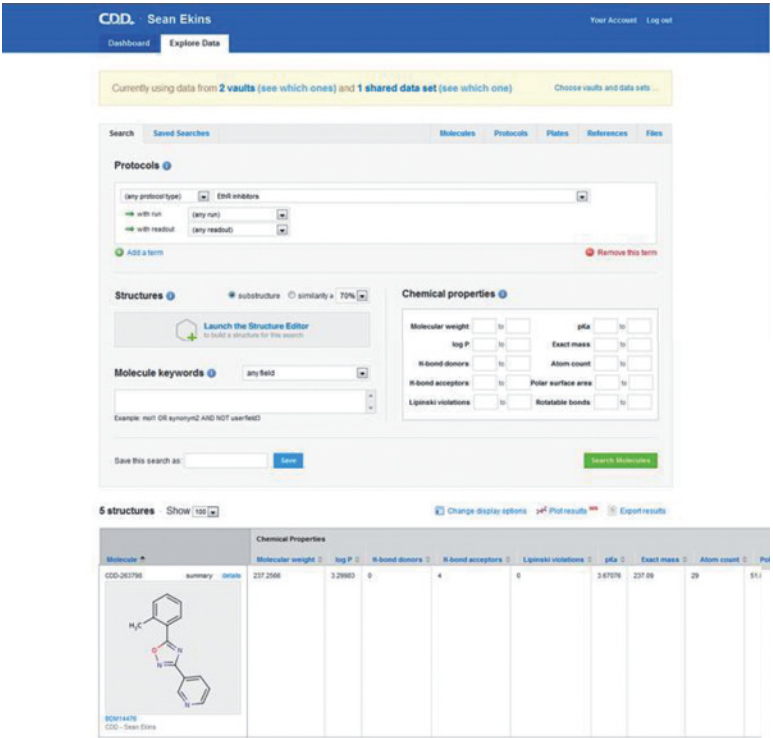
Figure 21.4 CDD vault and data set selection in preparation for data mining: (a) vault selection; (b) Public data set selection.

related gene (hERG) log IC_{50} (half maximal inhibitory concentration) [30–32].

21.3.1 Advantages of Cloud-Based Applications

As the software is hosted on a remote server, it lowers the cost of software distribution and updates while providing easier software evolution and is preferred for instantaneous collaborations. It should also be noted that all stored data are backed up automatically so the user does not need to do this. The database is also scalable, which enables researchers globally to use the software easily. We use a username/password protected group which ensures secure IP protection for private data.

Users without a subscription can upload an unlimited amount of data at no cost if they do not restrict access to that data. Also anyone, upon registration (http://www.collaborativedrug.com/register), can have free read-only access to the public data sets. If access is restricted through user-controlled privacy



Currently using data from 2 vaults (see which ones) and 1 shared data set (see which one)

(any protocol type) EthR inhibitors
with run (any run)
with readout (any readout)

● substructure ○ similarity ≥ 70%

any field

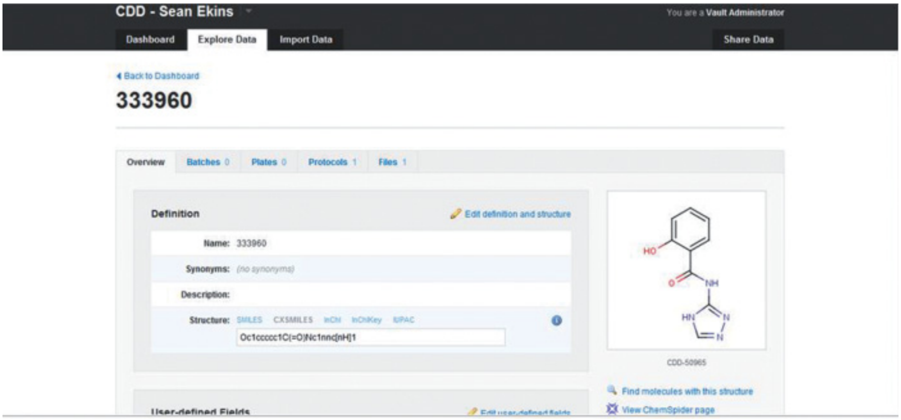
Example: mol1 OR synonym2 AND NOT userfield3

Save this search as:

5 structures · Show 100

	Chemical Properties								
Molecule	Molecular weight	log P	H-bond donors	H-bond acceptors	Lipinski violations	pKa	Exact mass	Atom count	Pol
CDD-263798 summary details BDM14476 CDD - Sean Ekins	237.2566	3.29983	0	4	0	3.67076	237.09	29	51.(

Figure 21.5 Example of exploring data in CDD.



333960

Definition

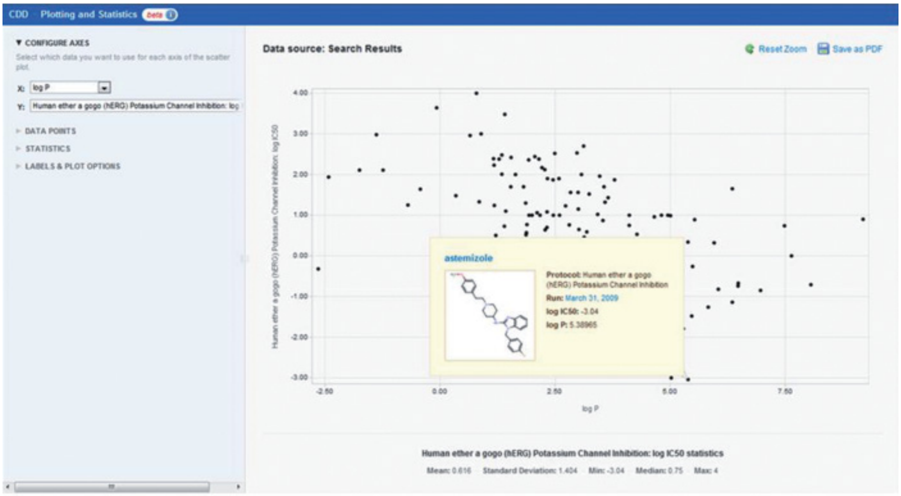
Name: 333960

Synonyms: (no synonyms)

Description:

Structure: Oc1ccccc1C(=O)Nc1nnc[nH]1

Figure 21.6 Example of CDD molecule overview page showing how it can readily link to other databases like ChemSpider.



Select which data you want to use for each axis of the scatter plot.

X: log P

Y: Human ether a gogo (hERG) Potassium Channel Inhibition: log

Data source: Search Results

Protocol: Human ether a gogo (hERG) Potassium Channel Inhibition

Run: March 31, 2009

log IC50: -3.04

log P: 5.38965

Human ether a gogo (hERG) Potassium Channel Inhibition: log IC50 statistics

Mean: 0.616 Standard Deviation: 1.404 Min: -3.04 Median: 0.75 Max: 4

Figure 21.7 Example of using the plot function in CDD with a hERG data set (log IC_{50} vs. log *P* data calculated with ChemAxon tools). Browsing over plot points shows molecule image and data.

settings, a subscription fee is charged. Therefore CDD provides data sets of interest to the scientific community for free while providing a comprehensive database tool for those who want to use a sophisticated and secure data storage and sharing environment. This in itself may represent a unique approach to collaborative drug discovery.

21.4 IMPACT ON NEGLECTED DISEASES

The CDD platform supports the full range of collaborations and the current community includes leading researchers working on neglected, developing world infectious diseases like malaria [33–35], *Mycobacterium tuberculosis* (Mtb) [36], Chagas disease [37], leishmaniasis [38], African sleeping sickness [39], and others as well as drug discovery projects on more commercial targets of interest to big pharma. Initially we have focused on the neglected disease community as an increase in the amount and depth of collaboration would speed the progress of research and help prevent premature deaths due to these diseases. There is relatively little funding for neglected disease research, so researchers must collaborate to achieve the level of efficiency and cost effectiveness required to rapidly produce new therapies. There are currently numerous pharmaceutical companies that are involved in neglected disease research, including Pfizer, Eli Lilly, Novartis, Glaxo Smith Kline (GSK), J&J, and so on, working in collaboration with academics or biotechs, and for any of them to use such a tool the technology has needed to undergo rigorous evaluation.

Even for neglected disease drug discovery collaborations, prepublication or prepatent data are still often treated sensitively. However, due to relatively relaxed commercial considerations, neglected disease researchers are more open to selective, appropriate sharing. CDD possesses novel functions that can be used to bring together neglected disease and other researchers from usually separate areas to collaborate and share compounds and drug discovery data with major pharmaceutical companies in the research community, which we hope will ultimately result in long-term improvements in the research enterprise and health care delivery.

Researchers working on different neglected diseases rarely coordinate their activities on shared compound libraries or reagents and generally only share positive results in publications often without providing their underlying data sets that would benefit others working on other neglected diseases. We envisage a paradigm shift from the limited private networks (or data silos) that are predominant today toward a future vision of interconnected, more open, and more collaborative, scientific networks across neglected diseases facilitated by intuitive scientific networking software and a series of rewards that will incentivize data import and selective sharing. Tangible incentives will include advanced data analysis, data visualization, and collaborative intelligence. As more researchers adopt the collaborative paradigm, additional researchers will take notice of the enhanced productivity enjoyed by the early adopters, creating a classical self-reinforcing growing network. However, one of the rate-limiting steps here is reinforcement from the funding community. Until the National Institutes of Health (NIH), National Science Foundation (NSF), and not for profits really start to actively encourage scientific collaboration (and direct their monies to scientists that collaborate), the current situation will not change. We have seen the European framework grants catalyze academic collaboration (http://cordis.europa.eu/fp7/home_en.html) and impact research on diseases such as tuberculosis (http://www.nm4tb.org/).

21.4.1 Malaria and Mtb

Drugs that are active against malaria and Mtb are urgently needed [40]. The rapid and pervasive emergence of resistance to antimalarial drugs has led to a reemergence of the disease. Of particular concern are chloroquine-resistant (CQR) *Plasmodium* strains. Similarly, the reemergence of Mtb and drug-resistant strains is of grave concern globally. Recent estimates suggest that over 2 billion individuals are infected with Mtb [41] with over 1.7 million deaths per year (latest figures for 2008 from the World Health Organization) or approximately one person every 8 s. Malaria infects ~200 million people and causes over 1 million deaths *per year*, disproportionately claiming African children under the age of five.

Discovery of anti-infectious agents is complex and incredibly difficult for a number of reasons that could be addressed if experts and organizations worked closer together, including:

1. Relatively low hit rates from resource-intensive high-throughput and secondary screens with disproportionately few lead candidates that maintain efficacy in humans.
2. Insufficient awareness among the researchers of the need to obtain very domain specific biologically relevant chemical diversity [42].
3. Academic and other nonprofit laboratories focused on neglected disease research tend to be distributed across the globe and in third world/developed nations with limited opportunities or methods for collaborations.
4. Independent efforts, while providing significant contributions, often lack the project management, data handling, decision gates, and pipeline integration functions that are critical to efficient drug development.
5. Pharmaceutical company contributions are significant but rarely shared publicly.

A new approach is needed to speed up the drug discovery and development process for neglected diseases. This could save millions of lives per year. Foundations such as the Medicines for Malaria Ventures (MMV), the Worldwide Anti-Malarial Resistance Network (WWARN), and TB Alliance have recently become aware that a new paradigm of increased collaboration is needed if new anti-infectives for neglected diseases are to become a reality. These groups are effectively virtual drug discovery units that outsource all aspects of preclinical and clinical research and act as a central point of organization. For example, MMV manages over 50 antimalarial projects in collaboration with over 80 pharmaceutical, academic, and endemic-country partners in 44 countries. TB Alliance has a similar role for TB drug discovery with over 20 on-going programs.

21.4.2 CDD TB DB and CDD Malaria DB as Examples of Community Data Sharing and Evaluation

We have created separate networks of researchers in tuberculosis (well over 100 including TB Alliance and major laboratories in the United States and Europe), in conjunction with the Bill and Melinda Gates Foundation (BMGF) funding, and malaria (dozens) and kinetoplastids (dozens) from many of the top laboratories. In an attempt to obtain a greater understanding of the chemical space of molecules tested against Mtb, we have created a Collaborative Drug Discovery Tuberculosis Database (CDD TB) for related molecular libraries of compounds from the literature [43]. CDD has collated at least 15 public data sets on Mtb-specific data sets representing well over 300,000 compounds derived from patents, the literature, and high throughput screening (HTS) data (Table 21.1). In addition, many major individual academic, nonprofit, and commercial groups have used this Web-based database system [19] to facilitate their own research and store and share their private data. To date

TABLE 21.1. CDD Publicly Available Mtb Data Sets

Database Name/ Source	Description	Molecules
TAACF	Antibacterial activity of a publicly available library compound against *Mycobacterium tuberculosis* (H37Rv) in Alamar blue whole-cell assay	812
Ballel	Tuberculosis SAR data compiled in a survey of agents active against *M. tuberculosis*, including those with both known and unknown modes of action [75]; updated April 17 with TubercuList/TBDB/other target links and improved references	49
MIC Prathipati GVKbio	SAR MIC data from a recent publication by Prathipati et al. at Novartis [76]; consists of a data set culled from the GVKbio database published as supplemental information at the journal website	2,880
MIC Prathipati NIAID	Literature TB MIC SAR data from a recent publication by Prathipati et al. at Novartis [76]; consists of a data set culled from the NIAID website published as supplemental information on the journal website	3,748
MLSMR	A diverse collection tested by the Southern Research Institute against Mtb H37Rv [77], most active compounds have dose response and cytotoxicity data	214,507
Efficacy data from literature	TB efficacy data from over 300 published literature sources; data include PubMed citations, targets, cells and organisms testes, MIC, % Inhibition, EC_{50}, IC_{50}, etc.	6,771
Toxicity data from literature	TB toxicity data from published literature sources; SAR data from PubMed references; data include PubMed citations, targets, cells and organisms testes, cell viability, LD_{50}, CC_{50}, MNTD, etc.	638
Pharmacokinetic data	TB pharmacokinetic data from published literature sources; data include PubMed citations, targets, cells and organisms tested, bioavailability, V_m, V_d, C_{max}, etc.	28
Absorption data	TB absorption data from Gupte et al. [78]	24
TAACF- NIAID-CB2	Results of screening a commercial compound library by the Southern Research Institute to inhibit the growth of *M. tuberculosis* strain H37Rv [79]	102,634

(*Continued*)

TABLE 21.1. (***Continued***)

Database Name/ Source	Description	Molecules
EthR inhibitors	Druglike inhibitors of transcriptional repressor EthR; molecules and data from Willand et al. [80]	5
Sacchettini review	First- and second-line anti-TB agents from Tables 1 and 2 in Sacchettini et al. [81]	14
Makarov et al., NM4TB consortia	SAR data for 1,3-benzothiazin-4-ones (BTZ); data obtained from Makarov et al. [44] at NM4TB consortia	32
Small-molecule patent data	Structures and patent information regarding TB research from the U.S. Patent and Trademark Office, European Patent Office, and World Intellectual Property Organization	20,775
Sacchettini review additional nonapproved anti-TB drugs	Nonapproved anti-TB agents from Figure 1 in Sacchettini et al. [81]	18
Novartis TB data	Aerobic MTB activity (MIC_{50}), anaerobic MTB ATP activity (IC_{50}), and cytotoxicity (CC_{50}) data	283

we have developed a unique community with over 20 pilot groups in the field of Mtb, including groups in the European Union (EU)–funded NM4TB initiative [44] and groups funded by the BMGF tuberculosis accelerator project. Our analysis of the public data sets [43] provided insights into molecular properties and features that are determinants of activity in whole cells [43, 45]. This database has also been used to build novel computational machine learning and pharmacophore models that could be used to filter other libraries of molecules to rapidly identify potential inhibitors [43, 45, 46].

CDD has also developed and deployed a robust preliminary public antimalarial database from five sources which hosts data on approximately 16,000 public compounds (Table 21.2). The growth of this database has fostered several key antimalarial discovery collaborations between CDD users [19]. For example, a substructure search for the known chemosensitizer substructure led to the identification of hundreds of compounds for laboratory evaluation by the laboratories of Dr. Peter Smith in Cape Town to overcome the resistance to chloroquine [19]. Leading candidates were identified and sent from collaborators for evaluation of efficacy in assays using the resistant African malarial parasite strains in human red blood cells. This process shaved months off a project timeline relative to synthesizing new compounds from scratch. Eighteen compounds were identified from a set of U.S. Food and Drug Administration (FDA)–approved drugs using substructure searching and half

TABLE 21.2 CDD Publicly Available Malaria Data Sets

Database Name/ Source	Description	Molecules
U.S. Army survey	Extensive collection of antimalarial drug animal SAR data, including structures and bioactivity, published originally by U.S. Army in 1946	12,318
St. Jude Public Data	Open-access malaria/trypanosome results from Kip Guy's laboratory, including HTS of bioactives against malaria and *Trypsanoma brucei*	2,426
Malaria natural products (NPPDB)	Antimalarial database of flavone natural products, including antimalarial and cytotoxicity data (University of Mississippi, National Center for Natural Products Research)	426
Malaria PlasmoDB	PlasmoDB of malaria inhibitors compiled from the literature, including chemical structure, PlasmoDB gene identifier, target gene name, and references against *Plasmodium falciparum*, *P. vivax*, *P. berghei*, *P. yoelii*, *P. chabaudi*, *P. vinckei petteri*	120
Drexel public data	Results from an ongoing open data collaboration between Drexel (Ugi-4CC products) and UCSF (antimalarial screening); data set represents an example of how researchers can choose to publish selected results openly (By default, in contrast, all groups are private)	195
Johns Hopkins—Sullivan	Percent inhibition of approved drugs at 10 μM	2,693
St Jude Childrens Research Hospital	Supplemental data for *Nature article* [82]; structures tested in a primary screen, with additional data in eight protocols: Bland-Altman analysis, calculated ADMET properties, phylochemogenetic screen, sensitivity, synergy, and enzyme assays as well as a thermal melt analysis	1,524
Novartis Malaria	Data from *Nature* paper [83]; *Plasmodium falciparum* strains 3d7 (drug susceptible) and W2 (chloroquine, quinine, pyrimethamine, cycloguanil, and sulfadoxine resistant), obtained from the Malaria Research and Reference Reagent Resource Center (MR4), were tested in an erythrocyte-based infection assay for susceptibility to inhibition of proliferation by selected compounds	5,695

a dozen were purchased and shipped to Africa, and when tested in the assay, these known drugs were shown to almost completely reverse (seven-fold reversal) the resistance in human blood cells [19]. We have also worked with groups to facilitate computational modeling of malaria data using the public data in the CDD malaria database which was then used for further database screening in silico. Recently we have added data for compounds active against malaria from GSK, Novartis, and other groups [47].

We anticipate that as we add more data sets we will create a combined neglected disease database that will grow into a major resource to help utilize limited research and development (R&D) resources more effectively to accelerate the discovery of better treatments for these diseases. These databases can be searched in CDD alongside private data sets in secure vaults. Because CDD already has over 3 million unique molecules for humanitarian and commercial applications, one can be confident that the architecture and processes will scale well for any drug discovery applications.

These proof-of-concept studies illustrate how (1) CDD can create a community which fosters archiving of data into a database for selective sharing, (2) groups will share some of their data with the community at large, (3) these data can then be used for creation of computational models, and (4) the computational models can then be used for searching the other open data sets or private data sets deposited in CDD to discover new compounds for testing.

21.5 PHARMACEUTICAL COMPANIES CHANGING THEIR BUSINESS MODEL TO INCREASE COLLABORATION AND CROWDSOURCING

There is a new urgency within pharmaceutical companies to cut back on internal drug discovery and to rely more on external collaborations with smaller companies and academics to bring them leads. Besides crowdsourcing tools like Innocentive, there are also other novel approaches, for example, the Lilly Phenotypic Drug Discovery (PD^2) initiative, a website where scientists can securely submit their molecules for evaluation by Lilly prior to selection and legal processing, that precedes in vitro testing for various diseases. We are rapidly approaching a future for biomedical research where loose networks of researchers from companies, academics, or consultants can create aligned communities around shared interests to gather ideas and advance projects. This represents an example of crowdsourcing, where the wisdom of the many and their varied perspectives benefit community-based efforts [48]. Good examples include online databases such as PubChem, the Chemical Entities of Biological Interest (ChEBI) database, DrugBank, the Human Metabolome Database, and ChemSpider [24, 25, 49], in addition to commercial databases [50] and collaborative systems like CDD.

We think the PD^2 approach could be extended to neglected diseases. By providing an entry point in the CDD database, submitters could send their molecules for evaluation. The molecules would pass through various desirability filters and models for bioactivity. Individual research groups or foundations could be alerted if a molecule matching their desired criteria is submitted. The platform will then prompt whoever has opted in and has appropriate privileges to set up a collaboration between the parties involved and to prepare any legal confidential disclosure agreements (CDAs), for example, prior to receipt of the physical molecule, to arrangement for purchase, to synthesis, or to screening.

21.6 FUTURE DIRECTIONS OF CDD DATABASE

In the future we propose that Web 2.0 technology [24] will enable researchers that "opt in" to be made aware of other potential researchers with similar interests or similar compounds or vendors with similar compounds. We will also facilitate the linkage between academic and foundation researchers with a network of experienced medicinal chemistry, cheminformatics, and absorption, distribution, metabolism, excretion, and toxicity (ADME/Tox) collaborators or consultants who can assist them with advancing their compounds further. In this way CDD could become a central marketplace for collaboration and research in much the same way that www.assaydepot.com provides a convenient central location for clinical research organizations (CROs).

21.6.1 How to Build (Neglected) Disease Communities on the Cloud

(i) *Capture and Curate Large Amounts of Disease Assay Data Points* Compound structures can be readily harvested from the most relevant scientific journals, patents, and other available published sources preferably using manual and a combination of natural language processing and manual curation. Provide these data as a public-access read-only resource to benefit all researchers (community member/subscribers would be able to download any public data set).

(ii) *Data Resulting from Different Disease Research Projects Are Archived in CDD Database* To enhance and accelerate discovery research collaborations throughout the network (with the requisite incentives to collaborate while always giving data generators the choice). CDD generally starts with several key seed groups of researchers that pledge participation—generally prominent disease researchers and drug discovery groups.

(iii) *Work with CDD Members to Identify Collaborative Research Opportunities within Network and Support Integration of New Networked Members' Research Efforts* CDD scientific consultants aid in building collaborations around the following drug discovery cycle components: (1) model building, (2) data mining and virtual screening, (3) compound procurement, (4) compound profiling, and (5) research data interpretation.

(iv) *Develop New Custom Capabilities in Database for Users* Generally each group may need some custom features developed, which greatly facilitates their use of the database. This benefits the community as a whole and also moves certain development tasks up the prioritization list.

Several examples of capabilities we think will be essential for such databases as CDD in future are listed in the next sections.

21.6.1.1 Develop an Ontology to Facilitate Advanced Data Mining Across CDD Databases The biomedical research community, and specifically those involved in neglected disease research, is generating very large data sets facilitated through HTS [42, 44, 45], and this presents impending informatics challenges both for selection of hit compounds for follow-up studies as well as computational analysis of such data. A highly effective concept for ensuring such data have continued utility and accessibility is through a formal ontology; for example, the Gene Ontology [46, 47] and the OBO Foundry have successfully demonstrated the utility of this approach. The benefits of ontologies have been well articulated by others and used to enable network analysis [51], facilitate translational bioinformatics [52], and link diseases to animal models [53]. Bioassay data are particularly well suited for management within an ontology because they encompass a wide diversity of experimental designs but usually have a limited range of prescribed objectives. A functional ontological framework will allow new assay descriptions to be meaningfully integrated into the knowledge base with relative ease.

Adoption of an open-assay ontology will be a major milestone in converting volumes of assay *data* into machine-interpretable *knowledge* and finally human *insight.* Given CDD's unique position with an already engaged research community, the key prerequisites are in place to make the ontology widely adopted and therefore maximally useful.

Incorporation of an ontology which captures this complexity will allow for more precise information extraction and integration of the various structured and unstructured data sources. These ontologies will be incorporated into the fabric of CDD with strategic "push" and "pull" mechanisms to promote adoption. Once a first pass is made at automated assay annotation, a series of simple questions (Boolean or short list choices) will complete the entry and ensure accuracy of the automated procedures.

CDD will assist the community in annotation of its assay definitions and data within an open-assay ontology. Because of the large backlog of historical assay instances and the importance of demonstrating a benefit quickly to promote adoption, CDD will directly assist partners in assay annotation. Through this assistance, project partners will greatly accelerate the completion of the task and ensure accuracy and fidelity.

An ontology will facilitate collaboration between researchers in a public/ private data-hosting environment by enabling automated systems to alert researchers of potential collaboration opportunities. Another powerful data-mining strategy that only becomes possible within the context of a working ontology used within an environment that hosts both public and private data sets is the suggestion of similar data that may be relevant to the researcher.

A major outstanding problem is the lack of a coordinated strategy for non-ontologists (e.g., experimentalists) to be sufficiently incentivized to mark up data for ontological binning upon importing screening data into databases. We will take a two-pronged approach consisting of both pull and push mechanisms. On the pull side, CDD Collaborate allows researchers to compare its private data with public data and private data with collaborators' data and, with the new technology envisioned, has the option to "opt into" alerts for notification when others are working on similar/complementary compounds, targets, and so on. Adopting the ontology will be a pull incentive by rewarding scientists with potentially complementary data and new collaborators. On the push side, CDD already has required fields in the database such as the type of assay (enzyme, cell, animal, etc.) and requirements for selecting a date for a run of a screen. Similar requirements will be added to accurately annotate new definitions within an assay ontology, such as selecting a target from a preloaded ontology. The combined pull and push mechanism will rapidly lead to a very large set of ontology-compliant data, greatly facilitating both human and automated connections between public and private data.

21.6.1.2 Collect, Mine, and Share Multiple Types of Data in CDD To date the CDD database has solely focused on the small-molecule community. There is clearly an enormous opportunity to greatly increase the size of the researcher community using the collaborative software by expanding to larger molecules and biological materials. These larger molecules and biologics are equally important to finding treatments for neglected diseases as well as of broad commercial and academic interest. This will engage a greater percentage of the research community, for example, those doing fundamental biology research or working on vaccines. Furthermore, importing related chemical and biological data sets opens up the possibilities for evaluating combinations of small and large molecules.

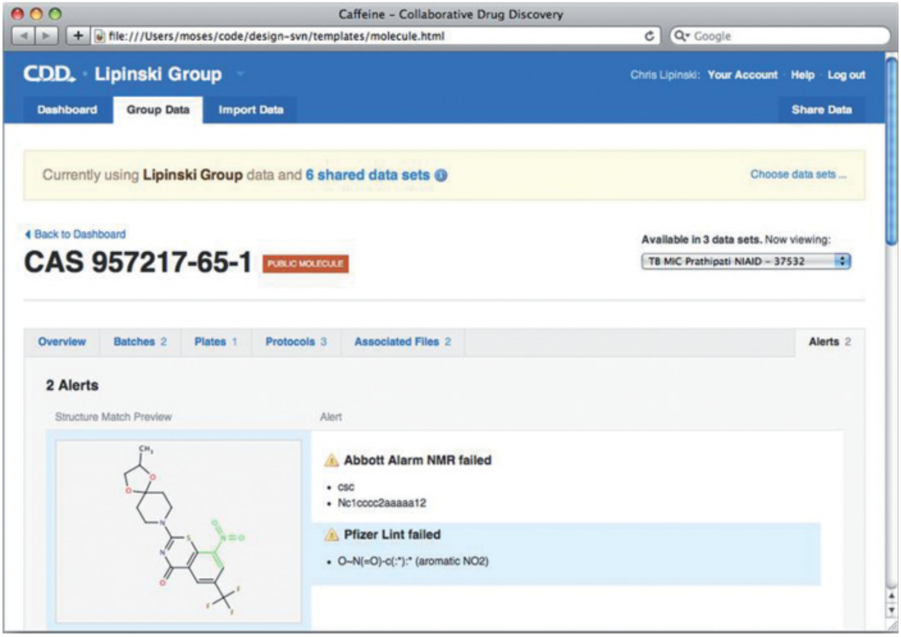
We will create the capacity to archive, mine, and collaborate with generic objects within CDD. We allow researchers to customize the database because they will be able to change the Molecule field to, say, a Sequence field. Simple renaming is not enough, so additional details will be engineered into domain-specific modules. We are aware of at least one pharmaceutical company that has developed a macromolecular structure notation, editing, and registration tool (Tianhong Zhang, personal communication, 2009) using the ChemAxon components (Marvin Sketcher, Marvin viewer, calculator plugin and Marvin API already integrated within CDD). The pharma application is currently unavailable to researchers in academia, other foundations, or companies and represents a significant investment.

21.6.1.3 Develop Alerts for Toxic/Reactive Functionality in Small and Large Molecules Hits or leads in rare, orphan, and neglected diseases can arise from phenotypic or mechanistic screening against commercially available libraries. Often these screening efforts arise in an academic setting. However, because of the disconnect between academic biology and expert medicinal chemistry, it is essential to carry out a time-consuming medicinal chemistry annotation of putative hits or leads before expenditure of significant drug discovery effort [18, 51]. Many companies have instituted filters (usually SMARTS [SMiles ARbitrary Target Specification] queries) to remove undesirable molecules, false positives, and frequent hitters from their HTS screening libraries or to filter vendor compounds. Examples include REOS from Vertex [54], filters from GSK [55], BMS [56], and Abbott [57–59]. An academic group in Australia has developed an extensive series of substructural features for removal of pan assay interference compounds (PAINS) from screening libraries [60]. There is as yet no coordinated or readily accessible automated method for filtering compounds or alerting users to reactivity issues or for that matter bringing the expertise of many medicinal chemists into a piece of software or database that would identify undesirable molecules for biologists. There is considerable need to influence the quality of hits and leads in public databases and prevent experimental repetition. An analogous approach could be applied to help optimize macromolecular properties; for example, immunogenicity can be an important obstacle to successful protein drug therapy as antibodies to a large-molecule drug may impact therapeutic function or pharmacokinetics or lead to severe undesirable adverse effects in vivo [59, 60]. Various *in silico* tools and databases (e.g., Immune Epitope Database (IEDB) [61] as well as commercial tools [62]) to identify potential for immunogenicity are available, representing an alternative to *in vitro* or *in vivo* immunogenicity assays [63].

We could integrate the ChemAxon toolkit SMARTS-based alerts and other rule bases to flag (potentially) problematic substructures within a molecule [52–56, 58]. We will use the state-of-the-art filters from references in the previous section and dynamically be able to add additional alerts requested from discussions with experienced medicinal chemists. We could then develop an intuitive alerts display (Fig. 21.8). Additionally, linking to external public databases such as PubChem and patent databases could help provide more useful information on a hit compound that could assist in deciding whether to pursue it.

21.7 DISCUSSION

We have described the development of the CDD database as a case study of how such a tool could be used for collaborations. The tool was developed using an agile development process which uses an integrated design–build–test process. In the space of six years this database has become a viable technology that has attracted many research foundations, academics, biotechs, and large-



Currently using Lipinski Group data and 6 shared data sets

CAS 957217-65-1

Available in 3 data sets. Now viewing: TB MIC Prathipati NIAID – 37532

2 Alerts

Abbott Alarm NMR failed

- csc
- Nc1cccc2aaaaa12

Pfizer Lint failed

- O~N(=O)-c(:*):* (aromatic NO2)

Figure 21.8 Example of highlighted alerts on molecules to avoid wasting money on compounds likely to fail.

pharma customers. In the process we have used it to provide new insights into the vast amounts of screening data being produced [43] as well as facilitate global collaborations [19] and provide a means for collaboration [9, 10, 23, 49]. As we see drug discovery become more reliant on networks of collaborators, we think the need for a cloud-based solution will become dominant.

It is feasible that other cheminformatics software solutions could be provided to CDD users in the same way, either separately or integrated into the current platform. For example, it may be of utility to integrate ADME/Tox models or other quantitative structure–activity relationships (QSARs) [64]. Software developed under the open-source model provides important visibility into the implementation of descriptors and algorithms, so that computational chemists can verify the algorithm and suggest or actually contribute improvements [65]. A number of open-source software packages exist that calculate descriptors [66, 67] or implement modeling algorithms (e.g., R). Some groups have used open descriptors and open modeling algorithms to build QSAR models [67–69] for mutagenicity, cytotoxicity, Caco-2 data, as well as some drug targets. The data sets used to date have been relatively small. While there are some toolkits for cheminformatics and bioinformatics [65, 70–72] as well as proposed Web services [73], no integrated toolkit exists that provides functionality for end-to-end QSAR training, validation, and prediction. We have recently used such open-source software with over 100,000 molecules with ADME data provided by a large pharmaceutical company to show that such models can be equivalent to those generated with commercial tools [64]. This sets the stage for using CDD as a selective ADME model building and sharing platform on the cloud. The rationale for this is that the existing computational ADME/Tox programs are limited in using the same very small data sets from the literature or combining data sets from different groups, which is suboptimal. These data sets also only cover a small region of chemical space, focused on druglike molecules that tend to be compliant with the Rule of Five [74]. Thus, there is a need for building models using data from various pharmaceutical and biotechnology companies and then securely sharing the models with collaborators or groups designated by the user. The advantage of using such data from pharmaceutical and biotech companies is that they have generally screened orders-of-magnitude more data (e.g., tens to hundreds of thousands of compounds under standardized conditions) than is in the public domain and thus have far better coverage of chemistry space. This could result in powerful models that will improve predictions for groups with compounds of interest but no idea of their ADME properties, for example, assisting neglected disease researchers.

New informatics tools that incorporate biology and chemistry with social networking technologies should enable a better, faster, and ultimately cheaper mechanism to discover and advance drug candidates in a collaborative manner, regardless of whether they are for neglected, orphan, or potential "blockbuster" diseases. We have found that many biotechs use CDD as their corporate database as it is cost effective, provides compound registration functions, and can handle their high-throughput screening data while having many other features. Though it is hard to predict the direction this or similar technologies could go, we have presented some ideas which we think are realistic. Challenges to a tool like CDD, which has a foot in the commercial sector while at the same time making data searchable to the community for free, come not only from other cheminformatics or database companies [e.g., Heos from Scynexis (http://www.scynexis.com/research_capabilities/heos_software.asp), which currently does not have both private and public sharing capabilities, and Ensemble from Artus Labs (www.artuslabs.com)] but also from the public–private partnership sector [e.g., the Innovative Medicines Initiative (http://imi.europa.eu/index_en.html), which recently had a call for development of Open Pharmacological Space]. It will be critical to continue to expand the CDD user community and integrate with other open and proprietary tools such as workflow software while at the same time showing demonstrable success in improving drug discovery by assisting in collaborations, speeding up the process, and finding new hits and leads. A pure database alone will only facilitate such results and ultimately it may come down to how it is used and how well it is exploited by the user. This will require integration of tools that can enhance the user experience and build on their own expertise. These requirements ultimately take away from the relatively elegant, yet straightforward and simple-to-use experience, so it will be important not to lose sight of this.

Collaborative R&D is undoubtedly the future of biomedical research and will require computational tools similar to CDD (and perhaps beyond) which are provided on the cloud. There is certainly plenty of opportunity and competition in the field that will likely result in some consolidation opportunities. One could also envisage that systems biology software tools (many of which are already cloud based) and pure bioinformatics software (e.g., NextBio) start to move more in the direction of chemistry databases, allowing different modes of data sharing, much as CDD has pioneered. It will also be important that such software be fully functional on any kind of device, for example, mobile phone, tablet, or other newer hardware from different manufacturers.

ACKNOWLEDGMENTS

The authors gratefully acknowledge our colleagues Justin Bradford, Krishna Dole, Anna Spektor, Kellan Gregory, David Blondeau, Sylvia Ernst, Orion Jankowski, David Ephron, Alph Bingham, Tony Williams (RSC), Chris Waller (Pfizer), Eric Gifford (Pfizer), and Rishi Gupta (Pfizer) as well as Dr. Chris Lipinski and the many researchers in the CDD community.

The CDD TB database is funded by the Bill and Melinda Gates Foundation (Grant No.49852, “Collaborative Drug Discovery for TB Through a Novel Database of SAR Data Optimized to Promote Data Archiving and Sharing”).

REFERENCES

1. Smalheiser NR, Perkins GA, Jones S. Guidelines for negotiating scientific collaboration. *PLoS Biolo* 2005;3:e217.
2. Vicens Q, Bourne PE. Ten simple rules for a successful collaboration. *PLoS Comput Biol* 2007;3:e44.
3. Hruby VJ. Organic chemistry and biology: Chemical biology through the eyes of collaboration. *J. Organic Chem* 2009;74:9245–9264.
4. Nwaka S, Ridley RG. Virtual drug discovery and development for neglected diseases through public-private partnerships. *Nat Rev Drug Discov* 2003;2:919–928.
5. Hunter AJ. The Innovative Medicines Initiative: A pre-competitive initiative to enhance the biomedical science base of Europe to expedite the development of new medicines for patients. *Drug Discov Today* 2008;13:371–373.
6. Barnes MR, et al. Lowering industry firewalls: Pre-competitive informatics initiatives in drug discovery. *Nat Rev Drug Discov* 2009;8:701–708.
7. Ekins S, Williams AJ. Precompetitive preclinical ADME/Tox Data: Set it free on the web to facilitate computational model building to assist drug development. *Lab on a Chip*, 2010;10:13–22.
8. Shah S, Federoff HJ. Drug discovery dilemma and Cura quartet collaboration. *Drug Discov Today* 2009;14:1006–1010.
9. Bingham A, Ekins S. Competitive collaboration in the pharmaceutical and biotechnology industry. *Drug Discov Today* 2009;14:1079–1081.

10. Ekins S, Williams AJ. Reaching out to collaborators: Crowdsourcing for pharmaceutical research. *Pharm Res* 2010;27:393–395.
11. Hunter J, Stephens S. Is open innovation the way forward for big pharma? *Nat Rev Drug Discov* 2010;9:87–88.
12. Melese T, Lin SM, Chang JL, Cohen NH. Open innovation networks between academia and industry: An imperative for breakthrough therapies. *Nat Med* 2009; 15:502–507.
13. Talaga P. Open innovation: Share or die. *Drug Discov Today* 2009;14:1003–1005.
14. Carpy AJ, Marchand-Geneste N. Structural e-bioinformatics and drug design. *SAR and QSAR Environ Res* 2006;17:1–10.
15. Ertl P, Jelfs S. Designing drugs on the internet? Free web tools and services supporting medicinal chemistry. *Current topics Med Chem* 2007;7:1491–1501.
16. Munos B. Can open-source R&D reinvigorate drug research? *Nat Rev* 2006;5: 723–729.
17. Ekins S, Iyer M, Krasowski MD, Kharasch ED. Molecular characterization of CYP2B6 substrates. *Current Drug Metab* 2008;9:363–373.
18. Ekins S, et al. Computational discovery of novel low micromolar human pregnane X receptor antagonists. *Mol Pharmacol* 2008;74:662–672.
19. Hohman M, Gregory K, Chibale K, Smith PJ, Ekins S, Bunin B. Novel web-based tools combining chemistry informatics, biology and social networks for drug discovery. *Drug Discov Today* 2009;14:261–270.
20. Lee S, Bozeman B. The impact of research collaboration on scientific productivity. *Social Stud Sci* 2005;35:673–702.
21. Katsouyanni K. Collaborative research: Accomplishments & potential. *Environ Health* 2008;7:3.
22. Guimera R, Uzzi B, Spiro J, Amaral LA. Team assembly mechanisms determine collaboration network structure and team performance. *Science (New York)* 2005;308:697–702.
23. Williams AJ, Tkachenko V, Lipinski C, Tropsha A, Ekins S. Free online resources enabling crowdsourced drug discovery. *Drug Discov World* 2009;10(Winter): 33–38.
24. Williams AJ. Internet-based tools for communication and collaboration in chemistry. *Drug Discov Today* 2008;13:502–506.
25. Williams AJ. A perspective of publicly accessible/open-access chemistry databases. *Drug Discov Today* 2008;13:495–501.
26. Radish J. More medicines for neglected and emerging infectious diseases. *Bull World Health Org* 2007;85:569–648.
27. Grace C. Developing new technologies to address neglected diseases: The role of product development partnerships and advanced market commitments. Department for International Development (DFID) Heath Research Center Report, 2006.
28. Wechsler J. Manufacturers tackle neglected diseases. Available: http://biopharminternationalfindpharmacom/biopharm/Article/Manufacturers-Tackle-Neglected-Diseases/ArticleStandard/Article/detail/439419, 2007.
29. Catteruccia F, Levashina EA. RNAi in the malaria vector, *Anopheles gambiae*. *Methods Mol Biol* 2009;555:63–75.

30. Aronov AM, Balakin KV, Kiselyov A, Varma-O'Brien S, Ekins S. Applications of QSAR methods to ion channels. In Ekins S, Ed. *Computational Toxicology: Risk Assessment for Pharmaceutical and Environmental Chemicals*. Hoboken, NJ: J Wiley, 2007, pp. 353–389.
31. Ekins S, Balakin KV, Savchuk N, Ivanenkov Y. Insights for human ether-a-go-go-related gene potassium channel inhibition using recursive partitioning, Kohonen and Sammon mapping techniques. *J Med Chem* 2006;49:5059–5071.
32. Crumb Jr WJ, et al. Effects of antipsychotic drugs on I_{to}, I_{Na}, I_{sus}, I_{K1}, and hERG: QT prolongation, structure activity relationship, and network analysis. *Pharm Res* 2006;23:1133–1143.
33. Chouteau F, Ramanitrahasimbola D, Rasoanaivo P, Chibale K. Exploiting a basic chemosensitizing pharmacophore hypothesis. Part 1: Synthesis and biological evaluation of novel arylbromide and bicyclic chemosensitizers against drug-resistant malaria parasites. *Bioorg Med Chem Lett* 2005;15:3024–3028.
34. Chiyanzu I, Clarkson C, Smith PJ, Lehman J, Gut J, Rosenthal PJ. Design, synthesis and anti-plasmodial evaluation in vitro of new 4-aminoquinoline isatin derivatives. *Bioorg Med Chem* 2005;13:3249–3261.
35. Weisman JL, Liou AP, Shelat AA, Cohen FE, Guy RK, DeRisi JL. Searching for new antimalarial therapeutics amongst known drugs. *Chem Biol Drug Design* 2006;67:409–416.
36. Gold B, Deng H, Bryk R, Vargas D, Eliezer D, Roberts J, et al. Identification of a copper-binding metallothionein in pathogenic mycobacteria. *Nat Chem Biol* 2008;4:609–616.
37. Ribeiro I, Sevcsik AM, Alves F, Diap G, Don R, Harhay MO, et al. New, improved treatments for Chagas disease: From the R&D pipeline to the patients. *PLoS Neglected Tropical Dis* 2009;3:e484.
38. Shukla AK, Singh BK, Patra S, Dubey VK. Rational approaches for drug designing against leishmaniasis. *Appl Biochem Biotechnol* 2010;160:2208–2218.
39. Nussbaum K, Honek J, Cadmus CM, Efferth T. Trypanosomatid parasites causing neglected diseases. *Curr Med Chem* 2010;17:1594–1617.
40. Zhang Y. The magic bullets and tuberculosis drug targets. *Annu Rev Pharmacol Toxicol* 2005;45:529–564.
41. Balganesh TS, Alzari PM, Cole ST. Rising standards for tuberculosis drug development. *Trends Pharmacol Sci* 2008;29:576–581.
42. Payne DA, Gwynn MN, Holmes DJ, Pompliano DL. Drugs for bad bugs: Confronting the challenges of antibacterial discovery. *Nat Rev Drug Disc* 2007;6:29–40.
43. Ekins S, Bradford J, Dole K, Spektor A, Gregory K, Blondeau D. A collaborative database and computational models for tuberculosis drug discovery. *Mol BioSyst* 2010;6:840–851.
44. Makarov V, Manina G, Mikusova K, Mollmann U, Ryabova O, Saint-Joanis B. Benzothiazinones kill *Mycobacterium tuberculosis* by blocking arabinan synthesis. *Science* 2009;324:801–804.
45. Ekins S, Kaneko T, Lipinksi CA, Bradford J, Dole K, Spektor A. Analysis and hit filtering of a very large library of compounds screened against *Mycobacterium tuberculosis*. *Mol BioSyst* 2010;6:2316–2324.
46. Ekins S, Williams AJ. Meta-analysis of molecular property patterns and filtering of public datasets of antimalarial "hits" and drugs. *MedChemComm* 2010;1:325–330.

47. Ekins S, Williams AJ. When pharmaceutical companies publish large datasets: An abundance of riches or fool's gold? *Drug Discov Today* 2010;15:812–815.

48. Tapscott D, Williams AJ. *Wikinomics: How Mass Collaboration Changes Everything*. New York: Portfolio, 2006.

49. Louise-May S, Bunin B, Ekins S. Towards integrated web-based tools in drug discovery. *Touch Brief Drug Discov* 2009;6:17–21.

50. Southan C, Varkonyi P, Muresan S. Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds. *J Cheminform* 2009;1:10.

51. Ma'ayan A, Jenkins SL, Goldfarb J, Iyengar R. Network analysis of FDA approved drugs and their targets. *Mount Sinai J Med New York* 2007;74:27–32.

52. Shah NH, Jonquet C, Chiang AP, Butte AJ, Chen R, Musen MA. Ontology-driven indexing of public datasets for translational bioinformatics. *BMC Bioinform* 2009;10(Suppl 2):S1.

53. Washington NL, Haendel MA, Mungall CJ, Ashburner M, Westerfield M, Lewis SE. Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol* 2009;7:e1000247.

54. Walters WP, Murcko MA. Prediction of "drug-likeness." *Adv Drug Del Rev* 2002;54:255–271.

55. Hann M, Hudson B, Lewell X, Lifely R, Miller L, Ramsden N. Strategic pooling of compounds for high-throughput screening. *J Chem Inf Comput Sci* 1999;39: 897–902.

56. Pearce BC, Sofia MJ, Good AC, Drexler DM, Stock DA. An empirical process for the design of high-throughput screening deck filters. *J Chem Inf Model* 2006; 46:1060–1068.

57. Huth JR, Mendoza R, Olejniczak ET, Johnson RW, Cothron DA, Liu Y. ALARM NMR: A rapid and robust experimental method to detect reactive false positives in biochemical screens. *J Am Chem Soc* 2005;127:217–224.

58. Huth JR, Song D, Mendoza RR, Black-Schaefer CL, Mack JC, Dorwin SA. Toxicological evaluation of thiol-reactive compounds identified using a la assay to detect reactive molecules by nuclear magnetic resonance. *Chem Res Toxicol* 2007; 20:1752–1759.

59. Metz JT, Huth JR, Hajduk PJ. Enhancement of chemical rules for predicting compound reactivity towards protein thiol groups. *J Comput Aided Mol Des* 2007;21: 139–144.

60. Baell JB, Holloway GA. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J Med Chem* 2010;53:2719–2740.

61. Bryson CJ, Jones TD, Baker MP. Prediction of immunogenicity of therapeutic proteins: Validity of computational tools. *BioDrugs* 2010;24:1–8.

62. De Groot AS, McMurry J, Moise L. Prediction of immunogenicity: In Silico paradigms, ex vivo and in vivo correlates. *Current Opin Pharmacol* 2008;8:620–626.

63. Tung CW, Ho SY. POPI: Predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties. *Bioinformatics* 2007;23: 942–949.

64. Gupta RR, Gifford EM, Liston T, Waller CL, Bunin B, Ekins S. Using open source computational tools for predicting human metabolic stability and additional ADME/TOX properties. *Drug Metab Dispos* 2010;38:2083–2090.

65. Guha R, Howard MT, Hutchison GR, Murray-Rust P, Rzepa H, Steinbeck C. The Blue Obelisk—Interoperability in chemical informatics. *J Chem Inf Model* 2006; 46:991–998.

66. Sykora VJ, Leahy DE. Chemical Descriptors Library (CDL): A generic, open source software library for chemical informatics. *J Chem Inf Model* 2008;48: 1931–1942.

67. Melville JL, Hirst JD. TMACC: Interpretable correlation descriptors for quantitative structure-activity relationships. *J Chem Inf Model* 2007;47:626–634.

68. Guangli M, Yiyu C. Predicting Caco-2 permeability using support vector machine and chemistry development kit. *J Pharm Pharm Sci* 2006;9:210–221.

69. Guha R. Flexible Web service infrastructure for the development and deployment of predictive models. *J Chem Inf Model* 2008;48:456–464.

70. Spjuth O, Alvarsson J, Berg A, Eklund M, Kuhn S, Masak C. Bioclipse 2: A scriptable integration platform for the life sciences. *BMC bioinform* 2009;10:397.

71. Spjuth O, Helmus T, Willighagen EL, Kuhn S, Eklund M, Wagener J. Bioclipse: An open source workbench for chemo- and bioinformatics. *BMC Bioinform* 2007;8:59.

72. Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, Willighagen EL. Recent developments of the chemistry development kit (CDK)—An open-source java library for chemo- and bioinformatics. *Curr Pharm Des* 2006;12:2111–2120.

73. Dong X, et al. Web service infrastructure for chemoinformatics. *J Chem Inf Model* 2007;47:1303–1307.

74. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Del Rev* 1997;23:3–25.

75. Ballel L, Field RA, Duncan K, Young RJ. New small-molecule synthetic antimycobacterials. *Antimicrob Agents Chemother* 2005;49:2153–2163.

76. Prathipati P, Ma NL, Keller TH. Global Bayesian models for the prioritization of antitubercular agents. *J Chem Inf Model* 2008;48:2362–2370.

77. Maddry JA, et al. Antituberculosis activity of the molecular libraries screening center network library. *Tuberculosis (Edinburgh, Scotland)* 2009;89:354–363.

78. Gupte A, et al. Inhibition of siderophore biosynthesis by 2-triazole substituted analogues of 5′-*O*-[*N*-(salicyl)sulfamoyl]adenosine: Antibacterial nucleosides effective against *Mycobacterium tuberculosis*. *J Med Chem* 2008;51:7495–7507.

79. Ananthan S. High-throughput screening for inhibitors of *Mycobacterium tuberculosis* H37Rv. *Tuberculosis (Edinburgh, Scotland)* 2009;89:334–353.

80. Willand N. Synthetic EthR inhibitors boost antituberculous activity of ethionamide. *Nat Med* 2009;15:537–544.

81. Sacchettini JC, Rubin EJ, Freundlich JS. Drugs versus bugs: In pursuit of the persistent predator *Mycobacterium tuberculosis*. *Nature Rev* 2008;6:41–52.

82. Guiguemde WA. Chemical genetics of *Plasmodium falciparum*. *Nature* 2010;465:311–315.

83. Gamo F-J. Thousands of chemical starting points for antimalarial lead identification. *Nature* 2010;465:305–310.

22

CHEMSPIDER: A PLATFORM FOR CROWDSOURCED COLLABORATION TO CURATE DATA DERIVED FROM PUBLIC COMPOUND DATABASES

Antony J. Williams

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.

22.1 INTRODUCTION

Accessing information about chemicals distributed across the Internet is, in many ways, too easy. Chemists simply type in the name of a chemical of interest into a search engine and then wade through the results hoping to find a result matching their query. Such approaches are limited to the whims of text-based matching, and it can be very time consuming to wade through pages of results attempting to segregate the various types of information retrieved. Many of these searches will, in any case, retrieve hits from public compound databases having variable quality, from those that manually curate each entry to those that are simply repositories of data. The identification of the chemical structure associated with a particular chemical can be almost intractable, and the quality of data associated with chemical compounds in online databases varies from questionable to valueless. Until recently there has been no real attempt to unify and integrate the public chemistry resources online and only one platform, ChemSpider, is taking on the challenge.

ChemSpider is a free online structure database developed with the intention of aggregating and linking chemical structure–based information and data across the Internet. Containing more than 25 million unique chemical entities and linked out to over 400 data sources, ChemSpider offers the ability to perform *both* text- and structure-based searches to resource information such as chemical vendors, properties, analytical data, patents, publications, and a myriad of other information [1, 2]. While enabling this broad form of searching for chemical data across the Internet, ChemSpider has also assumed a key role in allowing the community to expand and improve the online data by providing a platform for community deposition, annotation, and curation. As a result the ChemSpider website has become a crowdsourcing environment for chemists to expose their own activities to the community and participate in creating the richest single resource for chemistry-related information available online and, in keeping with the nature of the Web, for free. In 2009 the Royal Society of Chemistry (RSC) acquired ChemSpider to fulfill its objective of disseminating knowledge to the chemical community and advancing the chemical sciences. In partnership, RSC and ChemSpider will provide innovative services on a reliable infrastructure to service the chemical community and bring new, exciting opportunities for publishing and cheminformatics that will radically improve the online research environment.

22.2 PUBLIC COMPOUND DATABASES

Over the past few years efforts have been made to deliver "public compound databases" to the community to allow access to data relating to chemical compounds. These databases can contain from a few hundred to tens of millions of chemical structures with associated information and may be focused on drugs, metabolites, or pesticides or simply aggregate repositories of data with no specific focus. Databases built with a specific focus are generally quite small, a few hundred to thousands of compounds only, highly curated, painstakingly assembled, and developed with a particular class of chemists in mind. Data aggregators and repositories are commonly much larger, tens of thousands to millions of compounds, and are holders of data which are likely heavily contaminated with numerous errors and, while easy to search, can commonly deliver misleading results. As a result, the Internet hosts information that is hard to filter, difficult to segregate, and at best challenging to interpret in terms of quality. It is worth reviewing some of the databases available online prior to discussing some of the challenges, advantages, and approaches to linking together chemistry on the Web.

22.2.1 PubChem

The PubChem database [3] was launched by the National Institutes of Health in 2004 as part of a suite of databases to support its roadmap initiative [4]. PubChem archives and organizes information about the biological activities of chemical compounds and is intended to empower the scientific community to use low-molecular-weight chemical compounds in their research. PubChem consists of three databases (PubChem Compound, PubChem Substance, and PubChem BioAssay). As of August 2010 its content is approaching 72 million substances and 29 million unique structures but provides biological property information for only a fraction of these compounds, just over 450,000 in total. PubChem Substance contains records of substances from depositors into the system. These are publishers, chemical vendors, commercial databases, and other sources. It provides descriptions of chemicals and links to PubMed [5], protein three-dimensional (3D) structures, and screening results. PubChem BioAssay contains information about bioassays using specific terms pertinent to the bioassay. PubChem can be searched by alphanumeric text such as chemical names, property ranges, or structure, substructure, or structural similarity.

Such a source of data opens up new possibilities in regards to data mining and extraction and the system has an important role as a central repository for chemical vendors and content providers, enabling evaluation of commercial compound libraries. This saves biomedical researchers from the work associated with gathering and searching commercial databases, and the hit-to-lead decision-making process in drug discovery programs can certainly benefit from the ongoing annotation service provided by PubChem. PubChem is an example of collaboration between chemists and biologists as PubChem itself is only a repository platform for data and the data themselves need to be deposited onto the platform. These data come from national screening centers, chemical vendors, and other databases and, by depositing the data to a central resource pharmaceutical companies, universities and other organizations with an interest in mining, aggregating, and linking the data can download and reuse it. This is highly beneficial to the efforts to link together information, but the data must be treated with caution as there are no quality control processes in place and numerous scientists have commented regarding the quality of the data within PubChem [6–8]. Screening data are less rigorous than those in peer-reviewed articles and contain many false positives [9]. Deposited data are not curated, and so mistakes in structures, identifier units, and other characteristics can and do occur. The author of this chapter has frequently pointed to the accuracy of some of the identifiers associated with the PubChem compounds [10–12], and an example will be given later in this chapter. The problems arise from the quality of submissions from the various data sources. There are thousands of errors in the structure–identifier associations due to this contamination and this can lead to the retrieval of incorrect chemical structures. It is also common to have multiple representations of a single structure due to incomplete or total lack of stereochemistry for a molecule [13].

22.2.2 DrugBank

DrugBank [14] blends both bioinformatics and cheminformatics data and combines detailed drug (i.e., chemical) data with comprehensive drug target (i.e., protein) information. The database contains >4800 drug entries and >2500 protein or drug target sequences that are linked to these drug entries. Each DrugCard entry contains almost 100 data fields, with half of the information being devoted to drug/chemical data and the other half devoted to drug target or protein data. The database is fully searchable, supporting extensive text, sequence, chemical structure, and relational query searches. DrugBank has been used to facilitate in silico drug target discovery, drug design, drug docking or screening, drug metabolism prediction, drug interaction prediction, and general pharmaceutical education.

The group hosting DrugBank also hosts a series of other curated databases: the Human Metabolome Database [15] contains detailed information about small-molecule metabolites found in the human body and is used by scientists working in the areas of metabolomics, clinical chemistry, and biomarker discovery; FoodDB [16] is a comprehensive database providing information on over 1900 food components, the list being taken from the U.S. Food and Drug Administration (FDA) list of everything added to food in the United States. The author of this chapter has reviewed the data within DrugBank, and while efforts have been made to curate the data, there are numerous examples of inaccurate chemical structures associated with particular compounds and a distinct lack of expected stereochemistry for many of the chemical structures [13].

22.2.3 SureChem

SureChem [17] provides chemically intelligent searching of a patent database containing millions of U.S., European, and World patents. Using extraction heuristics to identify chemical and trade names and conversion of the extracted entities to chemical structures using a series of name-to-structure conversion tools, SureChem has delivered a database integrated to nearly 10 million individual chemical structures. The free-access online portal allows scientists to search the system based on structure, substructure, or similarity of structure as well as the text-based searching expected for patent inquiries.

22.2.4 Wikipedia

Wikipedia [18] is an unprecedented success story in the domain of community intellectual contribution and crowdsourcing. For chemistry it represents an important shift in terms of the future access of information associated with small molecules. A wiki is a type of computer software allowing users to easily create, edit, and link Web pages (see also Chapters 5 and 28). A wiki enables documents to be written collaboratively, in a simple markup language using a Web browser, and is essentially a database for creating, browsing, and searching information. For small molecules on Wikipedia each one generally has a drug box or a chemical infobox. The drug box shows a chemical structure, one or more chemical names or identifiers, links out to related resources, chemical and pharmacokinetic data, and therapeutic considerations. At present there are approximately 10,000 articles with a chembox or drugbox and more are added on a regular basis. The detailed information offered on Wikipedia regarding a particular chemical or drug can be excellent [19] or weak in the case of stub articles [20].

There are many dedicated supporters and contributors to the quality of the online resource. This community curation process makes Wikipedia a very important online chemistry resource whose impact will only expand with time. The author of this chapter is part of a dedicated team that has worked on validating and curating Wikipedia chemical compound pages for over two years [21], though this work is never complete, as will be shown later in this chapter. ChemSpider is the only online public compound database that directly provides a mash-up of the Wikipedia article into its compound pages, thereby making Wikipedia structure and substructure searchable via a ChemSpider search.

22.2.5 Community Wikis and Blogs

As described in detail in Chapter 5, an increasing number of scientists have an urge to communicate either their own science or science in general, commonly with the intention of educating others, proliferating data or opinions, or connecting with others for the purpose of collaboration or advice. There is an increasing interest in using Web-based software tools to speed communication. Both wikis and blogs are fast becoming chosen platforms for the exchange of information between many scientists [22].

A blog, or weblog, is a website where entries are written in chronological order and generally provide commentary or news on a particular subject (see

Chapter 5). A typical blog combines text, images and links to other blogs, web pages, and other media related to its topic. The ability for readers to leave comments and interact with the author is an important component of blogs. Similar approaches for comment posting are being adopted by publishers now [23, 24] as well as by the Royal Society of Chemistry in its deployment of ChemSpider SyntheticPages platform [25], a community resource of synthetic procedures (*vide infra*). The number of chemistry-related blogs continues to grow. There are blogs from members of the pharmaceutical industry, the cheminformatics world, the open-source chemistry software world, and other willing participants in the "blogosphere," specifically students. Some of these blogs are very rich in chemistry, for example, Org Prep Daily [26] and TotallySynthetic [27]. In addition, communities such as ResearchGate [28], Science3.0 [29], and Friendfeed communities such as LifeScientists [30] are increasingly becoming active communities for collaboration and communication.

The short list provided above is meant to be representative of the types of resources that are becoming increasingly available online as individuals, researchers, and organizations contribute to the data available via the Internet. These resources, as well as many more available online, are proving to be very valuable for collaboration. In terms of discoverability, chemistry in the form of chemical compounds, and reactions, searches are limited to text-based searches, and it is difficult to source information from a single search of the Internet linking these sources together. In order to provide a unified approach to searching across these multiple diverse resources via a single search engine, ChemSpider was developed.

22.2.6 ChemSpider

ChemSpider [31] was initially developed as a hobby project by this author and a small team of voluntary programmers simply to contribute to the chemistry community. ChemSpider is built primarily on commercial software using a Microsoft technology platform of asp.NET and SQL Server as this allowed ease of implementation and projected longevity and made best use of available skill sets. Following a short development cycle of just a few months ChemSpider was released to the public in March 2007 with the lofty goal of "building a structure centric community for chemists."

The database content, more than 25 million structures from over 400 data sources, has been aggregated as a result of contributions and depositions from chemical vendors, commercial database vendors, government databases, publishers, members of the Open Notebook Science community [32], and individual scientists [33, 34]. The database can be queried using structure/substructure searching and alphanumeric text searching of chemical names and both intrinsic as well as predicted molecular properties. Various searches have been added to the system to cater to various user personae, including mass spectrometrists and medicinal chemists. For example, mass spectrometrists in the field of metabonomics [35] would want to search the database using monoisotopic masses and specific data slices from the queries in order to search for metabolites and mass spectrometry instrument vendors have integrated to ChemSpider in order to query the database directly from the instrument software [36, 37]. Alternatively, a medicinal chemist investigating drug repurposing might want to search for chemicals that demonstrate affinity for binding to a particular target using in silico approaches. By layering on predicted LASSO values [38–41] describing ligand affinity relative to a set of targets, chemists are able to identify potentially active ligands for further analysis and investigation. These and other searches make ChemSpider very flexible in its applications.

22.2.6.1 Structure Quality Issues Following the deposition and aggregation of data from a multitude of data sources, it became obvious that one of the side effects of such an activity was that data of various levels of quality were being merged. The challenge is in distinguishing the quality of data in a particular collection. However, quality is difficult to define as in many cases it is based on assertions, experimentally determined data points, and ultimately the interpretation of data. A recent publication by this author discussed how many natural product chemical structures are incorrectly elucidated using analytical techniques and are initially reported in peer-reviewed publications [42]. When such an analysis is expanded to the analysis of public compound databases containing millions of chemical structures and associated data, the issues are exponentially more complex.

This author has invested many years in examining the primary assertions of structure–identifier relationships in order to produce disambiguation dictionaries which can be utilized for the purpose of entity extraction engines for the purpose of text mining chemistry-related articles and patents. Chemistry is a complex subject and the accurate representation of a chemical structure in an electronic format can be very difficult, especially when these are expected to encapsulate the bonding details of complex bonding systems such as organometallics. However, focusing only on small organic molecules of interest to the life sciences some of the most common issues identified include:

1. Chemical structures that are supposed to contain stereochemistry are commonly drawn without stereo bonds.
2. Chemical structures are drawn with inappropriate valences or with charge imbalance due to the absence of one or more expected counterions.
3. The relationship between a chemical compound and a particular chemical identifier is confused in a number of ways: (a) the name includes a counterion but it is absent; (b) the name defines specific stereochemistry but it is absent or partially present or is the opposite of the name; (c) the chemical names or registry number(s) are simply incorrectly associated; and many other variants.

4. Chemical structures associated with a particular asserted label can have timelines. An originally reported chemical structure for some newly extracted material may be assigned a particular label and enter the historical literature archive. Some period of time later the same compound may be freshly elucidated with newer experimental data and new structural details identified. The same asserted chemical name will now be associated with a new structure. This process can occur many times with the result that a single chemical may have a multitude of associated structures. One particular example is hexacyclinol, a natural product that generated a significant amount of blog discussion and resulted in two structures forever being associated with that chemical name [43–46].

As an example of the challenges of locating the "correct" chemical structure for what should be a well-known and easily locatable chemical compound, we will initiate a search for a well-known vitamin, vitamin K_1, commonly known as phylloquinone. A Google search will direct us to a number of resources and databases utilized by life scientists, and these include Wikipedia [18], PubChem [3], DrugBank [14], Chemical Entities of Biological Interest (ChEBI) [47], DailyMed [48], and ChemSpider [31], to name just a few. A review of the data concisely demonstrates the confusion that can exist online and the quality of available data. Figure 22.1 shows the images of the structures of vitamin K_1 extracted from a number of these databases. PubChem alone lists 10 different structures under the name vitamin K_1. It should be noted that there are differences in the structures shown, specifically in the stereochemistry and the *E*/*Z* orientation of the alkene bond in the phytyl side chain.

Some observations from Figure 22.1 include (1) the Wikipedia article and the Kyoto Encyclopedia of Genes and Genome (KEGG) database record contain no explicit stereochemistry, (2) DrugBank has ambiguous orientation around the alkene bond, and (3) ChEBI and ChemSpider are consistent with

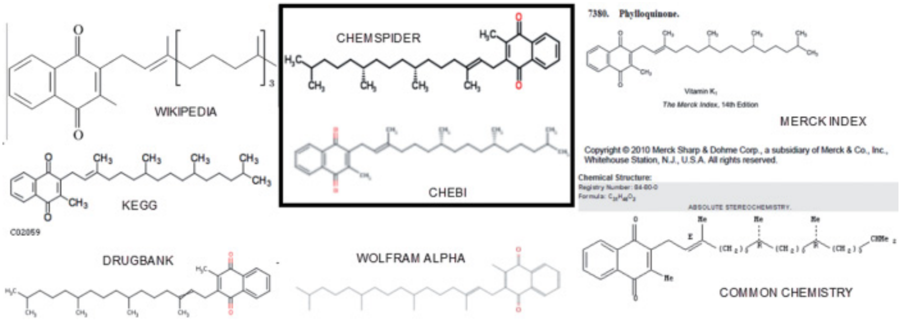

WIKIPEDIA
CHEMSPIDER
CHEBI
KEGG
C02059
DRUGBANK
WOLFRAM ALPHA
7380. Phylloquinone.
Vitamin K_1
The Merck Index, 14th Edition
MERCK INDEX
Copyright © 2010 Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc., Whitehouse Station, N.J., U.S.A. All rights reserved.
Chemical Structure:
Registry Number: 84-80-0
Formula: $C_{31}H_{46}O_2$
ABSOLUTE STEREOCHEMISTRY.
COMMON CHEMISTRY

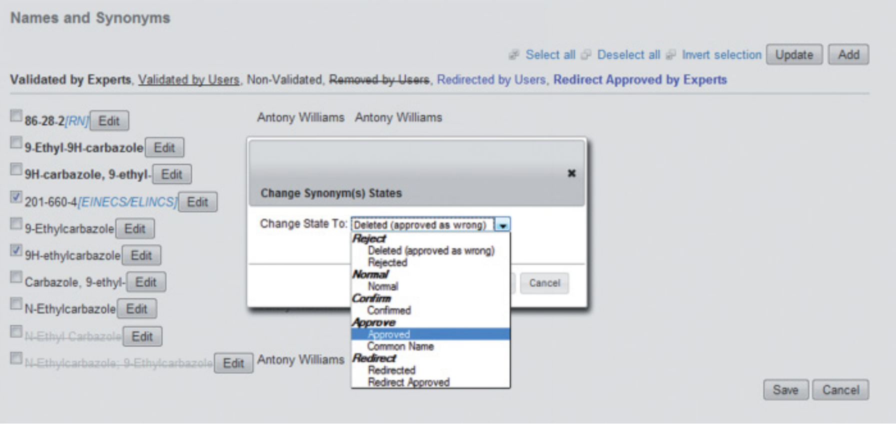
Figure 22.1 Images of chemical structures of vitamin K_1 extracted from series of databases labeled with name of associated database. The asserted structures of vitamin K_1 are surrounded by a bolded box and are consistent with those from the *Merck Index* and *Common Chemistry*.

defined *E* orientation and stereochemistry. Examining the PubChem records gives the following list of chemical names describing the side chain showing the distinct variation and confusion across the set of molecules labeled vitamin K_1 in the PubChem database.

- 2-Methyl-3-[(*E*,7*R*,11*R*)-3,7,11,15-tetramethyl-
- 2-Methyl-3-[(*E*,7*S*,11*R*)-3,7,11,15-tetramethyl-
- 2-Methyl-3-[(*E*,7*R*,11*S*)-3,7,11,15-tetramethyl-
- 2-Methyl-3-[(*E*,7*S*,11*S*)-3,7,11,15-tetramethyl-
- 2-Methyl-3-[(*E*,11*S*)-3,7,11,15-tetramethyl-
- 2-Methyl-3-[(*E*)-3,7,11,15-tetramethyl-
- 2-Methyl-3-(3,7,11,15-tetramethyl-
- 2-Methyl-3-[(*E*)-3,7,11,15-tetramethyl-

Attempting to declare the correct structure for vitamin K_1 from the data extracted from this series of public domain databases is clearly challenging. Two of the most highly respected and curated collections of structure-based data are the Merck Index [49] and the CAS Registry [50] from the Chemical Abstracts Service (CAS). CAS has made a subset of its tens of millions of structures available online as the Common Chemistry [51] public collection and fortunately vitamin K_1 is available in the database [52] as well as in the Merck Index. Figure 22.1 shows the structures from both sources, and they are consistent with an *E* orientation and *R*,*R* stereochemistry in the side chain. Of the public domain databases listed, only ChEBI [53] and ChemSpider [54] are consistent with the Common Chemistry and Merck Index structures. We will *assert* from these data that vitamin K_1 is the structure listed in ChEBI, ChemSpider, Common Chemistry, and the Merck Index. The errors found in Wikipedia, PubChem, KEGG, and DrugBank are representative of the quality of data online. This situation has been further exemplified in searches for the chemical structures of Taxol [55], vancomycin [56], and domoic acid [57]. While the structures of each of these compounds have now been ascertained and validated on ChemSpider, the structure of digitonin remains an issue and community involvement has failed to assert the structure as yet [58]. Public compound databases on the Internet are mixed quality and, in all cases, require ongoing validation. The reader should at all times use caution and not take the structures at face value from the Web. The validation of assertion-based data such as chemical names and identifiers contained within a chemical database is challenging enough. The validation of experimentally determined data for millions of compounds is essentially an impossible task without remeasurement.

With an intention to provide a trusted resource there was a clear and obvious need for data curation and validation of the data imported to ChemSpider. Since the ChemSpider team was both small and voluntary, there was no easy manner by which to perform data validation without engaging the



Names and Synonyms

Validated by Experts, Validated by Users, Non-Validated, ~~Removed by Users~~, Redirected by Users, Redirect Approved by Experts

□ 86-28-2[RN] Edit Antony Williams Antony Williams
□ 9-Ethyl-9H-carbazole Edit
□ 9H-carbazole, 9-ethyl- Edit
☑ 201-660-4[EINECS/ELINCS] Edit
□ 9-Ethylcarbazole Edit
☑ 9H-ethylcarbazole Edit
□ Carbazole, 9-ethyl- Edit
□ N-Ethylcarbazole Edit
□ ~~N-Ethyl Carbazole~~ Edit
□ ~~N-Ethylcarbazole; 9-Ethylcarbazole~~ Edit Antony Williams

Change Synonym(s) States

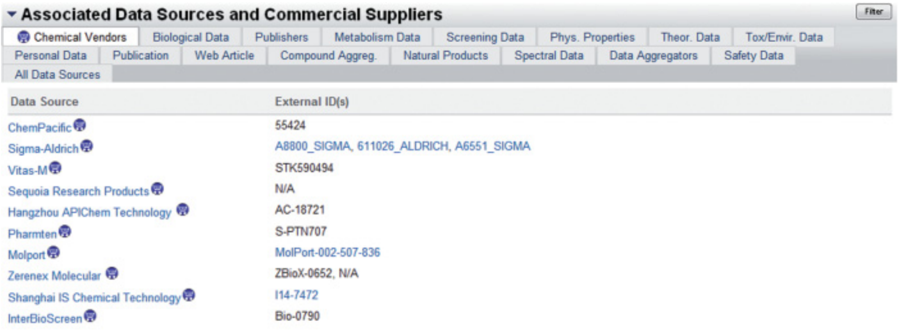
Change State To: Deleted (approved as wrong)

Figure 22.2 Curation interface for editing chemical identifiers associated with structure. Chemical identifiers can be added, deleted, and validated by any user. Master curators have additional curation capabilities.

community directly with a request to provide crowdsourced support of the project. A project was therefore undertaken to enable real-time curation of the data by providing a simple-to-use interface for adding, removing, and validating chemical identifiers associated with the chemical structures (see Fig. 22.2). In parallel with the community-based curation efforts, rules-based validation of the data was also undertaken and has resulted in the removal of hundreds of thousands of incorrect identifiers and the creation of a large validated name-to structure dictionary containing well over a million identifiers. Such a validated dictionary can be important to providing high precision for chemical name entity extraction, as reported by Hettne et al. [59].

Following the addition of community-based curation, facilities were then added to enable further annotation and expansion of the data. Features were added to allow real-time deposition of single or batches of chemical structures, transaction-based predictions of physicochemical data, and the deposition of analytical data associated with chemical structures, discussed in further detail below.

22.2.6.2 Data Sources Data on ChemSpider can be deposited by individuals or by organizations. Data sets can be limited to a single chemical compound deposited by a user simply to "register" it and receive a ChemSpider ID, or it can be a single compound with accompanying spectral data, a list of publications, measured experimental properties, and a set of chemical identifiers or a data collection (tens to millions of compounds) with links to other online resources. ChemSpider is a flexible host for data. All chemical compounds, whether singletons or collections, have a series of properties extracted or generated automatically at deposition. These include the molecular formula,



Data Source	External ID(s)
ChemPacific	55424
Sigma-Aldrich	A8800_SIGMA, 611026_ALDRICH, A6551_SIGMA
Vitas-M	STK590494
Sequoia Research Products	N/A
Hangzhou APIChem Technology	AC-18721
Pharmten	S-PTN707
Molport	MolPort-002-507-836
Zerenex Molecular	ZBioX-0652, N/A
Shanghai IS Chemical Technology	I14-7472
InterBioScreen	Bio-0790

Figure 22.3 Data sources can be segregated according to particular data slices at deposition. Each chemical record displays the types of associated data sources under a set of tabs. Where possible the external IDs associated with each data source are linked out to the website associated with the data source. The tabs displayed are associated with the chemical record for alprazolam [http://www.chemspider.com/2034].

molecular weight, nominal, average, and monoisotopic masses, isomeric simplified molecular input line entry specification (SMILES) string [60], InChIString, and InChiKey [61]. The chemicals are also passed through two separate property prediction suites, the ACD/Labs [62] and EPISuite [63] programs, to generate physicochemical properties such as log *P*, boiling point, and many others. While not exhaustive in terms of the general applicability of the algorithms to the entire structure space, for small organic molecules these algorithms generally provide excellent predicted values that can be used as fair estimates and good filters during searching that may be useful for understanding "lead or druglike" properties. As a result of the added value provided to every chemical record at deposition, every record can be a rich contributor to the overall data set that can then be used by scientists as they see fit.

Many of the online databases are focused in particular areas. For example, the Human Metabolome Database [15] is concerned specifically with molecule metabolites found in the human body while a number of the depositions are those of chemical vendors listed with an intention to provide access to their compound collection to the community and generate business opportunities. ChemSpider has segregated deposited data into slices using labels including chemical vendors, biological data, metabolism data, natural products, and so on, as shown in Figure 22.3. At the time of deposition of the data associated with a data source the particular segregation flags are defined by the data provider. There are no limits to the number of data slices with which a particular depositor can be associated. In the interface viewed by the user the data sources are displayed under tabs. The relevant data sources are listed together with the associated external identifier (generally a number) and, where possible, with a direct Uniform Resource Locator (URL) linking the external ID into the data source website. In this way a user can quickly navigate to information hosted on external sites, a feature which is particularly important when trying to source chemicals for purchase or garner additional information from a particular source.

22.2.6.3 Chemical Identifiers Chemical identifiers associated with chemical entities can include systematic names [generated using the International Union of Pure and Applied Chemistry (IUPAC) or other nomenclatures], trivial names, trade names, Chemical Abstracts registry numbers, international registration numbers, or database identifiers. Systematic, trivial, and trade names can be multilingual. As a result of the various series of identifiers which could possibly exist, there can be tens to hundreds of identifiers associated with just a single chemical entity. Chemical names in public databases are often of dubious quality and can often be ambiguous. For example, dichlorobenzene can be consistent with a dichloro-substituted benzene moiety, but because the positional substitution is not specified, the name is ambiguous. Many online databases are only available for text-based searching and chemical name–based searches are therefore used regularly. Since a chemical entity can be named in various ways, disambiguation dictionaries can lead to more complete result sets. A similar approach is used in Wikipedia for searching. For example, thalidomide is a well-known drug due to its well-publicized teratogenic side effects [64]. It exists under a number of trade names, including contergan and softenon, and searching on these names will produce the same result in Wikipedia of displaying the thalidomide Wikipedia page. The production of high-quality validated disambiguation dictionaries associated with the millions of chemical entities on ChemSpider has been one of the most successful aspects of the project and has produced a validated list of well over a million validated identifiers. The primary utility of these validated identifiers is then to use them as queries against one or more application programming interfaces (APIs) such as those for PubMed [65], Google Scholar [66], Google Patents [67], and the RSC Publishing platform [68] in order to retrieve hit lists from the queries. In this manner a single chemical record on ChemSpider will return hits from each of the platforms based on a query set from a validated disambiguation dictionary and, in general, provide a more complete result set than would be obtained with any single text query [69].

The production of a validated dictionary of chemical identifiers associated with the ChemSpider structure set has been produced using a combination of both robotic and manual curation. Since chemical names are introduced into the database by the deposition of data sets from various sources and with varying quality, it is necessary to apply ongoing filters to remove obvious errors. For example, it is quite common for chemical vendors to include only the primary component in their structure set yet leave the counterion in the chemical name. The result will be a mismatch between the represented chemical and the associated identifier. Many of these are easily recognized and removed using a set of simple filters. These include checking for “chloride” in the name and for “Cl” in the molecular formula and if there is no match remove the identifier. Similar approaches can be taken for many counterions and element lookups in the formula. Other approaches include checking for stereochemistry in the name but absences of stereochemistry in the structures and using name-to-structure conversion tools to convert names to structures and look for ambiguity collisions. Despite these automated approaches being of value for assisting in the validation of millions of identifiers, the most rigorous checks, especially in terms of trade names, are from visual inspection by users of the ChemSpider database and application of online curation tools.

ChemSpider users who wish to assist in curating the data are required to register on the system in order to police for potential vandalism of the data. Curators use intuitive approaches to approve and remove identifiers using a series of simple check boxes. Each such operation produces an e-mail into a centralized master curator inbox for further checking by one or more master curators who can further approve or disallow the suggested validations to the identifiers. A full tracking log of all such edits is maintained on the database. Such curations are made to the database on a daily basis, and the quality of the validated identifier dictionary improves incrementally as a result.

As soon as names are validated, they are used afresh to query against the integrated services associated with a chemical record so that new data will be retrieved from Pubmed, Google patents, Google scholar, and so on. An exemplar of this approach would be that a particular chemical record may have *no* associated hits from Pubmed initially, but approval of one or more identifiers would then trigger a lookup against the appropriate Web service and immediately retrieve a related hit list. There are risks with these approaches in that different chemicals can have the same associated identifiers and users should be cautious and check the associated data. This case is particularly challenging for abbreviations though procedures have been instituted to limit such issues as best as possible. The integration to search against external resources using identifiers will be discussed in further detail later in this chapter.

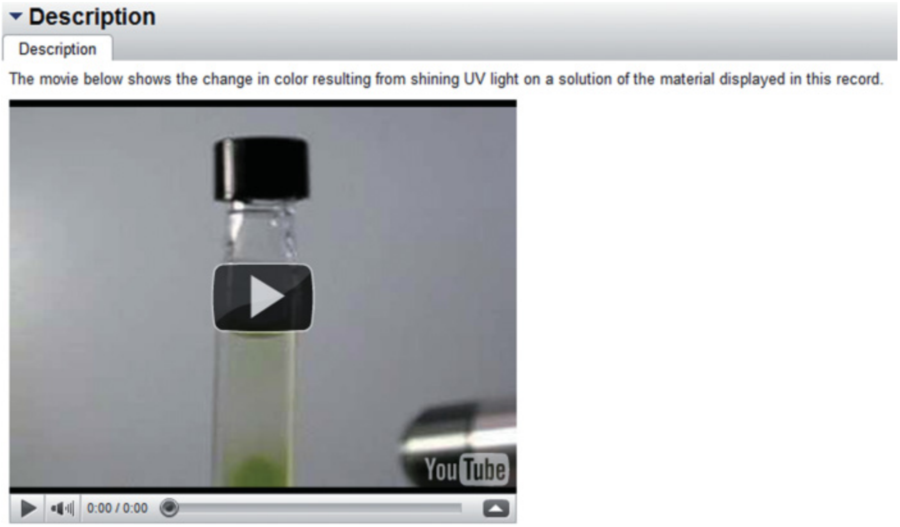
22.2.6.4 Physicochemical Data Physicochemical data play a defining role in the activity of chemical compounds through properties such as log *P*, log *D*, and aqueous solubility, to name only a few. The pharmaceutical industry uses such properties in their *in silico* screening approaches via the judicious application of the Lipinski Rule of Five [70] and other such filters. When such physicochemical data can be sourced as experimental data from databases, they are captured and listed against the chemical records. Where possible links are retained to the original sources of the data so that they can be investigated should there be any questions regarding the validity of the data.

The majority of the ChemSpider database does not have such properties measured and prediction algorithms are therefore used to predict them. The list of predicted properties includes boiling point, flash point, log *P*, log *D* (at two physiological pHs), number of rotatable bonds, number of proton donors, number of proton acceptors, and other related properties. The ability to search the entire database using such properties as filters has been enabled, and this is an excellent way to narrow a particular structure set from a query when, for example, a medicinal chemist may be investigating a particular area of structure space.

22.2.6.5 Analytical Data The value of analytical data is as reference data for comparing against other lab-generated data. Acquisition of a spectrum and comparison against a validated reference spectrum speeds up the process of sample verification without the arduous process of full data analysis. As a result of this general utility, ChemSpider has provided the ability to upload spectral data of various forms against a chemical record such that an individual chemical can have an aggregated set of analytical data to assist in structure verification. When uploaded to the database the depositor can choose whether or not to make the data open [71] and the majority of data have been deposited in this manner and are therefore available for download and reuse without restriction. As a result of contributions from scientists supporting the vision of ChemSpider as a valuable centralizing community-based resource for chemical data for chemists, over 2000 spectra have been added to ChemSpider in the past two years with additional data being added regularly [72]. These data include infrared, Raman, mass spectrometric, and nuclear magnetic resonance (NMR) spectra with the majority being ^{1}H and ^{13}C spectra. Spectral data can be submitted in JCAMP format [73] and displayed in an interactive applet [74] allowing zooming and expansion. Spectra can also be uploaded in various image formats so that two-dimensional (2D) NMR spectra can also be deposited to the database. These spectra are the foundation data for the development of a spectral game to assist in the teaching of NMR spectral interpretation [75].

22.2.6.6 Multimedia: Images, Videos, and Sound Files The vast majority of Web traffic today is consumed by the flow of data associated with image and video files as the Web continues to become a large multimedia distribution network. Chemistry ideally lends itself to multimedia content as it is rich in color, characters, and experimental details which can best be communicated through imagery and sound. ChemSpider has multimedia support via standard embedded content such as YouTube videos [76] and the hosting of sound files and images. Examples in the database include the impact of shining a laser on a colorless solution [77] (see Fig. 22.4) as well as entertaining interviews and videos regarding titanium [78].

22.2.6.7 Linking Scientific Literature and Online Resources The linking of scientific articles regarding a particular compound (or in some cases related compound) is clearly of value to a scientist searching for related information about that chemical. Through its curation and annotation layer, ChemSpider provides the ability for a user to associate one or more articles by linking directly to a particular URL, entering a PubMed identifier (PMID), and using the Pubmed programming interface [79] to automatically form the link or a digital object identifier (DOI) [80] and the Crossref resolver [81] to link



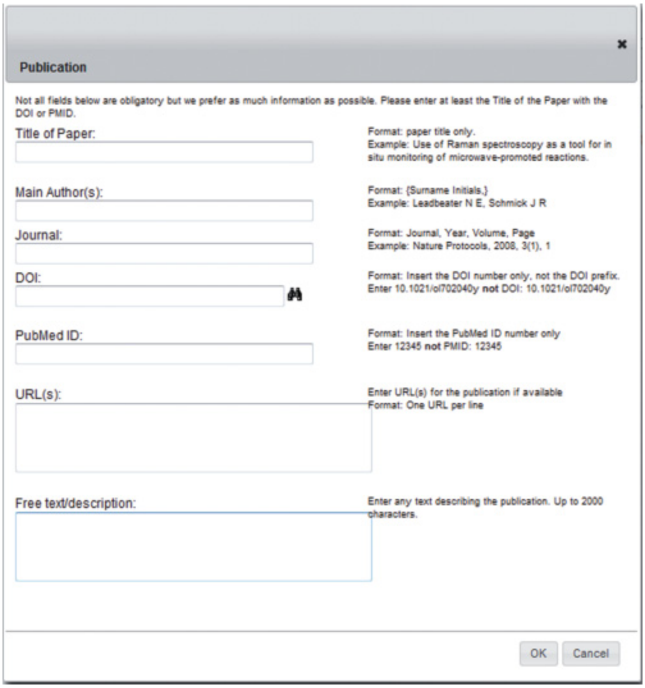
The movie below shows the change in color resulting from shining UV light on a solution of the material displayed in this record.

Figure 22.4 ChemSpider is a multimedia container supporting photographic images, MP3 files, and YouTube videos. An embedded video showing fast photochromism is displayed. The video is associated with the ChemSpider record http://www.chemspider.com/21230378#description.

directly to a publication. For those articles that cannot be linked using a Pubmed ID or DOI, a user can simply add the full reference details (see Fig. 22.5) and the data will be viewable directly in the record.

When ChemSpider joined RSC [82], a path was initiated to integrate RSC content into the database. RSC had previously developed an award-winning semantic markup project known as Project Prospect [83]. "Prospected" articles incorporate standard metadata within the full text of their articles and combine this with an intuitive on-screen manifestation of the advantages of including these metadata. For chemists this translates to a number of features, including the display of compound pages showing the chemical structure, various identifiers, and links to other online resources when hovering over a chemical name. Chemical structures which are prospected in the articles are now deposited directly into ChemSpider on an ongoing basis together with a direct link back to the associated article. This makes RSC articles more discoverable and provides direct benefits to the reader of the article as the compounds are linked into the ChemSpider database, thereby opening up access to an expansive set of data and links across the Internet.

Literature linking has been established in a more automated fashion taking advantage of freely available APIs and the ongoing curation work underway on the database to produce a validated dictionary of chemical names associated with the chemical records. Validated chemical names are used as the basis of a search against the Pubmed database searching *only* against the title and the abstract. In this way a search on cholesterol, for example, would only



Publication

Not all fields below are obligatory but we prefer as much information as possible. Please enter at least the Title of the Paper with the DOI or PMID.

Title of Paper:

Format: paper title only.
Example: Use of Raman spectroscopy as a tool for in situ monitoring of microwave-promoted reactions.

Main Author(s):

Format: {Surname Initials,}
Example: Leadbeater N E, Schmick J R

Journal:

Format: Journal, Year, Volume, Page
Example: Nature Protocols, 2008, 3(1), 1

DOI:

Format: Insert the DOI number only, not the DOI prefix.
Enter 10.1021/ol702040y **not** DOI: 10.1021/ol702040y

PubMed ID:

Format: Insert the PubMed ID number only
Enter 12345 **not** PMID: 12345

URL(s):

Enter URL(s) for the publication if available
Format: One URL per line

Free text/description:

Enter any text describing the publication. Up to 2000 characters.

Figure 22.5 Publications can be associated with a chemical record on ChemSpider via a DOI or PubMed ID. Alternatively, the data can be input directly into the input screen shown.

retrieve those articles with cholesterol in the title and abstract rather than the many tens of thousands of articles likely mentioning cholesterol in the body of the article. A similar approach has been taken to integrate to Google Scholar. This approach has been shown to be an effective manner to perform such an integration but is not without problems, especially when chemical compounds have trade names that have been validated but, unfortunately, are common English language words. Examples are “Advantage” for the chemical imidacloprid [84] that will return a number of false articles. Such issues are few and far between, however, and the approach does provide value and encourages participation of the community to continue to assist in the validation of chemical identifiers.

The approach of integrating to APIs on various websites to search *validated* chemical identifiers (systematic names, trivial names, registry numbers, etc.) using a text query provides access to fast and efficient searches providing direct links to the relevant data contained in the various databases. A chemist can now draw a structure on ChemSpider and retrieve books, articles, and patents served up by the world’s most well known search engine in just a couple of seconds by performing searches against Google Scholar, Google Patents, and Google Books. Most importantly the access to all of these data is free.

22.2.6.8 Patents The linking of chemical records to patents offers scientists direct access to information that may be of value to them in the scope of their investigations of a chemical. As well as the ability to provide a direct link to a patent via a URL, ChemSpider has been linked to the SureChem database [17] using their public programming interface against the database. An InChIKey [61] for the purpose of lookup retrieves a list of the first three associated patents and indicates the number of total records in their patent database (covering U.S., European, Japanese, and World patents) that can be accessed at their portal. The more indirect but nevertheless valuable approach of searching the Google patent database using validated chemical identifier lookup has also been implemented. This approach presently only provides access to U.S. based patents.

22.2.6.9 ChemSpider SyntheticPages What ChemSpider is to the delivery of information and data for chemical compounds, ChemSpider Synthetic Pages (CS|SP) [25] intends to provide to chemists in regards to reaction syntheses. There is one caveat however—the community is fully responsible for populating each record in the database as CS|SP is primarily a publishing platform for chemists. While there are many commercial reaction databases [85–87], there is no free database of synthetic routes that the community can comment on, populate, and expand. CS|SP is a derivative work of the original SyntheticPages project [88]. In a joint collaboration RSC–ChemSpider and the SyntheticPages team have delivered a new architecture for the hosting of synthesis procedures and enhanced the original data model such that the platform can now host multimedia content and spectral data and allow semantic markup and linking to the ChemSpider database and most importantly enhanced capabilities for the deposition of synthesis procedures and data by members of the community.

CS|SP is envisaged to be a manner by which chemists, and students specifically, can grow a professional online reputation for themselves as synthetic chemists. Each SyntheticPage has a single author, the chemist who performed the synthesis. The laboratory head or supervisor is credited via the association of the synthesis with a particular research group. Following submission a SyntheticPage proceeds through a review process by one or more members of the editorial board made up of five academic synthetic chemists [89]. Feedback is provided to the author if necessary and edits can be made online. When the SyntheticPage is published, the community can then provide direct feedback in terms of additional questions, comments regarding their own experiences of repeating the synthesis alternatives to the reported synthesis, and so on. In this way this community research becomes an engaging dialog between synthetic chemists as well as is representative of their skills and activities. Each SyntheticPage receives a DOI [80] and makes a valuable addition to a resume. As of August 2010 the database presently hosts almost 400 synthetic procedures with new submissions being made on a regular basis. The intention is to engage the community to participate in the further development of this rich resource for chemistry. At present this database is an additional resource for review for those accessing commercial databases but, with only 400 syntheses at present, it is far smaller than the commercial systems containing tens of thousands of reactions and is not a threat to the commercial systems as of yet.

22.3 FUTURE OF ONLINE CHEMISTRY RESOURCES

The expansion in scope, capability, and importance of the Internet as a source of scientific information, data, and contributions continues unabated. More scientists (accustomed to the Internet open-source model) are demanding free and open access to literature, patents, data, and algorithms. The open-source [90] model for software now underpinning a growing digital culture continues to flourish, and existing companies will need to reinvent themselves as participants within this changing industry or be relegated to lost leaders. Similarly, existing businesses generating revenues from chemistry databases likely perceive a risk from the increasing availability of free and open data online for scientists and chem/bioinformaticians to mash up into their in-house solutions. However, the primary advantage of commercial databases is that they have been in most cases manually curated, addressing the tedious task of quality data checking. The aggregation of data from multiple sources, both historical and modern, from multiple countries and languages and from sources not available electronically, offers greater coverage than what is available *via* an Internet search. However, how long will this remain an issue and when will the data available electronically, for free, offer a sufficient return on investment to start to negatively impact the commercial chemical database suppliers? Internet queries are increasingly favored by scientists and the chemistry community is likely to reap increasing benefits from the growing number of free-access services and content databases. Academics in particular are likely to have an increased focus on the use of free access databases. This will be further exaggerated in third world countries where free-access systems are the primary resources for information since commercial offerings have significant price barriers.

Librarians are believed to be retiring their print collections in favor of electronic repositories of chemical journals. Internet search engines are increasingly likely to be the first port of call for the majority of scientists for three simple reasons—they are fast, they are free, and they are available anywhere as the user is not tied to a physical location for the library. In terms of data quality issues, the Internet generation has already demonstrated a willingness to curate and enhance the quality of content as modeled by both Wikipedia and ChemSpider. With the improvements promised by the Semantic Web, if there are data of interest to be found, the search engines will facilitate it.

As discussed in more detail in Chapter 28, soon smart phones will become "genius phones" or tablets [91] and there will be an increasing number of mobile computing applications which will only further increase accessibility to information (See Chapter 28). Access to appropriate scientific databases via a hand-held device is already available and will increase in coverage [92, 93]. The promise of the Semantic Web will soon be delivered and an increasing number of public databases will become available. Integrated access to these data will be delivered soon after as resources such as ChemSpider mesh them into their services.

ChemSpider will likely continue to grow in importance as one of the primary free chemistry portals on the Internet. The number of compounds will continue to grow daily as additional publishers choose to participate in contributing to free structure-based discoverability by exposing their data. ChemSpider will expand from handling explicit chemical compounds to the support of compounds that cannot be represented by a specific connection table. As a result support will improve for organometallics, polymers, minerals, and other ambiguous compounds and generally expanding the coverage for this Internet portal for chemistry.

22.4 CONCLUSION

ChemSpider is probably one of the most successful examples of a project initiated by a small group of experts to address perceived issues with the integration and assimilation of masses of public data. As a result of out-of-the-box thinking and utilizing minimal resources other than intellect, willpower, and commitment to solve the problem, a small team innovated a solution to build a structure-centric database. Such a model could be readily applied elsewhere as an example of community collaboration for the benefit of all.

During the development of ChemSpider we were disparaged in the blogosphere [94] and had to respond accordingly [95]. Despite numerous grant applications to source funding to support ChemSpider development, we were unsuccessful. Nevertheless, ChemSpider flourished and received the active support of a community of users, depositors, and curators such that the system was soon responding to over 100,000 transactions per day for a system hosted out of a basement on a skeleton platform of three servers, two of them hand built, and via standard cable Internet. Now owned and hosted by the Royal Society of Chemistry, ChemSpider continues on its mission to provide one of the central Internet portals for chemistry, providing access to millions of chemical structures integrated to hundreds of online data sources. The true collaborative benefits of platforms such as ChemSpider will be felt as the multitude of online resources are integrated into federated searches and Semantic Web linking in a manner that single queries can be distributed across the myriad of resources to provide answers through a single interface. As these systems are established, the quality of results returned will become more important to reduce the sense of overwhelm and a few trusted resources will naturally become recognized for the quality of data provided. The active engagement of the community to provide crowdsourced filtering and validation of the data will likely establish such resources as the primary trusted platforms as is already in place for Wikipedia. With the recent shift in the life sciences to contribute data to the public domain [96–98] (see Chapters 5 and 21), we are likely to see in the very near future that they will provide additional precompetitive data [99] to systems that can underpin the development of federated systems. These in turn will allow pharmaceutical companies to link data across the abundance of life science databases that are already and will increasingly become available. As these data are made available to the community, we will see increasing usage of such information-rich resources to be used for quantitative structure–activity relationship (QSAR) modeling purposes, as has already been exemplified previously by Ekins et al., who have used ChemSpider as a source of validated chemical structures and matched against experimental properties [100–103]

In April 2010 ChemSpider was awarded a Best Practices Award by Bio-IT for its community service [104] and in June 2010 was awarded an *I-Expo* innovation award [105]. The ChemSpider team remains focused on delivering on the vision of developing a community portal for chemists to source data and information. The collaboration of the community, data source providers, and life science industry members in particular will be essential to ensuring the ongoing expansion of the data and expansion of the reach of ChemSpider. ChemSpider is likely to become one of the foundations of the Semantic Web for chemistry and, with an ongoing focus for enabling collaboration and integration for life sciences, will be an essential resource for future generations.

ACKNOWLEDGMENTS

ChemSpider is the result of the aggregate work of many contributors. All core ChemSpider development is performed by four developers: Valery Tkachenko (Chief Technology Officer, RSC-ChemSpider, Lead Developer: ChemSpider), Alexey Pshenichnov (Developer: ChemSpider), Jon Steele (Developer: ChemSpider), and Sergey Shevelev (Lead Developer: ChemSpider SyntheticPages). ChemSpider is supported by a dedicated team of IT specialists. The author acknowledges the support of the many data providers, curators, and users for their contributions to the development of the data content in terms of breadth and quality.

REFERENCES

1. Williams AJ. ChemSpider and its expanding Web: Building a structure-centric community for chemists. *Chem Int* 2007;30(1):30.
2. Pence H, Williams AJ. ChemSpider: An online chemical information resource. *J Chem Educ*, Article ASAP. Available: http://dx.doi.org/10.1021/ed100697w.
3. PubChem. http://pubchem.ncbi.nlm.nih.gov.
4. The National Institutes of Health Roadmap Initiative. http://nihroadmap.nih.gov/.
5. http://www.ncbi.nlm.nih.gov/pubmed.
6. How big is the challenge of curation and what is the structure of ginkgolide. http://www.chemspider.com/blog/how-big-is-the-challenge-of-curation-and-what-is-the-structure-of-ginkgolide-b.html.

7. ChemSpider presentation at a PubChem Public Meeting, August 2008. http://www.slideshare.net/AntonyWilliams/presentation-of-chemspider-at-pubchem-public-meeting-presentation.
8. Hacking Pubchem: Technology easy, quality difficult. http://www.chemspider.com/blog/hacking-pubchem-technology-easy-quality-difficult.html.
9. Baker M. Open-access chemistry databases evolving slowly but not surely. *Nature Rev Drug Discov* 2006;5:707–708.
10. The structure of Taxol. http://www.chemspider.com/blog/?p=64.
11. Stop counting the number of chemical entities in public compound databases and there are ghosts in the closet. http://www.chemspider.com/blog/stop-counting-the-number-of-chemical-entities-in-public-compound-databases-and-there-are-ghosts-in-the-closet.html.
12. Ugly organometallics and the challenge of structure depiction. http://www.chemspider.com/blog/ugly-organometallics-and-the-challenge-of-structure-depiction.html.
13. Antony Williams, Is this the future of linked chemistry on the Internet? http://www.slideshare.net/AntonyWilliams/chemspider-is-this-the-future-of-linked-chemistry-on-the-internet.
14. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. DrugBank: A comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* 2006;34:D668–672.
15. HMDB: The Human Metabolome Database. *Nucleic Acids Res* 2007;35:D521–526.
16. FooDB. http://www.foodbs.org/.
17. SureChem. http://www.surechem.org/.
18. Wikipedia. http://www.wikipedia.org/.
19. Cholesterol. http://en.wikipedia.org/wiki/Cholesterol.
20. Dihydroergocryptine. http://en.wikipedia.org/wiki/Dihydroergocryptine.
21. Antony Williams. Dedicating Christmas time to the cause of curating Wikipedia. The ChemConnector Blog. http://www.chemconnector.com/chemunicating/dedicating-christmas-time-to-the-cause-of-curating-wikipedia.html.
22. Williams AJ. Internet-based tools for communication and collaboration in chemistry. *Drug Discov Today* 2008;13:502–506.
23. Nature Precedings. http://precedings.nature.com/.
24. The Public Libraries of Science ONE journal. http://www.plosone.org/.
25. ChemSpider Synthetic Pages. http://cssp.chemspider.com/.
26. The OrgPrepDaily blog. http://orgprepdaily.wordpress.com/.
27. The TotallySynthetic blog. http://totallysynthetic.com/blog/.
28. The ResearchGate Scientific Community. http://www.researchgate.net/.
29. Science 3.0. http://www.science3point0.com/.
30. Friendfeed. http://friendfeed.com/the-life-scientists.
31. ChemSpider. http://www.chemspider.com/.
32. Bradley JC. Open notebook science using blogs and wikis. *Nature Proc* Available: http://precedings.nature.com/documents/39/version/1, 2007.
33. Data source details for David Sharpe. http://www.chemspider.com/DatasourceDetails.aspx?id=265.

34. Data source details for Joerg Kurt Wegner. http://www.chemspider.com/DatasourceDetails.aspx?id=106.

35. Wikipedia, Metabonomics. http://en.wikipedia.org/wiki/Metabonomics#Metabonomics.

36. Agilent Newsletter, May 2010. http://www.chem.agilent.com/en-US/Newsletters/accessagilent/2010/may/pages/qtof_library.aspx.

37. Integrating ChemSpider to Thermo's SIEVE software. http://www.vastscientific.com/sieve/63141_SIEVE_BR_121610.pdf.

38. *LASSO: Ligand Activity in Surface Similarity Order*. Toronto, Canada: SioBioSys. Available: http://www.simbiosys.ca/ehits_lasso/index.html.

39. Reid D, Sadjad BS, Zsoldos Z, Simon A. LASSO—Ligand activity by surface similarity order: A new tool for ligand based virtual screening. *J Computer-Aided Mol Design* 2008;22(6/7):479–487.

40. Integrating ChemSpider to SimBioSys' LASSO. http://www.chemspider.com/blog/announcing-the-chemspider-ligand-activity-project-partnering-with-simbiosys.html.

41. An overview of LASSO. http://www.simbiosys.ca/science/presentations/2010-08-acs/ACS2010_LASSO.pdf.

42. Elyashberg ME, Williams AJ, Blinov K. *Structural revisions of natural products by Computer-Assisted Structure Elucidation (CASE) systems*. Available: http://dx.doi.org/10.1039/c002332a.

43. Presentation, Adam Hoye, the Hexacyclinol Incident. http://ccc.chem.pitt.edu/wipf/Current%20Literature/Adam_3.pdf.

44. Structure revision of hexacyclinol. http://totallysynthetic.com/blog/?p=110.

45. Williams AJ, et al. Applying computer-assisted structure elucidation algorithms for the purpose of structure validation: Revisiting the NMR assignments of hexacyclinol. *J Nat Prod* 2008;71(4):581–588.

46. Hexacyclinol? Or not? http://pipeline.corante.com/archives/2006/06/05/hexacyclinol_or_not.php.

47. Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcántara R, Darsow M, Guedj M, Ashburner M: ChEBI: A database and ontology for chemical entities of biological interest. *Nucl Acids Res* 2008;36:D344–D350.

48. DailyMed. http://dailymed.nlm.nih.gov/dailymed/about.cfm.

49. Wikipedia, Merck Index. http://en.wikipedia.org/wiki/Merck_Index.

50. *CAS registry numbers*. Columbus, OH: Chemical Abstract Services, 2008. Available: http://www.cas.org/expertise/cascontent/registry/regsys.html.

51. Chemical Abstracts' Common Chemistry Database. http://www.commonchemistry.org/.

52. Vitamin K1. http://www.commonchemistry.org/ChemicalDetail.aspx?ref=84-80-0&terms=vitamin%20K1.

53. Vitamin K1. http://www.ebi.ac.uk/chebi/searchId.do?chebiId=CHEBI:18067.

54. Vitamin K1. http://www.chemspider.com/Chemical-Structure.4447652.html.

55. Will the Structure of Taxol Please Stand Up? http://www.chemspider.com/blog/?p=64.

56. Collaboration, Community and Quality in Chemistry Databases. http://www.chemspider.com/blog/collaboration-community-and-quality-in-chemistry-databases.html.

57. Where does C&E News source its chemical structures? http://www.chemspider.com/blog/where-does-ce-news-source-its-chemical-structures.html.

58. A request for a crowdsourced investigation of digitonin. http://www.chemspider.com/blog/a-request-for-a-crowdsourced-investigation-of-digitonin.html.

59. Hettne KM, et al. Automatic vs. manual curation of a multi-source chemical dictionary: The impact on text mining. *J Cheminform* 2010;2:3.

60. SMILES notation. http://en.wikipedia.org/wiki/SMILES.

61. International Chemical Identifier. http://www.iupac.org/inchi/.

62. Advanced Chemistry Development press release, Announcing a collaboration with ChemSpider. http://www.acdlabs.com/company/media/pr/100429_chemspider.php.

63. Adding EPISuite predicted properties to ChemSpider. http://www.chemspider.com/blog/adding-epa-episuite-properties-to-chemspider.html.

64. Thalidomide. http://en.wikipedia.org/wiki/Thalidomide.

65. PubMed. http://www.ncbi.nlm.nih.gov/pubmed.

66. Google Scholar. http://scholar.google.com/.

67. Google Patents. http://www.google.com/patents.

68. RSC Publishing. http://pubs.rsc.org/.

69. Antony Williams, Enhancing discoverability across RSC content by integrating ChemSpider. http://www.slideshare.net/AntonyWilliams/enhancing-discoverability-across-royal-society-of-chemistry-content-by-integrating-to-chem-spider-an-online-database-of-chemical-structures-3551795.

70. Lipinski's Rule of Five. http://en.wikipedia.org/wiki/Lipinski%27s_Rule_of_Five.

71. Open Data. http://en.wikipedia.org/wiki/Open_data.

72. Spectral data. http://www.chemspider.com/spectra.aspx.

73. JCAMP, the Joint Committee on Atomic and Molecular Physical data spectral exchange format. http://www.jcamp-dx.org/.

74. Lancashire RJ. The JSpecView Project: An open source Java viewer and converter for JCAMP-DX, and XML spectral data files. *Chem Central J* 2007;1:31.

75. Bradley JC, Lancashire RJ, Lang ASID, Williams AJ. The spectral game: Leveraging open data and crowdsourcing for education. *J Cheminform* 2009;1:9.

76. YouTube. http://www.youtube.com/.

77. A fast photochromism video on ChemSpider. http://www.chemspider.com/blog/fast-photochromism-and-an-amazing-video.html.

78. ChemSpider structure ID 22402. http://www.chemspider.com/Chemical-Structure.22402.html#description.

79. The Entrez query interface. http://www.ncbi.nlm.nih.gov/entrez/query/static/esoap_help.html.

80. Digital Object Identifier. http://en.wikipedia.org/wiki/Digital_object_identifier.

81. The CrossRef Resolver. http://dx.doi.org/.

82. Royal Society of Chemistry press release, The acquisition of ChemSpider. http://www.rsc.org/AboutUs/News/PressReleases/2009/ChemSpider.asp.

83. *Project Prospect*. RSC Publishing, 2008. Available: http://www.rsc.org/Publishing/Journals/ProjectProspect/.
84. Imidacloprid. http://en.wikipedia.org/wiki/Imidacloprid.
85. The Reaxys database. http://www.info.reaxys.com/about_overview.
86. SPRESI database. http://en.wikipedia.org/wiki/SPRESI_database.
87. Chemical Abstracts CASReact Reaction Database. http://www.cas.org/expertise/cascontent/casreact.html.
88. SyntheticPages. http://en.wikipedia.org/wiki/SyntheticPages.
89. The ChemSpider SyntheticPages editorial board. http://cssp.chemspider.com/Editors.aspx.
90. Open Source software. http://en.wikipedia.org/wiki/Open-source_software.
91. Mobile Chemistry, 2010. http://www.rsc.org/chemistryworld/Issues/2010/May/MobileChemistryChemistryHandsFace.asp.
92. Announcing the release of ChemSpider Mobile. http://www.chemspider.com/blog/chemspider-mobile-goes-live.html.
93. PubChem Mobile. http://appspace.com/apps/view/250552/pubchem-mobile/.
94. Peter Murray Rust, Aggregated chemistry and quality. http://wwmm.ch.cam.ac.uk/blogs/murrayrust/?p=261.
95. A response to Peter Murray-Rust. http://www.chemspider.com/blog/another-response-to-constructive-feedback-from-peter-murray-rust.html.
96. Gamo FJ, et al. Thousands of chemical starting points for antimalarial lead identification. *Nature* 2010;465:305–310.
97. Gagaring K, et al. *Novartis-GNF malaria box. In ChEMBL-NTD*. Available: www.ebi.ac.uk/chemblntd.
98. Ekins S, Williams AJ. When pharmaceutical companies publish large datasets: An abundance of riches or fool's gold. *Drug Discov Today.* Available: http://dx.doi.org/10.1016/j.drudis.2010.08.010.
99. Ekins S, Williams AJ. Meta-analysis of molecular property patterns and filtering of public datasets of antimalarial "hits" and drugs. *Med Chem Commun* 2010;1:325–330.
100. Ekins S, Iyer M, Krasowski MD, Kharasch ED. Molecular characterization of CYP2B6 substrates. *Curr Drug Metab* 2008;9:363–373.
101. Ekins S, Kholodovych V, Ai N, Sinz M, Gal J, Gera L, Welsh WJ, Bachmann K and Mani S, Computational discovery of novel low micromolar human pregnane X receptor antagonists. *Mol Pharmacol* 2008;74:662–672.
102. Khandelwal A, Bahadduri PM, Chang C, Polli JE, Swaan PW, Ekins S. Computational models to assign biopharmaceutics drug disposition classification from molecular structure. *Pharm Res* 2007;24:2249–2262.
103. Ekins S, Williams AJ, Xu JJ. A predictive ligand-based Bayesian model for human drug-induced liver injury. *Drug Metab Dispos* 2010;38:2302–2308.
104. Bio-IT press release, ChemSpider Community Service Award. http://www.bio-itworld.com/BioIT_Article.aspx?id=100728.
105. Royal Society of Chemistry press release. http://www.rsc.org/AboutUs/News/PressReleases/2010/ChemSpider.asp.

23

COLLABORATIVE-BASED BIOINFORMATICS APPLICATIONS

BRIAN D. HALLIGAN

23.1 INTRODUCTION

Until recently, most computing was done with local resources. Users ran applications on their desktop computer and stored files on this computer or on a server in a machine room nearby. There has been a shift to a paradigm in which users, instead of doing computing locally, use their desktop computer as a client to access resources outside their office or even institution. Often users have no idea where the computer running their application or their data is physically located, so it can be thought of being up in the clouds somewhere.

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.

The phrase “cloud computing” is often applied to both remote network storage and computer virtualization. Cloud computing can help to overcome some of the barriers that have inhibited the collaborative sharing of tools and methods in bioinformatics. Although many bioinformatics projects are open-source, the difficulties in replicating analysis platforms often prevent the use of these open source resources.

Many academic bioinformatics projects suffer from a common problem: Support for many academic bioinformatic projects is often sporadic. A project may be undertaken under the fixed term funding of a grant or as part of the training program of a graduate student or postdoctoral fellow. During the grant or training period, there are resources to develop and maintain the resource. This often culminates in the release of the resource to the research community. Over time, the research community discovers the utility of the resource and comes to incorporate it into their research workflow and thus becomes dependent on it. Meanwhile, one or more of the following may occur: the grant funding ends, the graduate student completes his or her degree or the postdoc obtains a job, and the project ends while the resources that it generated become orphaned. Often projects are left to coast until software incompatibility creeps up or hardware issues or budget/space constraints lead to the server hosting the resource to be taken offline. At that point there is often a lack of will and/or resources to repair or re-create the tool and the research community is left without access to what has become an important part of their workflow.

This has been one of the strong arguments for the release of these types of tools as open-source projects. In theory this means that the end users could download the source and replicate the resource on their own. Functionally this is not always easy. Many of these tools are not a simple executable file but rather a core of code enmeshed in a web of Web servers, databases, and other system resources. Although these dependencies usually are also open source as well, the integration of them is often difficult and not well described or documented. It is sometimes the case that tools require specific versions of these components and will not work when the components are upgraded. If the project is still active and with the detailed knowledge of the person who developed the system still present, these problems can be overcome. In contrast, rebuilding this “house of cards” from the ground up can be nearly impossible even for individuals with significant IT backgrounds. A possible solution to maintaining analysis infrastructure is to use cloud computing systems such as Amazon’s Web Services (AWS) that allow users to save snapshots of virtual computers as Amazon machine images (AMIs). These AMIs can be made publicly available and can be used by anyone with an Amazon AWS account. The end user can call the preconfigured computer into existence for the amount of time required and shut it off when done. Like a fly caught in amber, the AMI is a snapshot of the working system and remains in a static state and is never tied to a particular piece of hardware. Since the effort required to generate an AMI is not prohibitive and the storage costs for these images are low, it is feasible to generate images representing each version of the system. This can be critical for a long-running project that collects samples and data over the course of months or years. It is important that all of the data from the project be analyzed in precisely the same manner. By only using a single AMI version to carry out the analysis, the user can be sure that all of the results are comparable.

Another advantage of using a publicly available stored AMI is that it allows multiple groups to have access to precisely the same analysis platform, allowing it to be a standard for comparison of results between the groups. Other groups can save a copy of the AMI to their own Amazon (Simple Storage System) S3 storage area or even download it to local storage, thereby removing all dependence on other groups. They can then make changes to their copy of the AMI and either keep these changes private or return them for public use, while the public version remains unchanged. Running a tool as an Amazon virtual computer also has desirable security features. Some groups may be reluctant to upload their data to a third-party website for analysis. This could be because the data are of a proprietary nature or may be human patient related. In essence, the virtual computer created from the AMI is the property of the group that instantiated it, not the group that developed it. If the AMI was properly created, then the group that developed it does not have any access to the data analyzed by the AMI. Since it is possible to encrypt all data uploaded and downloaded to the AMI and the private S3 storage area of the user, the data being analyzed should be as secure as if the analysis was taking place in the user's home data center.

There are other advantages to cloud-based analysis compared to analysis carried out in local data centers. If a group required a number of different analysis tools as part of its research data workflow, then it might have to set up and maintain a separate server for each tool. This can require a significant investment of time and resources for tools that may only be used sporadically. Similarly, the requirement for each tool may only occur sporadically, but when needed there may be a large quantity of data to analyze. This could require either a significant time lag in obtaining the results or the use of multiple computers to carry out the work. With local machines, these extra servers have to be prepared and maintained in advance. With virtual computers, the number of nodes required to carry out the work can be easily instantiated. Since billing is done by node hours used, it costs the same to carry out an analysis for 100 hours on 1 node as for 1 hour on 100 nodes. This gives even small groups access to large-scale computing resources on demand.

Data loss is also less likely with cloud-based storage. Locally stored data are subject to disk failure. Usually this can be addressed using tape backup, but for this to be effective the tape backups have to be tested and stored offsite. Cloud-stored data are multiply replicated and stored in different geographic locations and even across multiple continents. This ensures not only that it is protected against loss but also that timely access to the data will not be interrupted by a single point of failure. Since users can set access policies for the data stored in S3, it is possible to limit access to the data from a particular computer, range of Internet Protocol (IP) addresses (providing access for multiple users from the same institution), or publicly. This last option can be used to address the growing requirement for public access to data imposed by publication and funding agencies. Publication in some journals is contingent on public access to the underlying data. Many journals are not prepared to host the data on their own systems, and for some fields such as proteomics, public repositories capable of handling the data have not yet become available. Several large bioinformatics data sets have been made available on the AWS system such as the Annotated Human Genome Data provided from ENSEMBL [1] and UniGene provided by the National Center for Biotechnology Information [2].

Another advantage of hosting data and tools in the cloud as publicly available data or AMIs is that the cloud helps maintain institutional data security. By moving these resources off-site, it prevents them from becoming points of attack. Since the AMI functions outside the institutional firewall, there is no opportunity for access to it to be a security hole, regardless of what ports they require.

23.2 CLOUD COMPUTING RESOURCES

The basic philosophy of cloud computing is to divorce the service (storage or computation) from a physical resource and have it available on demand like electricity from a wall socket. Remote network storage is storage that is kept on remote servers and accessed using Web browsers, File Transfer Protocol (FTP), or other clients. One of the main features that distinguishes it from local network storage is that the physical location of the storage is often unknown to the end user and the data are often redundantly distributed across physical locations and sometimes even across continents. This redundant distribution of data across widely scattered resources offers protection from a single point of failure. An individual drive failure does not cause data loss and does not necessarily even take the data offline. Data security has been a concern with many businesses using cloud computing. Unlike credit card or social security numbers, in the case of bioinformatics applications this is not as important as the individual pieces of data are not in themselves very subject to abuse. In order to make the data useful, the full data set is usually required as well as an understanding of the experiment that generated it. Data storage structures can vary with the need of the investigator. Data can be stored in the format of individual pieces of data such as e-mails on Hotmail, photos on Flickr, or documents on Google Docs or as individual files with utilities such as iDisk and DropBox. Amazon offers enterprise class data storage in its S3 system. Files are stored in “buckets” that are owned by individual users. The files can be made public allowing collaborators to view them or kept private. The key idea for cloud computing is that the user does not know where or how the resources they want to use exist or what physical form they take, only that they want to use a particular resource.

Network storage of data allows for computer virtualization. A machine image or, in the case of Amazon, an AMI is stored in S3. A user can invoke the image that causes it to be run as a virtual computer. Since the image can be prebuilt with different uses in mind, it is possible to have different images for different specialized purposes.

23.2.1 Amazon Web Services (AWS)

One of the most popular providers of cloud computing resources to the bioinformatics community is Amazon. This is because its AWS is both easy to use and inexpensive. AWS is composed of a number of different parts with different functionality. Data are stored in S3 as objects that can range in size from 1 byte to 5 Gb. Data objects are stored in the equivalent of folders called buckets. Retrieval of each object requires a developer to assign a key and data are secured from unauthorized access. A bucket can be located in one of four geographical regions and the data objects in the bucket are replicated across multiple servers within the region. This protects the data from loss due to disk failure or local disaster in up to two data centers concurrently. For increased data security, versioning is available which will store previous versions of files rather than overwriting, which allows users to roll back changes and correct errors. For data that are also being stored locally, Amazon also offers a lower cost alternative, reduced redundancy storage (RRS), that has an expected data loss of 0.01% per year as compared to the 10^{-9}% expected loss with S3. Currently data storage for the first 50 Tb of data is US $0.15/Gb for S3, and US $0.10/Gb for RRS, and transfer into S3 is currently free. For projects that require very large volumes of data to be uploaded, Amazon offers a service, AWS Import/Export, which allows a user to physically ship a disk to Amazon and it will download the data directly. In order to distribute data Amazon offers Amazon Cloud Front, which allows users to have a Web interface to give public access to objects in the user's S3 buckets. This could be used to give collaborators access to data or it could be used to make published data sets publically available. The publisher of the data only pays for the amount of data downloaded so there are very low upfront costs to hosting the data.

Cloud computing resources are available from Amazon through its Elastic Compute Cloud (EC2) system. EC2 makes virtual computers of different sizes available on demand on an hourly basis. Virtual computers are available from small compute instances with a single 32-bit compute unit with 1.7 Gb of memory and 160 Gb of local storage to extra large high-CPU instances with twenty 64-bit compute units with 7 Gb of memory and 1.6 Tb of local storage. Additionally, compute clusters with 33.5 compute units and 23 Gb of memory are also available. The user pays hourly for the amount of time that the instance is alive. There are a number of other additional services that can be added to EC2. Elastic Block Storage is a persistent data store that can be attached to an instance and read from and written to much like an external hard drive on a physical computer making data transfer easier than to S3. Like S3, EC2 instances can be located in different geographical locations to decrease network latency. When running, instances can be accessed by using a network address provided when the instance is launched. Additionally, an IP address can be assigned to an instance and an instance can include a Web server to allow for Web access to programs running on the instance. For additional security, a virtual private cloud using virtual private network (VPN) technology can be created. This allows institutions to connect to the AWS cloud as though it was part of the institution's network. In order to monitor and control running instances, Amazon has the Cloudwatch monitoring service, elastic load balancing, and autoscaling.

An important computational resource for bioinformatics is the Amazon Elastic MapReduce service. MapReduce is built on the Hadoop framework. Hadoop is a system that creates a compute cluster from a collection of virtual instances. It supports data-intensive distributed applications by creating a distributed file system that allows individual nodes to share data and job tracker and task tracker functions that oversee the analysis of the data by the individual instances. The MapReduce service takes problems that can be broken down to smaller elements and automates their analysis. These so-called embarrassingly parallel problems are characterized by having data elements that can be analyzed independently from the entire data set. A good example from proteomics is the peptide identification from mass spectra. A mass spectroscopy run can be broken down into individual spectra. Each of the spectra can be compared to the peptide sequence database to find the best match in the database, and the results from the individual searches can be combined to produce the final search results. MapReduce automates the splitting of the data, the "map" function, the establishment and oversight of the worker Hadoop cluster instances, and the combination of the results produced by the individual workers, the "reduce" function.

There are other AWS services available that can be used in concert with EC2 and S3. These include message management services such as Amazon Simple Queue Service (Amazon SQS), which allows instances to exchange messages and coordinate the parallel analysis of data, and Amazon Simple Notification Service (Amazon SNS), which allows running instances to send messages to other instances, servers, or end users that subscribe to the messages from the instances. This allows workflows composed of AWS instances to respond to events. Additionally AWS offers two database services. Amazon SimpleDB is a simple nonrelational database that provides easy access to data with a high degree of availability and scalability. For more demanding needs, AWS also offers a relational database service, Amazon Relational Database Service (Amazon RDS), which provides a cloud-based relational database equivalent to MySQL and is compatible with applications that use MySQL.

EC2 images are stored as AMI in S3 and can be private or public. Essentially the AMI is a snapshot of the boot disk of the virtual computer. The user launches an instance of an AMI, then selects the size of the instance to launch. Before launching the instance, the user generates a key pair file to use in accessing the file. With this file, the user can use SSH to log onto the instance.

23.2.2 Other Cloud Computing Resources

Although Amazon has taken a very prominent position in the cloud computing industry, there are other providers that also provide similar services. Microsoft Azure is similar to Amazon's AWS but is organized into Web and worker roles rather than instance types. One of the complications of cloud computing is that applications developed for one system cannot be easily transferred to a competing system. This has caused a fear that users would be locked into a single vendor. Recently an open-source cloud system has been developed. Open Stack [3] is software developed for cloud computing that is distributed on the Apache license model. Open Stack is not a provider of services, but the software can be used by an institution to establish its own private cloud.

23.3 EXAMPLES OF BIOINFORMATICS CLOUD COMPUTING RESOURCES

23.2.1 Proteomics

Proteomics is a good test case for the use of cloud computing in bioinformatics. Most high-throughput proteomics experiments are carried out using what is referred to as bottom-up shotgun proteomics [4]. What this means is that, instead of measuring the abundance of proteins directly, the proteins in the samples are first converted to peptides by digestion with proteases, usually trypsin, and then the peptide composition of the sample is analyzed and the protein composition is inferred from the peptides present in the sample. This is done because the peptides are smaller and less complex biochemical entities that behave better in chromatography and mass spectrometry systems than whole proteins. Peptides are ionized and charged peptide ions are transported through the instrument in proportion to the ratio of the peptides' mass and charge (m/z). Smaller or higher charged ions move faster than larger or less charged ions. At an intermediate stage in the instrument, the ions can be fragmented by interaction with inert gas or other means. The pattern of fragments roughly corresponds to the amino acids that compose the peptide. Peptides are identified by comparing these MS/MS fragmentation spectra to theoretical spectra in which the assumption is made that fragmentation occurs at each peptide bond in the peptide. There are a number of available algorithms that accomplish this. Some of these are commercial, such as Sequest [5] and Mascot [6], and some are open source or public domain, such as X!Tandem [7] or OMSSA [8].

Peptide identification from mass spectrometry data is amenable to cloud computing in that the data set consists of tens of thousands of individual fragmentation spectra and the peptide identification process is more or less independent from spectra to spectra. This allows the use of a MapReduce-like strategy in which worker nodes can be assigned packets of spectra to search, and they can return their results to a common area for integration when all the searches are completed. This works well because the majority of the computation effort is expended in the individual searches rather than in splitting the data or combining the results.

To allow for high-throughput analysis of proteomics data, we have developed the Virtual Proteomics Data Analysis Cluster (ViPDAC) system. ViPDAC is based on the AWS EC2 and S3 systems and relies on the use of open-source algorithms and programs for peptide identification and open-source software developed for ViPDAC to distribute spectra, manage worker nodes, and summarize the results. ViPDAC is available as a public AMI that can be launched by anyone having an AWS account. The ViPDAC AMI includes an integrated Web server so that interactions with the end user and the ViPDAC head node occur through the use of a familiar Web interface. Through this interface, the end user can choose data sets and analysis parameters and add or terminate worker nodes. Raw data are first uploaded to the end user's S3 storage area and results are returned to the user's S3 or through a download link on the website.

Since ViPDAC was developed before the MapReduce function of AWS was available, it uses its own facilities to distribute spectra to the worker nodes, manage the nodes, and collect the results. When the end user launches the initial ViPDAC instance, this instance configures itself to be the head node and controls the distribution and retrieval of data. When subsequent instances are launched by the same user, they recognize that the head node exists and configure themselves as worker nodes. Worker nodes then make requests to the head node for packets of spectra to search. The head node then responds with a message to the worker node, informing it of the location of the compressed file containing multiple spectra, parameters, and database to use for the search. When the searches are complete, the worker informs the head node that the packet has been completed and the data are collected. If the head node does not receive a message that the searches are complete within the specified time, the head node then considers that the worker node has failed and returns the packet of spectra back to the queue for analysis by a different node. An issue with this system is that the amount of time a search uses can vary greatly due to the complexity of the spectra, the size of the database, and the parameters chosen. For this reason, it is important for the end user not to choose a timeout value too short to complete a given set of spectra. This parameter can also be adjusted by changing the size of the spectra packet. For particularly difficult searches, such as those with no peptide specificity (unconstrained searches), it is advisable to use a spectra packet size of one spectra.

23.3.2 Other Bioinformatics Tools

The J. Craig Venter Institute (JCVI) has produced an AMI preconfigured with many of the standard bioinformatics tools that they have titled JCVI Cloud Bio-Linux. The instance is based on 64-bit Ubuntu Linux and contains the Celera Assembler [9], the European Molecular Biology Open Software Suite [10], BLAST [11], ClustalW [12], Glimmer [13], GeneSpring [14], HMMER [15], PHYLIP [16], and RasMol [17]. The goal of the project is to produce a platform with which groups could use to set up and distribute bioinformatics analysis systems and data. The hope is to overcome the difficulties in installing and setting up bioinformatics tools.

23.3.3 Next-Generation DNA Sequencing

One of the most significant challenges of bioinformatics is the analysis of the huge volume of data generated by next-generation DNA sequencing efforts [18]. This process produces millions of short sequence reads which must be aligned and merged to produce the final sequence. As the rate of sequencing has accelerated, the data storage requirements have moved from megabytes to gigabytes to terabytes and soon to petabytes. The computational time to process these data has similarly increased. To address this, systems using cloud computing have been developed. One of the uses of next-generation sequencing is the mapping of genomes and identification of single-nucleotide polymorphism (SNPs). The CloudBurst application (described below) uses AWS MapReduce and Hadoop to generate a cluster of computers to process the alignment of reads from next-generation sequencing instruments [18]. The algorithm is based on aligning reads to a reference genome and then extending the alignment by adding additional reads. This is expedited by the hosting of Ensembl and GenBank genomic data in S3. This makes the required reference genome data available with low latency and no cost for transfer and storage.

The Crossbow system for DNA sequence alignment and SNP discovery developed at Johns Hopkins University uses cloud computing to align high-throughput DNA sequencing reads and find individual polymorphisms [19]. It combines Bowtie [20] to align short reads and SoapSNP [21] to call genotypes. It is based on MapReduce and uses Hadoop to parallelize the computational load across multiple AWS instances. According to the developers, it can analyze over 35 times coverage of a human genome in 3 hours for about $85 using a 40-node, 320-core cluster rented from Amazon Web Services.

A similar program, also developed at Johns Hopkins University, is Myrna [22]. Myrna also uses Bowtie and Hadoop, but rather than assemble entire genomes, it measures gene expression by analyzing RNA-seq data sets. Like

Crossbow, it provides a graphical user interface to make constructing and running the virtual cluster easier.

Many of the available sequence aligners are based on having a reference genome to compare the individual short reads. In cases where the reference genome is not available, de novo assembly must be carried out. Contrail is an example of a de novo DNA assembly program that uses cloud computing to merge similar small reads into larger assemblies [23]. Contrail uses Hadoop to divide the work among multiple worker nodes and an innovative algorithm to represent the graph structures on disk rather than in memory, allowing the method to be scaled to larger genomes.

CloudBurst is another program for short-read DNA mapping using cloud computing [24]. Based on the RMAP short-read program, CloudBurst also uses MapReduce and Hadoop to create and manage parallel instances to speed the analysis of next-generation high-throughput sequencing.

Written by the University of Maryland, Quake uses Hadoop to make error corrections to high-throughput sequencing results by examining *k*-mer frequencies present in the short reads [25]. By examining these frequencies, it determines the most likely sequencing errors and how to correct them and achieve greater accuracy.

Other next-generation sequencing programs have used a similar approach using the Microsoft Azure cloud system to analyze next-generation sequencing data. The Azure system takes a different approach to cloud computing. Rather than focusing on running instances, Azure runs applications in either Web mode or worker modes. The Web mode applications are exposed to the outside through normal Hypertext Transfer Protocol (HTTP) methods such as REST or Simple Object Access Protocol (SOAP) and the worker nodes communicate directly with the Web application. The Azure system monitors and creates more worker nodes as necessary to carry out the task without intervention from the user. The virtual machine (VM) instances communicate with each other through queues and other technologies that are part of the Windows Azure fabric.

23.4 SUMMARY

Although many bioinformatics tools are distributed as free and open-source software, they are often difficult to install and have significant dependencies such as Web servers and database systems. Often this setup and configuration are not well documented and can require significant expertise and experimentation to craft a fully functional system. By setting analysis platforms in the cloud, they can be saved as machine images that can then be publicly shared. This allows the creator of bioinformatics applications to distribute their work in a form that can be used without investments in setup and architecture. Groups can now collaborate and use the same analysis programs with far less cost and higher levels of security than previously.

REFERENCES

1. Amazon. Amazon Web Services Developer Community: Ensembl Annotated Human Genome Data—for MySQL. Seattle, WA, 2010. Available: http://developer.amazonwebservices.com/connect/entry.jspa?externalID=2315.
2. Amazon. Amazon Web Services Developer Community: Unigene. Seattle, WA, 2010. Available: http://developer.amazonwebservices.com/connect/entry.jspa?externalID=2283.
3. Clark R. OpenStack Open Source Cloud Computing Software. San Francisco, CA, 2010. Available: http://www.openstack.org/.
4. Wu CC, Maccoss MJ. Shotgun proteomics: Tools for the analysis of complex biological systems. *Curr Opin Mol Ther* 2002;4:242–250.
5. Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 1994;5:976–989.
6. Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999;20:3551–3567.
7. Craig R, Beavis RC. TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics* 2004;20:1466–1467.
8. Geer LY, et al. Open mass spectrometry search algorithm. *J Proteome Res* 2004;3: 958–964.
9. Denisov G, et al. Consensus generation and variant detection by Celera Assembler. *Bioinformatics* 2008;24:1035–1040.
10. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet* 2000;16:276–277.
11. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410.
12. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994;22: 4673–4680.
13. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* 1999;27:4636–4641.
14. Chu L, Scharf E, Kondo T. GeneSpring: Tools for analyzing microarray expression data. *Genome Inform* 2001;12:227–229.
15. Eddy SR. Profile hidden Markov models. *Bioinformatics* 1998;14:755–763.
16. Felsenstein J. PHYLIP—Phylogeny inference package (version 3.2). *Cladistics* 1989;5:164–166.
17. Sayle RA, Milner-White EJ. RASMOL: Biomolecular graphics for all. *Trends Biochem Sci* 1995;20:374.
18. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol* 2008;26: 1135–1145.
19. Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL. Searching for SNPs with cloud computing. *Genome Biol* 2009;10:R134.

20. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;10:R25.

21. Li R, et al. SNP detection for massively parallel whole-genome resequencing. *Genome Res* 2009;19:1124–1132.

22. Langmead B, Hansen K, Leek J. Myran. Baltimore, MD: Johns Hopkins Bloomberg School of Public Health, 2010. Available: http://bowtie-bio.sourceforge.net/myrna/index.shtml.

23. Schatz DS, Kelley D, Pop M. Contrail. College Park, MD: Center for Bioinformatics and Computational Biology, University of Maryland, 2010. Available: http://sourceforge.net/apps/mediawiki/contrail-bio/index.php?title=Contrail.

24. Schatz MC. CloudBurst: Highly sensitive read mapping with MapReduce. *Bioinformatics* 2009;25:1363–1369.

25. Kelley DR, Schatz MC, Salzberg SL. Quake. College Park, MD: UMD Computer Science, 2010. Available: http://www.cbcb.umd.edu/software/quake/index.html.

24

COLLABORATIVE CHEMINFORMATICS APPLICATIONS

Rajarshi Guha, Ola Spjuth, and Egon Willighagen

24.1 INTRODUCTION

Cheminformatics is the science of chemical data and computation. The origin of the data is generally from the wet laboratory, thereby making collaboration an intrinsic part of cheminformatics: Data aggregation, preprocessing, and

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.

review involve many other scientists. Moreover, as most cheminformatics approaches find patterns and relationships at various levels of detail, using methodologies which are typically mathematical and are aimed at identifying correlations rather than physical and chemical cause–effect relations, the domain finds itself once again collaborating with the experimental scientist to validate the derived models and predictions against new experimental data.

However, as the field has matured, it has become more specialized, and with more specialization, the nature of those collaborations have changed. A study could have started as a laboratory scientist doing cheminformatics as a side project, while the domain later evolved to people specializing in cheminformatics while collaborating with other scientists in the same group and then later collaborating with other research groups. Nowadays, it is even common to collaborate with scientists around the world as even the same university may no longer share the same specialty. The evolution of Internet technologies resulted in an era of online science.

Fortunately, cheminformatics is at an advantage compared to other sciences. Data exchange, processing, and analysis are all done electronically, making it suitable to be scaled up to online science. This chapter will define and detail the various tools cheminformaticians have at hand to simplify this online collaboration.

Cheminformatics deals with the aggregation, handling, processing, and analysis of chemical data. The nature of the chemical data is in principle not important. However, the history of the field and its separation from quantum chemistry fields bias the domain toward small organic molecules. The patterns in collaborative applications are, however, independent of the exact nature of the data. Therefore, we will focus in this chapter on another dimension to discuss the aspects of collaborative cheminformatics applications: code development, knowledge handling and data exchange, and collaborative computation.

The first section will focus on methods involved in the collaborative development of cheminformatics software and will discuss the tools modern scientists have to perform this task. The next section will describe recent changes in which we handle chemical data and knowledge, in particular how the Internet is changing the way communities create new knowledge bases. The third section will focus on the aspects of collaborative computing in cheminformatics. Finally, the fourth section will describe social aspects of collaborative projects, in particular how collaboration is managed in projects with only loosely defined roles for the various partners.

24.2 COLLABORATIVE CODE DEVELOPMENT

Collaborative code development is a common approach for large software vendors. For scientific software, however, it is less common: New software is typically started as a Ph.D. or M.Sc. project with a single developer. There are, however, two areas where collaborative code development in cheminformatics flourishes: One situation is where a piece of software has become large and successful and multiple people have interest in contributing to the project; the second situation is where the project is fairly small and no single developer can or wants to lead the project, as the topic is not core to their research. An example of the latter situation is the JChemPaint project [1]. There are existing similar projects, making continued development out of the scope of cheminformatics research. However, in 1998 Steinbeck et al. showed that a collaborative project can lead to an ecosystem where such software can still be developed.

Central to collaborative code development is the sharing of source code. Particularly, it is the pipelining of how patches are shared and applied. While some cheminformatics projects still share source code as source distributions, the adoption of source code repositories has emerged as the golden standard. There are various open-source repository technologies available, including the Concurrent Versions System (CVS) [2], Subversion (SVN) [3], Mercurial [4], Bazaar [5], and Git [6]. CVS is the oldest and mostly replaced by the newer technologies. Subversion is still abundant but increasingly replaced by the last three technologies. The reason for this is that the latter three systems are distributed technologies, allowing for server redundancy. Moreover, because of the distributed nature of Mercurial, Bazaar, and Git, branching and merges of branches are often easier. However, the increased functionality also introduces further complexity, which is particularly the case for Git, and leads to a steeper learning curve.

As these tools are open source, anyone is able to set up a local, possibly private server, but open-source projects can take advantage of service providers that host free and public code repositories. Table 24.1 provides an overview of various larger service providers, but there are many alternatives.

24.2.1 Licensing

A second aspect that simplifies collaboration is to use an open-source license. Such a license ensures that potential contributors know that whatever work they invest in the source code is not lost: It will always be available to that contributor under those license terms. There are various open-source licenses available, each with different characteristics. A discussion of the differences

TABLE 24.1 Overview of Code-Sharing Technologies and Service Providers That Provide Free Online Hosting

Technology	Provider(s)
Subversion	SourceForge [7], Google Code [8]
Mercurial	Google Code
Bazaar	LaunchPad [9]
Git	GitHub [10], Gitorious [11], SourceForge

and details is well beyond the scope of this chapter. The reader is encouraged to read the book *Open Source Licensing* by Lawrence Rosen [12] to gain additional information in this area. Popular open-source licenses include the GPL licenses [13, 14] and the Massachusetts Institute of Technology (MIT) [15] and Berkeley Software Distribution (BSD) licenses [16]. A full overview of open-source licenses is maintained by the Open Source Initiative [17].

24.2.2 Peer Review

While sharing the source code is the primary channel of collaborative code development, communication is a close second. Designs need to be discussed and solutions proposed. Peer review is part of this and plays an important part in collaborative code development. A full discussion of communication will be discussed later in this chapter as part of project management, while here we will focus on peer review only.

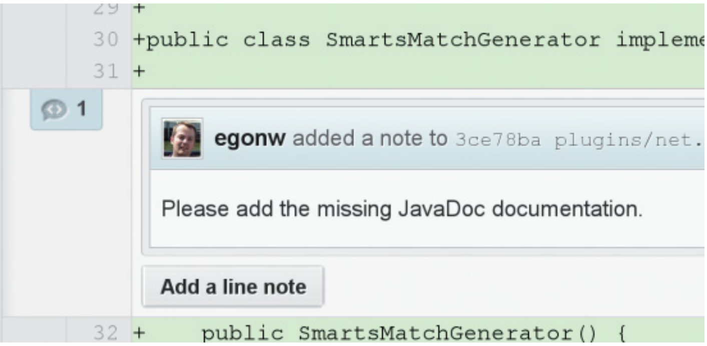
While the role of peer review in scientific publishing is well recognized as an important path for communication, it is rarely applied to scientific programming. Nevertheless, the advantages and disadvantages are clearly recognized for source code development as well, as detailed in the famous article titled "The Cathedral and the Bazaar" [18]: "It's one thing to observe in the large that the bazaar style greatly accelerates debugging and code evolution. It's another to understand exactly how and why it does so at the micro-level of day-to-day developer and tester behavior."

The article continues to detail why the bazaar model with many people looking at the code actually works. Interestingly, this is where it may not directly apply to cheminformatics as the number of potential people looking at the code is actually relatively low. Moreover, developers from competing projects may have additional reasons not to review the work from other projects, further reducing the number of reviewers. It should be noted that this problem is general to science and that peer review of publications too typically requires a more formal approach. Peer-reviewed code development has the advantage that static source code checkers are available that can detect common problems, such as PMD [19]. These tools check, for example, for unused variables, dead code, and so on, often highlighting sources of problems.

Manual peer code review is increasingly simplified by new tools. For example, GitHub provides the functionality to comment on changes [20], increasing the communication between developers and the social pressure to write better source code. Figure 24.1 shows an example comment made via the GitHub website.

24.3 COLLABORATIVE KNOWLEDGE BASES

Cheminformatics is, of course, very much about information. This information is embedded in knowledge bases, such as relational databases. The knowledge



```
+
+public class SmartsMatchGenerator impleme
+
```

egonw added a note to 3ce78ba plugins/net.

Please add the missing JavaDoc documentation.

```
+    public SmartsMatchGenerator() {
```

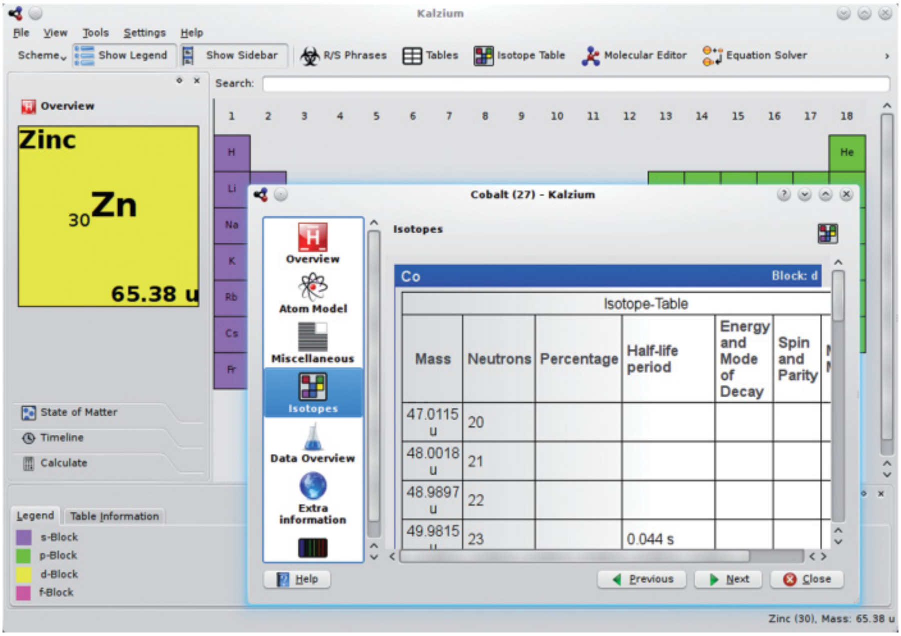
Figure 24.1 Screenshot of GitHub Webpage providing code review functionality. The comments show a reviewer question regarding a parameter description missing in the JavaDoc (see http://github.com/bioclipse/bioclipse.cheminformatics/commit/3ce78ba).

in these information resources is derived from experimental data. For example, force fields, which are used to calculate energies for molecular conformations, are based on common patterns, such as average bond lengths, the angles between two bonds that have one atom in common, and so on.

The collaborative building of knowledge bases is now well established, with popular examples in chemistry, including compound databases like PubChem [21] and ChemSpider [22], where people can deposit chemical structures. Social sites to share open data include the NMRShiftDB [23] (licensed under a GNU FDL license), ChemPedia [24] (available under a CC0 license), and the Blue Obelisk Data Repository, which is a collaborative project initiated by a number of cheminformatics tools [25], including Kalzium [26] (see Fig. 24.2), the Chemistry Development Kit (CDK), and others. However, it should be noted that these collaborative, open-data resources are small in size and the collaborating community is not, as yet, of critical mass.

In contrast, Wikipedia has an active development community and collaboration within the WikiProject Chemistry [27]. This project has many contributors and keeps track of the chemistry-related pages in Wikipedia and has frequent discussions on how to improve the chemistry on these pages. Collaboration is organized via a project wiki page that can be edited by all contributors (see Fig. 24.3).

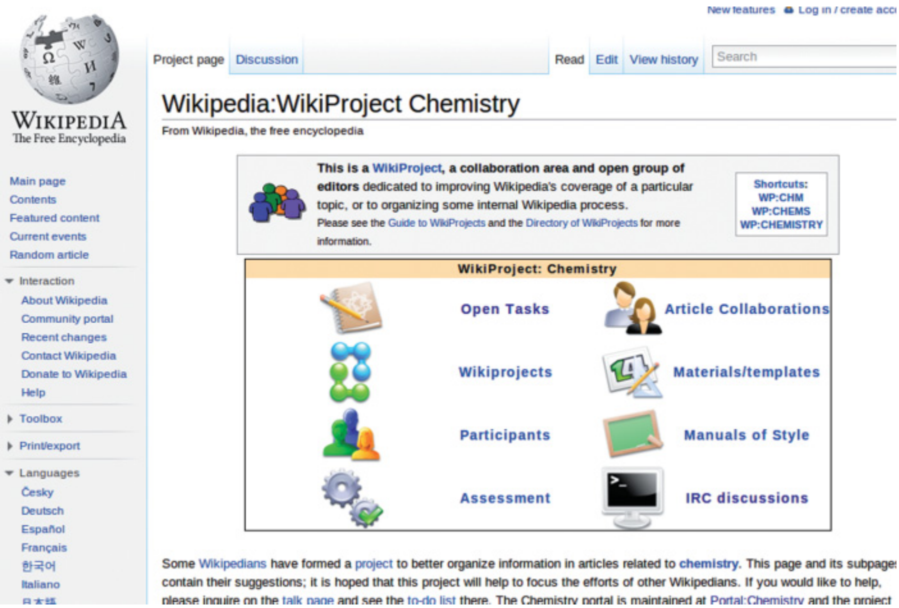
The toxicology community has bootstrapped a community effort to share knowledge around the OpenTox Open Standard [28]. OpenTox provides an interoperable standard for the support of predictive toxicology, including data management, and the specification of algorithms, modeling, validation, and reporting. OpenTox takes advantage of other open standards for data representation, interfaces, vocabularies, and ontologies: Functionality is provided as RESTful services [29], and replies are provided in various formats, including the resource description framework [30], the underlying technology of the Semantic Web [31].



Isotope-Table

Mass	Neutrons	Percentage	Half-life period	Energy and Mode of Decay	Spin and Parity
47.0115 u	20				
48.0018 u	21				
48.9897 u	22				
49.9815 u	23		0.044 s		

Figure 24.2 Screenshot of Kalzium, part of KDE Software Compilation, showing information on isotopes of cobalt using Blue Obelisk Data Repository.



Wikipedia:WikiProject Chemistry

From Wikipedia, the free encyclopedia

This is a WikiProject, a collaboration area and open group of editors dedicated to improving Wikipedia's coverage of a particular topic, or to organizing some internal Wikipedia process. Please see the Guide to WikiProjects and the Directory of WikiProjects for more information.

Some Wikipedians have formed a project to better organize information in articles related to chemistry. This page and its subpages contain their suggestions; it is hoped that this project will help to focus the efforts of other Wikipedians. If you would like to help, please inquire on the talk page and see the to-do list there. The Chemistry portal is maintained at Portal:Chemistry and the project

Figure 24.3 Home page of WikiProject Chemistry, organizing editing of chemical articles on Wikipedia.

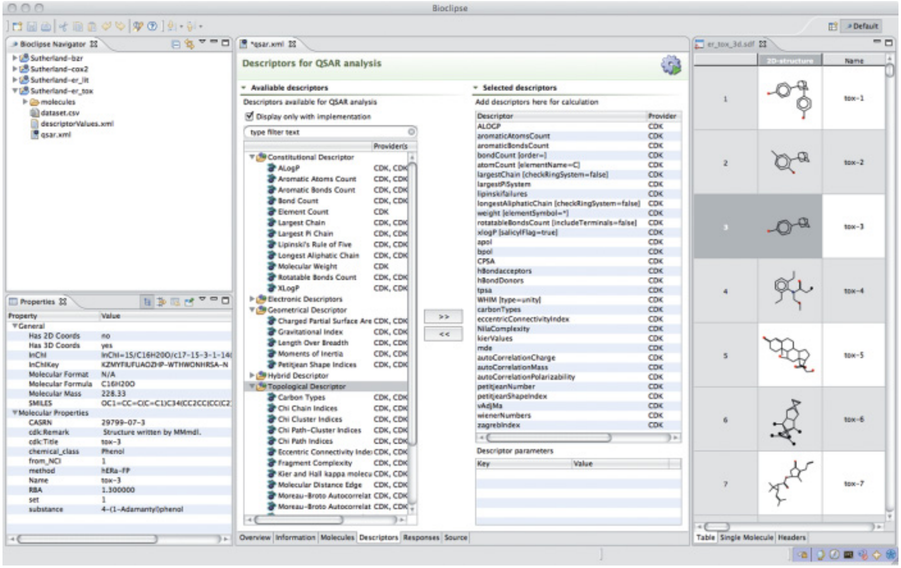
24.3.1 Data Standardization and Interoperability

Sharing of information and data requires well-developed standards and exchange formats as discussed in Chapter 13. An example in cheminformatics is the extensible Chemical Markup Language (CML) [32], which is an approach to manage primarily molecular data, which has been extended to also comprise other entities, including reactions and spectra. Another example is the Human Proteome Organization–Proteomics Standards Initiative (HUPO-PSI) molecular interaction format for the representation of molecular interaction data [33]. The advantage of standardized file formats is that applications can share information without loss of data, and it is becoming increasingly common that data must be deposited in public repositories in open exchange formats prior to publication in scientific journals.

The cheminformatics community is slowly moving toward a more classical standardization of knowledge: the use of ontologies (see also Chapter 12). Ontologies are formal representations that are used to define concepts and their relationships in a specific domain. By explicitly defining what a term means, it defines how it should be used. Likewise, knowledge expressed with terms defined in ontologies is more precise as others can then look up what the exact meaning is.

There are various levels of detail in an ontology and, in its most simple form, the ontology is a controlled vocabulary. An example is the International Union of Pure and Applied Chemistry (IUPAC) Gold Book [34], which specifies chemical terminology. More detailed ontologies, such as those used by the knowledge management community, define terms in much more detail, identifying classes, the hierarchy of classes (e.g., used by the Gene Ontology [35, 36]), and relationships between classes. For example, a chemical ontology can specify what a molecule is, that it is a subclass of chemical entities, that a molecule can have a boiling point, and that a boiling point is a physical property of a chemical entity. These are representative of the type of facts that are expressed in domain ontologies. Ontologies have been used in chemistry since at least the 1980s [37] but have received renewed interest lately [38–40], possibly triggered by the open Extensible Markup Language (XML) [41] and the Web Ontology Language (OWL) standards [42].

Ongoing community efforts to define ontologies related to cheminformatics include the OpenTox API mentioned earlier, the Blue Obelisk Descriptor Ontology, and the Chemical Information Ontology, a cheminformatics-oriented ontology [43]. Another ontology recently introduced to simplify building knowledge bases is the open exchange format QSAR-ML [44], which aims at representing data sets for quantitative structure–activity relationships (QSARs) in an open and completely reproducible way. In QSARs, chemical structures are described by numerical vectors (known as descriptors), and QSAR-ML makes use of the Blue Obelisk Descriptor Ontology for uniquely defining these descriptors. Also included is support for multiple, alternative implementations of these descriptors, which could be available on the local



Property	Value
General	
Has 2D Coords	no
Has 3D Coords	yes
InChI	InChI=1S/C16H20O/c17-15-3-1-14(
InChIKey	KZMYFIUFUAOZHP-WTHWONHRSA-N
Molecular Format	N/A
Molecular Formula	C16H20O
Molecular Mass	228.33
SMILES	OC1=CC=C(C=C1)C34(CC2CC(CC(C2
Molecular Properties	
CASRN	29799-07-3
cdk:Remark	Structure written by MMmdl.
cdk:Title	tox-3
chemical_class	Phenol
from_NCI	1
method	hERa-FP
Name	tox-3
RBA	1.300000
set	1
substance	4-(1-Adamantyl)phenol

Descriptors for QSAR analysis

Available descriptors

Descriptors available for QSAR analysis

☑ Display only with implementation

type filter text

	Providers
Constitutional Descriptor	
ALogP	CDK, CDK
Aromatic Atoms Count	CDK, CDK
Aromatic Bonds Count	CDK, CDK
Bond Count	CDK, CDK
Element Count	CDK
Largest Chain	CDK, CDK
Largest Pi Chain	CDK, CDK
Lipinski's Rule of Five	CDK, CDK
Longest Aliphatic Chain	CDK, CDK
Molecular Weight	CDK
Rotatable Bonds Count	CDK, CDK
XLogP	CDK, CDK
Electronic Descriptors	
Geometrical Descriptor	
Charged Partial Surface Are	CDK, CDK
Gravitational Index	CDK, CDK
Length Over Breadth	CDK, CDK
Moments of Inertia	CDK, CDK
Petitjean Shape Indices	CDK, CDK
Hybrid Descriptor	
Topological Descriptor	
Carbon Types	CDK, CDK
Chi Chain Indices	CDK, CDK
Chi Cluster Indices	CDK, CDK
Chi Path-Cluster Indices	CDK, CDK
Chi Path Indices	CDK, CDK
Eccentric Connectivity Inde	CDK, CDK
Fragment Complexity	CDK, CDK
Kier and Hall kappa molecu	CDK, CDK
Molecular Distance Edge	CDK, CDK
Moreau-Broto Autocorrelat	CDK, CDK
Moreau-Broto Autocorrelat	CDK, CDK

Selected descriptors

Add descriptors here for calculation

Descriptor	Provider
ALOGP	CDK
aromaticAtomsCount	CDK
aromaticBondsCount	CDK
bondCount [order=]	CDK
atomCount [elementName=C]	CDK
largestChain [checkRingSystem=false]	CDK
largestPISystem	CDK
lipinskifailures	CDK
longestAliphaticChain [checkRingSystem=false]	CDK
weight [elementSymbol=*]	CDK
rotatableBondsCount [includeTerminals=false]	CDK
xlogP [salicylFlag=true]	CDK
apol	CDK
bpol	CDK
CPSA	CDK
hBondacceptors	CDK
hBondDonors	CDK
tpsa	CDK
WHIM [type=unity]	CDK
carbonTypes	CDK
eccentricConnectivityIndex	CDK
NilaComplexity	CDK
kierValues	CDK
mde	CDK
autoCorrelationCharge	CDK
autoCorrelationMass	CDK
autoCorrelationPolarizability	CDK
petitjeanNumber	CDK
petitjeanShapeIndex	CDK
vAdjMa	CDK
wienerNumbers	CDK
zagrebIndex	CDK

Descriptor parameters

Key	Value

	2D-structure	Name
1		tox-1
2		tox-2
3		tox-3
4		tox-4
5		tox-5
6		tox-6
7		tox-7

Figure 24.4 Screenshot showing creation of QSAR data set in Bioclipse, where a set of molecules is aggregated and molecular descriptors are selected, creating a numerical representation suitable for statistical modeling.

computer or via remote Web services. QSAR-ML also comes with a reference implementation for the Bioclipse workbench [45, 46], which provides a graphical interface for setting up QSAR data sets, as shown in Figure 24.4. Prominent features include adding and normalizing chemical structures in various formats, cherry-picking local and remote descriptor implementations, adding responses and metadata, and finally performing all calculations and exporting the complete data set in QSAR-ML. Standardized QSAR opens up new ways to store, query, and exchange analysis and makes it easy to join, extend, combine, and work collectively with data.

24.3.2 Linking Knowledge Bases

One of the keys to collaboration is also the sharing of knowledge. A prominent role here is in the linking of various databases, thereby allowing their integration and, where appropriate, their federation. Traditionally, linking databases is done by using shared identifiers. Well-known identifiers for chemical structures include database-specific identifiers such as the CAS registry number [47], the PubChem compound identifier [48], and the ChemSpider identifier [22]. When these are shared, they can be used to connect databases. Alternatively, one could use an identifier which can be calculated from the object itself. For a wide set of small, organic molecules the InChI [49] fulfills this role [50].

An interesting proposal was made with the resource description framework (RDF), [30] which suggests that a universal resource identifier (URI) [51] is

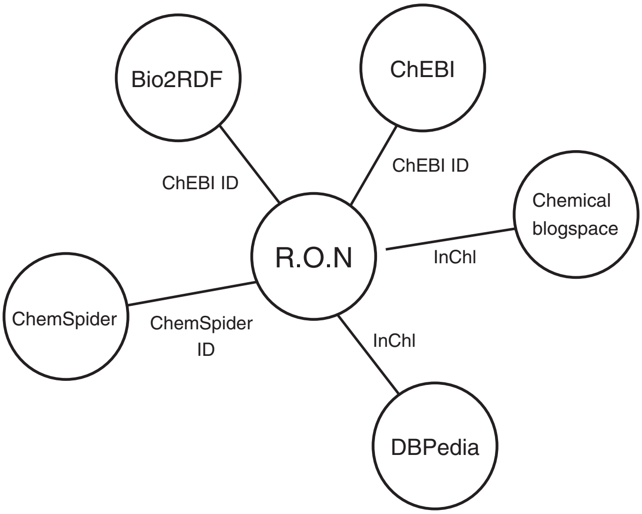


Figure 24.5 The rdf.openmolecules.net website operates as a hub in the Semantic Web for small molecules by using the InChI to create unique molecular URIs to provide a referenceable RDF resource for any compound.

used to identify information in a database. Common URIs include Web addresses, such as http://www.chemspider.com/. The URI specification is an open standard that formalizes the structure of identifiers: an identifier consists of a scheme (http), followed by a : and then an authority prefixed with // (//www.pharmbio.org) and finally a path (/). An example of a URI that does not contain an authority is the URI used in Web pages to link to an e-mail address, such as mailto:cdk-users@lists.sf.net, where the scheme is mailto, and the path is the e-mail address. Other URIs include the life science identifier (LSID) [52]. For example, The LSID for the 1AFT protein in the PDB database is urn:lsid:pdb.org:pdb:1aft. The urn scheme in the LSID refers to the uniform resource name (URN) specification, which is based on the URI but aimed at being location independent [53].

Using these URIs, it is possible to create a linked open-data network, linking together different databases on the Semantic Web. A subsection of this linked data network is created by the website http://rdf.openmolecules.net/, which takes advantage of the International Chemical Indentifier (InChI) to create a network for small molecules. Figure 24.5 shows how this is used to create links to the Chemical Entities of Biological Interest (ChEBI) [54], ChemSpider, NMRShiftDB [55], and other resources [56]. By taking advantage of such identifiers, the various participants can easily collaborate but work quite independently at the same time. Other large initiatives in this area include the Semantic Web for Health Care and Life Sciences Interest Group of the World Wide Web Consortium [57], Bio2RDF [58], and Chem2Bio2RDF [59].

Userscripts also present an approach to linking resources with a very low barrier to entry. The idea of a userscript was first made available by the

Greasemonkey Mozilla extension [60] and later by the Ubiquity Mozilla extension [61]. A userscript is simply a small JavaScript program that runs on the client browser and can modify a given Web page before it is displayed to the viewer. In other words, given permission to do so, a userscript can completely rewrite a Web page in any way deemed appropriate and necessary to the task at hand. This opens up exciting possibilities in annotating Web content and linking Web content to arbitrary data sources. A variety of userscripts for cheminformatics have been described by Willighagen et al. [62]. For example, when viewing a Web page describing chemistry, a userscript can be written to take the text and run it through a chemical entity recognition tool (such as OSCAR3 [63]) and then highlight terms that were recognized. Such a script can be further enhanced by not only highlighting recognized terms but also inserting hyperlinks to chemical databases such as PubChem or ChemSpider. Another userscript application described is to display the three-dimensional (3D) structures of molecules when browsing PubChem Web pages. Previously, PubChem had not provided 3D structure information, while Indiana University had separately generated a single low-energy conformer for 99% of PubChem and stored them in an independent database. A userscript was implemented that when run on a PubChem compound page would identify the compound ID and retrieve the corresponding 3D structure (if available) from the Indiana University database and then display it in a Jmol window. Key to the functionality of many cheminformatics userscripts is the use of freely accessible cheminformatics Web services (Section 24.3.1) and databases.

24.4 COLLABORATIVE COMPUTING

The development of the necessary infrastructure and tools to link knowledge bases is a fundamental requirement for efficient collaborations. The previous sections have highlighted a variety of efforts in these areas. While it is true that collaborative efforts (in any field) are primarily a function of social interactions, it is important to remember that collaborations need not be directly between individuals. Rather, they can also be mediated by software. From this point of view, there have been a number of developments in the last few years that allow individuals unrelated to each other in terms of formal collaborative agreements to interact with each others' resources. But such interactions do not necessarily have to involve remote resources. Instead, a collaboration could also be in the form of shared specifications. That is, individuals could collaborate on the specification of a process or program, which would then be run locally using each collaborator's own resources. Finally, to achieve these types of collaborative efforts, technologies to support these are necessary. Many of these are well established, including mailing lists and chat systems, whereas a number are more recent and include service registries. The following sections discuss these facets of collaborative computing in more detail.

24.4.1 Shared Computing Services

Distributed computing has gone through many phases starting with remote procedure calls (RPCs) [64] in the 1970s–1980s to the Common Object Request Broker Architecture (CORBA) [65] in the 1990s. More recently, distributed computing in the form of so-called Web services have emerged. Because these services expose methods, like RPC and CORBA, but also data, we consider Web services to be a generalization, allowing one to provide access to both data (contained in databases) as well as algorithms.

There are many protocols on which Web services can be based. These include Simple Object Access Protocol (SOAP) [66], Representational State Transfer (REST) [29], and Extensible Messaging and Presence Protocol (XMPP) [67]. Most Web service protocols are designed to work with multiple types of transport layers [Hypertext Transfer Protocol (HTTP), User Datagram Protocol (UDP), etc.], but the majority of Web service protocols currently in use work over HTTP. This approach results in significantly easier deployment and access of services since HTTP is ubiquitous across the Internet. For a more detailed review of Web service technologies the reader is referred to Curcin et al. [68] and Fielding et al. [69].

In this section we provide a brief overview of cheminformatics Web services and some use cases highlighting the collaborative potential of such Web services. Indiana University has developed a number of cheminformatics Web services [70] that provide access to core cheminformatics methods (fingerprints, 2D depiction, and various molecular descriptors), statistical techniques (using R [71] as the backend), and chemical database access methods. Most of these services are implemented in Java using the Chemistry Development Kit [72] and SOAP as the underlying protocol. The services have been used in a variety of scenarios. For example, the fingerprint and statistical model services were employed by Indiana University to develop models to predict the activity of user-supplied compounds against the NCI60 cancer cell lines [70]. Interestingly, the final predictive model was itself converted to a Web service and thus accessible by any SOAP client, remote or local. Note that, while the model service was also located at Indiana University, it could easily make use of fingerprint services located at other sites, anywhere across the world.

Recently, these Web services have been forked and reimplemented as REST-based services. While SOAP services can be "hidden" behind a REST interface, the reimplementation avoids the extra complexity imposed by SOAP by directly exposing the functionality via the REST interface. These services can be found at http://rguha.net/rest. Currently, the services are hosted at multiple locations and include Drexel University in the United States and Uppsala University in Sweden and are utilized by a number of independent applications. An example is the use of these services to enhance an Open Notebook Science (ONS) project [73], as reviewed elsewhere (see Chapter 25). The ONS Solubility Challenge is an ongoing project in which a number of research groups have experimentally determined the solubility of a variety of solutes in a variety of solvents. All of the data generated as part of this project is publicly hosted on a GoogleDocs spreadsheet and outsiders are encouraged to explore and mine the data, with the hope that their results also will be open. At the time of writing the project has seen contributions from a number of people, including chemists, mathematicians, and programmers. As the project has grown, the number of measurements now numbers in the hundreds. The spreadsheet contains alphanumeric identifiers for solutes and solvents along with Simplified molecular input line entry specification (SMILES) representations and solubility data. In a number of cases, external references are also included. While the use of Google spreadsheets is a very simple way to share data, the nature of the data makes it unwieldy to explore. In general, while numeric solubility data are useful, it is more appropriate to explore it from a chemical point of view—that is, in terms of structures and substructures.

As a result, a simple Web page interface was developed [74] that extracts data from the Google spreadsheet via the Google-provided data application programming interface (API) and presents views of the data or filtered subsets of the data (based on solute or solvent identifiers, substructures, or solubility ranges). The key feature of this application is the incorporation of chemical intelligence by making use of cheminformatics Web services hosted at Uppsala University. By making use of these services, the SMILES strings [75] stored in the spreadsheet could be filtered by the presence or absence of substructures (specified via SMILES Arbitrary Target Specification [SMARTS] [76]). In addition, the services were also employed to provide 2D structure depictions of the results matching satisfying the query. From a collaborative point of view, this application is interesting as the developer had no role in the gathering of the solubility data and did not create the online spreadsheet. Instead, the application made use of public data APIs provided by Google and public Web services hosted at another, remote location to extract and present data that satisfied the requirements of another, external group (i.e., the experimentalists making the measurements).

A related application was developed by another researcher to explore the chemical space via descriptor calculations followed by principal-component analysis [77]. This application also made use of the cheminformatics services hosted at Drexel University as well as visualization services provided by Google, allowing users to generate principal-component plots of the solubility data and thereby understanding the extent of the chemical space occupied by the current set of chemicals.

These applications highlight the fact that distributed software resources were key in allowing multiple, unrelated parties to collaborate on a publicly available data set. While this type of collaboration could certainly be achieved using traditional software resources (i.e., locally installed libraries and programs), the presence of freely accessible Web services (for both cheminformatics as well as data and visualization) allows arbitrary individuals or groups to develop novel applications that were not considered by the original researchers. Furthermore, the free and distributed nature of the resources allows such developers to free themselves of software installation and management and onerous licensing conditions.

24.4.2 Sharing Computation Specifications

A simple example of a "computation specification" is the source code of a program or even an input file for a program. In both cases, the document fully specifies what is required to run a program to achieve a desired result. Clearly there have been many mechanisms to share such specifications. However, with a few exceptions these sharing mechanisms tend to simply collect source code, for example, in a central repository and let users access it from there. Usually, there is no metadata attached or associated with the source code, and thus exploring related cases is difficult.

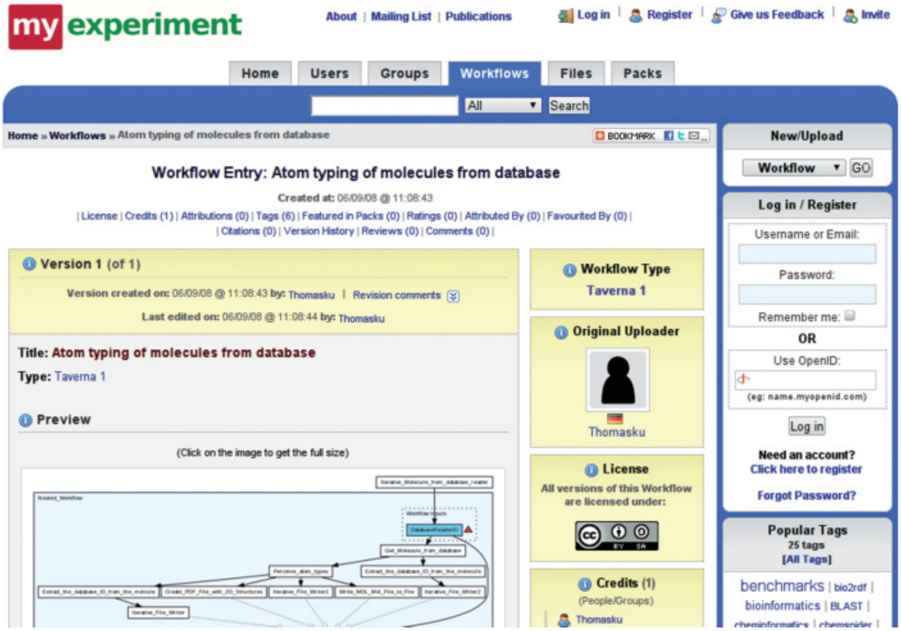
Recently, there has been a significant amount of effort devoted to the development of "workflow" or "pipelining" tools for chem- and bioinformatics. These tools encapsulate core cheminformatics tasks such as reading in SMILES, evaluating descriptors, and so on, in simple graphical elements (usually boxes and connectors). A user can then arrange sequences of such elements to perform a task. The graphical user interface (GUI) approach, coupled with the fact that in most cases such workflow tools require no programming knowledge, makes such tools very attractive to experimentalists and other users with limited programming experience. A number of such tools are available, including Pipeline Pilot [78], KNIME [79], and Taverna [80].

Another development around workflow tools is the sharing of workflow specifications (or "programs"), exemplified by the MyExperiment website [81], which allows sharing of scientific workflow of various formats, including Taverna [80, 82] (see Fig. 24.6) and Bioclipse [45, 46]. MyExperiment is a social website where scientists can share workflow specifications, tag them, categorize them, or modify and upload new, improved versions. This makes it easy to reproduce analyses and potentially improve scientific progress.

24.4.3 Online Cheminformatics Computation

It should be noted that these online workflow sharing services do require download of the workflows to a local desktop, where they can then be run. Both Taverna and Bioclipse provide the means to download a workflow shared on MyExperiment.org from within the desktop software, but it is not possible to run the computation at a remote server unless the workflow specifies that. However, the ongoing evolution of collaborative cheminformatics is making this possible and various companies are appearing that develop software that make it possible to design and run workflows via websites.

One example of this approach is Wingu Elements [83], by the Boston start-up Wingu. This product provides a cloud-based platform for research teams to define workflows, make them available within the collaboration, and then share the results as shown in Figure 24.7.



Home » Workflows » Atom typing of molecules from database

Workflow Entry: Atom typing of molecules from database

Created at: 06/09/08 @ 11:08:43

| License | Credits (1) | Attributions (0) | Tags (6) | Featured in Packs (0) | Ratings (0) | Attributed By (0) | Favourited By (0) |
| Citations (0) | Version History | Reviews (0) | Comments (0) |

Version 1 (of 1)

Version created on: 06/09/08 @ 11:08:43 by: Thomasku | Revision comments

Last edited on: 06/09/08 @ 11:08:44 by: Thomasku

Title: Atom typing of molecules from database

Type: Taverna 1

Preview

(Click on the image to get the full size)

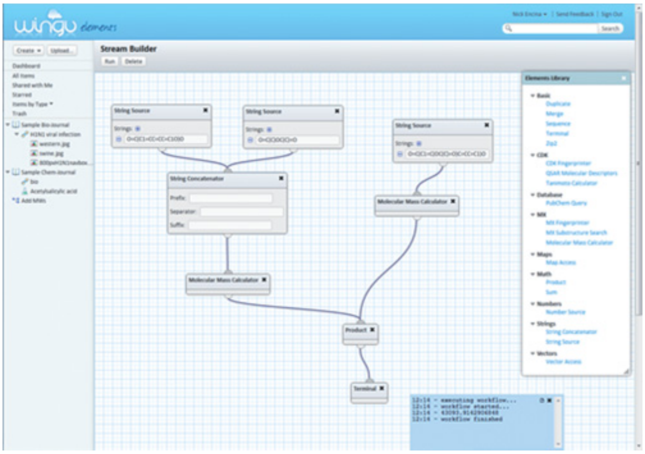
Workflow Type: Taverna 1

Original Uploader: Thomasku

License: All versions of this Workflow are licensed under:

Credits (1) (People/Groups): Thomasku

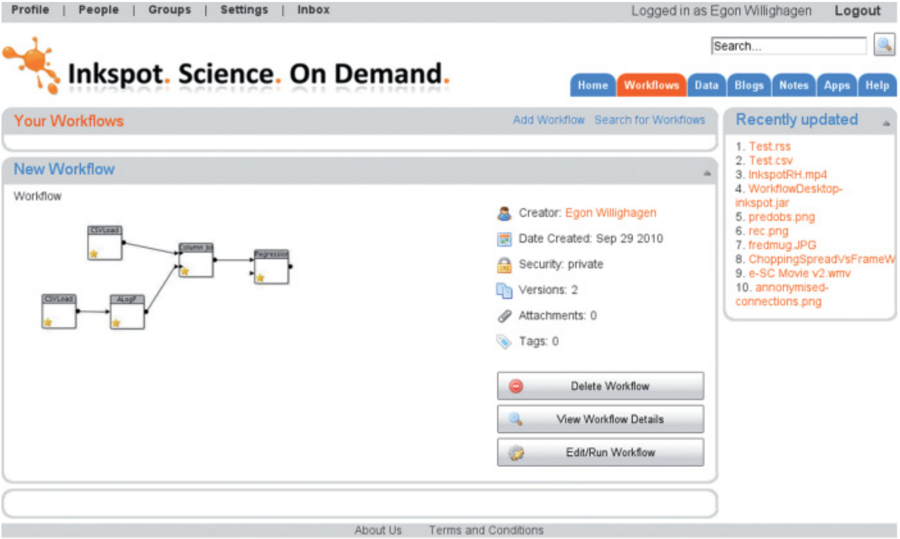
Figure 24.6 Screenshot of MyExperiment.org Web page for CDK–Taverna workflow (http://www.myexperiment.org/workflows/389.html).



```
12:14 - executing workflow...
12:14 - workflow started...
12:14 - 43093.9142906848
12:14 - workflow finished
```

Figure 24.7 Screenshot of Wingu Elements, an online scientific computing platform with cheminformatics functionality.

A second online collaborative cheminformatics platform is the Inkspot Platform by Inkspot Science [84]. On this platform workflows can be designed, run, and shared (see Fig. 24.8). An extra dimension is given here by the company to provide hosting, thereby allowing the creation of small communities or research projects, much like open-source projects take advantage of hosting, services like the aforementioned SourceForge and GitHub.



Creator: Egon Willighagen

Date Created: Sep 29 2010

Security: private

Versions: 2

Attachments: 0

Tags: 0

1. Test.rss
2. Test.csv
3. InkspotRH.mp4
4. WorkflowDesktop-inkspot.jar
5. predobs.png
6. rec.png
7. fredmug.JPG
8. ChoppingSpreadVsFrameW
9. e-SC Movie v2.wmv
10. annonymised-connections.png

Figure 24.8 Screenshot of Inkspot platform for collaborative cheminformatics analyses.

24.5 MANAGING COLLABORATIVE PROJECTS

We have provided a brief overview of various technologies that enable and enhance collaborative projects in the field of cheminformatics. Yet, one aspect still remains open to discussion. While tools are available to share source code and experimental data, how should such collaborative projects be managed? When the collaboration is between two individuals, project management is not a significant problem, but it becomes much more problematic when the number of collaborators grows larger.

The first aspect is communication between members within the collaboration. A number of mechanisms are available, including mailing lists and online messaging systems. Both of these methods have a long history and mailing lists are a useful way to broadcast messages to all members of a group [85]: By sending an e-mail to the mailing list address it is automatically distributed across all of the participants. This has an associated downside as members may not desire such broadcasted messages which may not be relevant to them. Furthermore, e-mail is relatively inefficient at handling multiple conversations, though threading does alleviate this. Modern e-mail systems such as Google Mail have provided a number of enhancements to improve the handling of multiple conversations in a mailing list, such as automatic filtering of messages to mailing lists directly into a folder, instead of the main inbox.

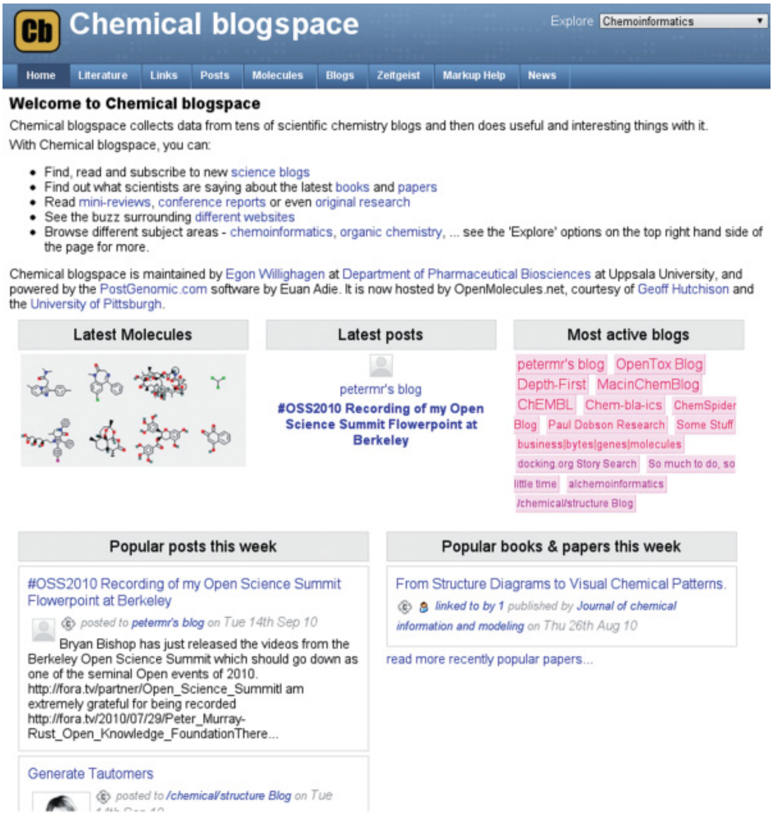
One important aspect of mailing lists is that they are not real time or interactive. On the other hand, messaging systems such as Internet Relay Chat (IRC) [86] or instant messengers (Yahoo Chat, AOL, etc.) offer real-time interactivity between participants. These systems allow for direct interactions between multiple members and are extremely useful for immediate problem solving and discussions. Of course, since these technologies are text based, it can be slower to communicate problems than it would be using phone or video conferencing. On the other hand, these systems are very light on resources and quite efficient on slow Internet connections. The use of IRC is particularly prominent among open-source projects. For example, Bioclipse, OpenBabel [87], and CDK developers use, respectively, the #bioclipse, #openbabel, and #cdk channels on the FreeNode.net network [88]. While it is common to use dedicated IRC clients (see Wikipedia [89]), these channels can also be accessed via a Web interface at http://webchat.freenode.net/.

More recently, weblogs, or blogs, have become a useful mode of communication. This approach allows a degree of interactivity between the producer of the blog and readers but is primarily a vehicle for an individual or group to provide updates. Of course, by allowing multiple people to post on the blog, it can be a useful way for a collaborative group to provide updates and information on the project. Blogs are also useful from the consumer's point of view since they are a pull technology. That is, the consumer (i.e., reader) will usually read the blog via an RSS reader and thus, rather than receive updates from the blog, will read new posts when desired.

One interesting aspect is that increasingly this communication is becoming more open and no longer limited to one project as is often the case for mailing lists. Many open source developers have started using blogs, where they discuss algorithms, theories, and so on, Among those blogs are those of two of the authors of this chapter [90, 91], but other blogs include the excellent one of Gilleain Torrance [92], Noel O'Boyle [93], and Tim VanderMeersch [94].

Blog planets and aggregators play an important role here. Various cheminformatics projects have blog planets, where the blogs from developers and users from the community around that project are aggregated. For example, the Chemistry Development Kit and Bioclipse have planets at, respectively, http://pele.farmbio.uu.se/planetcdk/ and planet.bioclipse.net. Aggregators also aggregate blogs, but not necessarily around a specific developer or user community. One such website that aggregates cheminformatics blogs is Chemical Blogspace (see Fig. 24.9).

A similar role is played by question-and-answer websites, a new type of communication channel popularized by StackOverflow [95]. This communication concept is used, for example, by the Blue Obelisk eXchange at http://blueobelisk.shapado.com/ (see Fig. 24.10), where people can ask questions on how to use a particular cheminformatics library or how to solve a particular problem. These sites essentially extend the concept of frequently asked question (FAQ) sites, except that they are grown and maintained by a community rather than by a single person. In addition, novel mechanisms such as merit points and badges and the ability to up vote (or down vote) for answers provide a "social incentive" to the users of such sites to engage in the community (as opposed to simply taking information, also called leeching).



Welcome to Chemical blogspace

Chemical blogspace collects data from tens of scientific chemistry blogs and then does useful and interesting things with it.

With Chemical blogspace, you can:

- Find, read and subscribe to new science blogs
- Find out what scientists are saying about the latest books and papers
- Read mini-reviews, conference reports or even original research
- See the buzz surrounding different websites
- Browse different subject areas - chemoinformatics, organic chemistry, ... see the 'Explore' options on the top right hand side of the page for more.

Chemical blogspace is maintained by Egon Willighagen at Department of Pharmaceutical Biosciences at Uppsala University, and powered by the PostGenomic.com software by Euan Adie. It is now hosted by OpenMolecules.net, courtesy of Geoff Hutchison and the University of Pittsburgh.

Latest Molecules

Latest posts

petermr's blog

#OSS2010 Recording of my Open Science Summit Flowerpoint at Berkeley

Most active blogs

petermr's blog OpenTox Blog Depth-First MacinChemBlog ChEMBL Chem-bla-ics ChemSpider Blog Paul Dobson Research Some Stuff business|bytes|genes|molecules docking.org Story Search So much to do, so little time alchemoinformatics /chemical/structure Blog

Popular posts this week

#OSS2010 Recording of my Open Science Summit Flowerpoint at Berkeley

posted to *petermr's blog* on Tue 14th Sep 10

Bryan Bishop has just released the videos from the Berkeley Open Science Summit which should go down as one of the seminal Open events of 2010. http://fora.tv/partner/Open_Science_SummitI am extremely grateful for being recorded http://fora.tv/2010/07/29/Peter_Murray-Rust_Open_Knowledge_FoundationThere...

Generate Tautomers

posted to */chemical/structure Blog* on Tue

Popular books & papers this week

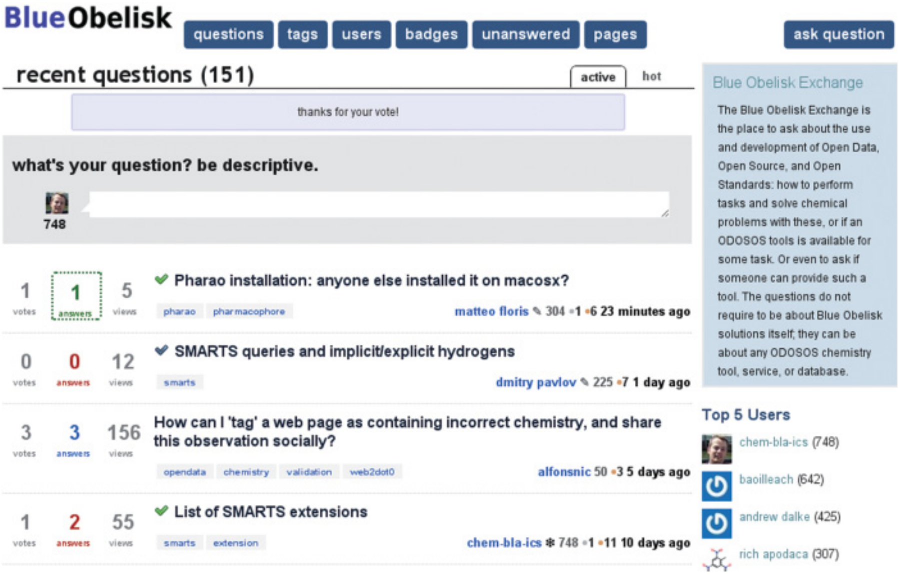
From Structure Diagrams to Visual Chemical Patterns.

linked to by 1 published by *Journal of chemical information and modeling* on Thu 26th Aug 10

read more recently popular papers...

Figure 24.9 Cheminformatics section of chemical blogspace aggregates blog posts from cheminformaticians, providing a platform for discussion of algorithms and theory within a larger community than that of a single project.

The technologies discussed so far have focused on communication between members of a collaboration and other interested parties. But another vital aspect of collaborative projects is the development of documentation—ranging from API documentation to tutorials and policy documents. While one could exchange documents via e-mail, one very quickly runs into the problem of keeping everybody's editions synchronized. Collaborative document editing systems have recently been developed that directly address this problem. One example is Google Docs, which is an online resource that allows one to create documents, spreadsheets, and presentations and then share them between other users. Each authorized user can edit the document and, more importantly, multiple users can simultaneously work on these documents. The service automatically tracks the edits by each user and provides an intuitive view of the document history, allowing one to view the edits made by each user. The documents can be exported to a variety of common formats, allowing one to introduce such documents into the traditional workflow.



recent questions (151)

thanks for your vote!

what's your question? be descriptive.

748

votes	answers	views	question	tags	user
1	1	5	✓ Pharao installation: anyone else installed it on macosx?	pharao, pharmacophore	matteo floris ✎ 304 •1 •6 23 minutes ago
0	0	12	✓ SMARTS queries and implicit/explicit hydrogens	smarts	dmitry pavlov ✎ 225 •7 1 day ago
3	3	156	How can I 'tag' a web page as containing incorrect chemistry, and share this observation socially?	opendata, chemistry, validation, web2dot0	alfonsnic 50 •3 5 days ago
1	2	55	✓ List of SMARTS extensions	smarts, extension	chem-bla-ics ✻ 748 •1 •11 10 days ago

Blue Obelisk Exchange

The Blue Obelisk Exchange is the place to ask about the use and development of Open Data, Open Source, and Open Standards: how to perform tasks and solve chemical problems with these, or if an ODOSOS tools is available for some task. Or even to ask if someone can provide such a tool. The questions do not require to be about Blue Obelisk solutions itself; they can be about any ODOSOS chemistry tool, service, or database.

Top 5 Users

- chem-bla-ics (748)
- baoilleach (642)
- andrew dalke (425)
- rich apodaca (307)

Figure 24.10 Blue Obelisk eXchange question-and-answer website is an open platform where people can ask questions regarding the use of cheminformatics software or ask for solutions to a problem they post.

Google Docs is a useful solution to the problem of collaborative editing of traditional documents. Wikis provide another approach that is more free-form. The fundamental idea of a wiki is that it is a collection of pages linked via hyperlinks and authorized users can edit, add, to or delete these pages arbitrarily. Most wikis will also keep a history of the edits made to each page, allowing one to track who did what and when.

Additionally, it is possible to add supplementary files to a wiki, which allows one to record and track the entire state of a project over time (cf. the ONS Solubility Challenge on Wikispaces [96]). While wikis are useful, they are not necessarily the best solution for all cases. For example, in software development projects, keeping an associated wiki up to date with the state of the project can be tedious, especially if the development is very rapid. While the wiki might be useful, for material such as tutorials and so on, it is rarely a good solution for API documentation. Many would argue that even usage information should be a part of the API documentation and should be written in line with the code. Such inline documentation can easily be extracted and formatted using tools such as Doxygen [97] or Sphinx [98] and linked to from the wiki. This last point highlights one advantage of a wiki type of system—it is a very easy way to aggregate resources via linking rather than include them directly in the wiki itself. This leads to significantly lower efforts in maintaining the wiki.

In addition to these systems, tools that have traditionally been focused on software development can be usefully applied to other scenarios. A good example is the use of a bug tracking systems to keep track of feature requests or new ideas. These systems allow one to keep a list of these issues and possibly assign them to one or more people for followup.

It should be noted that the software systems to support the ideas described here can be obtained in a variety of forms. Each component described here can be obtained individually, require the collaboration to set them up whereas more comprehensive solutions also exist that couple wikis, and feature tracking systems and collaborative editing systems in a single package. In addition, one has a choice of open-source or commercial solutions from which to choose.

We have discussed a variety of technologies that facilitate collaborative project management. But a vital component of this is the management of people within such projects. There are many approaches to this, ranging from committees to dictatorships (benevolent or otherwise). In this chapter we do not discuss the merits or demerits of any given approach, save to say that efficient personnel management is vital to the successful completion of a collaborative project.

24.6 CONCLUSION

In this chapter we have discussed the various aspects of collaborative cheminformatics. We have outlined tools available for software development, building knowledge bases, and tools for collaborative computing. We have also outlined the roles that open source, open data, and open standards have in building successful collaborations, including the roles of licensing as a social contract between collaborators and communication channels to discuss ideas. Similarly, we have shown how open standards make it easier to build collaborative knowledge bases as they provide a unified and clear-to-all interface to the data. We have further described how collaborative computing, both locally and remotely, can be established and how scientists can share high-level workflow specifications.

As technologies continue to improve over the next few years, these kind of collaborative cheminformatics uses will become easier and easier. The cheminformatics applications outlined here merely present a view on what will be possible.

REFERENCES

1. Krause S, Willighagen E, Steinbeck C. JChemPaint—Using the collaborative forces of the Internet to develop a free editor for 2D chemical structures. *Molecules* 2000;5(1):93–98.
2. CVS—Concurrent Versions System. Available: http://www.nongnu.org/cvs/.
3. Apache Subversion. Available: http://subversion.apache.org/.

4. Mercurial SCM. Available: http://mercurial.selenic.com/.
5. Bazaar. Available: http://bazaar.canonical.com/en/.
6. Git—Fast Version Control System. Available: http://git-scm.com/.
7. Sourceforge. Available: http://www.sourceforge.net/.
8. Google code. Available: http://code.google.com/.
9. Launchpad. Available: http://www.launchpad.net/.
10. Github. Available: http://www.github.com/.
11. Gitorious. Available: http://www.gitorious.org/.
12. Rosen L. *Open Source Licensing*. 2004 Available: http://www.rosenlaw.com/oslbook.htm.
13. GNU Lesser General Public License, Version 3. Available: http://www.gnu.org/licenses/lgpl-3.0?-standalone.html, 2007.
14. GNU General Public License, Version 3. Available: http://www.gnu.org/licenses/gpl-3.0-standalone.html, 2007.
15. The MIT License. Available: http://www.opensource.org/licenses/mit-license.html.
16. The BSD License. Available: http://www.opensource.org/licenses/bsd-license.php.
17. Open Source Licenses—Open Source Initiative. Available: http://www.opensource.org/licenses/.
18. Raymond ES. The cathedral and the bazaar. O'Reilly, 2001. Available: http://catb.org/~esr/writings/homesteading/cathedral-bazaar/.
19. PMD. Available: http://pmd.sourceforge.net/.
20. Willighagen EL. How to use GitHub for [CDK—Bioclipse] code review. Technical report, 2010. Available: http://chem-bla-ics.blogspot.com/2010/05/how-to-use-github-for-cdkbioclipse-code.html.
21. The PubChem Project. Available: http://pubchem.ncbi.nlm.nih.gov/.
22. ChemSpider—Database of Chemical Structures and Property Predictions. Available: http://www.chemspider.com/.
23. Steinbeck C, Kuhn S. NMRShiftDB—Compound identification and structure elucidation support through a free community-built web database. *Phytochemistry* 2004;65(19):2711–2717.
24. Apodaca R. ChemPedia Substances. Available: http://chempedia.org/substances.
25. Guha R, Howard MT, Hutchison GR, Murray-Rust P, Rzepa H, Steinbeck C, Wegner J, Willighagen EL. The Blue Obelisk—Interoperability in chemical informatics. *J Chem Inf Model* 2006;46(3):991–998.
26. The KDE Education Project—Kalzium. Available: http://edu.kde.org/kalzium/.
27. WikiProject Chemistry. Available: http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Chemistry.
28. Hardy B, Douglas N, Helma C, Rautenberg M, Jeliazkova N, Jeliazkov V, Nikolova I, Benigni R, Tcheremenskaia O, Kramer S, Girschick T, Buchwald F, Wicker J, Karwath A, Gutlein M, Maunz A, Sarimveis H, Melagraki G, Afantitis A, Sopasakis P, Gallagher D, Poroikov V, Filimonov D, Zakharov A, Lagunin A, Gloriozova T, Novikov S, Skvortsova N, Druzhilovsky D, Chawla S, Ghosh I, Ray S, Patel H, Escher S. Collaborative development of predictive toxicology applications. *J Cheminform* 2010;2(1):7+.

29. Fielding R. Architectural styles and the design of network-based software architectures. PhD thesis, University of California, Irvine, 2000.

30. Carroll JJ, Klyne G. Resource description framework (RDF): Concepts and abstract syntax. W3C recommendation. Available: http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/, 2004.

31. Berners-Lee T, Hendler J, Lassila O. The semantic web. *Sci Am* 2001;284(5): 34–43.

32. Murray-Rust P, Rzepa HS. Chemical markup, XML, and the Worldwide Web. 1. Basic principles. *J Chem Inf Model* 1999;39(6):928–942.

33. Orchard S, Kerrien S. Molecular interactions and data standardisation. *Methods Mol Biol* 2010;604:309–318.

34. *IUPAC Compendium of Chemical Terminology—the Gold Book*. Available: http://goldbook.iupac.org/.

35. The Gene Ontology Consortium. The gene ontology project in 2008. *Nucl Acids Res* 2008;36(Suppl 1):D440–444.

36. Gene Ontology Consortium. The gene ontology in 2010: Extensions and refinements. *Nucl Acids Res* 2010;38(Database issue):D331–335.

37. Gordon JE, Brockwell JC. Chemical inference. 1. Formalization of the language of organic chemistry: Generic structural formulas. *J Chem Inf Computer Sci* 1983; 23(3):117–134.

38. Feldman H, Dumontier M, Ling S, Haider N, Hogue C. Co: A chemical ontology for identification of functional groups and semantic comparison of small molecules. *FEBS Lett* 2005;579(21):4685–4691.

39. Dumontier M, Villanueva-Rosales N. Towards pharmacogenomics knowledge discovery with the semantic web. *Brief Bioinform* 2009;10(2):153–163.

40. Sankar P, Alain K, Aghila G. Model tool to describe chemical structures in xml format utilizing structural fragments and chemical ontology. *J Chem Inf Model* 2010;50(5):755–770.

41. Bray T, Paoli J, Maler E, Yergeau F, Sperberg-McQueen CM. Extensible markup language (XML) 1.0 (fifth edition). W3C recommendation. Available: http://www.w3.org/TR/2008/REC-xml-20081126/, 2008.

42. W3C OWL Working Group. OWL 2 web ontology language document overview. Technical report. W3C. Available: http://www.w3.org/TR/2009/REC-owl2-overview-20091027/, 2009.

43. Chemical information ontology. Available: http://code.google.com/p/semanticchemistry/.

44. Spjuth O, Willighagen EL, Guha R, Eklund M, Wikberg JES. Towards interoperable and reproducible QSAR analyses: Exchange of datasets. *J Cheminform* 2010;2(1):5.

45. Spjuth O, Alvarsson J, Berg A, Eklund M, Kuhn S, Masak C, Torrance G, Wagener J, Willighagen EL, Steinbeck C, Wikberg JE. Bioclipse 2: A scriptable integration platform for the life sciences. *BMC Bioinform* 2009;10(1):397+.

46. Spjuth O, Helmus T, Willighagen EL, Kuhn S, Eklund M, Wagener J, Murray-Rust P, Steinbeck C, Wikberg JE. Bioclipse: An open source workbench for chemo- and bioinformatics. *BMC Bioinform* 2007;8;59.

47. CAS REGISTRY and CAS registry numbers. Available: http://www.cas.org/expertise/cascontent/registry/regsys.html.

48. PubChem compound identifier. Available: http://pubchem.ncbi.nlm.nih.gov/search/help_search.html#Cid.

49. Stein SE, Heller SR, Tchekhovski D. An open standard for chemical structure representation—The IUPAC Chemical Identifier. In *Nimes International Chemical Information Conference Proceedings*. 2003; pp. 131–143. Available: http://www.iupac.org/inchi/Stein-2003-ref1.html.

50. Coles SJ, Day NE, Murray-Rust P, Rzepa HS, Zhang Y. Enhancement of the chemical semantic web through the use of InChI identifiers. *Org Biomol Chem* 2005; 3(10):1832–1834.

51. Berners-Lee T, Fielding R, Masinter L. Uniform Resource Identifier (URI): Generic syntax. RFC 3986 (Standard), January 2005. Available: http://tools.ietf.org/html/rfc3986.

52. Clark T, Martin S, Liefeld T. Globally distributed object identification for biological knowledge bases. *Brief Bioinform* 2004;5(1):59–70.

53. R. Moats. URN syntax. RFC 2141 (Standard), 1997.

54. Degtyarenko K, De Matos P, Ennis M, Hastings J, Zbinden M, Mcnaught A, Alcántara R, Darsow M, Guedj M, Ashburner M. ChEBI: A database and ontology for chemical entities of biological interest. *Nucl Acids Res* 2008;36(Suppl 1): D344–D350.

55. Steinbeck C, Krause S, Kuhn S. Nmrshiftdb—Constructing a free chemical information system with open-source components. *J Chem Inf Computer Sci* 2003; 43(6):1733–1739.

56. Willighagen EL, Alvarsson J, Andersson A, Eklund M, Lampa S, Lapins M, Spjuth S, Wikberg JES. Linking the resource description framework data to cheminformatics and proteochemometrics. *J Biomed Semantics* 2011;2(Suppl 1):S1–S6.

57. Semantic Web for Health Care and Life Sciences Interest Group. Available: http://esw.w3.org/HCLSIG.

58. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J. Bio2rdf: Towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform* 2008;41(5):706–716.

59. Chen B, Dong X, Jiao D, Wang H, Zhu Q, Ding Y, Wild DJ. Chem2bio2rdf: A semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC Bioinform* 2010;11(1):255.

60. Greasemonkey. Available: https://addons.mozilla.org/en-US/firefox/addon/748/.

61. Mozilla Labs—Ubiquity. Available: https://addons.mozilla.org/en-US/firefox/addon/9527/.

62. Willighagen E, O'Boyle N, Gopalakrishnan H, Jiao D, Guha R, Steinbeck C, Wild D. Userscripts for the life sciences. *BMC Bioinform* 2007;8(1):487.

63. Corbett P, Murray-Rust P. High-throughput identification of chemistry in life science texts. *Lecture Notes in Computer Science*, 2006;4216:107–118.

64. A high-level framework for network-based resource sharing. RFC 707 (Standard), 1976. Available: http://tools.ietf.org/html/rfc707.

65. History of CORBA. Available: http://www.omg.org/gettingstarted/history_of_corba.htm.

66. Gudgin M, Hadley M, Mendelsohn N, Lafon Y, Moreau J, Karmarkar A, Nielsen HF. SOAP version 1.2 part 1: Messaging framework (second edition). W3C recommendation. Available: http://www.w3.org/TR/2007/REC-soap12-part1-20070427/, April 2007.
67. Wagener J, Spjuth O, Willighagen E, Wikberg J. XMPP for cloud computing in bioinformatics supporting discovery and invocation of asynchronous web services. *BMC Bioinform* 2009;10(1):279.
68. Curcin V, Ghaem M, Guo Y. Web services in the life sciences. *Drug Discov Today* 2005;10(12):865–871.
69. Fielding RT, Taylor RN. Principled design of the modern web architecture. *ACM Trans Internet Tech* 2002;2(2):115–150.
70. Dong X, Gilbert KE, Guha R, Heiland R, Kim J, Pierce ME, Fox GC, Wild DJ. Web service infrastructure for chemoinformatics. *J Chem Inf Model* 2007;47(4): 1303–1307.
71. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2008.
72. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E. The Chemistry Development Kit (CDK): An open-source java library for chemo- and bioinformatics. *J Chem Inf Computer Sci* 2003;43(2):493–500.
73. Bradley JC, Guha R, Lang A, Lindenbaum P, Neylon C, Williams A, Willighagen EL. *Beautifying data in the real world*. Sebastopol: O'Reilly Media, 2009, Chapter 16.
74. ONS solubility queries. Available: http://toposome.chemistry.drexel.edu/~rguha/jcsol/sol.html.
75. Weininger D. Smiles, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 1988;28(1):31–36.
76. SMARTS—A language for describing molecular patterns. Available: http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.
77. Descriptor chemical space. Available: http://old.oru.edu/cccda/sl/descriptorspace/ds.php.
78. Pipeline Pilot, Accelrys' scientific informatics platform. Available: http://accelrys.com/products/pipeline-pilot/.
79. Berthold MR, Cebron N, Dill F, Gabriel TR, Kotter T, Meinl T, Ohl P, Thiel K, Wiswedel B. KNIME—The Konstanz information miner: Version 2.0 and beyond. *SIGKDD Explor Newsl* 2009;11(1):26–31.
80. Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, Carver T, Glover K, Pocock MR, Wipat A, Li P. Taverna: A tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* 2004;20(17):3045–3054.
81. Goble CA, Bhagat J, Aleksejevs S, Cruickshank D, Michaelides D, Newman D, Borkum M, Bechhofer S, Roos M, Li P, De Roure D. myExperiment: A repository and social network for the sharing of bioinformatics workflows. *Nucl Acids Res* 2010;38 (Suppl):W677–682.
82. Kuhn T, Willighagen EL, Zielesny A, Steinbeck C. CDK-Taverna: An open workflow environment for cheminformatics. *BMC Bioinform* 2010;11(1):159.
83. Wingu Elements. Available: https://wingu.com/.

84. Inkspot Science. Available: http://inkspot.co/.
85. Mailing list—Wikipedia, the free encyclopedia. Available: http://en.wikipedia.org/wiki/Mailing_list.
86. Oikarinen J, Reed D. Internet Relay Chat Protocol. RFC 1459 (Standard), 1993. Available: http://tools.ietf.org/html/rfc1459.
87. OpenBabel. Available: http://openbabel.org/.
88. FreeNode. Available: http://freenode.net/.
89. Comparison of Internet Relay Chat clients—Wikipedia, the free encyclopedia. Available: http://en.wikipedia.org/wiki/Comparison_of_Internet_Relay_Chat_clients.
90. Guha R. So much to do, so little time. Available: http://blog.rguha.net/.
91. Willighagen EL. chem-bla-ics. Available: http://chem-bla-ics.blogspot.com/.
92. Some stuff—An online research notebook. Available: http://gilleain.blogspot.com/.
93. O’Boyle N. Noel O’Blog. Available: http://baoilleach.blogspot.com/.
94. Vandermeersh T. OB, Avogadro and molecular modelling. Available: http://timvdm.blogspot.com/.
95. StackOverflow. Available: http://stackoverflow.com/.
96. ONSChallenge—EXP034. Available: http://onschallenge.wikispaces.com/EXP034.
97. Doxygen. Available: http://www.stack.nl/~dimitri/doxygen/.
98. Sphinx—Python Documentation Generator. Available: http://sphinx.pocoo.org/.

PART IV

THE FUTURE OF COLLABORATIONS

25

COLLABORATION USING OPEN NOTEBOOK SCIENCE IN ACADEMIA

Jean-Claude Bradley, Andrew S. I. D. Lang, Steve Koch, and Cameron Neylon

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.

25.1 INTRODUCTION

Technology has a profound effect on how scientists can communicate with each other. This affects how quickly science can progress and what kinds of collaboration are possible. Although the printing press and the subsequent establishment of scientific journals dramatically increased the ability of researchers to disseminate their results and ideas, close collaborations between geographically separated individuals had to await the availability of telecommunication technologies, particularly the development of the Internet.

Today, the ubiquity of sophisticated and easy-to-use tools to exchange information is enabling the creation of a "shared presence" between people, regardless of their geographical location. Researchers can share not only their data but also details regarding how they processed their data, their interpretation of their results, and their future plans. However, the *ability* to share only translates into actual sharing if there is a motivation to do so. In this chapter we will provide examples of what is possible when researchers choose to share their experimental work in progress. The chapter presents a chronological timeline of some key events in the history of these examples.

25.2 OPEN NOTEBOOK SCIENCE

The term *open notebook science* (ONS) was introduced in 2006 to enable an unambiguous discussion of open collaboration in science [1, 2]. The term *open science* is too broad and nebulous while *open-source science* has been used inconsistently, sometimes referring to open-source software in science. ONS specifically refers to the public sharing of the entirety of one's laboratory notebook, including all associated raw data files. The default assumption is that all experiments from a project are shared in near real time. This allows others to contribute quickly since it can be assumed that, if an experiment is not reported, it has not yet been done [3]. Forms of partial ONS, where there is either a significant delay or selective sharing, can be made explicit by the use of logos [4].

There are some interesting consequences to ONS with respect to collaboration. Since the entire content is shared, not only do others know what has been done in a lab, they can also infer what has not yet been attempted. Potential collaborators can then confidently carry out needed experiments without worrying that they are unnecessarily duplicating work. If they choose to replicate an experiment, then they can do so with the prior knowledge of what happened in all previous attempts.

25.3 USEFULCHEM PROJECT

25.3.1 Platforms

The UsefulChem project was initiated in the Bradley laboratory at Drexel University in the summer of 2005. The concept was to discover and work on urgent problems in chemistry and report on the progress of the project in a transparent way. The project started with the UsefulChem blog on the free and hosted Blogger service provided by Google [5]. A wiki [6] was later established to organize collective information by linking to relevant blog posts or other resources. Wikispaces was chosen as the platform for this purpose because it provided a free hosted service for public wikis and afforded an intuitive visual editor, simplified wikitext, and convenient back-up and alerting capabilities [7].

This model of providing specialized services for free as long as data remain open has been widely exploited for diverse applications on the Web. For example, on Wikispaces, only private accounts require payment. This is a mutually beneficial situation for the client who enjoys free services and for the service provider, where the public accounts provide free examples and testimonials which can serve as a form of advertising for the pay services. Many of these services also monetize the free versions by displaying ads. The first laboratory experiments were recorded on a new blog—UsefulChem Experiments [8, 9]—and information about relevant molecules was collected in posts on the UsefulChem-Molecules blog [10, 11]. The first comment on the Experiments blog from a researcher outside the existing group came from a researcher at the University of Sydney [12], Mat Todd, and provided valuable insight. This contribution was reciprocated later by promoting the Todd group's open project on the chemical praziquantel [13]. Other scientists would continue to periodically comment on later blog posts [14].

By June 2006 it became clear that a blog was not providing the necessary functions for a laboratory notebook, mainly because version control was not available [15, 16]. The plan at this time was to record the laboratory notebook information on the wiki, then copy it over to the Experiments blog when the experiment was finalized. However, in practice, the concept of a "finalized experiment" proved difficult to judge and the wiki was simply used as the actual laboratory notebook. This way errors discovered at any time could be corrected on the wiki with proper version tracking to determine who contributed what and when. The use of a wiki for a laboratory notebook also made it very convenient for mentors to communicate with students by commenting directly on specific sections of a page. The availability of e-mail alerts for any changes on the wiki facilitated very rapid communication.

With the accumulation of data, more effort was invested into providing tools for searching. It was deemed important that both the blogs and wikis be quickly indexed on major search engines. Google Co-op Search allowed for a very simple way of performing a federated search of all of the UsefulChem platforms [17] and was also used later for other multiple ONS resources [18]. Google applications would prove to be key for other sophisticated search and retrieval tools that would evolve over time.

In March 2007 UsefulChem compounds were hosted as part of the eMolecules collection, thereby permitting additional sophisticated services such as substructure searching [19]. The use of Google spreadsheets in UsefulChem for data storage and manipulation proved to be another powerful example of leveraging free hosted resources. Free Google and Sitemeter services also facilitated the discovery of UsefulChem content via license filtering and visitor tracking, respectively [20]. In August 2007 Collaborative Drug Discovery (CDD) provided UsefulChem with a free account to store and share assay results [21]. Neylon's laboratory used another free hosted database application, Dabble, to list people involved in ONS [22].

At the end of March 2007, ChemSpider was first used to manage UsefulChem molecules [23]. A full transition to ChemSpider was completed in June 2007 with the demonstration of substructure searching and the use of the UsefulChem-Molecules blog was discontinued [24]. This free and hosted online chemical database would prove to be integral to many projects. The ability to provide experimental and predicted properties was one of the first essential functionalities exploited. UsefulChem acquired a subdomain on ChemSpider in April 2008 and students were encouraged to upload nuclear magnetic resonance (NMR) spectra of reagents and purified products as open data [25].

In July 2007 a mailing list was created to work through details of challenges related to the UsefulChem project [26]. This was done to capture discussions taking place by e-mail. Collaborators outside of the core UsefulChem team seemed to prefer e-mail over the wiki or blog to communicate and this was done to keep the discussions public.

In April 2008 FriendFeed was first investigated as a collaboration platform for UsefulChem [27]. Its basic function is to aggregate all relevant feeds from various social networking sites for a user to a single account. For example, feeds for all blogs, SlideShare, LinkedIn, Google Reader, YouTube, SciVee, and so on, can be aggregated to one uniform resource locator (URL) [28]. Whenever the user generates a new entry in any of the source accounts, a FriendFeed post is automatically made and reported to all subscribers. Discussions can then take place on FriendFeed itself instead of on the original blog or other type of post. This is particularly convenient since extended discussions on FriendFeed around a post can be referenced with a short URL. Since the open-science community is well represented on FriendFeed, much of the discussion and activity related to UsefulChem now takes place on this platform. This was later detailed in an article in *Chemical and Engineering News* [29].

25.3.2 Medicinal Chemistry: Collaborations Between Synthetic Chemists, Computational Chemists, and Biochemists

The UsefulChem project started with searches on Google Scholar and Scirus in the chemistry category for phrases like "there is a pressing need for," "what is needed now," and "needs to be synthesized." A need for new antimalarial compounds proved to be a recurrent theme [30, 31]. An example of early collaboration spontaneously arose, with renowned blogger David Bradley suggesting to vary the spelling of "synthesize" to the British version of "synthesise" [30].

A deeper collaboration followed the identification of Find-A-Drug as a source of virtual libraries for HIV protease inhibitors [32] and malarial enoyl reductase inhibitors [33]. Find-A-Drug provided a virtual library of diketopiperazines and three-dimensional (3D) docking information of a sample member onto malarial enoyl reductase [34].

The Drexel group started to perform docking calculations using the THINK software [35]. With the intention of adhering to the concept of ONS, the docking runs were recorded using a similar format to wet laboratory experiments so that other researchers would be able to reproduce the computational results and conclusions based on the information provided in the notebook.

A response to an open request for docking collaborators changed the course of the UsefulChem project [36]. A member of the bioinformatics group at Nanyang Polytechnic in Singapore attempted to dock the Ugi product precursors to the diketopiperazine targets and determined that some of them docked onto enoyl reductase. As a result, the problematic cyclization step (see synthesis section below) was abandoned and all subsequent libraries focused on the Ugi products themselves. This is advantageous from a synthetic standpoint since these can be prepared in only one step from readily available starting materials.

In April 2007 Zaharevitz from the National Cancer Institute (NCI) discovered the UsefulChem project through the network of open scientists and offered free testing of compounds for antitumor activity [37]. The first Ugi product was submitted shortly thereafter [38], and in May 2007 the compound was submitted to a tumor inhibition prediction service. Although predicted to be inactive (as was later confirmed [39]), it demonstrated for the first time the "closing of the open-science loop" for drug discovery—where hypothesis formation, docking, synthesis, and assay results were performed openly in real time [40]. This strategy was extended by prioritizing synthetic targets from a virtual library of Ugi products based on the predicted ability to inhibit tumor cell lines. Naphthyl fragments showed up disproportionately in the products with high predicted activity [41]. Zaharevitz further assisted by inviting one of us (JCB) to a National Institutes of Health (NIH) workshop on drug development in January 2008 [42]. Synthetic focus was directed to Ugi product libraries and Guha initiated a malarial enoyl reductase docking study on a 500,000-compound virtual library based on starting materials that could be obtained cheaply and quickly [43]. The most highly ranked compounds from this study were prioritized for synthesis via the Ugi reaction.

Assay results were hosted on CDD and catalyzed the initiation of a new collaboration with the Rosenthal group at the University of California—San Francisco (UCSF), which agreed to run malaria assays for UsefulChem compounds at no charge. The Rosenthal group had previously discovered the malarial enzyme falcipain-2, and it was convenient for them to run an inhibitory assay against that protein, in addition to red blood cell assays to measure the inhibition of infection by the malarial parasite [44]. The focus of the work thus shifted from enoyl reductase to falcipain-2. With a crystal structure and known binding site, docking calculations were performed and two Ugi products in the top 1000 from Guha's docking results were synthesized and shipped to the Rosenthal lab in December 2007 [45]. Activity at the micromolar range against both the enzyme and the infection of red blood cells by the parasite was reported in January 2008 [46]. By August 2008, 4 of the 17 Ugi products tested showed similar results for inhibition of the enzyme and infection [47], clearly a very impressive proof of principle.

25.3.3 Chemical Synthesis Strategy: Collaborations Between Synthetic Chemists, Both Locally and Remotely

A general synthesis was proposed to generate the putative malaria inhibitors suggested by Find-A-Drug, which were based on a 2,5-diketopiperazine scaffold [48]. Further literature searching revealed some examples of the diketopiperazine synthesis on solid support [49, 50]. Finally, in December 2005, a more convenient solution was found based on a Ugi reaction followed by a cyclization [51].

One of the required starting materials for an Ugi-related synthesis strategy to many of the members of the diketopiperazine library was the compound known as DOPAL (3,4-dihydroxyphenylacetaldehyde). This compound could not be purchased and the synthesis of DOPAL therefore became a primary synthetic focus. A question arose as to whether the presence of a phenolic group would interfere in the Ugi reaction [52]. This concern was greatly diminished by a contribution from ChemRefer, where an article reported that an electron-withdrawing group on the aromatic ring is necessary for the phenol to participate in the Ugi reaction [53]. Feedback from an expert in the field supported this assessment [54].

Synthetic methods to prepare DOPAL were discussed on the blog [55] and the synthesis was finally solved in October 2006 [56]. Progress was slowed by errors in the literature. However, a report [57] detailing these errors and linking to the "failed experiments" that uncovered their discovery demonstrated that ONS could be useful for providing more transparency in science and saving time in the future for anyone attempting to repeat the synthesis.

Unfortunately, DOPAL and similar aldehydes proved to be too susceptible to side reactions, and other more stable compounds were used to try to understand the behavior and kinetics of the Ugi reaction first [58, 59]. Research work often has to deviate from initial plans due to unexpected problems. However, the nature of those problems is not usually communicated in sufficient detail (or at all) via traditional channels. In a sense, making the details of these problems easily indexed on major search engines is a type of collaboration with future researchers who may run into similar problems and benefit from the details provided.

25.3.4 Cheminformatics: Collaborations Between Chemists and Programmers

The representation, manipulation, and communication of chemical information in an open-science environment is not a trivial challenge. One of the earliest cheminformatics tasks consisted of converting the format of the first malaria virtual library to one that was easier to share publicly. One of the Find-A-Drug volunteers contributed by providing the library as a simplified molecular input line entry specification (SMILES) list [60], a text-based format consisting of a string of characters that can be easily manipulated in spreadsheets [61]. The discovery of open-source software such as Open Babel [62] would also prove to be critical for the cheminformatics needs of the project. An ecosystem of open science related to cheminformatics evolved over time. Projects with overlapping objectives naturally interlinked at a convenient level. For example, the UsefulChem project had a presence on The Synaptic Leap for the purpose of finding collaborators [63], including a suggestion for a free docking program [64]. Several key individuals with overlapping interests started blogging about their cheminformatics work. Willighagen started CML Explained [65], Apodaca blogged on Depth First [66], Murray-Rust started the CML blog [67] and PeterMR's blog [68], and Williams maintained the ChemSpider blog [69]. Simply following each other's blogs turned out to be a fairly efficient way for the community to collaborate on shared interests. An excellent example of this took place in September 2006 when Willighagen suggested that the Open Source JSpecView applet could be used to view NMR spectra in JCAMP-DX format [70], and this became immensely important to sharing and manipulating raw data from UsefulChem [71], including the monitoring of reactions [72]. The ability to use JSpecView to overlay NMR spectra was particularly useful for monitoring reactions [73] and measuring kinetics when integrated with Excel VBA [74, 75].

Chemical Markup Language (CML) represented a promising way of openly sharing chemical information. As we attempted to create CML really simple syndication (RSS) feeds from our molecules blog, Willighagen and Murray-Rust shared their expertise [76]. It may be better to characterize this type of interaction as a metacollaboration since it did not involve project-specific objectives so much as general ways of representing and manipulating chemical information publicly. Lessons learned in this space would prove to be valuable for quickly starting other chemistry open notebook projects, including the conversion of laboratory notebook pages describing a Ugi synthesis into CML [77].

Willighagen proposed a method of introducing tags into blog posts to make the molecules discussed machine readable [78]. This was experimented with for the UsefulChem blog and provided a new means for potential collaborators to find information via the Chemical Blogspace, an aggregation service for chemistry blogs. Further collaborations ensued. In June 2007 Guha created a public Web service to generate a combinatorial list of all Ugi products resulting from lists of starting materials represented as SMILES [79]. Shattuck created a Web service to deconvolute NMR spectra using JCAMP-DX files as input [80], and in August 2007 an account was created on MyExperiment to attempt to process and organize Web services related to the UsefulChem project [81]. However, productive use of this system would have to await the involvement of new collaborators after the creation of the ChemTaverna project in 2010 [82].

In November 2007 enough data were being generated in the laboratory notebook that it made sense to start abstracting Ugi reaction information into a Google Spreadsheet to compile a CombiUgi Master Table [83]. Since each record points to the corresponding laboratory notebook page, information is not lost, but the abstraction allows for semantic querying of the data set. Attempts were also made to convert the workflows into organized machine-readable formats, involving a discussion between others (Williams, Willighagen, and Murray-Rust) interested in overlapping objectives [84, 85]. In April 2008, Guha created the first version of a model to predict precipitation from methanol based on molecular descriptors of the products [86].

In March 2010 the reactions recorded in the UsefulChem notebook were abstracted into a machine-readable format as part of the Reaction Attempts (RA) database [87]. In April 2010 the first edition of the RA book was published in conjunction with the first archive of the UsefulChem laboratory notebook and associated raw data files [88]. The RA database also started to abstract reactions from other open notebooks, like the one shared by the Todd group on the Synaptic Leap [89]. The usefulness of sharing the abstracted information from open notebooks became clear in June 2010 when attempted reactions revealed an overlap between the Bradley and Todd groups, allowing for a very efficient collaboration and sharing of details about challenges beneficial to both groups and anyone else with an interest [90]. A Web-based Reaction Attempts explorer was also created to allow searching by reactant or product drop-down menus or substructure [91].

25.3.5 Second Life

Long-lasting collaborations can spring from some unusual places. While exploring the virtual world Second Life as another channel to communicate open notebook information, a contact was made between two of the authors, Bradley (JCB) and Lang (AL) [92]. The first collaborative project involved improving the 3D Second Life molecule rezzer developed by AL so that it could be used easily by simply supplying a SMILES in the chat box to generate the molecules with a realistic 3D conformation [93]. This permitted a display of Ugi products, enoyl reductase, and slides from a recent presentation, all hyperlinked to either blog posts or laboratory notebook pages for further details [94]. An effort was then made to index molecules in Second Life on a wiki [95].

In June 2007 a collaboration between an extended team resulted in a 3D animation demo of the docking of a Ugi product into the binding pocket of enoyl reductase via four hydrogen bond sites [96]. An interactive 3D animation of the formation of imine—the first step in the mechanism of the Ugi reaction—was displayed in Second Life in August 2007 [97]. These are powerful demonstrations of how sophisticated representations of research within an open notebook can be leveraged from the contribution of expertise from several individuals brought together for rapidly implemented applications.

25.3.6 Requesting Collaboration

In March 2006, JCB requested help with disabled instrumentation on the UsefulChem blog [98]. It is interesting to note that most specific open requests of this type were not directly answered. Most of the collaborations to arise from the project did so based on a shared overlap of interests, and this often caused a shift in project direction. Flexibility is of paramount importance when embarking upon these types of collaboration—all parties need to benefit. This experience suggests that open-science platforms primarily based on very specific tasks and questions may find it difficult to thrive. For example, the discussion forum ChemUnPub did answer one of our questions but did not result in a long-term collaboration [99]. A question on the OrgList mailing list was also helpful for a specific laboratory cleaning procedure [100].

25.3.7 Sharing Drafts of Papers and Proposals

In April 2007 drafts for a paper [101], a thesis [101], and a proposal [102] related to UsefulChem were started on the wiki. Being quickly indexed on major search engines, these documents represent a new way to share research work as it is being organized and planned. This is especially the case for proposals, which are rarely made public at any point. Nature Precedings, which provides a platform with an easily citable format including an author list and DOI [103], was used to publish another proposal for the project in January 2008 [104]. In June 2008, Nature Precedings no longer accepted proposals and so a proposal to the Gates Foundation was made public on Harel's S.C.I.E.n.C.E wiki, set up for this purpose [105], and Scridb [106].

As for drafts of papers, not all instances of started drafts end up as submissions to journals in a rapid and straightforward way. If the drafts are always public and indexed in search engines, there is a chance for someone to make use of even partial information from the very start. For existing or potential collaborators, this information can facilitate a deeper understanding and more efficient exchange of ideas, especially when the proposals or drafts of papers reference experiments in open notebooks. Writing a paper on a wiki essentially is a form of preprint, and journal guidelines should be consulted for subsequent submission for peer-reviewed publication [107]. Reports about other students writing up at least a part of their thesis openly started to appear [108].

25.3.8 Media Coverage: Collaborations with Journalists and Authors

By definition, a collaboration involves any situation where two or more parties work together to the benefit of all involved. In the case of ONS, journalists and authors of review articles in both the popular media and the peer-reviewed literature turned out to be important collaborators. The journalists obtained material for their pieces on the changing dynamics of scientific collaboration and the open-science movement and projects like UsefulChem received a significant amount of coverage that often led to new collaborations with other scientists as a result. News coverage also proved to be critical to lending legitimacy to the effort allowing the Wikipedia entry on ONS to be accepted in October 2008 [109, 110].

25.3.9 Other Open Notebook Science Projects

The foundation work established in Bradley's work has catalyzed a number of other ONS projects, including a platform to share research proposal ideas in natural product synthesis which became the wiki TotallyRetrosynthetic [111], a laboratory notebook for Faith [112], and a blog and wiki for the Rosania team based on the UsefulChem template to track his work on subcellular drug transport [113, 114]. The Rosania group also extended the reach of sharing their experimental results on Second Life [115].

25.3.10 Other Types of Collaboration

The UsefulChem project experimented with another form of collaboration: guest blogging. David Bradley reported on open access in chemistry [116, 117] and arsenic remediation projects [118]. On occasion a student would submit a post, but over time the UsefulChem blog evolved to a single-author modality. An unexpected collaboration arose involving the interaction of students in the humanities with the UsefulChem project [119]. The UsefulChem Writing Partners program required students from the Ritter–Guth group at Lehigh Carbon Community College to write less technically about UsefulChem themes, especially malaria [120]. This was beneficial for both the humanities students to understand how science is done and for the chemistry students to try to explain their research to a wider community.

In July 2006 an anonymous commenter brought up the issue regarding whether patents can help or hinder humanitarian applications [121]. This is an example of a type of collaboration originating from working openly, where larger issues and concerns can be addressed early on. We also found that "accidental collaboration" was occasionally very useful. For example, by monitoring search terms on Sitemeter, we discovered that water was a viable and potentially better solvent for Ugi reactions [122].

In November 2006 an offer was made to provide compounds on a "copyleft" basis, the concept being that samples of products made in the lab that could be spared would be provided freely—as long as the research done with those compounds was made open immediately [123]. Thus far no requests for this type of collaboration have been made.

In May 2008 another opportunity to collaborate with a company arose. Mettler-Toledo lent Drexel a liquid-handling robot to carry out Ugi reactions using more automation [124, 125]. An optimization study was done and the problems encountered with the use of such a parallel strategy were reported [126]. In addition, the *Journal of Visualized Experiments* (JoVE) contributed by sending a cameraman to record a video to document the execution of the reaction [127]. The JoVE article, composed of a conventional text portion and a video, was published in November 2008 [107].

A first attempt was made to allow collaboration via a specific page for anyone to request experiments to perform [128]. No requests were made from this attempt, although this strategy was successful for requests of solubility measurements. For example, a request for the solubility of pyrene in acetonitrile was made from a group in Israel to assess soil contamination, and the Drexel group provided an answer within days. A Google spreadsheet was set up to collect all such solubility requests from either people or autonomous agents [129]. Early on, related projects were discovered. These include the e-malaria project at Southampton University [130] and the Synaptic Leap [131]. A connection with Southampton University and the Synaptic Leap would eventually intersect with UsefulChem in important ways. A collaboration between JCB and Mesa Analytics and Computing via a Small Business Innovation Research (SBIR) award demonstrated that it is possible for academia and industry to work openly on a drug discovery software project [132]. X-ray crystallographer Matthias Zeller from Youngstown State University contributed to the UsefulChem project by providing a crystal structure for one of the Ugi products [133]. In June 2008 Richard Stephenson from Southampton University further contributed by setting up an eCrystals repository for Drexel where UsefulChem crystal structures could be shared openly [134]. A collaboration with Brent Friesen at Dominican University was initiated involving the synthesis of new Ugi products in his sophomore organic chemistry teaching laboratory [135, 136]. He incorporated the ONS Solubility Challenge as the first week of his laboratory [137]. The Spectral Game was made available on Second Life [138] and then on the Web [139]. It was reported a few months later in a paper in the *Journal of Cheminformatics* [140]. This was only possible because of the contributions of NMR spectra by chemists as open data when uploading to ChemSpider. This includes spectra that are routinely obtained as part of the UsefulChem project and demonstrates that, by making data open, collaborative projects not initially imagined at the time of submission can quickly arise.

The usefulness of reporting raw laboratory notebook data was demonstrated in the summer of 2009 when an article with surprising results appeared in the *Journal of the American Chemical Society* [141], specifically the observation of an oxidation by a reducing agent. Social networks such as FriendFeed spread the information very quickly and to enough chemists that two groups (UsefulChem and Totally Synthetic) attempted to reproduce some of the experiments and another found a precedent from the literature explaining the phenomenon. In this case it was critical for the two groups to produce the raw NMR data which could be unambiguously interpreted by the chemistry community. Simply reporting without proof that the experiment had not worked would not have been unequivocal.

25.4 OPEN NOTEBOOK SCIENCE SOLUBILITY CHALLENGE COLLABORATIONS

25.4.1 Crowdsourcing Solubility Measurements

In September 2008 the ONS Challenge was announced to attempt to crowdsource the measurement of nonaqueous solubility using open notebooks based on the same Wikispaces and Google spreadsheet platforms as the UsefulChem project [142]. There are currently about 200 specific solubility queries per day utilizing the results of the Challenge, originating mainly from Google and

Wikipedia [143]. The query results ultimately lead to the relevant lab notebook pages and raw data for anyone who wants to track the ultimate provenance of the data [144].

25.4.2 Sponsorship

Sigma-Aldrich sponsored the Challenge by contributing chemicals on an as-needed basis [145]. The first shipment was sent in February 2009 [146]. In October 2008 Submeta sponsored the ONS Challenge with ten $500 prizes to be awarded approximately once a month to students in the United States and the United Kingdom [147]. In November 2008 Nature committed one year subscriptions to the *Nature* journal for the first three winners of the ONS Challenge [148]. The first winner was announced at the end of November 2008 [149]. The Royal Society of Chemistry sponsored another five prizes in March 2010 [150].

25.4.3 Gaining Experience by Laboratory Rotations

In June 2009 the collaboration evolved to include face-to-face interaction when a student, David Bulger (February 2009 Submeta Award winner), spent a few weeks at Drexel with JCB, then with Cameron Neylon (CN) at Southampton University, to learn laboratory techniques before returning to Oral Roberts University with AL [151]. The stay at Drexel helped resolve some issues about the reliability of using NMR to measure solubility [152].

25.4.4 Solubility Modeling and Visualization

Several programmers collaborated with chemists to provide interfaces to the solubility data set as well as build models. This included a way to intuitively navigate the data over a Web browser [153, 154] or Second Life [155] and allow substructure searching [156]. The data set was also converted to resource description framework (RDF) to extend its use to a broader group [157]. A solubility model based on Abraham descriptors was made freely available [158]. Ultimately, both the UsefulChem Reaction Attempts and the ONS Solubility Challenge databases were combined to generate a Solvent Selection service that could be used in principle for any reaction where high solubility of reactants and low solubility of the product are desired [159].

25.4.5 ChemTaverna and MyExperiment

Recently, it was demonstrated that the solubility and reaction Web services created for the UsefulChem and ONS Solubility Challenge can be integrated into ChemTaverna and shared on MyExperiment [82]. By putting these tools into the hands of a vibrant community already using this platform for bioinformatics, it is hoped that future collaborations in the area of cheminformatics will be greatly facilitated.

25.5 OPEN NOTEBOOK SCIENCE IN UNDERGRADUATE PHYSICS LABORATORY HOSTED ON OPENWETWARE

25.5.1 Overview

For the four fall semesters of 2007–2010, ONS has been carried out by physics students enrolled in a junior laboratory course at the University of New Mexico (2007 [160], 2008 [161], 2009 [162], 2010 [163]). The experiences have been summarized in blog posts [164, 165]. The fully public electronic notebooks are hosted by OpenWetWare (OWW), a service initiated in 2005 by students in the Endy and Knight laboratories at the Massachusetts Institute of Technology (MIT) [166]. OWW currently has over 8000 members and its primary funding is through a grant from the National Science Foundation [167]. All student work is recorded and presented on the public wiki, and almost all instructor feedback is presented on the same wiki pages [168]. The only nonpublic information is the letter grade for the work. From 2007 to 2010, approximately 60 students have participated in the ONS course, with most of them majoring in physics, astronomy, math, or a combination of those. No effort was made to formally track the students, but the instructor knows that at least six students from the 2007 and 2008 semesters have since enrolled in Ph.D. programs. Two students from those semesters have begun teaching high school physics. Many of the students continued to use OWW after completing the junior laboratory course for a variety of purposes, including other lab courses [169] and undergraduate research [170].

25.5.2 Description of How Students and Instructor Carry Out ONS

There are three types of work that junior lab students record in their public pages on OWW: a primary laboratory notebook [168], informal laboratory summaries [171], and one formal research report [172]. For the purposes of this report, we will focus on the primary laboratory notebook in the context of one laboratory "cycle." The students complete six individual laboratories throughout the semester, and they are free to work alone or in groups of two. We will describe typical workflow for one of these cycles.

25.5.2.1 Preparation and Safety After choosing a laboratory to work on for the subsequent two three-hour lab sessions (three hours in week 1, three hours in week 2), students are required to do background reading so that they have a good understanding of what their goals will be, what kind of equipment they will need, and especially what the safety hazards will be. When they feel they are fully prepared, they will ask the instructor or the teaching assistant to carry out their "safety quiz." The instructor or TA asks the students to explain the work and to identify the main personal safety hazards and then potential hazards to the equipment. This exam is carried out orally. Many students will record safety issues in their primary laboratory notebook, with by far the most common safety hazard being electrical shock [173].

25.5.2.2 Primary Work (Equipment Setup, Data Acquisition, Data Analysis) Each student has a designated area on the wiki for recording electronic notes detailing his or her work while setting up an experiment, taking data, and analyzing data. Students have leeway as to how they organize their work, but the default method is a chronological system based on OWW's lab notebook with one-click setup [174]. Primary notebooks of students from prior weeks or semesters have become the de facto laboratory manual for the course. When performing background research for the laboratory work, the instructor has observed that most students refer to other students' lab notebooks in combination with a lab manual from the prior instructor [175]. This behavior is encouraged, as is citing and linking to those resources.

The instructor has observed that students' primary notebooks have converged on a structure that is a mix of chronological and topical recording of notes. A general structure that has emerged is for the primary notebook to have the following sections: title, purpose/overview, equipment and setup, data, data analysis and code, results/link to results summary, discussion of errors, and acknowledgments. This is not a rule and students are free to record their information in a variety of formats provided sufficient information is recorded. A guiding principle that the instructor dictates is that the main purpose of the electronic laboratory notebook is "reproducibility." For the purpose of the junior laboratory, "reproducibility" is defined as the ability for the same student to replicate the experiment one year later using only his or her own laboratory notebook as a guide. Students should imagine whether they would be able to obtain measurements with similar amounts of random and systematic errors after their memory has faded over the course of a year. Anecdotally, this appears to be an understandable goal for the students.

25.5.2.3 Equipment Setup Students are required to record the make and model number for all the equipment used during their experiments. They are also required to record how the equipment are set up and detailed procedures for obtaining data. From 2007 to 2010, there has been a marked increase in the percentage of students who have smart phones in the laboratory. This has correlated with an increase in the usage of digital photographs to describe the setup of the experiment. This behavior is strongly encouraged by the instructor.

25.5.2.4 Data Acquisition Students are required to record their data electronically and to display the data and detailed notes about how the data were acquired in their public notebooks. A common problem with any electronic notebook is difficulty in capturing information and data without disrupting the experimenters' ability to work. In particular, for the junior laboratory, it takes some effort to record data in the wiki, especially tabular data. Uploading images also requires many manual steps. Finally, light from computer screens is sometimes too bright for use next to an experiment with a dim signal (such as during optical spectroscopy by eye). OWW is run on a MediaWiki engine, the same engine used for Wikipedia [176]. This allows for many extensions and widgets. Flanagan, lead developer for OWW, has implemented many of these widgets, many of which attempt to make it easier to capture information into a laboratory notebook. One of these allows easy embedding of a Google Docs spreadsheet [177]. Junior laboratory students are encouraged to innovate and try out different ways of using their laboratory notebook. In 2007 students struggled with wiki or Hypertext Markup Language (HTML) tables for recording data. By 2009, the majority of students used Google Docs spreadsheets for recording data. A major advantage of this is easy recording of information that is autosaved and easy to share with the world. A drawback is that the information in the spreadsheets is currently not archived by OWW, so an electronic laboratory notebook is not a self-contained entity.

Many students record data directly into Google Docs, into the wiki, or into another electronic resource, such as Evernote [178]. However, it should be noted that even in 2010, with the ubiquity of smart phones and Web-based resources, a number of students resort to recording notes on paper and then transferring them to electronic form later. One simple reason for this is that some laboratories require dark-adjusted eyes which are not achievable when using even a smartphone. Another reason is that some students continue to find pencil and paper the fastest, easiest, and/or most comfortable means of recording information. These are the anecdotal observations of the instructor, and in his opinion it remains a problem with ONS or electronic lab notebooks more generally. The instructor does not require students to discontinue use of paper, provided they subsequently copy their notes to the primary electronic notebook.

25.5.2.5 Data Analysis Students are required to record their data analysis procedures and results in their primary laboratory notebook. It is stressed that this information is an important component for reproducibility, including the type of software used, spreadsheets, and code. For example, students will embed or link to their spreadsheets (typically Microsoft Excel or Google Doc) or they will upload their original Matlab code [179]. Important functions used for processing the data (such as LINEST) are highlighted.

25.5.2.6 Informal Lab Summary For most laboratories, in lieu of a formal laboratory report in the style that would be submitted to a typical peer-reviewed journal, students instead produce short, informal laboratory summaries that are separate from their primary laboratory notebook [171]. As described below, the students produce one formal report that includes a rough draft with extensive instructor feedback. The informal summaries are on separate wiki pages from the primary laboratory notebook. In the summaries, the students give a brief overview of the laboratory, report their final results, and discuss any discrepancies with accepted values, sources of systematic and random error, and ideas for improving future measurements. They link to their primary laboratory notebooks as the underlying resource for any readers wanting to reproduce the work or understand it better. On either the informal laboratory summary page or the primary laboratory notebook, students are required to acknowledge and link to helpful resources they relied on to carry out the work, such as other students, the laboratory manual, Wikipedia, and so on, to helpful resources, either on a summary page or primary laboratory notebook.

25.5.2.7 Instructor Feedback When students have completed their laboratory summary, they "hand in" their work by sending the instructor an e-mail with a link to the summary. The instructor puts comments, compliments, and criticism "in the margins" of the work by directly editing the wiki. For the one formal report, instructor feedback on the rough draft is extensive, attempting to explicitly point out all deficiencies that need to be worked on. Thus, almost all of the student work and instructor feedback is publicly visible to the world, with only the letter grade and perhaps other e-mail communication kept private. As the semester progresses, the volume of instructor feedback diminishes greatly, and usually feedback for the final informal summaries is not given. The instructor's impression is that feedback earlier in the semester is more valuable. Furthermore, student work seems to improve greatly after feedback is given for the first summary.

25.5.2.8 Formal Report For one of the laboratories of their choosing, the students are required to prepare a formal report in the style of a typical peer-reviewed publication. A rough draft of this report is due in approximately week 12 of the 16-week semester. This report is "handed in" to the instructor by e-mailing a link to the wiki page. All feedback by the instructor is put in "the margins" as with the other feedback. This feedback tends to be more extensive than with informal summaries [172]. A letter grade for the rough draft is e-mailed to the students privately. This letter grade is typically a D or C to indicate the amount of improvements needed, but students receive full credit as long as they hand in the rough draft on time and with sufficient effort having been made. During the final week of the semester, the students work in the lab to repeat the experiment for which they're writing their formal report. The goal is to implement some of their ideas for improving their measurements after having thought deeply about the work while writing the formal report. The final draft of the formal report is due at the end of the semester [180].

25.5.3 Outcomes (Anecdotal)

An effort has not been made to scientifically track the impact of ONS in this laboratory course. In order to do so, measurable goals would need to be defined and students would need to be tested before and after the course. Instead, the instructor has so far relied on anecdotal observations. He has observed many positives from the use of ONS. Students routinely read each others' laboratory notebooks and give credit for the assistance. Building upon prior work and citing it properly are fundamental aspects of science, and it appears that ONS strongly promotes learning this skill. It appears that the quality of measurements and sophistication of data analysis methods have improved every year. One experiment to follow through the years is the Millikan Oil Drop experiment, where students attempt to measure the value of the electron's charge, *e*. An example from 2007 is found from Le's primary notebook entries [181, 182]; another from 2008 is provided by Osinski [183, 184] and one from 2009 is provided by Callow [185–188]. Interestingly, Callow's work drew some interest from other scientists on a FriendFeed thread [189].

This would be an expected outcome of students building upon prior students' work. Another positive aspect of ONS in this course is that students can implement ONS in their future research careers. Some students have already done so [190]. It is hypothesized that a positive ONS experience in this undergraduate laboratory course will increase the likelihood of carrying out ONS in the future, especially after having become a principal investigator who can dictate laboratory notebook policies. Finally, another positive result has been the transfer of ONS techniques from the teaching laboratory to the instructor's research laboratory, for example, use of embedded Google Docs and other techniques to increase the ease of capturing information in the lab. The positives appear to have far outweighed the negatives, which are difficult to find. One negative could be that ONS has reduced the effort students need to exert to get an experiment to work. So, it is plausible that they are developing less hands-on skills than students who start "from scratch." Another possible negative is that students can balk at presenting their work publicly, and their creativity and performance could suffer significantly. While plausible, the instructor has not yet detected this outcome. Overall, feedback from students has been overwhelmingly positive—this comes from direct communication as well as from anonymous end-of-semester course evaluations.

25.5.4 Future Work and How to Replicate

What is needed for other instructors to carry out ONS in their own courses? As long as an electronic platform for ONS is available in the laboratories, extensive planning is not required. In the case study described here, the instructor simply decided that students should do ONS, provided them with accounts on OpenWetWare, and set them loose. While somewhat chaotic at first, the outcome was delightful. If there are resources for planning the course, there are some things that could be carried out better, especially in terms of assessment. As mentioned above, pre- and posttesting of students are essential to know with certainty that ONS is impacting desired outcomes. Similarly, mechanisms should be developed to keep track of alumni of these courses in order to assess whether ONS in the undergraduate teaching lab affected their future research behaviors or opinions toward ONS or other open-science ideas.

25.6 LABORATORY BLOGGING: FRAMEWORK FOR SMALL-SCALE COLLABORATION

The laboratory notebook system developed at the University of Southampton in Frey's group [191] has formed the basis for the primary laboratory record for one of us (CN) for nearly five years [192]. Over time this has been used in a range of different ways and with different organization schemes [193], but here we will focus on its role in supporting collaborations, particularly geographically distributed ones. The system is similar to most blog engines in being organized into posts, usually presented in reverse chronological order, an ability to comment, including people other than the post author, and the generation of RSS feeds of posts. These main features, which are relevant to the discussion of collaboration per se, are common to almost all blog engines. Most of the other technical capacities of the system are not relevant to this discussion, but one difference is important. Posts within the Lab Blog system cannot be deleted by the user, consistent with best practice in retaining a permanent record of the research process. Where changes are made to a post, a full version history is maintained, effectively enabling a final version of the record to be presented by default but providing the complete detail of changes or mistakes to be available if required.

25.6.1 One-to-One Collaborations

The most successful collaborative projects that have been supported by the blog system have been largely one-to-one interactions. In the first, the supervision of a student based at Southampton by CN was effectively supported by the system after he had moved to a new site [194, 195]. The system enabled a close interaction on a daily basis with the details of the experimental work. The details of experimental protocols and results could be discussed in close to real time despite the geographical distance. From a technical perspective this was achieved through the monitoring of the RSS feed for the student's blog in Google Reader. This functioned mainly as a notification system as Google Reader did not display many elements of the rendered post correctly, due to the loss of formatting information in the XML of the RSS feed. Commenting and communication would occur back on the blog system rather than through any third-party service. This pattern has been more or less repeated in subsequent collaborations, both those taking a completely open approach and exposing the record freely on the web and cases where the interaction has been through a closed, password-protected blog.

25.6.2 Failures

There have also been a number of attempts to utilize the system to support collaborations that have failed. On the surface these have many characteristics of the successful examples: geographical dispersion, an acceptance of the value of Web-based record keeping, and a desire to maintain a high-quality record. In some cases these efforts have even involved groups on the same site. However, a common feature of all the failures is that the use of the blog was for only a portion of the work being undertaken. In some cases this was due to multiple projects being run, only one of which was being recorded in this manner. In other cases problems arose due to the confidentiality of portions of a project that could be entrusted to the level of security available within the system.

The clear end result is that where the record becomes split the nontraditional record, usually the one that for either technical or social reasons requires more effort to keep up to date, suffers and falls behind. Once the record keeping falls behind or is temporarily recorded in some other form, it rarely catches up again. This is not by any means a specific characteristic of the blog-based notebook and is likely to be true of the use of any new system. However, the lack of geographical colocation and consequent lack of "nagging" available to encourage use as well as the limitations of the user interface for the blog system exacerbated these issues. The lack of peer pressure that resulted from primarily one-to-one as opposed to wider group collaborations was also a contributing factor.

25.6.3 Scaling the Collaboration

It appears that a blog-based system, where posts and comments are clearly attributed to one author, provides a somewhat more personal space that is more suited to one-to-one collaborations. The splitting of each person's record into individual blogs also seems to encourage this, making it less likely that community members will directly contribute to or edit each other's material. In comparison, the Wiki-based systems used in the UsefulChem and ONS Challenge projects provide a single unified space, where direct editing of content is supported and encouraged, but commenting less so. In the wiki systems an approach of commenting in line has been adopted, due largely to the need for comments to be closely associated with the relevant text. The more modular nature of the way the blog system has been used means that separate comments do not drift away from the relevant text as much as is the case with the talk page on the wiki system. These different approaches to commenting, in separate comments in the blog and directly in the text of the wiki, may mean that the wiki provides less of a sense of personal ownership of the text. By comparison the blog system supports a back-and-forth conversation in the comments that may be felt to be more personal. There is a balance to be struck here between the need to give people space to feel comfortable to write and the need to support effective communication. The system as it currently exists requires some form of account to comment or contribute. This has limited direct contact with external users.

Supporting larger-scale collaboration in the context of the blog system will require careful attention to the integration of notification schemes, of both posts and comments. The personal nature of posts may assist in making people more comfortable with describing their work in a public space by avoiding the additional fear of having their text edited. However, at the same time it does not encourage the more direct interaction that appears to be supported by wiki-based systems. A key aspect for both systems is effective notification of the community when new content has been created. There is a significant technical infrastructure available that supports this for blogs, including RSS feed manipulation tools like Yahoo Pipes and collaborative RSS reading environments such as Google Reader where content can be shared, tagged, and commented on. The configuration of this for specific projects will require care. Both blogs and wikis suffer from a problem in notification where important changes are made to a post or page. In the case of most blogs modifications are not posted to the feed, whereas for wikis in general the feed contains all committed changes. Neither of these extremes is helpful, and in addition the useful display of changes to a preexisting document remains a challenge. The effective notification of significant or important changes is an important technical challenge for the effective use of collaborative online tools for recording research.

25.7 CONCLUSION

Collaboration on many levels can be facilitated by ONS and other open-science projects. However, getting things done generally requires a specific person to champion a specific subset of tasks [196]. Fortunately, there have been enough collaborators during the past few years in the open-science community with enough shared goals between projects to enable useful tools and resources to emerge.

Concerning collaborative platforms, for UsefulChem, an evolution took place over the course of the project. Initially blogging and commenting on blogs was a significant means of public communication. A blog was tried initially to host the actual laboratory notebook, but limitations quickly led to migration to a free hosted wiki on Wikispaces and raw numerical data stored in public Google spreadsheets. A mailing list was in use for a brief time to facilitate public communication with collaborators. However, in the latter half of the ONS projects at Drexel and much of the open-science community, FriendFeed became a very important mode of public communication.

In the case of using OpenWetWare for teaching laboratory applications, the flexibility of ONS allows implementation without excessive planning. The ability for students to view each other's work and the ease with which the instructor can provide specific feedback are strong assets to this approach.

The Laboratory Blog system has demonstrated that a blog-style framework is a useful way of generating an online research record. It seems particularly effective at supporting the small scale, particularly one-to-one collaborations and monitoring of student work. The use of one blog per person and a lack of integrated notification frameworks make it more difficult to scale these collaborations using this system. Successful implementation of these systems requires an all-or-nothing approach. Mixed record keeping always favors the incumbent system.

As long as there is an intent to share and be open, the platforms of communication can continue to change—as they have in the past—without the risk that conversations will stop. The trend has been toward tools that make it easier to collaborate and discuss—and the use and combination of multiple platforms to leverage what each does best. This redundancy is beneficial in a world where it is not possible to predict which technologies and services will be dominant or even available a few years down the road.

REFERENCES

1. http://usefulchem.blogspot.com/2006/09/defining-what-usefulchem-does.html.
2. http://en.wikipedia.org/wiki/Open_Notebook_Science.
3. http://usefulchem.blogspot.com/2009/02/open-notebook-science-reproducibility.html.
4. http://usefulchem.blogspot.com/2009/02/open-notebook-science-claims-and-logos.html.
5. http://usefulchem.blogspot.com.
6. http://usefulchem.wikispaces.com.
7. http://www.wikispaces.com.
8. http://usefulchem-experiments1.blogspot.com.
9. http://usefulchem.blogspot.com/2006/02/experiments-blog.html.
10. http://usefulchem-molecules.blogspot.com.
11. http://usefulchem.blogspot.com/2005/11/new-blog-for-molecules.html.
12. http://usefulchem-experiments1.blogspot.com/2006/02/exp-003.html.
13. http://usefulchem.blogspot.com/2006/02/praziquantel-synthesis-request.html.
14. http://usefulchem.blogspot.com/2006/05/auto-oxidation-of-catechol.html.
15. http://usefulchem.blogspot.com/2006/06/new-members.html.
16. http://usefulchem.blogspot.com/2006/08/following-usefulchem-experiments.html.
17. http://usefulchem.blogspot.com/2006/11/google-co-op-for-usefulchem.html.
18. http://usefulchem.blogspot.com/2008/05/google-custom-search-for-open-notebook.html.
19. http://usefulchem.blogspot.com/2007/03/usefulchem-on-emolecules-and-structure.html.
20. http://usefulchem.blogspot.com/2007/03/searching-for-open-access-chemistry-on.html.
21. http://usefulchem.blogspot.com/2007/08/usefulchem-on-cdd.html.
22. http://usefulchem.blogspot.com/2007/10/ons-friendly-labs.html.
23. http://usefulchem.blogspot.com/2007/03/chemspider.html.

24. http://usefulchem.blogspot.com/2007/06/inchimatic-chemspider-and-usefulchem.html.
25. http://usefulchem.blogspot.com/2008/04/usefulchem-on-chemspider.html.
26. http://usefulchem.blogspot.com/2007/07/usefulchem-mailing-list_04.html.
27. http://usefulchem.blogspot.com/2008/04/scholar2scholar-tomorrow.html.
28. http://friendfeed.com/jcbradley.
29. http://usefulchem.blogspot.com/2009/02/c-news-article-on-ons-and-friendfeed.html.
30. http://usefulchem.blogspot.com/2005/07/sample-search-phrases.html.
31. http://usefulchem.blogspot.com/2005/10/antimalarial-compounds.html.
32. http://usefulchem.blogspot.com/2005/11/find-drug.html.
33. http://usefulchem.blogspot.com/2005/11/anti-malaria-compounds.html.
34. http://usefulchem.blogspot.com/2006/04/diketopiperazine-in-pocket.html.
35. http://usefulchem.blogspot.com/2006/11/wanted-docking-collaborator.html.
36. http://usefulchem.blogspot.com/2007/04/open-source-science-expands-with-tan.html.
37. http://usefulchem.blogspot.com/2007/04/nci-usefulchem-link.html.
38. http://usefulchem.blogspot.com/2007/04/first-dtp-nci-compound-submission.html.
39. http://usefulchem.blogspot.com/2007/08/first-nci-results.html.
40. http://usefulchem.blogspot.com/2007/05/combiugi-and-closing-open-science-loop.html.
41. http://usefulchem.blogspot.com/2007/06/combiugi-says-order-2-naphthyl.html.
42. http://usefulchem.blogspot.com/2008/01/crowdsourcing-drug-development.html.
43. http://usefulchem.blogspot.com/2007/07/combiugi-time-for-synthesis.html.
44. http://usefulchem.blogspot.com/2007/08/falcipain-collaboration.html.
45. http://usefulchem.blogspot.com/2007/12/first-falcipain-2-targets-shipped.html.
46. http://usefulchem.blogspot.com/2008/01/we-have-anti-malarial-activity.html.
47. http://usefulchem.blogspot.com/2008/08/fall-2008-acs-meeting-ends.html.
48. http://usefulchem.blogspot.com/2005/11/proposed-synthesis-of-anti-malarials.html.
49. http://usefulchem.blogspot.com/2005/12/diketopiperazine-synthesis-found.html.
50. http://usefulchem.blogspot.com/2005/12/malaria-solid-support-dkp-synthesis.html.
51. http://usefulchem.blogspot.com/2005/12/ugi-dkp-synthesis-for-malaria.html.
52. http://usefulchem.blogspot.com/2005/12/aldehyde-problem.html.
53. http://usefulchem.blogspot.com/2006/01/chemrefer-answers-ugi-phenol-question.html.
54. http://usefulchem.blogspot.com/2006/01/chris-hulme-on-ugi-synthesis.html.
55. http://usefulchem.blogspot.com/2006/06/dopal-nmr-and-phosphoric-acid.html.
56. http://usefulchem.blogspot.com/2006/10/isonitrile-problem.html.
57. http://usefulchem.wikispaces.com/dopal.
58. http://usefulchem.blogspot.com/2007/01/anatomy-of-ugi-reaction.html.

59. http://usefulchem.blogspot.com/2006/12/imine-understanding.html.
60. http://usefulchem.blogspot.com/2005/12/smiles-lookup-table-for-malaria37.html.
61. http://en.wikipedia.org/wiki/Simplified_molecular_input_line_entry_specification.
62. http://usefulchem.blogspot.com/2005/12/openbabel-20-release.html.
63. http://usefulchem.blogspot.com/2006/08/synaptic-leap-presentation.html.
64. http://usefulchem.blogspot.com/2006/09/arguslab.html.
65. http://usefulchem.blogspot.com/2006/08/cml-explained-blog.html.
66. http://usefulchem.blogspot.com/2006/08/depth-first-blog.html.
67. http://usefulchem.blogspot.com/2006/09/murray-rust-blog.html.
68. http://usefulchem.blogspot.com/2006/09/final-thoughts-on-acs.html.
69. http://www.chemspider.com/blog.
70. http://usefulchem.blogspot.com/2006/09/spectra-in-jcamp.html.
71. http://usefulchem.blogspot.com/2006/10/nmr-spectra-on-browser-with-jcamp.html.
72. http://usefulchem.blogspot.com/2006/11/jspecview-demo.html.
73. http://usefulchem.blogspot.com/2006/12/nmr-overlay-with-jspecview.html.
74. http://usefulchem.blogspot.com/2007/01/automated-reaction-kinetics-using.html.
75. http://usefulchem.blogspot.com/2007/03/automated-reaction-kinetics-using-excel.html.
76. http://usefulchem.blogspot.com/2006/03/cmlrss-attempt.html.
77. http://usefulchem.blogspot.com/2008/08/usefulchem-and-cml-in-cambridge.html.
78. http://usefulchem.blogspot.com/2007/02/molecules-on-chemical-blogspace.html.
79. http://usefulchem.blogspot.com/2007/06/combiugi-web-service.html.
80. http://usefulchem.blogspot.com/2007/07/spectral-deconvolution-with-usefulchem.html.
81. http://usefulchem.blogspot.com/2007/08/combiugi-on-myexperiment.html.
82. http://usefulchem.blogspot.com/2010/08/chemtaverna-workflows-of-ons-web.html.
83. http://usefulchem.blogspot.com/2007/11/combiugi-update-master-table.html.
84. http://usefulchem.blogspot.com/2008/01/modularizing-results-and-analysis-in.html.
85. http://usefulchem.blogspot.com/2008/01/tracking-results-with-workflow-tables.html.
86. http://usefulchem.blogspot.com/2008/04/ugi-precipitation-predictions.html.
87. http://usefulchem.blogspot.com/2010/03/reaction-attempts-on-chemspider.html.
88. http://usefulchem.blogspot.com/2010/04/reaction-attempts-book-edition-1-and.html.
89. http://usefulchem.blogspot.com/2010/05/synaptic-leap-experiments-on-reaction.html.
90. http://usefulchem.blogspot.com/2010/06/use-of-ons-to-protect-open-research.html.
91. http://usefulchem.blogspot.com/2010/06/reaction-attempts-explorer.html.

92. http://usefulchem.blogspot.com/2007/04/chemistry-on-nature-island-in-second.html.
93. http://usefulchem.blogspot.com/2007/09/inchi-rezzer-in-second-life.html.
94. http://usefulchem.blogspot.com/2007/05/usefulchem-on-drexel-island.html.
95. http://usefulchem.blogspot.com/2007/07/indexing-molecules-in-second-life.html.
96. http://usefulchem.blogspot.com/2007/06/molecule-docking-in-second-life.htm.
97. http://usefulchem.blogspot.com/2007/08/chemical-reactions-in-second-life.html.
98. http://usefulchem.blogspot.com/2006/03/status-of-experimental-work-month-2.html.
99. http://usefulchem.blogspot.com/2006/05/chemunpub-forum.html.
100. http://usefulchem.blogspot.com/2006/11/praise-for-orglist.html.
101. http://usefulchem.blogspot.com/2007/04/wiki-paper-experiment-started.html.
102. http://usefulchem.blogspot.com/2007/04/funding-usefulchem.html.
103. http://usefulchem.blogspot.com/2007/06/nature-precedings.html.
104. http://usefulchem.blogspot.com/2008/01/chemistry-crowdsourcing-pre-proposal.html.
105. http://usefulchem.blogspot.com/2008/06/gates-submission-on-ons-crowdsourcing.html.
106. http://usefulchem.blogspot.com/2008/08/scribd-as-repository-for-proposals-and.html.
107. http://usefulchem.blogspot.com/2008/11/from-ons-to-peer-review-our-jove.html.
108. http://usefulchem.blogspot.com/2007/06/thesis-on-wiki-interest.html.
109. http://usefulchem.blogspot.com/2008/10/open-notebook-science-on-wikipedia.html.
110. http://en.wikipedia.org/wiki/Open_Notebook_Science.
111. http://usefulchem.blogspot.com/2007/05/totally-retrosynthetic.html.
112. http://usefulchem.blogspot.com/2007/07/jeremiahs-open-notebook.html.
113. http://usefulchem.blogspot.com/2007/12/subcellular-drug-transport-usefulchem.html.
114. http://usefulchem.blogspot.com/2007/12/rosania-lab-open-notebook-science-wiki.html.
115. http://usefulchem.blogspot.com/2008/01/rosania-blog.html.
116. http://usefulchem.blogspot.com/2006/03/interview-with-steve-bryant.html.
117. http://usefulchem.blogspot.com/2006/04/interview-with-wikichems-martin-walker.html.
118. http://usefulchem.blogspot.com/2005/11/hellish-water.html.
119. http://usefulchem.blogspot.com/2006/05/usefulchem-writing-partners.html.
120. http://usefulchemwritingpartners.blogspot.com/.
121. http://usefulchem.blogspot.com/2006/07/patents-in-open-source-science.html.
122. http://usefulchem.blogspot.com/2006/09/ugi-reaction-in-water.html.
123. http://usefulchem.blogspot.com/2006/11/copylefting-compounds.html.
124. http://usefulchem.blogspot.com/2008/05/usefulchem-automation-trial-with.html.
125. http://usefulchem.blogspot.com/2008/05/minimapper-in-lab.html.

126. http://usefulchem.blogspot.com/2008/06/ugi-reaction-on-minimapper-trial-1.html.
127. http://usefulchem.blogspot.com/2008/07/jove-shoot-at-drexel.html.
128. http://usefulchem.blogspot.com/2006/12/remote-controlled-labs.html.
129. http://usefulchem.blogspot.com/2009/06/crowdsourcing-solubility-requests-from.html.
130. http://usefulchem.blogspot.com/2005/12/e-malaria-project.html.
131. http://usefulchem.blogspot.com/2005/12/synaptic-leap.html.
132. http://usefulchem.blogspot.com/2007/10/drug-design-on-open-web.html.
133. http://usefulchem.blogspot.com/2007/12/x-ray-crystallography-collaborator.html.
134. http://usefulchem.blogspot.com/2008/06/drexel-ecrysals-repository.html.
135. http://usefulchem.blogspot.com/2008/03/expanding-usefulchem-collaboration-to.html.
136. http://usefulchem.blogspot.com/2008/04/were-gonna-ugi-all-night.html.
137. http://usefulchem.blogspot.com/2009/01/ons-solubility-challenge-in-teaching.html.
138. http://usefulchem.blogspot.com/2009/02/nmr-game-on-second-life.html.
139. http://usefulchem.blogspot.com/2009/02/web-based-spectra-game.html.
140. http://usefulchem.blogspot.com/2009/07/spectral-game-paper-live-on-journal-of.html.
141. http://usefulchem.blogspot.com/2009/08/our-attempt-to-reproduce-oxidation-by.html.
142. http://usefulchem.blogspot.com/2008/09/open-notebook-science-challenge.html.
143. http://usefulchem.blogspot.com/2009/02/maintaining-solubility-data-provenance.html.
144. http://usefulchem.blogspot.com/2008/10/there-are-no-facts-my-position-at-nsf.html.
145. http://usefulchem.blogspot.com/2008/09/sigma-aldrich-first-official-sponsor-of.html.
146. http://usefulchem.blogspot.com/2009/02/sigma-aldrich-ships-ons-solubility.html.
147. http://usefulchem.blogspot.com/2008/11/submeta-open-notebook-science-awards.html.
148. http://usefulchem.blogspot.com/2008/11/nature-sponsors-open-notebook-science.html.
149. http://usefulchem.blogspot.com/2008/11/first-submeta-open-notebook-science.html.
150. http://usefulchem.blogspot.com/2010/03/rsc-sponsors-open-notebook-science.html.
151. http://usefulchem.blogspot.com/2009/06/david-bulgers-drexel-visit.html.
152. http://usefulchem.blogspot.com/2009/06/recent-insights-about-solubility.html.
153. http://usefulchem.blogspot.com/2008/11/ons-solubility-web-query.html.
154. http://usefulchem.blogspot.com/2008/11/google-visualization-api-on-ons.html.
155. http://usefulchem.blogspot.com/2009/01/interactive-visualization-of-ons.html.
156. http://usefulchem.blogspot.com/2009/02/substructure-searching-on-ons.html.

157. http://usefulchem.blogspot.com/2008/10/rdf-triples-for-open-notebook-science.html.

158. http://usefulchem.blogspot.com/2010/07/general-transparent-solubility.html.

159. http://usefulchem.blogspot.com/2010/08/reaction-attempts-solvent-selector.html.

160. http://openwetware.org/index.php?title=Physics307L_F07:People&oldid=234756.

161. http://openwetware.org/index.php?title=Physics307L_F08:People&oldid=377898.

162. http://openwetware.org/wiki/Physics307L_F09:People.

163. http://openwetware.org/index.php?title=Physics307L:People&oldid=454625.

164. http://stevekochteaching.blogspot.com/2008/12/get-em-while-theyre-young-open-science.html.

165. http://stevekochteaching.blogspot.com/2009/09/open-notebook-science-unm-physics-round.html.

166. http://openwetware.org/wiki/OpenWetWare:FAQ.

167. http://nsf.gov/awardsearch/showAward.do?AwardNumber=0640709.

168. http://openwetware.org/wiki/User:Paul_V_Klimov/Notebook/JuniorLab307L/2008/10/13#Electron_Diffraction.

169. http://openwetware.org/wiki/Polarization_Study_Data_Record.

170. http://openwetware.org/wiki/User:Linh_N_Le/Notebook.

171. http://openwetware.org/wiki/Physics307L_F08:People/Klimov/eDiffraction.

172. http://openwetware.org/wiki/Physics307L_F08:People/Klimov/Electron_Diffraction.

173. http://openwetware.org/wiki/User:Alexandra_S._Andrego/Notebook/Physics_307L/2009/10/12.

174. OWW Lab Notebook system developed by William Flanagan, Ricardo Vidal, Jason Kelly, and others. Available: http://openwetware.org/wiki/Lab_Notebook.

175. http://openwetware.org/wiki/Image:Gold's_Junior_Lab_Manual.pdf.

176. http://www.mediawiki.org/wiki/MediaWiki.

177. http://openwetware.org/wiki/User:Thomas_S._Mahony/Notebook/Physics_307L/2009/10/12.

178. http://www.evernote.com.

179. http://openwetware.org/wiki/User:John_Callow/Notebook/Junior_Lab/Final_Formal_Report#Matlab_Code.

180. http:/openwetware.org/wiki/Physics307L_F08:People/Klimov/Electron_Diffraction_Final.

181. http://openwetware.org/wiki/Physics307L_F07:People/Le/Notebook/070910.

182. http://openwetware.org/wiki/Physics307L_F07:People/Le/Notebook/070917.

183. http://openwetware.org/wiki/Physics307L_F08:People/Osinski/Millikan.

184. http:/openwetwareorg/wiki/User:Boleszek/Notebook/Physics_307l,_Junior_Lab,_Boleszek/2008/11/03.

185. http://openwetware.org/wiki/Physics307L:People/Callow/millikanoildrop.

186. http://openwetware.org/wiki/User:John_Callow/Notebook/Junior_Lab_307/2009/10/14.

187. http://openwetware.org/wiki/User:John_Callow/Notebook/Junior_Lab/Formal_Report.

188. http://openwetware.org/wiki/User:John_Callow/Notebook/Junior_Lab/Final_Formal_Report.
189. http://friendfeed.com/stevekoch/4d252b4e/john-callow-fun-mathematical-analysis-of?embed=1.
190. http://openwetware.org/wiki/User:Linh_N_Le/Notebook.
191. http://blogs.chem.soton.ac.uk.
192. http://blogs.chem.soton.ac.uk/beta_glu/month/1162339200.
193. http://cameronneylon.net/blog/evolving-usage-patterns-on-the-southampton-lab-blog-book.
194. http://blogs.chem.soton.ac.uk/beta_glu.
195. http://blogs.chem.soton.ac.uk/neutral_drift.
196. http://usefulchem.blogspot.com/2009/05/leaders-and-pushers-in-open-science.html.

26

COLLABORATION AND THE SEMANTIC WEB

Christine Chichester and Barend Mons

26.1 INTRODUCTION

There is no shortage of experimental data, derived information, and knowledge in the life sciences. However, it is siloed in databases, the scientific literature, and the minds of scientists. Any locally performed reasoning process, either computationally by computers or conceptually by humans, on one silo of information will miss potentially relevant data, making serendipitous findings less likely. Numerous projects have tried to address this problem by data integration but have only marginally succeeded. With the advent of Semantic Web technologies and standards, we are now able to realistically develop a fundamentally different approach to enhanced knowledge management and

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.

collaborative intelligence. Semantic Web technologies can now generate an interoperable, interdisciplinary catalog of unique scientific assertions assembled from previously documented data. This chapter will cover the requirements and the future of collaborative intelligence systems as we move toward realizing the potential of the Semantic Web. Specifically we will address the generation of computationally inferred assumptions and the role of human computation in reviewing and annotating assertions as a method to mine knowledge directly from the minds of the scientific community at large.

26.2 SPRINGBOARD FOR COLLABORATIVE SEMANTIC WEB TECHNOLOGIES

Several pioneering projects based on community annotation, such as WikiProteins [1], WikiPathways [2], Wikigenes [3], and ChemSpider [4], have provided a clear view of the impact of community annotation as well as some of the pitfalls. It is important to emphasize that most of the pioneering applications mentioned here still exist and gradually continue to pick up speed. In fact, Huss et al. concluded: “With the explosion in biological wikis, it is clear that the community intelligence model resonates with the biology and scientific community” [5, p. D637]. However, the authors suggest that the problem of the unavoidable expansion of community annotation, namely, too many wikis, is on the horizon. An overabundance of editable sites will dilute the impact of the collaborative intelligence movement. Scientists want to have one or a very limited number of places run by a “trusted party,” where they can go for community annotation. The trusted party obligation was clearly verbalized at the Scientific Interest Group on Bio-ontologies at Intelligent Systems for Molecular Biology (ISMB) 2009 (Stockholm); some early wiki installations were run by companies, which appears to invoke suspicion in certain scientists. However, probably, the most often heard objection to contributing to collaborative systems is the lack of recognition and reward for the submitted work. Why would competitive and busy scientists share their views in public places run by private companies or entirely nonscientific communities like the WikiMedia Foundation and not receive any formal credit for it?

Technically, the classic problems deal with heterogeneous data and making entire collections interoperable while ensuring that any annotation, which includes the recognition-and-reward system of scientific publishing, needs to fit into a seamless process beginning to end. The challenge is to create a system that manages heterogeneity and to provide interoperability using open and extensible standards and methodologies. The ultimate goal is to create a sustainable future for a large-scale, community-editable store of disambiguated scientific assertions, exemplifying a new paradigm in life sciences data accumulation. This will only be done by drawing from the mental resources of an extended scientific community in an innovative and complex, yet “daily practice,” manner that promises a profound impact on our ability to use existing data to generate new knowledge with the maximum conceivable serendipity.

However, technical aspects are only one part of the challenge. Additionally, it is necessary to build a community of people with a common purpose who are extraordinarily enthusiastic about the collaborative offer. This community will invest the necessary energy in pursuit of creating something unique. They submit original ideas and content as well as remix each others' material to produce solutions that will earn them respect, status, acceptance, reputation, as well as rewards in the form of microcitations. In other words, they are competing to get the credit for the best result. At another level, there is the larger crowd that is participating on a much lower level of activity and involvement. They tag, recommend, rate, vote, send e-mail links to colleagues, and sometimes make an occasional annotation. This interaction is therefore quite shallow compared to the passionate annotators. There is however a great wisdom to be gathered from all of this grassroots activity; their carefully elicited input helps organize the solutions and understand their worth. Thus, they introduce value to the community knowledge as they confirm the relevance and importance of the best material produced.

An awareness of the challenges does not mean that the previously outlined pitfalls are automatically mitigated. Using Semantic Web technologies brings us one step forward from the early developments of the "million minds" approach [1]. A few of the major bottlenecks can now conceivably be solved: the interoperability issue (related to "too many wikis") with the adaptation of Semantic Web and its standards and the "busy scientist syndrome" with microcitation credit and ease of use of mobile technologies.

26.3 SEMANTIC WEB APPROACH

The Semantic Web can benefit all producers and consumers of information by providing improved mechanisms for organizing information on a global scale. Based on the rapidly expanding role of Web-based resources, the Semantic Web technologies offer critical support to the life sciences by (1) unique identifiers that are supported by the Semantic Web uniform resource identifiers (URIs); (2) coordination and management of terminologies and ontologies; (3) model database conversions of life sciences data; (4) account and channel access for scientists to store and share annotations based on the Semantic Web; (5) tools and viewers conversant in the resource description framework (RDF); and (6) inference and reasoning to produce theories, hypotheses, and models.

The key to harmonizing the diverse life scientific information via the Semantic Web is based on a data structure called concept triples or assertions, which when coupled to their provenance data are called nanopublications [6, 7]. Each concept triple represents the scientific assertion of a fact, an observation, or an inference. For example, <lovastatin> <inhibits> <3-hydroxy-3-methylglutaryl-coenzyme A reductase (*Homo sapiens*)> is a triple that can be encoded using RDF. The RDF represents data as a set of directed graphs; URIs are assembled into triples composed of a subject URI, a predicate URI, and an object URI. The predicates of RDF triples are similar to hyperlinks;

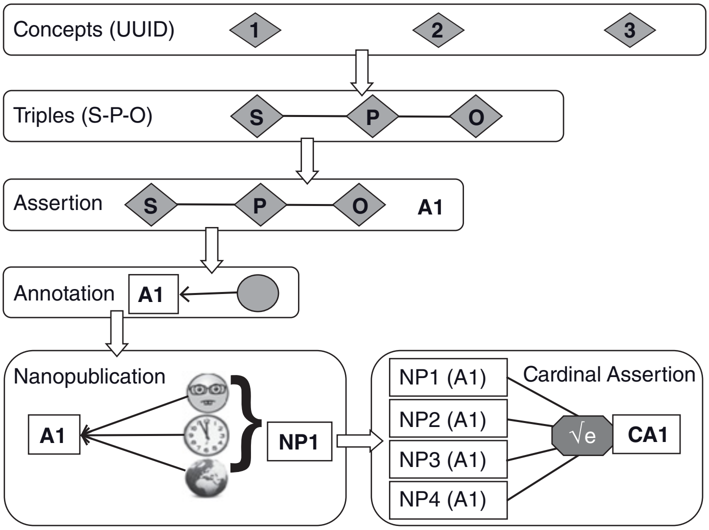


Figure 26.1 Evolution from concepts to triples to nanopublications to cardinal assertions.

however, the advantage of RDF triples over Hypertext Markup Language (HTML) hyperlinks is that the links are explicitly labeled. The semantics of the relationship between the two entities is computationally accessible through URI resolution. Finally, key to the efficient use of triples is based entirely on syntactic matching of the URI strings; each concept in the triple is uniquely identified, which leads to each triple itself being uniquely identified as a cardinal assertion (Fig. 26.1). In this way it is possible to remove redundancy and ambiguity, and generate the cardinal assertions, from the huge amount of harvested assertions principally via linking to standardized concepts.

Although the task of capturing and disambiguating billions of potential triples appears overwhelming at first, the efficient acquisition of triples can be achieved through a judicious combination of automated technologies. Ideally, the Semantic Web provides the platform technologies to generate assertions, extract assertions from existing literature, and finally share them in a way that will allow computational agents to discover, aggregate, and interpret these assertions. For example, a named graph [8] is a simple extension to RDF that provides the capability for assigning a URI to a given RDF graph. Named graphs were specifically designed to support the tracking of provenance data during aggregation and the description of the context for a particular graph. Using the named graph technology, all annotations belonging to a nanopublication can be collected and should facilitate the collection of fine-grained scientific information across the Web. Next, a key role for aggregator technologies will be to find, filter, and combine all the evidence for an assertion from a variety of nanopublications to determine the certainty of an assertion.

Even though the Semantic Web's key enhancement to the current Web is its improved computational accessibility, the success of semantically enabled browsers in connecting human users with information integrated through RDF will likely prove pivotal in exactly the same way that browsers are vital to the adoption and rapid growth of the original Web. To effectively make use of the triples, efficient searching and creative reasoning with massive quantities of concept triples must be supported. This requires a reasoning algorithm, a reasoning engine, a way to express queries, and a way to browse results. The mechanisms for navigating through RDF graphs are starting to be realized in a variety of emerging Semantic Web browsers [9–17].

Clearly, using the identical identifiers across the board for concepts and assertions makes data collection and aggregation simpler, but the key limitation is that there is no requirement to do so. Indeed, any Semantic Web resource can be used, although nanopublication publishers should follow initiatives like Linked Open Data [18] and the "shared names initiative" of the Semantic Web Health Care and Life Sciences (HCLS) Interest Group [19].

26.4 CONCEPT WEB ALLIANCE AND CONCEPTWIKI

A major impetus to achieving the potential of the shared identifiers for the semantic technologies is the collaborative mindset represented by the Concept Web Alliance (CWA). The CWA is a recently chartered nonprofit organization whose mission is "to enable an open collaborative environment to jointly address the challenges associated with high volume scholarly and professional data production, storage, interoperability and analyses for knowledge discovery" (from the CWA declaration, available at http://www.nbic.nl/about-nbic/affiliated-organisations/cwa/declaration). Together, the CWA partners will enhance existing information exchange by developing open-platform protocols, data formats, workflow tools, and semantic integration to overcome existing legacies and information bottlenecks. In this capacity, the CWA has demonstrated a leadership role in obtaining service- and application-level agreements for public data and established consensus expertise in the area of semantic frameworks. Both aspects represent key factors necessary for establishing standards that will sustain processes required for mediating large-scale data interoperability. Drawing on its diverse membership in academia and private enterprise, the CWA is uniquely positioned as a trusted agent to mediate this unprecedented confederation of existing public and private information. As an initial venture, the CWA will provide ConceptWiki as a repository of uniquely identifiable and unambiguous uniform resource locators (URLs) for concepts.

ConceptWiki (http://www.conceptwiki.org) is an open-access system that accepts essentially unlimited numbers of synonyms, in multiple languages, and then maps all the terms correctly back to one unique concept identifier, alleviating problems of vocabulary and identifier differences. ConceptWiki is built on the earlier established Wiki predecessors, Omegawiki [20] and WikiProfessional [1]. It is a Web-based system containing the biomedical terminology of the Unified Medical Language System (UMLS, levels 0 and 1) mapped where appropriate to the protein terminology from SwissProt. In the near future, the ConceptWiki repository will be expanded to incorporate the chemical terminology from ChemSpider [4] for biologically relevant chemical molecules. Each concept in ConceptWiki is annotated with one or more semantic types and basic information like a definition. Users can view and edit information through a uniform interface. The information in the system is stored and edited in a highly structured way, as triples (e.g., <concept A> <has synonym> <term B>). The WikiData backend has been designed to support the storage of concepts in a very generic form, thereby trying to avoid as much as possible the exclusion of potential valuable information sources. This compatibility with our other information storage systems enables higher level applications to easily query, summarize, and mine the knowledge. In line with recommendations from the CWA, identifiers in WikiData are completely opaque; they have no inherent structure and no information can be derived from them. An opaque identifier is a robust identifier as there will never be a need to change the identifier when underlying information changes. WikiData uses universally unique identifiers (UUIDs) [21]. Additionally, WikiData keeps a complete history containing every change made to the concepts. Changes can be analyzed via the history page of each concept or the global chronological transaction log. Transactions can be rolled back partially or completely. Using the ConceptWiki interface scientists with no background in programming can directly map, merge, or integrate individual concepts. ConceptWiki supports the distinction between "authority" and "community" data [1] and permits general editing only on the community branch of the data. This distinction is the highly innovative aspect that convinces authorities that it is prudent to donate and integrate their data into the system. Additionally, the comparison between the authority branches and the community branches allows personal value judgments of displayed data.

ConceptWiki has an API for a thesauri extractor, an application for freely downloading terminology systems for specific purposes or domains. The downloaded thesauri can be used to identify concept-denoting tokens in text and databases so that individual indexers can be linked to the concepts to create a linked open-data system [22]. ConceptWiki is in the public domain, indicated by the Creative Commons so-called CC Zero Waiver + SC Norms, indicating that copyrights are waived but that adherence to scientific community norms regarding attribution and citation is expected. In this way, the community can be regarded to "own" ConceptWiki and the UUID concept reference identifiers in it.

26.5 AUTHORSHIP OF SCIENTIFIC ASSERTIONS

Citability and credit to authors are of prime importance to the way the scientific publishing system works. In a scientific context, publications are only publications if they are citable and appropriate credit is given to the authors. There is no intrinsic reason why such publications need necessarily be full-length papers. Published contributions could be as short as single statements that interpret data and yet be equally valuable to scientific progress and understanding. If scientific assertions (unique concept triples) could be properly attributed and credited, the incentive to publish them would increase, and quite conceivably the speed of dissemination of useful research results. Scientific assertions that are both citable and creditable to the authors are the nanopublications that can be easily created in ConceptWiki. This opens up many possibilities for new publication avenues. For instance, those authors who are not in the position to have their papers published in prestigious journals, because they live and work in countries that do not have the research infrastructure to facilitate top-level science, can still build up a public record of their contributions to science. Especially for scientists in the developing world, this may be a welcome addition to the possibilities they have for sharing their knowledge and insights in a structural way. Additionally, the acknowledgment of individual contributions made during the review process of new scientific assertions will make initial publishing efforts more accessible to groups that are currently underrepresented groups in the life sciences.

Authors publishing in PubMed have been imported into ConceptWiki. When possible, each author has been disambiguated through a series of steps based on their publications (shared title words, journal name, co-authors, medical subject headings, language, and affiliation) as well as distinctive features of their names [23]. The unambiguously assigned publications are recorded on each ConceptWiki author page. Using these specific publications, we have constructed a list of prominently figuring concepts for each author (i.e., concepts of interest). The list of concepts of interest is displayed on the ConceptWiki author page and can be edited as appropriate. The concepts shown in the concept of interest lists can be matched to concepts found in scientific triples. From these matches, it is possible to actively solicit comments and review from the ConceptWiki registered users on assertions that appear important to them. Using this method, we explicitly target users within established scientific domains to contribute data, relying on social accountability and self-interest in maintaining a positive reflection within the specific domain. Rather than gathering information common to all Web-enabled humans, we directly target information that is known and verifiable only by a defined group, which improves the quality of the annotations. The assertions can be sent by e-mail or via mobile device or posted on the ConceptWiki author page.

26.6 CULTIVATING COMMUNITIES OF PRACTICE

Given that the establishment of the life science Semantic Web will depend primarily on the will and participation of its consumers [24], gentle processes must be developed to bring regular life scientists into this domain. Recent rapid advances in both information and communication technologies are creating a new revolution in scientific discovery and learning applications. The focus of these new approaches is effective handling of data, specifically the ability to manage orders-of-magnitude more data than ever before, the ability to provide these data directly and immediately to a global community, the ability to use algorithmic approaches to extract meaning from massive volumes of data, and the ability to seamlessly collect community contributions and put these to use. For example, Wikipedia has created an entirely new model by capturing enormous volumes of data and making them freely available in useful ways, transforming how people find and make use of information on a daily basis. Similar but more semantically supported technologies can help us in an era of e-science where data can be made available to everyone, not only to professional scientists but also to students, patients, and teachers. Beyond new scientific discoveries, we are at the dawn of a revolution in collective learning due to these Web-based information and communication technologies. New applications will give users a way to explore and understand a vast and rapidly changing world of interoperable scientific data. Increasing by small increments in complexity will make users feel comfortable so that they can effortlessly see the benefit of these applications; thus, they will gain the necessary insight into the processes required to make future data interoperable. In particular, the interactive and Web-based annotation tools will be valuable learning aids. These systems will promote familiarity with the underlying formalisms and technologies necessary for enabling the Semantic Web. Life science scientists will become capable of spotting subtle differences in the semantics of seemingly similar concepts in their fields. Although ConceptWiki and associated applications specifically target the biological and chemical domain, the software and methods are intended to be reusable for any science moving from heterogeneous data to a shared, global, collaboratory system.

To visualize and identify concepts and assertions found in data repositories, we have developed a recognition system that provides a graphical interface for displaying the result from running the text through an indexing system. The knowledge enhancer is an in-text semantic support application that exploits the data contained in ConceptWiki to recognize concepts on the fly in any website text. Recognized concepts are highlighted with colors specific to the different semantic types. A variety of different functionalities can be invoked when a highlighted concept-denoting term in the text is clicked. For example, each unambiguous term detected by the knowledge enhancer is directly linked to the concept it denotes in ConceptWiki, and therefore all information accessible in ConceptWiki can be displayed in a popup or in the left-hand panel frame of the browser. Another function available in the knowledge enhancer popup constructs a query with a concept and all its synonyms, which are then submitted to multiple search engines. The results are more comprehensive than only using one possible synonym of the concept. The knowledge enhancer application may also be developed as a browser plugin application to allow scholars logged in through university systems to semantically browse text that is only available behind their firewalls.

Scientists will be encouraged to identify familiar concepts in text that are not automatically recognized by the knowledge enhancer. When an unrecognized concept is discovered, the scientist should go through the process of determining whether it is a synonym of a concept already present in ConceptWiki or whether it is entirely new. If it is a new term representing an existing concept, then the term is added as a synonym. If it is a new concept, then the concept plus its definition and semantic type should be added to ConceptWiki. For addition of new concepts, the knowledge enhancer will act as an editor application that can be used to highlight any new word group representing a concept or new term (n-gram) found. Highlighting a new n-gram will automatically copy the n-gram to an editor window and allow the addition of definition and semantic-type data. This facilitates the very easy incorporation of newly coined concepts from recently published scientific articles. New rules will be added to the knowledge enhancer to help determine whether the term identified for addition to ConceptWiki is ambiguous. It is important to acknowledge the editors/annotators of concept-denoting terms as well as those users who add precision to ambiguous terms; therefore, names of registered community editors will be connected to each edit. The contributor names will be displayed to add creditability to the edit and provide acknowledgment for the contribution. Contribution metrics will establish the productivity of every participant, so that appropriate credit for all annotations can be shown, as this has been found to be a motivating technique for increasing annotations.

The intentions of mobile users are typically more immediate and goal directed than the interests of users of fixed or desktop devices. Mobile users often aim at finding out specific, context-relevant information [25], resembling microblogs, when accessing the Internet. Lengthy documents or browsing are typically of less interest due to mobile devices' ergonomics. Therefore, an application available to registered users that actively pushes scientific assertions featuring concepts that appear to be of interest to them is a way to stimulate participation. The mobile assertion verifier or mobile nanopublication reviewer is an application that alerts users to new triples and gives them the opportunity to attach an annotation. Users will be able to judge the validity of new triples using a very accessible dedicated user interface. The primary annotation will be a simple "endorse or deny" choice. The immediate notification of potential new scientific assertions has a twofold effect of positive reinforcement for the scientist; they are the first to know the new assertions entering the scientific record concerning their concepts of interest and they are able to amass nanopublications (all annotations are credited) during periods when other scientific processes are not feasible (e.g., during delays in public transport). It also results in a positive effect for the store of scientific assertions as it builds the scientific validity of the assertions through the endorsements or denials by registered experts.

26.7 ADOPTION OF TECHNOLOGIES IN OPEN PHARMACOLOGICAL SPACE

Drug discovery is data intensive, requiring all major pharmaceutical companies to maintain extensive in-house instances of public biomedical and chemical data alongside internal data. Analysis and hypothesis generation for drug discovery projects require careful assembly, overlay, and comparison of data from many sources, requiring shared identifiers and common semantics. For example, expression profiles need to be overlaid with gene and pathway identifiers and reports on compounds in vitro and in vivo pharmacology. The utility of data-driven research goes from virtual screening, high-throughput screening analysis, via target fishing and secondary pharmacology to biomarker identification. The alignment and integration of internal and public data and information sources are significant efforts and the process is repeated across companies, institutes, and academic laboratories. This represents both significant waste and an important opportunity at the same time.

To address these challenges, the Open PHACTS consortium, comprised of 14 European core academic and small and medium enterprises, will partner with leading pharmaceutical companies to develop an open-source, open-standards, and open-access innovation platform, the Open Pharmacological Space (OPS), via a Semantic Web approach. OPS will bring together the data, vocabularies, and infrastructure needed to accelerate drug-oriented research. This semantic integration hub will address key bottlenecks in small-molecule drug discovery—disparate information sources, lack of standards, and shared concept identifiers—and be guided by well-defined research questions assembled from the participating drug discovery teams (Fig. 26.2). Vocabularies, or simple terminology systems, contain multiple symbols referring to each concept contained in the vocabulary. These symbols are essentially synonyms of each concept. Each symbol used by the OPS will be recognized and mapped to the correct concept UUID, as done in ConceptWiki. This is not always trivial, due to vast ambiguity in symbols used. Many symbols have multiple meanings in the sense that they can refer to multiple concepts. For instance, BSE is an ambiguous symbol as it can refer to many concepts, including bovine spongiform encephalopathy and breast self-examination. To map ambiguous symbols, disambiguation algorithms are used, usually based on context. Several partners in Open PHACTS are among the leading research groups developing and exploiting high-performance disambiguation algorithms in the life sciences and chemistry.

For the OPS, the growing and curated vocabularies containing eventually all symbols for all relevant concepts will be based on the principle that each concept has a UUID, and all symbols known to refer to that concept will be listed in an identity resolution system, presently represented by ConceptWiki. Active resolution of symbols to UUIDs will enable specialized concept taggers to map symbols in text and databases to the correct concepts. Highly ambiguous symbols can be actively discouraged by immediate exposure of the pos-

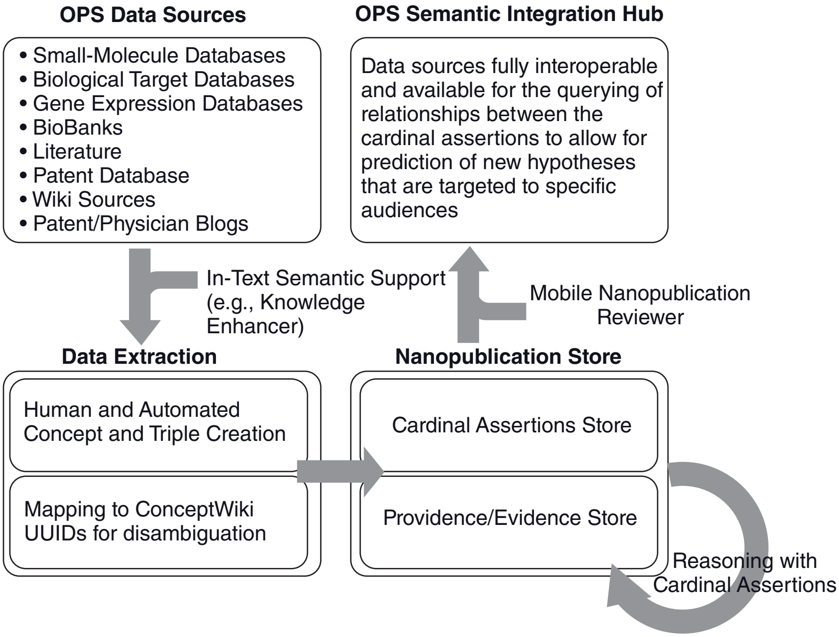

OPS Data Sources
• Small-Molecule Databases
• Biological Target Databases
• Gene Expression Databases
• BioBanks
• Literature
• Patent Database
• Wiki Sources
• Patent/Physician Blogs
OPS Semantic Integration Hub
Data sources fully interoperable and available for the querying of relationships between the cardinal assertions to allow for prediction of new hypotheses that are targeted to specific audiences
In-Text Semantic Support (e.g., Knowledge Enhancer)
Mobile Nanopublication Reviewer
Data Extraction
Human and Automated Concept and Triple Creation
Mapping to ConceptWiki UUIDs for disambiguation
Nanopublication Store
Cardinal Assertions Store
Providence/Evidence Store
Reasoning with Cardinal Assertions

Figure 26.2 The OPS process moving from disparate data sources to an interoperable, searchable resource.

sible homonyms, but of course the use of these will never disappear completely due to legacy literature. Thus, human disambiguation for notoriously problematic symbols will be an integral part of the OPS, aided by in-text semantic support systems, such as the knowledge enhancer.

As described earlier, concept triples or assertions typically have the format of three concepts, namely, an object, a predicate, and a subject. In classical RDF triples, the predicate is not necessarily seen as a concept. However, in the OPS, the predicate will be defined as a concept and have a UUID. Based on this approach, each unique assertion can be defined as a unique three-UUID combination of subject, predicate, and object. The OPS interoperability layer is in fact a very rich triple store or nanopublication store, ideally containing all relevant assertions in the pharmacological space, richly annotated with provenance metadata. Obviously, many nanopublications may make the same basic assertion and only differ in metadata. This is due to the fact that many assertions in the biomedical and chemical space are frequently repeated over and over in articles or database records subsequent to their original publication. Mapping of all identical subject–predicate–object combinations in the OPS to create unique assertions is a crucial step to enable applications that are suitable for retrieval, browsing, and reasoning. It is important to keep track of the frequency of the repetition of the assertions, as it will be reflected in an evidence score catalogued in the OPS nanopublication store.

Assertions that are confirmed by the community via frequent citation or by recognized biomedical and chemical resources should be given the highest evidence score. At the other end of the evidence spectrum are inferred assertions, such as computer predictions based on reasoning algorithms. New nanopublications internally generated by inferencing algorithms over the updated triple stores or through other forms of server side reasoning will have a lower evidence score and will need human verification. For massive distributed triple inference, the Large Knowledge Collider (LarKC) consortium uses the Massive RDF Versatile Inference Network (MaRVIN) [26], which emphasizes scalability through parallelization of the execution of an open set of software components. The system works as a scalable workflow engine for reasoning tasks. In each workflow, there are several components (plug-ins), which are responsible for diverse processing tasks, for example, identifying relevant data, transforming data, selecting data, and reasoning over data. The execution of the workflow is overseen by a decider plug-in [27]. New complex semantic relationships can be queried and discovered through traversing a sequence of links among the entities of interest. For the OPS goals, it will be necessary to include the integration of weightings within the inference rules, to reflect the reliability of the source data. In this way, both false-positive and false-negative relationships can be mitigated by considering only higher confidence or multiple layers of evidence. A mechanism will translate the internal reasoning into some unifying representation language. The calculated reliabilities are kept in a separate store; scientists will be free to use the automatically calculated evidence scores or to calculate their own evidence scores with measures more suitable for their purposes. As well as the necessity for trustworthiness, the system must be able to react quickly as triples may be undergoing constant change through daily or even hourly updates. When new evidence arrives for any assertions, all linked assertions must be refamiliarized with the existing knowledge to interpret the latest findings. The ability to continuously compare and revisit hypotheses is crucial. The final result will allow exploratory querying supporting investigations where one does not initially know precisely what one is looking for but rather uses approaches that permit discovery.

Early estimates by the Open PHACTS consortium members, based on the experience of LarKC and the current size of the Linked Life Data store [18], are that the current number of nanopublications in candidate OPS resources is of the order of 10^{14} while the removal of redundancy may reduce this amount to roughly 1–200 billion unique assertions. With these numbers, the benefit of a massive reasoning system is clear; due to the fact that our conceptualizations of biology have grown in size and complexity, even experts cannot have a wide enough overview of known relationships to be able to make inferences over potentially different disciplines without an automated system. Since the OPS will include an extraction service and in-text semantic support to generate new content in nanopublication format, while guarding the provenance data to enable proper citation and linking back to the original source, the added value of nanopublications generated from traditional texts and database records in computer-readable and interoperable format is easily demonstrated. This added value alone should encourage publishers to offer their full text for nanopublication extraction.

For the OPS project to succeed and more importantly become sustainable into the future, it will be critical to engage the widest possible community of researchers, data, and infrastructure providers worldwide. Engagement of this kind will ensure that the benefits are maximized for the general community; obviously, adoption of the OPS by a wide, global community is directly correlated with its long-term sustainability.

26.8 CONCLUSION

The Semantic Web approach can deliver an in-text semantic support system, a database of scientific assertions capable of supporting massive reasoning, and innovative community annotation tools. The information from a host of experiments such as genomics, proteomics, metabolomics, pharmacokinetics, and pharmacodynamics as well as peer-reviewed literature spanning diverse disciplines can be made available in small packets resembling a microblogging service, thereby transforming how people find and make use of data on a daily basis. With appropriate recognition and traceability of the assertions via the nanopublication scheme, this will enable an entirely different way of scholarly communication, much more adapted to the current rate of data production. The OPS is one of the first realizations of this potential.

REFERENCES

1. Mons B. Calling on a million minds for community annotation in WikiProteins. *Genome Biol* 2008;9:R89.
2. Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C. WikiPathways: Pathway editing for the people. *PLoS Biol* 2008;6:e184.
3. Hoffmann R. A wiki for the life sciences where authorship matters. *Nat Genet* 2008;40:1047–1051.
4. Royal Society of Chemistry's ChemSpider. Available: http://www.chemspider.com.
5. Huss JW. The Gene Wiki: Community intelligence applied to human gene annotation. *Nucl Acids Res* 2010;38:D633–639.
6. Mons B, Velterop J. Nano-Publication in the e-science era. Workshop on Semantic Web applications in scientific discourse. In Clark T, Luciano JS, Marshall MS, Prud'hommeaux E, Stephens S, Eds. *Semantic Web Applications in Scientific Discourse 2009: Proceedings of the Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009),* October 26, 2009, Washington, DC. CEUR-WS.org, 2009, vol. 523.
7. Groth P, Gibson A, Velterop J. The anatomy of a nanopublication. *Inf Serv Use* 2010;30:51–56.

8. Carroll JJ, Bizer C, Hayes P, Stickler P. Named graphs, provenance and trust. In *WWW '05: Proceedings of the 14th International Conference on World Wide Web*, May 10–14, 2005. Chiba, Japan. ACM Press, 2005, pp. 613–622.
9. Bizer C, Gauss T. Disco-Hyperdata Browser: A simple browser for navigating the Semantic Web. Available: http://www4.wiwiss.fu-berlin.de/bizer/ng4j/disco/.
10. SIMILE. Semantic Interoperability of metadata and information in unlike environments. Available: http://simile.mit.edu/.
11. Quan D, Huynh DF, Karger, D. Haystack. A platform for authoring end user Semantic Web applications. In Fensel D, Sycara K, Mylopoulos J, Eds. *The SemanticWeb—ISWC 2003*. Heidelberg/Berlin: Springer, 2003, pp. 738–753.
12. Steer D. XML.com: BrownSauce: an RDF browser. Available: http://www.xml.com/pub/a/2003/02/05/brownsauce.html.
13. Scerri S, Abela C, Montebello M. SemantExplorer: A Semantic Web browser. In Nunes MB, Isaías P, Eds. *Proceedings of the IADIS International Conference WWW/Internet 2005*, October 19–22, 2005. Lisbon, Portugal. IADIS Press, 2005, pp. 35–42.
14. Huynh DF, Mazzocchi S, Karger D. Piggy Bank: Experience the Semantic Web inside your Web browser. *Lect Notes Comput Sci* 2005;3729:413–430.
15. Dzbor DM, Domingue DJ, Motta PE. Magpie, towards a semantic web browser. *Lect Notes Comput Sci* 2003;2870:690–705.
16. Quan D, Karger, D. How to make a semantic web browser. In Feldman SI, Uretsky M, Najork M, Wills CE, Eds. *WWW2004: Thirteenth International World Wide Web Conference*, May 17–22, 2004, New York. ACM Press, 2004, pp. 255–265.
17. Neumann EK, Quan D. Biodash: A Semantic Web dashboard for drug development. *Pacific Symp Biocomput Proc* 2006;11:176–187.
18. Linked Open Data. Available: http://linkeddata.org.
19. W3C Semantic Web Health and Life Sciences Interest Group. Available: http://www.w3.org/blog/hcls?cat=85.
20. OmegaWiki. Available: http://www.omegawiki.org/Meta:Main_Page.
21. Open Source Software Project. Universally Unique Identifier (UUID). Available: http://www.ossp.org/pkg/lib/uuid/.
22. Berners-Lee T, Hendler AJ, Lassila O. The Semantic Web. *Sci Am* 2001;284:34–43.
23. Torvik VI, Weeber M, Swanson DR, Smalheiser NR. A probabilistic similarity metric for Medline records: A model for author name disambiguation. *AMIA Annu Symp Proc* 2003:1033.
24. Good BM, Wilkinson MD. The Life Sciences Semantic Web is full of creeps! *Brief Bioinform* 2006;7:275–286.
25. W3C. Mobile web best practices 1.0—Basic guidelines 2006. Available: http://www.w3.org/TR/2008/REC-mobile-bp-20080729.
26. Oren E, Kotoulas S, Anadiotis G, Siebes R, ten Teije A, van Harmelen F. MaRVIN: A platform for large-scale analysis of Semantic Web data. *Web Semant* 2009;7:305–316.
27. Fensel D, et al. Towards larkc: A platform for web-scale reasoning. In *Proceeding in the International Conference on Semantic Computing (ICSC)*, August 4–7, 2008, Santa Clara, CA. IEEE Computer Society, 2008, pp. 524–529.

27

COLLABORATIVE VISUAL ANALYTICS ENVIRONMENT FOR IMAGING GENETICS

ZHIYU HE, KEVIN PONTO, AND FALKO KUESTER

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.

27.1 MOTIVATION: THE LARGER CHALLENGE

The annual amount of information the world produces has been at a continuously increasing pace over the past decade. Varian and Lyman found that in the early part of the 2000s the world produced between one and two exabytes (one billion gigabytes, or 10^{18} bytes) of unique information per year [1]. A white paper by the International Data Corporation (IDC), sponsored by EMC, found that this number had jumped to 161 exabytes by 2006 [2]. This paper projected that by the year 2010 this number could jump to 988 exabytes per year. This information overload has left many researchers scrambling to find ways to analyze this kind of massive amount of data. While well-defined tasks can be delegated to computer processing, oftentimes, researchers may desire to approach data analysis from a more exploratory perspective. In this way, researchers may attempt to find characteristics and/or patterns of interest in the available information. Similar to the overall data explosion across the fields of science and engineering, medicine and biomedical engineering have seen their own data avalanche through the introduction of new imaging, sampling, and modeling techniques. As a result, domain experts are pressed to create new methodologies, techniques, and tools to manage and most of all harness this wealth of information.

The field of visual analytics attempts to approach these problems through human-centric visualization. Thomas defines visual analytics as the “science of analytical reasoning facilitated by interactive visual interfaces” [3, p. 4]. The goal of visual analytics is to present data in such a way that the human mind is able to efficiently process information, combining the benefits of machine and human analysis. Thomas proposes that visual analytics has the ability to “detect the expected and discover the unexpected” [3]. The goal of this paradigm is to allow researchers to visually explore data sets without explicitly knowing what they are looking for initially. For general-purpose pattern recognition, the human brain can outperform machine-based algorithms [4]. It only takes the human brain a little over a tenth of a second to identify and classify an object in a complicated environment [5]. Furthermore, the human brain can find patterns and differences even when the differences seen in objects are not easily quantifiable, that is, the symbol grounding problem [6].

The challenge for visual analytics is that data must be organized and presented in a meaningful way to be effective for analysts. For example, the visual analytics paradigms useful for large image collection will most likely be different from those for detecting intruders on a network. Visual analytics techniques need to be customized for the data being analyzed as well the users of the system. This is to say, there is no “one size fits all” solution. As such, this chapter focuses on a high-level overview with focus on large-scale multidimensional data analysis in environments suitable for team-based visual analytics. Commonly encountered data in biomedical research consist of multidimensional, layered two-dimensional data and three-dimensional volumetric data sets combined with multivariate records in alphanumeric form. While these types of data represent just a subset of the overall data avalanche, the fields for which the associated data analysis techniques are applicable are rather broad.

In order to reap the full benefit of the human computation component, it is necessary to break away from the traditional single-user mouse-and-keyboard paradigms. These standard setups present users with a few megapixels to view and manipulate. For large-scale data sets, this interface allows a narrow window through which users can analyze data. A new generation of ultra-high-resolution tiled display systems creates new opportunities to present information, which can greatly increase the “bandwidth” to the human visual system [7–10]. The human retina can process approximately 10 one-million-point images per second [4]. As tiled-display systems are able to render hundreds of megapixels simultaneously, these systems are capable of fully saturating the human visual system. These systems have been shown to be much more effective for data analysis compared to the standard pan-and-zoom environments. The advantage of these systems comes not only from the immense amount of pixel real estate available but also from the user’s ability to physically interact with the space. Furthermore, the large display size allows multiple users to interact in the workspace collaboratively.

While collaborative workspaces have been shown to have great utility, developing these environments effectively is a challenging endeavor [11]. In order to keep users engaged, working environments must present users with the ability to see and interact with content. As group sizes grow to more than a few users, this requirement presents challenges for standard interface modalities. Unfortunately, traditional means to interface with these types of systems generally do not work well in multiuser paradigms as interactions are often controlled from standard single-user mouse-and-keyboard interfaces. In particular, multiuser interface modalities developed for ultra-high-resolution display spaces have generally provided specialized solutions requiring users to learn new paradigms before interfacing with the system.

Multitouch devices present new opportunities for interfacing with data sets. While standard interface devices limit speed of interaction, creating a bottleneck in data analysis [12], multitouch systems leverage the user’s real-world experiences as part of the computer interface. Recent advancements in consumer electronics have generated an explosion of multitouch capable devices. Multitouch devices have been shown to be a natural interface as they leverage users’ previous experiences. Most of all, users carry many of these devices on their person (e.g., cell phones, tablets), providing a familiar and readily usable hardware interface for use in collaborative digital workspaces. Huang showed that gestures are physical expressions of mental concepts [13]. Multitouch interfaces enable simple gestures which can convey complicated interactions [12]. Furthermore, natural interfaces allow for greater accessibility for multiple users to work on a single data set, thereby allowing for more eyes and brains to examine the data.

27.2 PREVIOUS WORK RELATED TO COLLABORATIVE TECHNOLOGIES

A brief review of the literature with respect to digital collaborative workspaces, tiled display environments, multitouch systems, and multiuser interaction techniques is presented next to better outline the capabilities, techniques, and tools that may be harnessed for the creation of new collaborative technologies and environments for biomedical research.

27.2.1 Collaborative Digital Workspaces

Digital colocated collaborative workspaces have shown great potential for enabling a multiuser approach for scientific analysis [14]. Heer and Agrawala outline the strengths and challenges for multiuser environments in their paper, "Design Considerations for Collaborative Visual Analytics" [11]. "The office of the future" project demonstrated a collaborative virtual working environment for the workplace. Churchill et al. examined the history of these types of spaces, the technical issues and challenges these systems present, and the types of applications enabled through this technology [14]. Taesombut et al. explored real-time collaborative efforts to analyze earth science data on tiled-display systems [10]. Scott et al. examined how to develop colocated collaborative workspaces for tabletop environments [15]. Shen et al. continued in the same direction and developed a software architecture for multiuser table interactions [16]. Peltonen et al. studied how multiple users interacted on a multitouch-enabled display wall [17], analyzing how users negotiated their interaction space and how conflicts between users were resolved.

27.2.2 Large-Scale Multitile Display Environments

Large-scale, ultra-high-resolution display environments have emerged as viable solution for coping with some aspect of the data avalanche, allowing sizable data collections to be colocated to literally help with presenting the "big picture" in conjunction with being able to explore varying levels of details intuitively and interactively by traversing through the data with power-of-10 capabilities. Rendering content on tiled-display systems is not a straightforward task though. Various paradigms have been derived in order to present a uniform visual environment across the entire display space. Each of these paradigms presents unique advantages and challenges for domain experts (users) and developers alike. The advantages and disadvantages of the various approaches is briefly outlined below since they define possible use scenarios in the context of collaborative biomedical research.

27.2.2.1 Geometry Forwarding One way of controlling tiled-display walls is to create a virtual unified display as shown by Chromium [18] and OpenSG [19]. This method forwards each call made to the graphics card on the head node to the entire display environment and is advantageous as it works with most applications and may not require code to be recompiled. Unfortunately, this approach generally does not work with textures and shaders and has poor synchronization mechanisms, limiting its use when real-time data processing and filtering are desirable, which should be considered core capabilities of collaborative visual analytics spaces.

27.2.2.2 Pixel Streaming Another approach to drive tiled-display environments is to send pixel content to display nodes as outlined in the Scalable Adaptive Graphics Environment (SAGE) [20–22]. In this approach, one system renders content into a buffer that is mapped to the tiled-display environment. This buffer is segmented to match the configuration of the wall and is streamed out via the network. The advantage of this approach is that data have to be processed only once on the streaming node, being the only node required to have access to data files and applications. Consequently, the rendering nodes only need to have minimal computational power as they are only tasked with receiving and rendering this content. The drawback of this concept is that it requires a very low latency, high-bandwidth network. Larger buffer sizes or frame rates require increased network costs, removing the possibility for native resolution rendering on large tiled-display systems. Even for content, which fits inside of the network bandwidth requirements, the read-back and splitting operations also have performance costs associated with them, thereby increasing latency. Finally, as the burden of rendering content is left to a single machine for a uniform environment, the performance of this system sets the upper bound for the capabilities of the tiled-display environment.

27.2.2.3 Macroblock Forwarding Chen, interested in distributing the computational workload for playing large-scale media, derived a method for segmenting and forwarding compressed video information. As most compression schemes contain global motion vectors and progressive frame decoding, they do not work for region-of-interest decoding. In the MPEG2 standard, motion vectors are confined to macroblocks, allowing for the possibility of partial frame decoding [23, 24]. Unfortunately the MPEG2 standard only allows for video sizes up to 1920 × 1152 [25], meaning that encoding videos of greater resolution cannot be accomplished using common encoders. Furthermore, this approach requires a second level of nodes in between the head node and render nodes in order to negotiate macroblock forwarding. These routing nodes must receive and resend information, including header data, which incurs an additional 20% bandwidth cost. While this method is useful for ultra-large-resolution video data, it requires additional hardware, is limited in its playback ability, and still requires substantial network resources in order to operate.

27.2.2.4 Distributed Application The distributed application approach as shown in VRJuggler [26] and Cross-Platform Cluster Graphic Library (CGLX)

[27] allows for a tiled-display environment to also act as a distributed computer. In this approach, the same application is started on the tiled-display environment as well as the head node. The viewpoint is changed on each rendering node to match the overall tiled-display view with that of the head node. All user events such as key presses and mouse movements as well as graphical events such as buffer swaps are synchronized. The advantage of this approach is that it requires very little network bandwidth in order to operate and has the ability to scale well beyond the other approaches mentioned. Furthermore, this approach allows for computation to be distributed throughout the display environment and for graphic card shaders to run natively. The disadvantage of this approach is that code must be compiled for the tiled-display environment. Furthermore, each node needs to have access to the data to be rendered, either by distributing it to the render nodes beforehand or through a network file system or data stream.

27.2.3 Multitouch-Enabled Environments

While multitouch technology has only recently been adopted for consumer applications, the history of multitouch research has spanned multiple decades. In 1984, Lee wrote his master's thesis [28] on the use of multitouch and a year later continued his work with Buxton and Smith [29]. Unfortunately, while the interface technology was being developed, the computation power needed for its practical implementation was still lacking. In 2001, Westerman et al. published a paper on human–computer interaction discussing how multitouch could be used as an intuitive computer interface [30]. Later that year, DiamondTouch, a multiuser touch system was produced [31]. In 2004, Han [32] and Wilson [33] created significant interest in this technology, following their presentation of more accessible approaches to multitouch technology. Subsequently, Smith et al. [34] proposed creating low-cost pressure-sensitive surfaces in 2007. Today, multitouch devices are pervasively available and used and represent a unique interface for collaborative visual analytics environments.

27.2.4 Multiuser Interaction Techniques for Large-Display Environments

Several approaches have been developed for multiuser interactions in large-display environments and stress different types of modalities and metaphors. "Fluid Interaction with High-Resolution Wall-size Displays," by Frangois et al., showed how a pen interface could be useful for interfacing with display walls [35] and enabled user identification through handwriting analysis. "LumiPoint: Multi-User Laser-Based Interaction on Large Tiled Displays," by James Davis and Xing Chen, examined techniques for enabling multiuser interaction through laser pointers [36]. Laser pointer position and velocity were tracked through a central server and multiuser support enabled by adding laser pointers of differing colors. This approach was found to be scalable by adding additional cameras. Malik et al. explored the use of touch pads to interact with a display wall in "Interacting with Large Displays from a Distance with Vision-Tracked Multi-Finger Gestural Input" [37]. This system allowed two users to simultaneously use the same working environment. Cameras were used to determine hand placement and allowed the incorporation of hand-specific gestures, which could bridge the limited touchpad size to the wall size.

Cao and Balakrishnan [38] explored the use of a passive wand for interaction with tiled-display environments, allowing users to perform gestures to select, move, and rotate content. While this system worked well for a single user, it showed performance degradation with additional users. Ringel et al. explored a multiuser virtual whiteboard in "Barehands: Implement-Free Interaction with a Wall-Mounted Display" using diffuse illumination in order to track upward of 120 touches simultaneously [39]. Another whiteboard approach presented by Rekimoto proposed a system for allowing many users to interface with a whiteboard through personal digital assistants (PDAs) [40]. Rekimoto proposed content be transferred directly from each user's PDA to a virtual whiteboard environment.

27.3 A CYBER-COLLABORATORY FOR IMAGING GENETICS

In biomedical research and especially in the area of imaging-genetics research, it is commonplace that scientists from different disciplines work collaboratively on a variety of data sets. In schizophrenia research, for instance, brain imaging experts, geneticists, neuropsychiatrists, and clinical physicians in collaboration with statisticians will study a subject set focusing on different biological levels. Gathering the different data points and applied analysis techniques in a unified collaborative workspace and providing a platform that allows users to communicate in a compatible visual language can then provide a mechanism to jointly analyze complex relationships and propose and harden new hypotheses.

In order to create a visual analytics framework for the processing of large-scale hybrid biomedical data in a collaborative setting, the available display real estate has proven to be one critical factor, allowing the visual analytics pipeline to expose the data concurrently and at a scale suitable for use by large research teams. This progress led to the development of our room-sized HIPerWall and HIPerSpace visualization environment. These operate at 204 megapixels and 286 megapixels resolution, respectively, making them the highest resolution displays of their time and the first to break the 100- and 200-megapixel-resolution barriers. Both systems take advantage of the cluster graphics middleware, called CGLX [27], to seamlessly interconnect and control flexibly scalable, networked, multitile environments.

A case will be presented for which the developed visual analytics techniques are applied to explore and discover correlations between brain function, anatomy, and genetics. As part of this, scientists from various domains collaboratively worked on data analysis and synthesis while drawing from an enormous amount of multidisciplinary data available at various scales and levels of resolution. Case study results are derived from user experiences reported by neuroscientists, clinical physicians, statisticians, and computer scientists. The evaluation of these reports confirms that the proposed visualization and analytics framework is an efficient mechanism to detect and validate expected information but most importantly an instrument aiding with the discovery of unexpected information contained within the multiscale multimodality data. Within this context, it had to be possible to process and fuse multimodality and multiscale biomedical data at interactive rates at the genomic, proteomic, cellular synaptic, psychometric, and behavioral levels, requiring the development of a range of processing, visualization, and interaction techniques. Of particular importance were techniques for the representation of conventional structural and functional two- and three-dimensional (2D, 3D) imaging data as well as brain activation patterns, derived information such as brain tractography, genetics information in the form of results from SNP array runs conducted for each patient, and digitally published reference material including archival publications via PubMed and elsewhere. Once this was possible, domain experts were in the position to concurrently and collaboratively work toward extracting new insights while also significantly shortening data processing times when compared against traditional methods, which in one particular case led to a discovery of genetic markers for schizophrenia, previously considered years of work away.

27.3.1 Visualization Practices in Imaging Genetics

Scientific visualization is pervasive in many biomedical research areas. Visualization tools usually utilize 3D or 2D algorithms to render a single data modality such as computerized tomography (CT), magnetic resonance imaging (MRI), or microscopic data. Examples are volume rendering and isosurface rendering on a set of radiology data. Users are commonly domain experts that understand the particular data and thus can change the rendering parameters to visually interpret the region of interest and highlight important aspects to others. Further reasoning can be based on the visual data or from the actual raw data. One such example is the Visualization Toolkit (VTK) [41]. VTK provides a set of implementations of common visualization algorithms and programming interfaces are defined for customizing applications. Some other tools go one step further by integrating image processing algorithms. One such example is the Insight Segmentation and Registration Toolkit (ITK) [42]. This toolkit was designed to support the Visible Human Project and has become a platform for fundamental segmentation and registration algorithms. Another example is 3D Slicer [43], which is based on VTK and ITK to support visualization and image analysis capabilities for biomedical data. In bioinformatics, the enormous amount of data make it feasible to combine visualization and data-mining algorithms into one system. At the Erasmus University Medical Center in Holland [44], researchers have employed a combination of visualization and artificial intelligence techniques as an important tool for converting millions of tissue and gene records into straightforward visual information. In their research, visualization can provide a quick overview of a large amount of data and help the user narrow down the initial region of interest while the AI-aided quantitative analysis algorithm further analyzes the data with more accuracy. Zhou and Liu [45] developed a Java program for visually analyzing microarray data for gene expression by supporting microarray data visualization, quantative assessment, and data mining. Jean Pylouster et al. [46] described a Web-based tool for gene analysis. This Web tool performs statistical analysis on gene expression data and identifies the gene tags that are differentially expressed and presents plots for the final results. Boyle and his colleagues [47] developed a software package for exploring embryonic development using time-lapse confocal imaging and a tree structure to describe the location and relationship of each nucleus. Their system implements visualization and other algorithms to extract biological significance out of the data. The combination of visualization and analytical tools are widely used beyond the above-mentioned areas. In addition to existing algorithms, techniques, and tools, current data repositories exist that provide access to abundant community knowledge. For example, genecards.org provides access to an index database for gene symbols, and Human Genome Browser Gateway at the University of California—Santa Cruz [48] provides access to an image-enhanced database for human genomics.

27.3.2 Technical Approach

The presented visual analytics framework draws from two highly interactive parallelized display walls termed HIPerWall and HIPerSpace, first commissioned in 2005 and 2007, respectively, capitalizing on custom-developed middleware called CGLX. The visual analytics tools in turn use both of these to provide an intuitive and collaborative digital workspace.

27.3.2.1 Hardware Configuration HIPerWall utilizes fifty 30-in. displays with a resolution of 2560 × 1600 pixels each, configured in a 10 × 5 (width-by-length) layout, resulting in a combined resolution of 25,600 × 8000 pixels (204,800,000 pixels total). HIPerSpace in turn uses 70 display tiles in a 14 × 5 layout, resulting in 35,840 × 8000 pixels resolution (286,720,000 pixels total). The computing and rendering cluster of HIPerWall is based on 25 Power Mac G5 running OS-X, with a 2.7-GHz IBM PowerPC processor, 2 GB of RAM, with dual, dual-core processors, and NVIDIA Quadro FX 4500 graphics. Each one of these HIPerWall nodes drives two displays and is interconnected in a dedicated gigabit subnet. Data access is provided via a dedicated, nfs mounted storage node (HIPerStore), while a stand-alone control node serves as the front end. Similarly, HIPerSpace utilizes 18 machines running Linux,

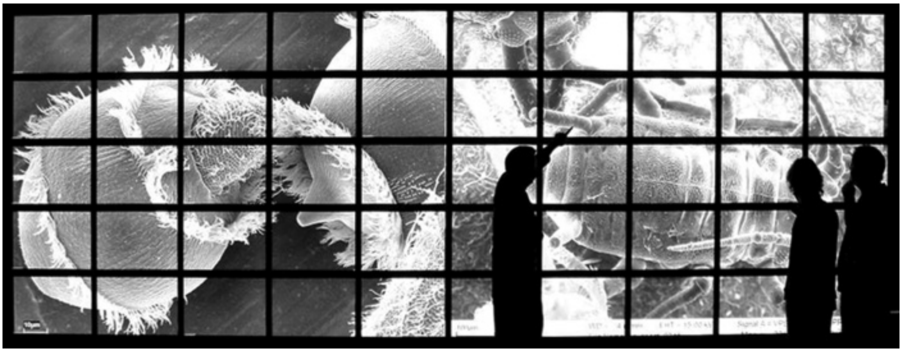


Figure 27.1 HIPerWall used to analyze microscopy data.

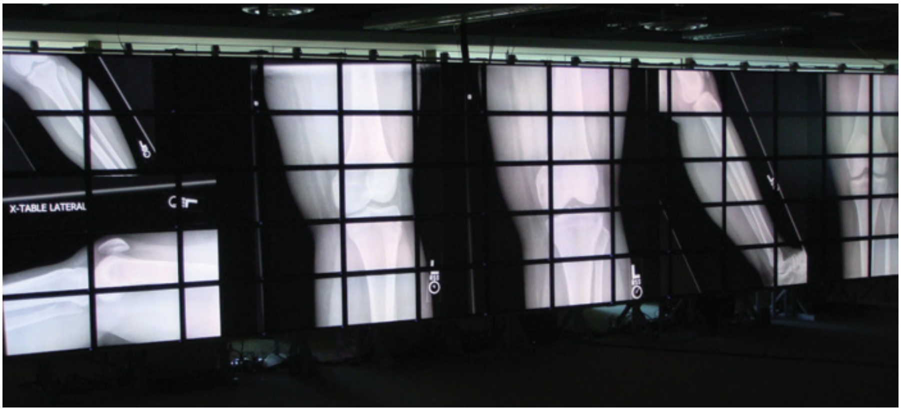


Figure 27.2 HIPerSpace used as a digital lightboard.

with 2.4-GHz quad core Intel processors, 4 GB of RAM, and NVIDIA Quadro FX 5600 graphics, interconnected with dedicated gigabit Ethernet and additional 10-Gbps uplinks into its network-centric data storage back end, while multiple auxiliary nodes may serve as its control front end. Notably, both systems can be seamlessly tied together into one, termed HIPerVerse, using the existing middleware. Figures 27.1 and 27.2 show HIPerWall and HIPerSpace respectively, while being used for the analysis of biomedical data.

27.3.2.2 Cluster Middleware The CGLX middleware is a flexible, transparent OpenGL-based graphics framework (Fig. 27.3) for distributed high-performance visualization systems in a master–worker setup. The framework was developed to enable OpenGL programs to be executed on heterogeneous visualization clusters and to maximize the achievable performance and resolution for OpenGL-based applications on such systems. To overcome performance- and configuration-related challenges in networked display environments,

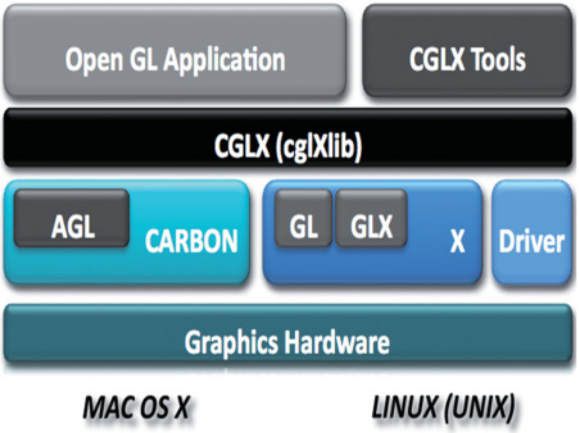
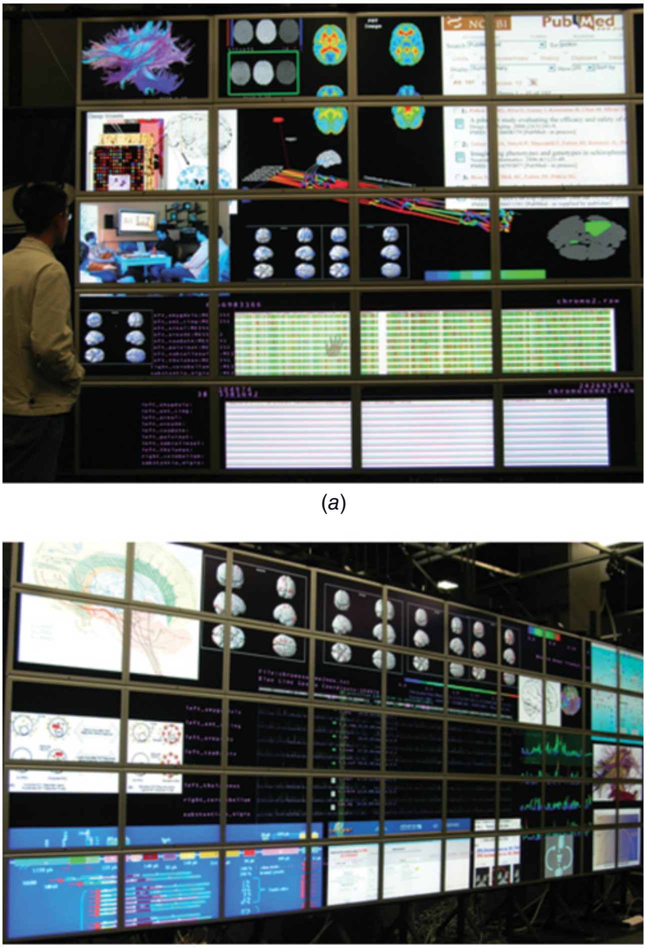


Figure 27.3 Architecture of CGLX, cross-platform, cluster-graphics middleware.

CGLX launches and manages instances of an application on all rendering nodes through a lightweight thread-based network communication layer. A GLUT-like (Open GL Utility Toolkit) interface is presented to the user which allows this framework to intercept and interpret OpenGL calls and to provide a distributed large-scale OpenGL context on a tiled display. CGLX provides distributed parallelized rendering of OpenGL applications with all OpenGL extensions that are supported through the graphics hardware.

27.4 VISUAL ANALYTICS APPROACH

By design, HIPerWall/HIPerSpace provides a large-scale workspace for massive visual data correlation at great detail. More precisely, it supports the integration of multiple imaging modalities from different biological scales and information domains. Data size and overall system configuration mandate a distributed approach, which allows individual nodes to access and visualize data while being properly synchronized via a control (head) node. The control node also manages most user interactions as the front end into the collaborative workspace. This approach takes full advantage of the computational and graphical power of each single node while reducing network traffic through an out-of-core, adaptive, and progressive data access paradigm. Since the framework was developed such that it scales flexibly from a single node with just one display tile to arbitrary cluster configurations, users may opt for using their personal laptop side by side with the large-scale visual real estate of the multitile system. Figure 27.4 shows a visual analytics session combining multiple image modalities consisting of structural, functional, and brain response data side by side with brain tractography, genetic, and archival publication data.



(*a*)

(*b*)

Figure 27.4 Data mash-ups during imaging genetics sessions.

Leveraging the existing system architecture, the visual analytics framework was compressed down to three major components managing data visualization, processing and synthesis, and user interaction. The visualization stage is responsible for selecting the appropriate rendering technique for the available image sources, geometric models, multidimensional data constructs, and alphanumeric data records. The processing-and-synthesis stage then provides access to content and context-specific interaction, allowing data records to be loaded, filtered, clustered, segmented, fused, and correlated against auxiliary information, with an immediate-mode data analysis paradigm in mind. Finally, the user interface for the workspace is accessible via two complimentary routes in the form of a graphical user interface on the workspace's head node(s) providing access to all available data assets and analysis tools and the workspace itself within which each visual can be directly interacted with and controlled via a hand-held interface device such as a gyroscopic mouse.

27.4.1 Imaging Genetics Case Study

The target audience in this case is an interdisciplinary group of neurologists, neuroanatomists, psychiatrists, imaging experts and geneticists, statisticians, and computer scientists collaboratively and concurrently working toward formulating new hypotheses in the domain of schizophrenia research. Traditionally, most of these researchers had worked on data sets using conventional data analysis techniques, methodologies, and settings commonly working alone within their specific domain of specialization coupled with occasional synchronization points with other team members. This traditional approach has two major limitations in that the massive amount of data generally cannot be visualized in a comprehensive format using standard techniques and devices, frequently falling back to studying individual images or genetics data at a time, captured in the form of slides, overheads, and printouts to organize the available material. The second limitation is that of restrictied interaction and lack of team-based, immediate-mode access to the data, making it difficult to develop and explore new hypotheses requiring transdisciplinary collaboration. Transdisciplinary data analysis relies heavily on efficient interactions and especially on how quickly domain knowledge can be exchanged and applied on common subjects.

A typical imaging genetics study in this digital collaborative workspace context is based on concurrent visualization and analytics tasks, targeting (1) chromosome, gene, and single-nucleotide polymorphism (SNP) visualization associated with selected brain regions; (2) gene selection, that is, screening by predefined patterns; (3) brain imaging; (4) brain tractography; (5) cross-disciplinary data analysis and linkage analysis; (6) online database search and cross-referencing; and (7) real-time data exchange and interaction.

The baseline data in this case consist of brain imaging, genetics, and clinical test results as well as multimedia data including related databases and publications. The latter are synthesized via a Web interface that creates active texture maps of the specified online resource, flexibly embedding active Web content into the collaborative workspace. This comes with the added benefit that any Web content can be fused with the collaborative workspace, allowing Web-editable documents, notebooks, and tele-presence sessions to be embedded. This data mash-up effectively allows the colocation and parallelization of tasks that traditionally require separate and sequential sessions. Figure 27.5 illustrates the ability to navigate between different functional modules seamlessly and independently between head node(s) and the collaborative workspace providing the shared view of all active data assets. Selected visualization strategies that are being utilized are described next.

27.4.1.1 Chromosome Visualization Analysis of the patient's chromosomes carrying a massive amount of genetic information and correlation

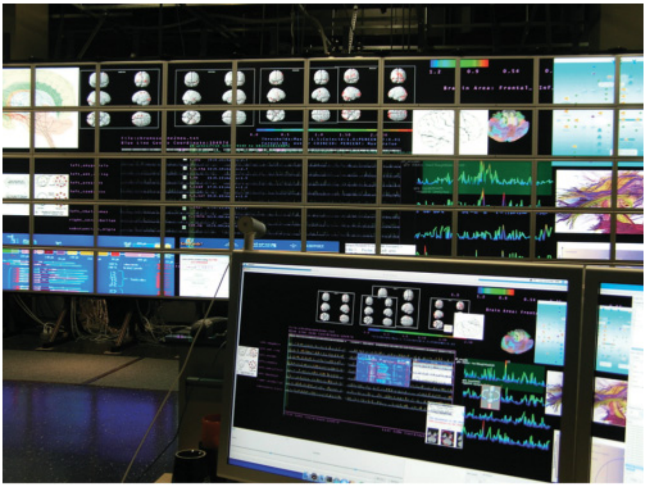


Figure 27.5 Scalable imaging genetics workspace allowing data to be processed through user-centric view and shared visual analytics space, the wall.

against results obtained with a variety of imaging techniques provides a vital step in studying brain-related diseases. Given 23 pairs of chromosomes with over 1 million SNPs on these chromosomes, the conducted research targeting the alleles (DNA sequences) is based on statistical tests with *p*-values on different genotypes (specific genomes, which include both the genes and the noncoding sequences of the DNA). Given the data size, an efficient data representation scheme is needed to highlight significant values and correlations. The developed visualization strategy utilizes a 2D and a 3D view, respectively, to visually combine SNP base-pair position with the associated *p*-value. The 2D view visually connects each SNP position to the respective base-pair location with the chromosome represented on the *x*-axis and the corresponding *p*-value mapped to the *y*-axis (Fig. 27.6). Users can then interactively query for *p*-values and the associated gene at arbitrary base-pair locations by either moving a search window across the graph or by specifying the boundary condition for the search space.

To further bracket values of interest, hardware accelerated filters can be applied to recolor the graph in accordance with a user-definable color look-up table. In order to correlate genetic information with the brain, data obtained for multiple brain regions is exposed concurrently. To provide further insight into a subrange of base coordinates, a user-definable search window can be placed on the 2D representation, serving as a magnifying glass into a 3D representation. This 3D window can be set to cover an arbitrary number of SNP base-pair positions or number of genes in accordance with user preferences. Information associated with a particular SNP position as well as statistics on a particular gene is displayed as the user browses across the chromosome (Fig.

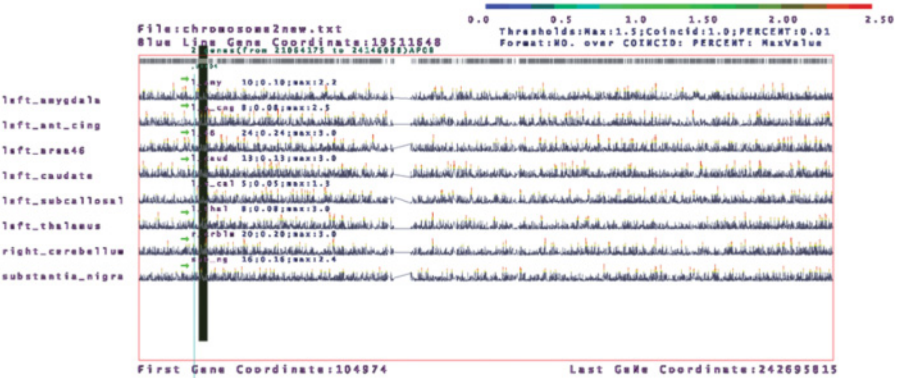


Figure 27.6 Two-dimensional visualization of SNP data obtained for multiple brain regions in the context of chromosome 2.

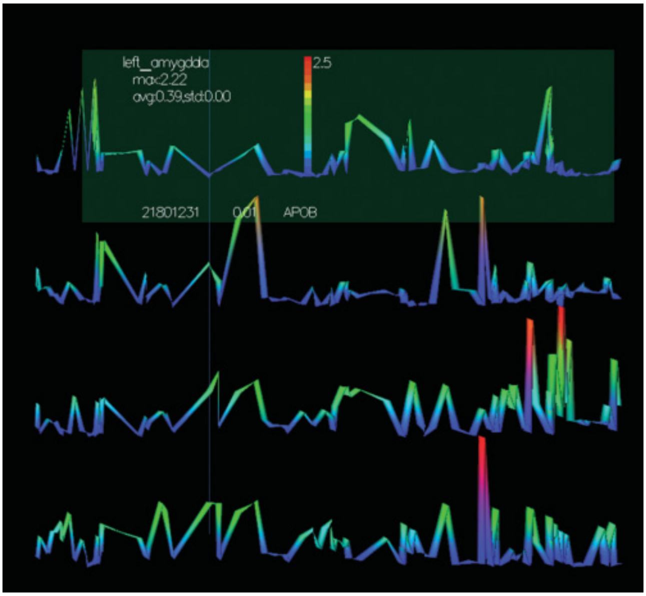


Figure 27.7 Three-dimensional visualization of SNPs as function of selected subregion within a targeted chromosome.

27.7). In practice, the ability to visually explore and align data in 3D by rotating the graphics turned out to be a powerful analysis modality.

To aid with fast detection of gene patterns, users can also specify search criteria such as maximum *p*-value and percentage over a threshold, which can be changed via textural input or provided GUI controls. The search result, which is the gene segment that satisfies the conditions, is then shown on both

2D and 3D visuals. This is referred to as the active search mode and proved invaluable for swift data analysis and helped with the identification of two new genetic markers for schizophrenia. Another search mode is the already mentioned passive mode, for which users can scan through the whole chromosome in all brain areas by manually dragging a search window. The search window size is variable and can be fit to a specified number of genes or SNP locations. Gene information associated with the browsing window, such as the gene symbol and *p*-value, is displayed as the user pans through the chromosome and is further augmented with real-time statistics, such as mean value and standard deviation. Furthermore, data query is also linked to an integrated Web interface, allowing searches to be applied against other published materials in standard data banks such as GeneCards or publications such as PubMed. Thereby, the Web interface provides access to a rich collection of knowledge as an integrative part of the visual analytics space.

27.4.1.2 Analysis of 2D and 3D Image Data Imaging results from CT, MRI, position emission tomography (PET), and traditional X rays can either be viewed in a slice-by-slice 2D mode or be rendered as volumetric 3D images with proper registration and visualization algorithms. In the context of the collaborative visual analytics space, all 2D data are directly shown as texture maps and 3D data as volumes with optional 2D cross sections freely movable in sagittal, coronal, and axial directions. Both vertex and fragment shader programs are applied to change display properties at run time. For example, different color look-up tables can be applied to flexibly highlight different core features, while the user-definable threshold can be applied to remove unrelated brain areas. Study on group subjects can be performed by arranging images based on the associated data values. For example, while studying a group of schizophrenia subjects, the underlying statistical tests associated with specific genes can be used to sort their functional images. This facilitates a quick detection of the pattern that correlates genes and brain activation areas. Statistical beta maps can be used to perform quantitative comparisons. Figures 27.8 and 27.9 show the digital equivalent of a light-board with the added advantage that all data can be interactively and intuitively manipulated.

27.4.1.3 Web Interface Internet resources, such as genecards.org and PubMed, contain a vast number of categorized research results. Another example is the UCSC gene browser [48], a Web-based, interactive database for genomics. These types of public domain resources can greatly enhance the expressiveness and effectiveness of visualization tools, and a Web portal engine was implemented as an extension to the WebVR platform [49]. WebVR provides a Web browser interface which seamlessly translates queried information into texture space for streamlined use in visualization environments. More specifically, Web queries are sent to WebVR, which retrieves the corresponding page, converts it into a tagged image, and subsequently returns a texture map to be used within the visualization environment. This active texture, also

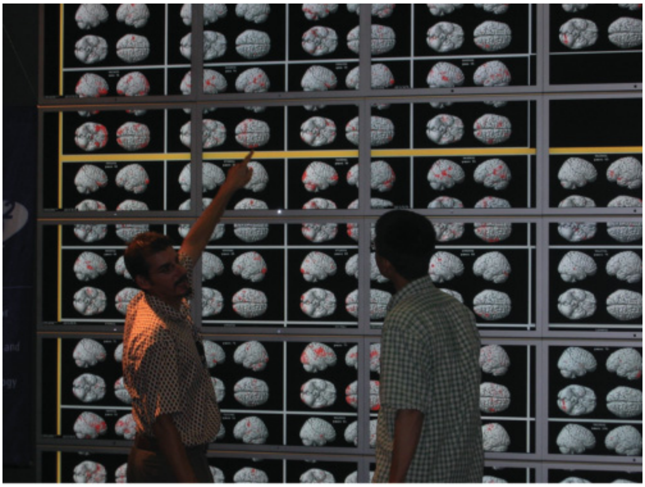


Figure 27.8 Collaborative correlation and clustering of PET data.

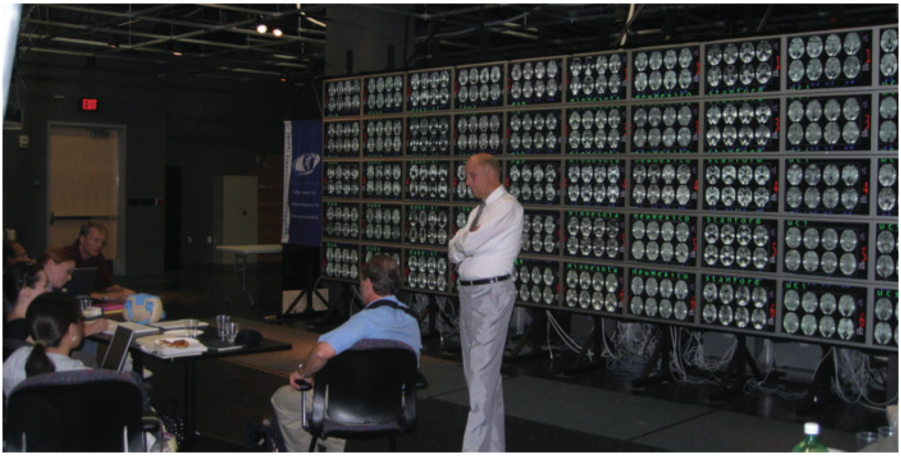


Figure 27.9 Collaborative analysis of MRI and CT data.

referred to as texture skin, inherits all traits of a regular Web page, allowing it to be used for further navigation and inquiries. That is, links remain active once the active texture skin is mapped to the workspace. Figure 27.10 shows the 2D chromosome viewer and the result of an automatically created query into gencards.org colocated with it that continuously updates as the user pans across base-pair coordinates.

Intriguingly, the same mechanism allows the user to associate any Web-based content with the collaborative workspace, allowing images, static or editable documents, and streamed media to be colocated. Figure 27.11

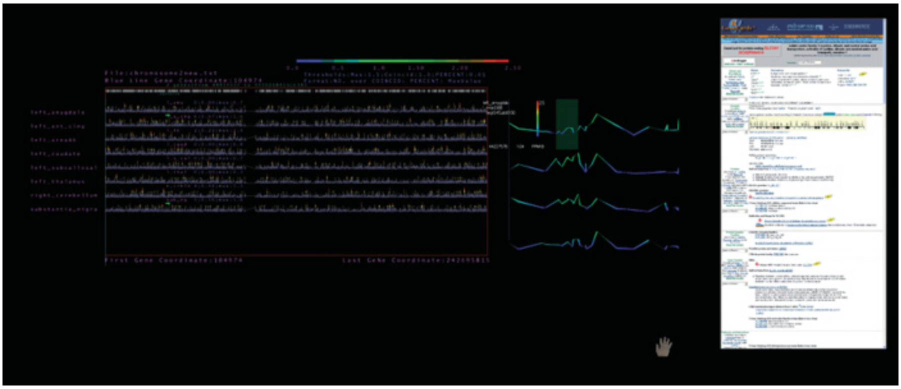


Figure 27.10 Interactive chromosome viewer colocated with associated Web query.

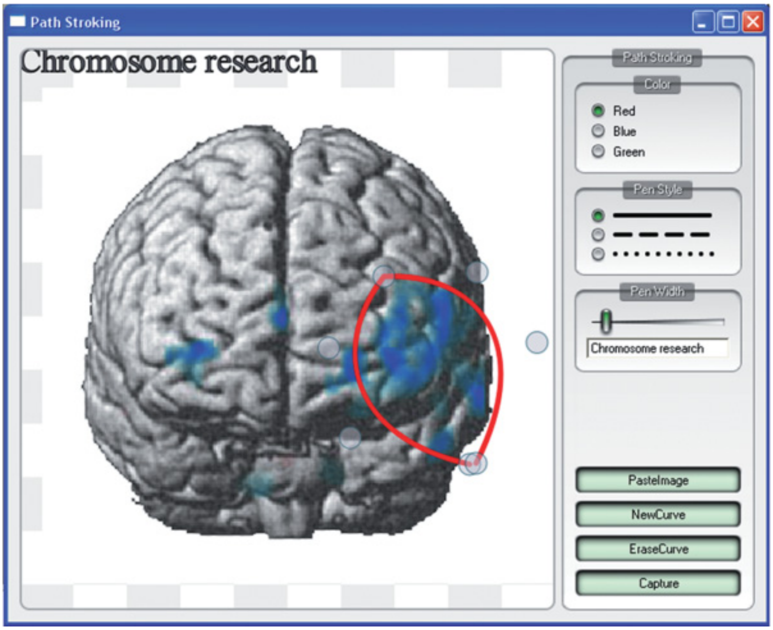


Figure 27.11 Web-centric Live-Notes imported as active texture skin.

illustrates this flexibility in the context of a Web-enabled data annotation and note-taking application.

27.4.2 Imaging Genetics Workspace

Drawing from the presented components, a unified workspace can be created, allowing domain specialists to work concurrently and collaboratively. For the imaging genetics case study, the discovery process starts at the brain imaging level and is infused with patient- and cohort-specific genetics information throughout the visual analytics session. Results from structural and functional brain imaging techniques such as CT, MRI, functional MRI (fMRI), and PET provide the starting point for initial qualitative evaluation. In this first step patient data can be visually segmented and clustered within the collaborative visualization environment with its wall-sized display canvas while serving as an input to further statistical processing. More specifically, as the user navigates and conceptually sorts through the data collection obtained via imaging, auxiliary patient data such as age, gender, and medical record can be linked and processed in the background to create a hypothesis about imaging genetics data relationships. For example, the collaborative analysis session may first conceptually identify candidate brain areas showing distinctive patterns by visually grouping the associated images. Concurrently, the genetic (SNP) data for the studied brain regions can be visually colocated such that users may apply search, threshold, or screening parameters. This approach, in particular, enabled the swift identification of important linkage relations between brain areas and particular genes. The genomic and brain pathway information is in place here to provide additional hints on the SNP. At the same time, the genetic information and brain areas can be assessed in a 3D brain atlas or 2D anatomic brain imaging, along with the metabolic pathway image. The brain atlas can then be interactively adjusted to highlight functional areas. Diffusion tensor imaging (DTI), subsequent computation of brain tractography, and interactive visualization of the resulting 3D model serve as additional aids (Fig. 27.12) to identify disease-related deterioration of pathways.

Finally, researchers verify any discoveries against pharmaceutical data and clinical trial results and discuss findings using the live notes tools. Throughout this analysis cycle, online libraries are accessible through the Web portal. Figure 27.13 provides an annotated overview of the more commonly used analytics tools that contributed to the collaborative workspace.

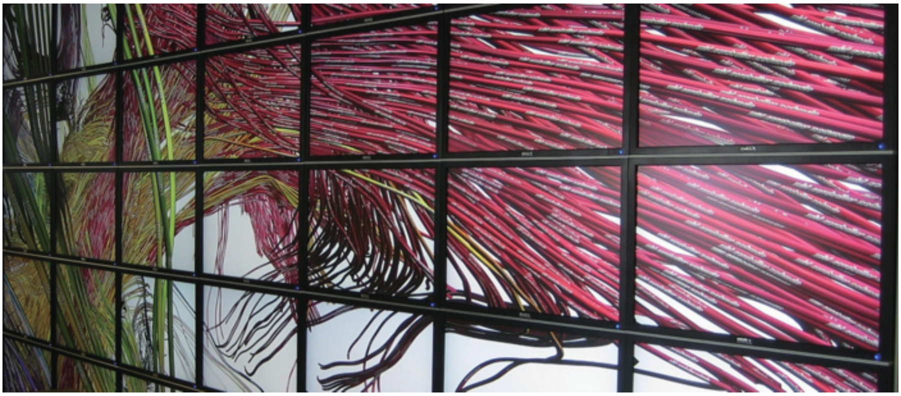


Figure 27.12 Visualization of brain tractography using annotated fiber bundles.

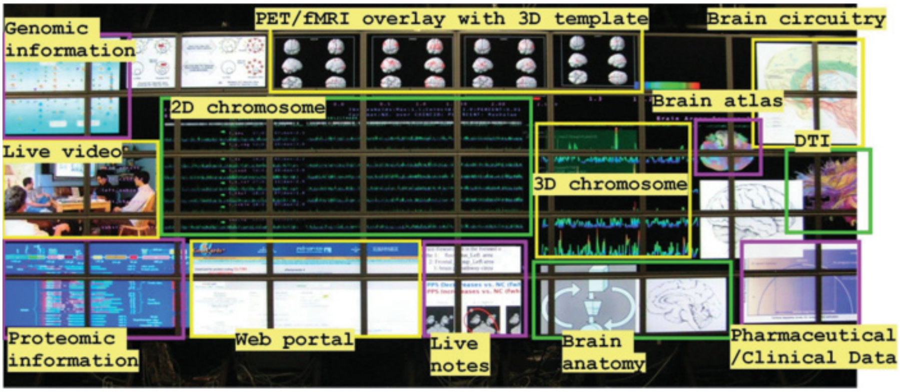


Figure 27.13 Overview and data analysis modalities synthesized within the workspace.

Another interactive discovery case involved a dichotomy of data and aimed at the detection of brain functional significance. In this case the analysis process starts with a collection of fMRI images with associated underlying characteristics such as clinical test results. Image data are then clustered based on the activation on anterior lobe versus posterior lobe, that is, the motor associated part against the sensor perceptive part. The result is then further clustered by the dorsal versus ventral half of the brain and refinement can continue until something significant is identified or termination criteria met. For the presented system, this simple dichotomy clustering is implemented as a drag-and-drop operation, which has shown to compress the process of multiple selection rounds to hours, a fraction of the time originally needed to sift through thousands of printouts. The latter is a process that can take days or weeks to complete, and researchers may be pressed to recall what they saw in a particular print-out viewed the day before, rendering the traditional discovery process inefficient. These types of analysis methodologies are exemplary for the targeted imaging genetics research.

27.5 CONCLUSIONS

A driving force that brought researchers initially together was the curiosity for an entirely different collaborative visual analytics space, which quickly was adopted as a productivity tool as part of weekly meetings. With the size of data sets growing, researchers found that they were able to sort and categorize their data faster and more accurately and in turn identify trends and subjects with distinctive characteristics. At the same time, the team-centric analysis pass allowed them to identify areas where data should be analyzed more thoroughly or additional data points would be desirable. The team identified five genes out of whole-length chromosomes, whose significance was previously overlooked and warrants further investigation.

REFERENCES

1. Varian HR, Lyman P. How much information. University of California at Berkeley, School of Information Management & Systems (SIMS), 2003.
2. Gantz JF, et al. The expanding digital universe: A forecast of worldwide information growth through 2010. IDC white paper, 2007. Available: http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf.
3. Thomas JJ, Cook KA. Illuminating the path: The research and development agenda for visual analytics. National Visualization and Analytics Center. Available: http://nvac.pnl.gov/agenda.stm, 2005.
4. Moravec H. When will computer hardware match the human brain. *J Evol Technol* 1998;1:1–14.
5. Riesenhuber M, Poggio T. Models of object recognition. *Nature Neurosci* 2000; 3:1199–1204.
6. Harnad S. The symbol grounding problem. *Phys* D *Nonlinear Phenomena* 1990; 42(1–3):335–346.
7. DeFanti TA, Leigh J, Renambot L, Jeong B, Verlo A, Long L, Brown M, Sandin DJ, Vishwanath V, Liu Q, Katz MJ, Papadopoulos P, Keefe JP, Hidley GR, Dawe GL, Kaufman I, Glogowski B, Doerr K-U, Singh R, Girado J, Schulze JP, Kuester F, Smarr L. The optiportal, a scalable visualization, storage, and computing interface device for the optiputer. *Future Gener Comput Syst* 2009;25(2):114–123.
8. Smarr L, Brown M, de Laat C. Special section: Optiplanet—The optiputer global collaboratory. *Future Gener Comput Syst* 2009;25(2):109–113.
9. Stolk B, Wielinga P. Building a 100 mpixel graphics device for the optiputer. *Future Gener Comput Syst* 2006;22(8):972–975.
10. Taesombut N, Wu X, Chien AA, Nayak A, Smith B, Kilb D, Im T, Samilo D, Kent G, Orcutt J. Collaborative data visualization for earth sciences with the optiputer. *Future Gener Comput Syst*, 2006;22(8):955–963.
11. Heer J, Agrawala M. Design considerations for collaborative visual analytics. *Inf Visualiz* 2008;7(1):49–62.
12. Pavlovic VI, Sharma R, Huang TS. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Trans Pattern Anal Machine Intelligence* 1997;19(7).
13. Sharma R, Huang TS, Pavlovi'c VI, Zhao Y, Lo Z, Chu S, Schulten K, Dalke A, Phillips J, Zeller M, Humphrey W. Speech/gesture interface to a visual computing environment for molecular biologists. *IEEE Comput Graphics Appl* 1996;20: 30–35.
14. Churchill EF, Snowdon DN, Munro AJ. Collaborative virtual environments: digital places and spaces for interaction. *Ed Technol Soc* 2002;5(4).
15. Scott SD, Grant KD, Mandryk RL. System guidelines for co-located, collaborative work on a tabletop display. In *ECSCW'03: Proceedings of the Eighth Conference*

on European Conference on Computer Supported Cooperative Work. Norwell, MA: Kluwer Academic, 2003, pp. 159–178.

16. Shen C, Vernier FD, Forlines C, Ringel M. Diamondspin: An extensible toolkit for around-the-table interaction. In *CHI '04: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York: ACM, 2004, pp. 167–174.
17. Peltonen P, Kurvinen E, Salovaara A, Jacucci G, Ilmonen T, Evans J, Oulasvirta A, Saarikko P. It's mine, don't touch!: Interactions at a large multi-touch display in a city centre. In *CHI '08: Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems*. New York: ACM, 2008, pp. 1285–1294.
18. Humphreys G, Eldridge M, Buck I, Stoll G, Everett M, Hanrahan P. Wiregl: A scalable graphics system for clusters. In *SIGGRAPH '01: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*. New York: ACM, 2001, pp. 129–140.
19. Voss G, Behr J, Reiners D, Roth M. A multi-thread safe foundation for scene graphs and its extension to clusters. In *EGPGV '02: Proceedings of the Fourth Eurographics Workshop on Parallel Graphics and Visualization, Aire-la-Ville, Switzerland, Switzerland*. Eurographics Association, 2002, pp. 33–37.
20. Jeong B, Jagodic R, Renambot L, Singh R, Johnson A, Leigh J. Scalable graphics architecture for high-resolution displays. Paper presented at the IEEE Information Visualization Workshop, Minneapolis, MN, 2005.
21. Renambot L, Jeong B, Jagodic R, Johnson A, Leigh J, Aguilera J. Collaborative visualization using high-resolution tiled displays. Paper presented at the ACM CHI Workshop on Information Visualization Interaction Techniques for Collaboration Across Multiple Displays, Montreal, Canada, 2006.
22. Jeong B, Renambot L, Jagodic R, Singh R, Aguilera J, Johnson A, Leigh J. High-performance dynamic graphics streaming for scalable adaptive graphics environment. In *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing (SC '06)*. ACM, New York, Article 108.
23. Chen H. A parallel ultra-high resolution mpeg-2 video decoder for pc cluster based tiled display system. In *Proc. Int'l Parallel and Distributed Processing Symp. (IPDPS)*. New York: IEEE Press, Abstracts and CD-ROM, 2002, pp. 15–22.
24. Chen H. Scalable and ultra-high resolution MPEG video delivery on tiled displays. PhD dissertation, Princeton University, 2003.
25. Mitchell JL, Pennebaker WB, Fogg CE, Legall DJ (Eds.). *MPEG Video Compression Standard*. London: Chapman & Hall, 1996.
26. Bierbaum A, Just C, Hartling P, Meinert K, Baker A, Cruz-Neira C. Vr juggler: A virtual platform for virtual reality application development. In *SIGGRAPH Asia '08: ACM SIGGRAPH ASIA 2008 courses*. New York: ACM, 2008, pp. 1–8.
27. Doerr K-U, Kuester F. CGLX: A scalable, high-performance visualization framework for networked display environments. *IEEE Trans Visualization Computer Graphics* 2011;17(3):320–332.
28. Lee S. A fast multiple-touch-sensitive input device. Master's thesis, University of Toronto, 1984.

29. Lee SK, Buxton W, Smith KC. A multi-touch three dimensional touch-sensitive tablet. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York: ACM, 1985, pp. 21–25.
30. Westerman W, Elias J, Hedge A. Multi-touch: A new tactile 2-d gesture interface for human-computer interaction. *Proc Human Factors Ergonom Soc* 2001;1: 632–636.
31. Dietz P, Leigh D. Diamondtouch: A multi-user touch technology. In *Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology*. New York: ACM, 2001, pp. 219–226.
32. Han JY. Low-cost multi-touch sensing through frustrated total internal reflection. In *Proceedings of the 18th Annual ACM Symposium on User Interface Software and Technology*. New York: ACM, 2005, pp. 115–118.
33. Wilson AD. Touchlight: An imaging touch screen and display for gesture-based interaction. In *Proceedings of the 6th International Conference on Multimodal Interfaces*. New York: ACM, 2004, pp. 69–76.
34. Smith JD, Graham TC, Holman D, Borchers J. Low-cost malleable surfaces with multi-touch pressure sensitivity. In *Horizontal Interactive Human-Computer Systems, 2007. TABLETOP'07. Second Annual IEEE International Workshop*, 2007, pp. 205–208.
35. Guimbretière F, Stone M, Winograd T. Fluid interaction with high-resolution wall-size displays. In *UIST '01: Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology*. New York: ACM, 2001, pp. 21–30.
36. Davis J, Chen X. Lumipoint: Multi-user laser-based interaction on large tiled displays. *Displays, IEEE* 2002;23(5):205–211.
37. Malik S, Ranjan A, Balakrishnan R. Interacting with large displays from a distance with vision-tracked multi-finger gestural input. In *UIST '05: Proceedings of the 18th Annual ACM Symposium on User Interface Software and Technology*. New York: ACM, 2005, pp. 43–52.
38. Cao X, Balakrishnan R. Visionwand: Interaction techniques for large displays using a passive wand tracked in 3d. In *SIGGRAPH '04: ACM SIGGRAPH 2004 Papers*. New York: ACM, 2004;23(3):729.
39. Ringel M, Berg H, Jin Y, Winograd T. Barehands: Implement-free interaction with a wall-mounted display. In *CHI '01: CHI '01 Extended Abstracts on Human Factors in Computing Systems*. New York: ACM, 2001, pp. 367–368.
40. Rekimoto J. A multiple device approach for supporting whiteboard-based interactions. In *CHI '98: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York: ACM Press/Addison-Wesley, 1998, pp. 344–351.
41. Kitware. Visualization toolkit : VTK. Available: http://www.vtk.org, 2010.
42. National Library of Medicine. Insight segmentation and registration toolkit (itk). National Institutes of Health. Available: http://www.itk.org, 2010.
43. 3D Slicer. National Alliance for Medical Image Computing. Available: http://www.slicer.org, 2010.
44. OmniViz, Inc. OmniViz. Available: http://www.omniviz.com, 2010.
45. Zhou Y, Liu J. AVA: Visual analysis of gene expression microarraydata. *Oxford J Bioinform* 2003;19(2):293–294.

46. Pylouster J, Senamaud-Beaufort C, Saison-Behmoaras TE. Websage: A web tool for visual analysis of differentially expressed human sage tags. *Oxford J Nucl Acids Res* 2005;33(1):693–695.
47. Boyle TJ, Bao Z, Murray J, Araya CL, Waterston RH. AceTree: A tool for visual analysis of *Caenorhabditis elegans* embryogenesis. *BMC Bioinform* 2006;7:275.
48. University of California, Santa Cruz. UCSC genome bioinformatics. Available: http://genome.ucsc.edu/cgi-bin/hgGateway, 2006.
49. Barsoum E, Kuester F. WebVR: An interactive web browser for virtual environments. *Proc* SPIE *Stereoscopic Displays and Virtual Reality Syst* XII *Int Soc Opt Eng* 2005;5664:540–547.

28

CURRENT AND FUTURE CHALLENGES FOR COLLABORATIVE COMPUTATIONAL TECHNOLOGIES FOR THE LIFE SCIENCES

Antony J. Williams, Renée J. G. Arnold, Cameron Neylon, Robin W. Spencer, Stephan Schürer, and Sean Ekins

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.

28.1 INTRODUCTION

Drug discovery and development is largely considered a linear process progressing through many steps such as from target discovery through registration and post approval, at which point the drug is on the market (Fig. 28.1). It is obviously not as straightforward as such simple graphical representations suggest. Through each of the steps there may be feedback subloops, blurring of the boundaries, and much finer levels of detail could be added. Needless to say it is a long and very expensive process and is clearly unsustainable for many, other than the very big pharmaceutical companies. It is widely recognized that the way drug discovery and development are carried out has to change. We have seen in recent years pharmaceutical companies significantly reduce their research and development (R&D) footprint to the point where most chemistry is performed relatively cheaply in Asia while clinical development is increasingly outsourced to contract research organizations (CROs). The big companies are splintering, and those scientists no longer employed by these multinational companies will reform new companies (micropharma) or loosely linked collaborative groups whether as consultants or virtual CROs. The now smaller "big" pharma can achieve their goals through leverage of a growing number of external relationships whether collaborative, precompetivive, partnerships, and so on.

This book has brought together the collective observations of a number of specialists who are engaged in supporting the paradigm change that is occurring as biomedical research rapidly moves toward a collaborative network of chemists and biologists. In the process, it will make both data and knowledge available to the masses, thereby enabling rapid sharing of information [1–4]. This new paradigm will present many opportunities for collaborative software and data-sharing tools to be further developed and is likely to result in new technologies to overhaul the drug discovery R&D process (Fig. 28.1). But there are many questions still to answer, such as how people need to be trained to collaborate and whether collaborations will truly replace the "great man theory" of science in which major discoveries are often attributed to one or more figurehead men or women rather than teams of scientists. Also many companies have iron firewalls which prevent linkage to common collaborative tools like GoogleDocs and therefore directly impede potential for collaboration. Such issues will need discussion but may be outside the scope of this chapter and book.

28.1.1 Gap Analysis for Drug Discovery and Development

One way to look for the opportunities for collaborative approaches is to understand the process and perform a gap analysis. The well-known drug discovery and development process is a good example onto which we have mapped those areas in the process that may be addressed with collaborative software and mobile computing efforts (Fig. 28.1 and Section 28.7). Mobile

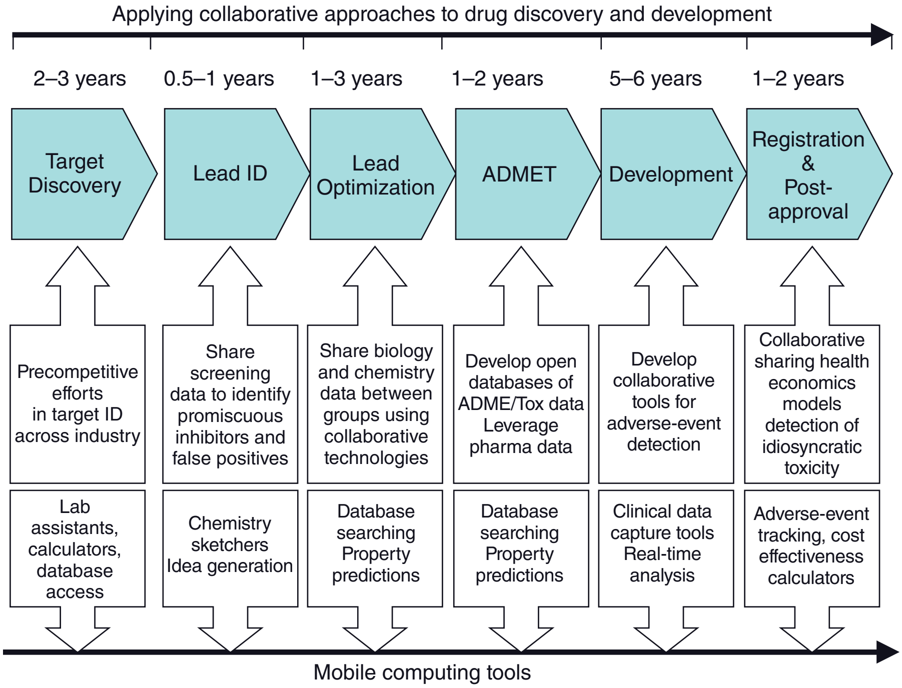

Applying collaborative approaches to drug discovery and development
2–3 years
0.5–1 years
1–3 years
1–2 years
5–6 years
1–2 years
Target Discovery
Lead ID
Lead Optimization
ADMET
Development
Registration & Post-approval
Precompetitive efforts in target ID across industry
Share screening data to identify promiscuous inhibitors and false positives
Share biology and chemistry data between groups using collaborative technologies
Develop open databases of ADME/Tox data Leverage pharma data
Develop collaborative tools for adverse-event detection
Collaborative sharing health economics models detection of idiosyncratic toxicity
Lab assistants, calculators, database access
Chemistry sketchers Idea generation
Database searching Property predictions
Database searching Property predictions
Clinical data capture tools Real-time analysis
Adverse-event tracking, cost effectiveness calculators
Mobile computing tools

Figure 28.1 Applying collaborative approaches and mobile computing to drug discovery and development. The schematic shows the linear process of drug discovery and development alongside areas where we think collaboration could be useful. We have also indicated where mobile computing tools could be implemented.

computing is certainly the “wave of the future” but in reality is arriving so fast that by the time this volume is printed it is likely to be established and in place in many organizations that will be embracing the newfound capabilities and advantages of tablets, slates, and Hypertext Markup Language (HTML 5).

A major limitation of drug discovery for those outside major pharmaceutical companies is the availability of biological information related to chemical structures. This is already starting to change via precompetitive collaborations between biomedical organizations (both industrial and academic) which may cover areas such as cheminformatics, toxicology, preclinical toxicology, and beyond. We have previously argued that absorption, distribution, metabolism, excretion, and toxicity (ADME/Tox) data are also precompetitive data and should be made freely available on the Web for all scientists [5]. Others such as the nonprofit and associated community SAGE Bionetworks (http://www.sagebase.org/) aim to make the whole of the biology of drug discovery a precompetitive space and they have initially focused on the systems biology of cancer. As public hosts of data continue to expand their content, for example, PubChem, ChEMBL, and ChemSpider, and as data-mining tools expand in their capabilities and performance, the integration of chemistry and biology databases is likely to offer even greater opportunities to benefit the process of drug discovery. Efforts to expand the existing structure-centric communities for biomedical researchers with key information relevant to drug discovery which is precompetitive will bring benefit in terms of access and discoverability of data. It will be very important to distinguish which precompetitive data can be of most value so that users are not swamped with data overload. We also need discovery tools to filter the data as an obvious consequence of making more data available is that it creates a potential filtering problem. New discovery mechanisms and tools will be needed to both identify the right data and critique its quality and relevance to a specific problem. While a natural response is to attempt to reduce or filter the data that get published, the long-term future must lie in applying the lessons from the wider Web in building effective search and discovery tools. The transition, however, is likely to be difficult and manual, and semi-automated data curation will play a big role in easing that transition. A major limitation of approaches to capturing the public information is that most data will be in publications and, until the publishers make these data semantically accessible, it will not be easily mined other than by manual extraction. While there have been, and continue to be, many efforts to improve the underlying mechanisms of scientific publishing to make data extraction easier, this is likely to be a slow process and large quantities of information will remain in the legacy literature. There is however already a considerable amount of data for drugs on the market that could be extracted from various online databases and that could be valuable for developing computational models, for example. Text-mining tools have already been developed that can be partially successful in aggregating these data, but it would be preferable if instead of harvesting these data out of publications and patents drug companies, researchers, and health authorities could supply the data in a homogeneous standardized format and in a coordinated fashion. International funding agencies are presently tendering for the development of systems that could facilitate these kinds of data-sharing opportunities as pharmaceutical companies acknowledge that the cost burden that they need to assume to aggregate these data is too high and, since it is precompetitive in nature, collaborative efforts across the life sciences should facilitate data access.

While much of the biological data used in drug discovery can be used to generate computational models in each company, this is also true for other data generated at different stages of drug discovery and development. Computational models reported in publications and in the public domain are hardly accessible in terms of testing the models against internal data sets. Similarly, models are rarely shared between companies or even between researchers, and there has been little research or efforts invested to facilitate this [6]. The following sections represent those areas we think are challenges that would likely also benefit from collaborative computational approaches.

28.2 COLLABORATIONS IN HEALTH ECONOMICS MODELING

The observations described above are also true for the health economics community where models built to analyze comparative effectiveness or cost-effectiveness are maintained in the developers' silos, rather than being made available to all potential stakeholders—policy makers, investigators, developers, industry sponsors, academicians, health authorities, and others charged with making decisions based on these models. Can we overcome the proprietary and technological challenges that might reduce the feasibility and desirability to use the "cloud" and other advances to enhance our future opportunities for collaboration?

Health economics and comparative-effectiveness questions are being increasingly answered using computational models in the hands of the stakeholders who have to make decisions using them [7]. With limited health care dollars, exploding health care costs, and confusion about which strategies result in the best patient outcomes, computerized models can help to objectify the complexities of comparative effectiveness and cost-effectiveness of different therapeutic options to aid in decision making by pharmaceutical/device manufacturers, health authorities, and health care practitioners regarding therapeutic guidelines, reimbursement/coverage, and overall patient health. Indeed, computational models are used to answer many questions such as determining therapy/market advancement/characterization [8, 9] through postmarketing surveillance [10, 11] and budget impact and policy decisions. Countless examples in other areas of health care demonstrate the enormous importance of modeling studies. However, there is an inherent complexity in modeling health care decisions and the relative isolation under which the work of modeling is often carried out; this likely needs to be overcome [12–15]. Many of these published models are developed using readily accessible commercial software; however, there is no way to easily share such computer-based models across organizations and make them available for reuse in the public domain to interested parties. Various researchers [16–18] and organizations have called for transparency and availability of models, "reused" with different data, and continually revised as new information becomes available. It has been proposed that what is needed is the creation of a collaborative Web-based tool [tentatively called Economic Collaborative for Health Outcomes (ECHO)] that would serve as a means to store and share health care models, allowing for the proprietary nature of some of these [19]. This could enable disparate modelers to create higher quality products by being more collaborative and sharing models and techniques. It would also provide a dynamic resource for interacting online with some of the thousands of published models so they would remain in use for a longer period [19] and shifting incrementally as more data are made available.

As stated above, potential challenges may include those of price, confidentiality, quality assurance, and overcoming the silo mentality of modelers. Price will have to be tailored to each stakeholder group and incentives offered to upload models. Also, simplified versions of models can be made available at a greatly reduced price and free for students/professors. To address issues of confidentiality, models can be made available via username and password, incorporating secure sockets layer (SSL) encryption mechanisms for users/modelers with this concern. To assure quality, models can be vetted by an expert advisory panel or expert users.

28.3 COLLABORATIVE ADVERSE-EVENT DETECTION AND DRUG SAFETY DATABASES

The institutes of the National Institutes of Health (NIH) have sponsored or executed innumerable clinical trial programs over the years and continue to do so. Nevertheless, one of the most important aspects of trial management is the detection, reporting, and analysis of adverse events (AEs), which have, unfortunately, not been adequately standardized with regard to definitions, reporting forms, processes, and treatment of the data. AE reporting is somewhat more standardized in these regards for the postmarketing environment. The U.S. Food and Drug Administration (FDA) has formal regulations and systems for dealing with serious medical reports in a somewhat more consistent fashion (e.g., manufacturers send AE reports to the FDA either on standardized paper forms or, more recently and increasingly, through electronic reporting under defined specifications; patients or health care professionals can submit reports through the MedWatch program) [20].

The postmarketing environment in the United States also includes the FDA Adverse Events Reporting System (AERS) database as a repository for all such reports. Still, the problem of timeliness of identification, risk assessment, reporting, and dissemination of information about AEs persists within postmarketing surveillance as well. Indeed, Sentinel Network is intended to be an integrated, electronic nationwide medical product safety network which is supposed to combine the efforts of both the public- and private-sector postmarketing safety surveillance tools and methodologies into one cohesive system. However, the FDA's progress with development of Sentinel Network has been slower than anticipated.

The lack of adequate standardization within the NIH was highlighted on the NIH's Clinical Research Policy Analysis & Coordination website [21]: "Tremendous diversity exists among AE reporting requirements promulgated by various federal agencies, as well as among the NIH Institutes. This heterogeneity is a challenge for investigators, institutional review boards, and sponsors, who may face multiple requirements regarding the content, format, and timing of reports that must be made to different agencies and oversight bodies." The lack of a systematic and consistent standard with regard to AEs within the NIH could have important implications: less than optimum protection of trial subjects, inability to merge and understand data across institutes and even across trials within the same institute, poor efficiency and increased costs due to disparate and outmoded methodologies, and, as a result, lack of the necessary speed with which certain medical situations should be addressed. A practical example of potential problems can be seen from several oncology studies in which inconsistencies were found between AE reporting/publication and the raw data from the clinical trial databases themselves [22, 23].

There appear to be many independent initiatives and already established systems throughout the federal government addressing aspects of AE reporting during clinical trials. Among them are new clinical trial guidelines on reviewing and reporting unanticipated problems and AEs that occur in clinical trials conducted or supported by the federal government [24]. Indeed, the Federal Adverse Event Task Force (FAET), an interagency body composed of representatives of the NIH, FDA, Office of Human Research Protections (OHRP), Centers for Disease Control and Prevention, Department of Veterans Affairs, Department of Defense, and Agency for Healthcare Research and Quality, has been established.

Moreover, to address the diversity of requirements solely within the NIH, a Trans-NIH Adverse Event Task Force has been established and charged with "proposing ways to harmonize the reporting policies of the agency's many Institutes and Centers" (http://oba.od.nih.gov/policy/policy_issues.html). In addition to the Department of Health and Human Services' (DHHS) 45 CFR part 46 and the OHRP's draft guidance [25, 26], the National Cancer Institute (NCI) Cancer Therapy Evaluation Program has established the Common Toxicity Criteria version 2.0 (CTC) and the Common Terminology Criteria for Adverse Events version 3.0 (CTCAE) in attempting to delineate adverse-event (AE) criteria. NCI CTC version 2.0 was the worldwide standard dictionary for reporting acute AEs in cancer clinical trials until August 9, 2006, when the CTCAE version 3.0 was published. The CTCAE version 3.0, a Web-based listing of AEs, includes AEs applicable to all oncology clinical trials regardless of chronicity or modality. Building on the oncology CTC, the Outcomes Measurement in Rheumatology (OMERACT) Drug Safety Working Group has focused on standardization of assessment and reporting of AEs in clinical trials and longitudinal and observational studies in rheumatology—the Rheumatology Common Toxicity Criteria (RCTC) [16, 27, 28].

Despite all of these initiatives there is still no uniform, practical working model that suits the needs of all parties. For example, the National Institute on Drug Abuse (NIDA) has a Web-based system for serious adverse events (SAEs), the Serious Adverse Event Tracking and Reporting System (SAETRS), which is currently being used by about 30% of principal investigators (PIs). The PIs are concerned about the security of the current SAETRS and the fact that it appears easier to send an SAE report manually (by fax) rather than logging into the system infrequently to report an SAE. The website AE system is "difficult to trust to collect confidential information/data ('afraid of losing data')." It is possible that if such a system were to be used for the more frequent occurrence of all AEs, rather than just for the rarer SAEs, PIs and other personnel may become more familiar with it and use it more frequently. In addition, if the system were to be more evidence based, using previous information to guide researchers in not just reporting AEs but also deploying clinical studies with inclusion and exclusion criteria that would be inherently more protective of vulnerable subjects, the system may be even more useful [20]. Thus, even within centers within the NIH, the silo mentality exists.

An additional area beyond formal reporting that has barely been explored is the potential of patient advocacy websites and communities to be valuable sources of information on AEs. Patients are likely to represent the most willing and richest source of information on adverse drug reactions, drug interactions, and environmental effects. Increasingly these communities are driving, funding, and even directing the drug discovery process and the development of treatment regimens. While there are obvious potential issues with reporting consistency and precision, this represents a vast untapped reserve of information on drug performance. In our modern connected age, failing to connect with the end user and to take on board their feedback has badly damaged many organizations. There is no reason to expect drug discovery and validation processes and organizations to be any different.

It is our opinion that AE systems could be developed that are more collaborative in nature such that they are in line with the other collaborative tools described above. It is also possible that such efforts could also be combined with the creation of drug safety databases and used as a crowdsourcing initiative, whereby anyone can contribute observations and data for a particular approved drug. Obviously there would need to be filters implemented for spurious data. Such systems should be harmonized throughout the drug approval process such that the structure–activity relationship (SAR) of the compound indicates the likelihood of an agent exhibiting a particular AE, the preapproval clinical trials track it, and the postmarketing surveillance quantifies it. This may be particularly important to capture information on molecules associated with idiosyncratic toxicity that may not be observed in relatively small clinical trials [29–31]. For example, the liver is a frequent site of toxicity of pharmaceuticals in humans, [32, 33] likely because of the physiological location and drug clearance function of the liver leading to higher exposure to drug than that being measured systemically [34]. Drug metabolism in the liver can also convert some drugs into highly reactive intermediates which, in turn, can adversely affect the structure and functions of the liver [35–38]. Drug-induced liver injury (DILI) is a major reason why drugs are not approved and why some of them were withdrawn from the market after approval [39]. Postmarketing data may help find additional molecules with this issue and help in alerting authorities to new drugs displaying such toxicities. The real issue here is providing a mechanism for patients and health practitioners to alert the FDA and other health authorities in a timely manner.

28.4 ONTOLOGIES AND COLLABORATIONS

The biomedical research community and specifically those involved in neglected disease research are generating very large data sets facilitated through high-throughput screening (HTS) [40–42]. Although large HTS data sets and low-throughput screening results have become available in the public domain (PubChem [40, 41], ChEMBL [42], Psychoactive Drug Screening Program (PDSP) [43, 44], ChemBank [45, 46], Collaborative Drug Discovery (CDD) [3, 47], and others), these data sets are not well standardized, experimental metadata are poorly annotated, and much of the relevant information is often only available as free text (particularly in PubChem). This presents impending informatics challenges for selection of hit compounds and follow-up studies as well as computational analysis of such data or the development of predictive models. The lack of established and formal standards to annotate the publicly available screening data also limits their integration with other structured data sources such as biological pathways, human disease, or adverse drug effects. A related challenge is knowledge and data representation.

One way that such data have continued utility and accessibility is through an ontology [43, 44] (see also Chapters 12 and 21). An ontology is a formal explicit description of a subject domain (a conceptualization) as classes, individuals, and their relationship and properties to represent static knowledge [48]. Ontologies are one of the cornerstones of Semantic Web technologies, which have been proposed as solutions to data integration problems because formally defined semantics and semantic knowledge representation make it possible to track data provenance across different data sources that typically use different descriptions and naming conventions [49]. The lack of a standardized terminology with clear definitions can even be a severe issue within an individual data source, for example, in the case of PubChem, where data from numerous organizations and various experiments are deposited and which typically vary by the details that are reported for any data set (screening experiment), the way the information is reported (how the data and the experimental details are organized), and the type of results that are reported. This is despite existing recommendations regarding the types of information that should be reported for HTS assays [50]. For example, there are thousands of unique endpoint names deposited in PubChem (as of December 2009 there were over 12,000), many of which are redundant. Although there are two endpoints which are required in all deposited assays (except summary assays), activity outcome and activity score, there is no agreed-upon definition of "active" or "inactive" or how the score is to be calculated. Instead, for each assay submission the depositor can define a "local" meaning of activity outcome and activity score. The lack of established standards and a semantic framework to describe the assay experiment and the reported endpoints poses severe limitations to computational analyses across multiple data sets and their integration with other data sources. However, as far as PubChem is concerned, this is a consequence of the openness and “flexibility” allowing data to be deposited from essentially any source and in a wide variety of formats. In that sense PubChem has more of the characteristics of an open data repository and less that of a data warehouse with defined terminology metadata. This was a conscious choice when PubChem was set up, because the fast pace of innovation in assay designs to interrogate complex biological processes using novel detection technologies limits the effectiveness of “static” relational database systems to capture and manage the diversity of screening experiments and their outcomes. To effectively address these limitations, a semantic framework with a bioassay ontology at its core is required.

During the last several years many biomedical “ontologies” have been developed with the goal of describing and integrating complex biological knowledge with existing databases [51] and advancing translational research [52]. Many biomedical ontologies are available in the Open Biological and Biomedical Ontologies (OBO) Foundry [53] with the most prevalent being Gene Ontology (GO) [54, 55]. However, only a few of the other OBO ontologies are widely used, and there has been criticism about the lack of international standards in many bio-ontologies from the Semantic Web community [56].

More recently the semantic integration and annotation of small-molecule data with existing biological databases have been reported [57,58]. However—until now—there is no comprehensive effort to develop an ontology to describe the increasing body of HTS experiments and the data these experiments produce. In particular, we are not aware of a standardized assay ontology that is accessible to the neglected disease community. Collaborative software for chemistry and biology data could have a direct impact on promoting the adoption of such an ontology for neglected disease bioassay data. Adoption of an open-assay ontology will be a major milestone in converting volumes of assay *data* into machine-interpretable *knowledge* and finally human *insight.*

With PubChem and several other accessible screening data repositories there is now a sizable publicly accessible corpus of screening experiments and their results. This makes it possible for the first time to develop a knowledge representation of HTS assays and screening results in an open effort. Because this corpus and its diversity are growing exponentially, the development of a clearly structured and standardized formal description of the concepts that are relevant to interpret HTS results is also very timely. To be successful in the long term, such an assay ontology needs to be maintained and kept up to date (much like GO), and there is also a need for ongoing bio-curation to systematically annotate the data sets.

28.4.1 BioAssay Ontology

BioAssay Ontology (BAO) [59] is an NIH-funded project to facilitate analysis of screening results from large numbers of diverse biological screens spanning various technologies (and originating from different sources). The BAO project develops standard terminologies and a formal, extensible, knowledge-based description of biological assays. Bio-curation is also an integral component of the BAO project to systematically annotate PubChem bioassays with standardized terminology describing assay concepts. The BAO project also develops software tools to query and explore a data set in the context of the ontology.

BAO follows an established ontology development methodology using a combination of a top-down domain expert-driven and bottom-up data-driven approach [60]. The scope of the ontology, knowledge acquisition, software requirements, and specifications were driven by use-case scenarios presented to domain experts (workshop given at the Society for Biomolecular Sciences, Phoenix 2010 [60]) and by derived competency questions. The ontology will unify knowledge-related HTS and other types of screening, including the concepts described below. One initial goal of the BAO ontology and software is to enable researchers to query the bioassay repositories and to retrieve relevant data. For example, in the case of PubChem, even seemingly trivial queries such as biochemical versus cell-based assays or luciferase reporter gene assays are currently not possible, because the assays do not have explicit annotations that capture this type of information. Other relevant queries can include the identification of nontoxic kinase inhibitors, promiscuous luciferase reporter gene compounds, or compounds that may interfere with fluorescence intensity assays. To enable researchers to retrieve quantitative information that is meaningful across multiple assays of interest also requires the standardization of the endpoints that are reported as the primary (most important) outcomes of screening experiments, for example, percent inhibition, IC_{50}, K_i, and so on. BAO defines the meaning of common endpoints and can relate different types of endpoints to other concepts in the ontology. A later goal of BAO is to semantically integrate screening data with other publicly available biological databases, such as biological pathways, human diseases, known toxicities, adverse drug reactions, and possibly also predictive models.

BAO will make use of existing ontologies where appropriate, for example, cell lines [61] or Gene Ontology [54, 55]. It will facilitate integration with other databases such as biological pathways via BioPAX [62, 63] and it will support inference. The software development component of the BAO project makes use of Semantic Web technologies, such as Jena [64] and Vivo [65]. The ontology is also being implemented using the ontology management application framework Protégé 4.1 [66] to support the design of the structure of the assay ontology. It is being developed using Web Ontology Language (OWL) 2.0 [48, 67], which is currently the most expressive description logic (DL) language. This is in contrast to most of the OBO ontologies.

BAO includes several high-level concepts related to biological screening, including perturbagen, metatarget, format, technology, analysis, and endpoint. Perturbagens deposited in PubChem and the other screening data sources mentioned above are mostly small molecules, but perturbagens can include various other perturbing agents that are screened in an assay. The metatarget component describes the protein target, pathway, biological process or event, and so on, targeted by the assay. Format includes biochemical, cell based, organism based, and variations thereof. Technology describes the assay methodology, assay design, and implementation (including detection method) of how the perturbation of the biological system is translated into a detectable signal. Analysis describes how the raw signals are transformed into reported endpoints. Endpoints are the final HTS results as they are usually published (such as IC_{50}, percent inhibition, etc.). BAO also captures other assay properties and relationships, such as assay purpose or how assays are related in campaigns. BAO is designed to handle multiplexed assays. All main BAO components include multiple levels of subclasses and specification classes, which are linked via object property relationships forming a knowledge representation.

All these aspects of the BAO project described above will facilitate collaborations among scientists of various disciplines, including screening biologists, chemical biologists, medicinal chemists, cheminformaticians, and modelers. The bioassay ontology will enable scientists to readily compare screening results and to evaluate screening outcomes in the context of the existing large public data sets—for example to distinguish artifactual hits from desired ones. It will make it much easier to retrieve quantitative outcomes and activity profiles relevant for medicinal chemists. BAO will also enable modelers to generate consistent quantitative data sets that are related to a distinct biological process or mechanism of action, for example, a protein target, a pathway, an assay technology, and so on. Because ontologies define semantics using text expression, BAO can also make domain-specific information accessible to nonexperts, for example, chemical structural information to cell biologists or assay technologies to cheminformaticians. With an ontology of sufficient detail and appropriate software tools, one could imagine being able to express complex queries covering several concepts by simple text, for example, "promiscuous inhibitors of luciferase reporter gene assays."

This BAO will also facilitate collaboration between researchers in a public/private data-hosting environment by enabling automated systems to alert researchers of potential collaboration opportunities (see Chapters 12 and 21). For example, the owner of a private assay instance might opt to have an introduction made to researchers that meet certain criteria based on assay data (as opposed to structure or other alerts). This can be done in an automated fashion, without either party viewing confidential information directly. An assay ontology makes this possible because it allows considerable flexibility in defining the terms of the alert and can make it very easy to define such an alert (i.e., by textual expressions). The private assay instance owner might then opt to suggest collaboration only with other researchers who meet certain affiliation requirements and have agreed up-front to a standard confidentiality agreement (e.g., available from Science Commons or elsewhere). Once both parties acknowledge the collaboration, the data set (and means of direct communication) could be shared within whatever database software is used. This will enable a mechanism of suggestive collaboration that does not currently exist for biomedical researchers and requires assay definitions provided by the ontology.

28.5 WILL WIKIS AND ONLINE COLLABORATION CHANGE THE WORLD?

The word “wiki” has become a general-purpose term sometimes used to cover a very wide range of online collaborative authoring environments from Wikipedia, to blogs, and to community forums and news sites (see Chapter 5). Many of these differ widely in their interfaces, communities, and approaches to editing and publishing information. Wikipedia is clearly changing the world, especially when it comes to engaging the masses to collaborate, contribute, and develop a new form of encyclopedia. The Media Wiki platform (http://en.wikipedia.org/wiki/Media_wiki) is free and open-source software and has been downloaded and deployed many times to set up corporate wikis, for example, Pfizerpedia (http://pubs.acs.org/email/cen/html/090207084512.html). Wikis are showing up everywhere. WikiGenes (http://www.wikigenes.org/), WikiProteins (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2441475/), WikiPathways (http://www.wikipathways.org/index.php/WikiPathways), and the WikiProfessional Concept Web (http://www.wikiprofessional.org/about.php), to name a few, are already contributing to knowledge sharing, management, and development of the Semantic Web for the life sciences.

Many blogs cover areas of the drug development process from biology through chemistry and compound analysis and validation. While these rarely deal with the day-to-day details of an ongoing drug discovery and development process, they do play an important role in education, spreading best practice and identifying poor and unethical practices. Blogs and their associated commenting communities are becoming a strong component of the self-regulation and analysis of the drug discovery and development community. At the same time blogs and similar websites and community forums are also providing a platform for those critical of the processes, organizations, and individuals involved in the advance of modern medicine. In some areas these communities and their websites overlap. This form of criticism and analysis, whether constructive or not, is likely to continue and the corporate, academic, and individual response to it will be an increasingly important area for community engagement. The necessity for this engagement is another way in which the social Web is changing the world of drug development.

Between the word processor document on a shared disk and the fully open and editable wiki, or a blog accepting comments, there is a wide range of collaborative authoring tools and publication mechanisms. GoogleDocs, EtherPad, Wave, and Microsoft Office Live Workspace as well as wikis and other Web-based content management systems offer shared spaces where documents can be prepared and edited with one other person, defined communities and groups, or the whole world. In most of these environments switching between completely private and public is a matter of pressing a button, making it possible to author in private but easily publish to the Web. Authoring and securing criticism prior to "publication" are both much easier and potentially much more effective. However the potential of these tools remains unexploited while most authors prefer to e-mail around Word documents.

In the future wikis and other social software are likely to continue their growth and prominence as knowledge-sharing environments with the software platforms continuing to expand in functionality based on the needs of the user organizations. Google SideWiki (http://www.google.com/sidewiki/intl/en/index.html) already allows anyone to leave comments about pages as they surf the Web, thereby further enabling the community to participate in wiki'ing the Web. The challenges for wikis will be whether they can be seen as nearly equivalent to traditional peer-reviewed publishing to gain further acceptance from the scientific community. Until their credibility increases, individuals will be less motivated to participate compared with other modes of communication. Will wikis change the world? They already are and could be exploited further to impact biomedical research.

28.6 COLLABORATIVE SYSTEMS BIOLOGY

In a similar way to wikis (described above), we presently take for granted a systems-level understanding of the linked networks of multiple interacting genes, gene products, and metabolic processes that determine phenotype. Systems biology is considered an interdisciplinary methodology incorporating collaborations between experimental biology, physiology, physics, engineering, mathematics, and computer science. Systems biology emphasizes combining high-quality, quantitative data from multiple levels with computational modeling to develop mechanism-based models of how networks of individual genes and proteins interact. For many years the functional organization of a biological system was described in terms of pathways which were relatively small linear chains of biochemical reactions or signaling interactions. In recent years, as biology has used high-throughput methods for determining protein–protein interactions, it has also required the development of pathway databases and natural language processing algorithms for automatic extraction of pathway information. Now we realize that biology is enormously complex and molecular processes can be linked to very large, highly interconnected networks [68]. We have seen the availability of software for visualizing complex gene networks become very commonplace, and this, in many ways, has been the underpinning of research on systems biology [41–49]. For example, combining comprehensive databases, powerful analytical and network building tools have resulted in the development of commercially available integrated high-throughput data-mining suites like Pathway Assist™ (Ingenuity), PathArt™ (Jubilant Biosys), Pathways Studio™ (Ariadne), MetaCore™, MetaDrug™

(GeneGo), and others [69]. Noncommercial tools, including PATIKAweb (http://www.cs.bilkent.edu.tr/~patikaweb/), are also available [70]. These tools can enable visualization of global cellular mechanisms and use curated content on human physical protein–protein interactions, allowing different levels of cellular functionality captured as maps of current biological knowledge or custom-built interaction networks. There are even significant open approaches to building pathway databases (http://www.wikipathways.org/index.php/WikiPathways) [71, 72] and systems biology platforms [73] which perhaps have persuaded the commercial companies to make some of their pathways freely available (http://www.genego.com/mapbrowse.php).

28.6.1 Facilitating Collaborations in Systems Biology Community

There need to be active methods to facilitate collaborations between scientists and systems biologists if science is to advance. Currently these types of interactions predominantly occur in established institutes. We posit the question about what could happen if they could be more open. For example, Web-based social networking technologies that could enable any scientist to find and collaborate with another scientist that could apply systems biology to the project in question, or simply for a biologist to share data with systems biologists (so that computational models could then be sent back and tested by the researcher), would potentially advance the science in more than an additive fashion.

There have been some efforts to develop software that can be used to build biomedical Web communities using a semantically aware content management system. One example is the science collaboration framework that has been used to create open-access online communities such as StemBook (http://www.stembook.org/) and PD online (http://www.pdonlineresearch.org/), which deal with stem cell research and Parkinson's disease, respectively [74]. These communities provide clear attribution for content as well as editorial review. It would not be too much of a step if these were extended to connect scientists and their data in addition to their publications. What if scientists with a systems biology background were a part of these communities or such communities were interconnected?

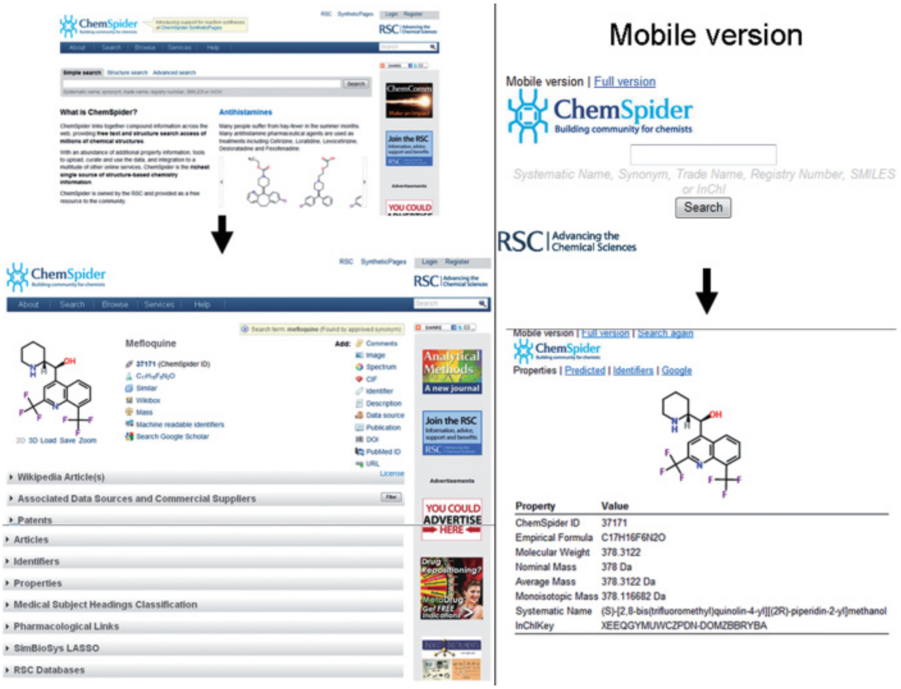
28.7 MOBILE COMPUTING AND ITS IMPACT ON COLLABORATIONS

In the last decade we have seen not only the fusion of mobile telephones with music players and Web browsing but also the introduction of less expensive, smaller devices like netbooks, tablet computers, and smart phones that are likely to become increasingly prevalent and more powerful as microprocessors and memory modules become even cheaper. There has been parallel development of programming languages and standards leading to software applications for mobile devices which in the past would only have been possible on large computers. This has resulted in a growing number of medical applications of smart phones that assist physicians in accessing test results to evaluate, for example, blood pressure records, electrocardiograms, blood glucose, and pulmonary function or to monitor a fetal heart rate. Obstetricians can even remotely monitor the heart rate of the fetus as well as that of the mother. Such real-time monitoring enables tracking and intervention in cases of chronic health conditions, improving patient outcomes. The simple use of camera phones for imaging could also have a transforming effect of bringing medical care to remote areas [75].

In biomedical research, chemists in particular may move from accessing data in their electronic notebooks at the desktop computer to their tablet PCs or their smart phones/devices. These mobile devices offer a window into how scientists will operate in the future—such devices will further enhance the provision of collaborative software to biomedical scientists, an existing limitation being how much information one can show on current screen real estate.

Already there are numerous applications that run on smart phones and tablet computers like the iPad. There has been a focus on chemistry tools (both applications and mobile, browser-based access), many of which can be downloaded at no charge or purchased very cheaply. Applications bring kudos and marketing value with dedicated platform functionality and allow offline usage. Examples include laboratory assistants (stoichiometry calculators, equation balancing, elemental formulas to mass), educational tools (periodic tables, flip cards, questions and answers, study aids), structure drawing and viewing tools for both small molecules and biomolecules, and look-up tools (e.g., chemical reactions, Wikipedia, chemistry database searching) [76]. Scientific publishers also have made the move to mobile computing as another outlet for their content, as the American Chemical Society already provides mobile feeds from magazines and journals and nature.com presents highlights from the journal. Structure drawing also serves as a starting point for calculations (formula, mass, physicochemical properties) and database look-up (Internet for latest content, on device—currently space is not a problem). Symyx ChemMobi and Mobile ChemSpider are both examples of providing access to large online databases, while the latter presents a very simple interface and limits the amount of data returned compared with the full Internet version, thus maximizing the visibility of structures and data on the small screen (Figs. 28.2 and 28.3).

In the short term it is likely that a handful of key science applications will dominate, although it is unclear at present what these will be. From the collaboration side so far there have been no real technologies developed of which we are aware. The cloud-based management of publications (e.g., Papers and Mendeley) is probably the most obvious benefit of mobile computing today, but there is a key gap here in the provision of tools for sharing data or finding collaborators via mobile computing. In the midterm the popularity of tablet computers will likely increase as a result of the in-vogue nature of the iPad, and the upcoming availability of multiple other tablet devices, and this will



ChemSpider
Building community for chemists

RSC SyntheticPages Login Register

RSC | Advancing the Chemical Sciences

About | Search | Browse | Services | Help

Simple search | Structure search | Advanced search

Search

Systematic name, synonym, trade name, registry number, SMILES or InChI

What is ChemSpider?

ChemSpider links together compound information across the web, providing **free text and structure search access of millions of chemical structures.**

With an abundance of additional property information, tools to upload, curate and use the data, and integration to a multitude of other online services, ChemSpider is the **richest single source of structure-based chemistry information.**

ChemSpider is owned by the RSC and provided as a free resource to the community.

Antihistamines

Many people suffer from hay-fever in the summer months. Many antihistamine pharmaceutical agents are used as treatments including Cetirizine, Loratidine, Levocetirizine, Desloratadine and Fexofenadine.

Advertisements

Search term: **mefloquine** (Found by approved synonym)

Mefloquine

37171 (ChemSpider ID)
$C_{17}H_{16}F_6N_2O$
Similar
Wikibox
Mass
Machine readable identifiers
Search Google Scholar

2D 3D Load Save Zoom

Add: Comments, Image, Spectrum, CIF, Identifier, Description, Data source, Publication, DOI, PubMed ID, URL

License

- Wikipedia Article(s)
- Associated Data Sources and Commercial Suppliers
- Patents
- Articles
- Identifiers
- Properties
- Medical Subject Headings Classification
- Pharmacological Links
- SimBioSys LASSO
- RSC Databases

Mobile version

Mobile version | Full version

ChemSpider
Building community for chemists

Systematic Name, Synonym, Trade Name, Registry Number, SMILES or InChI

Search

RSC | Advancing the Chemical Sciences

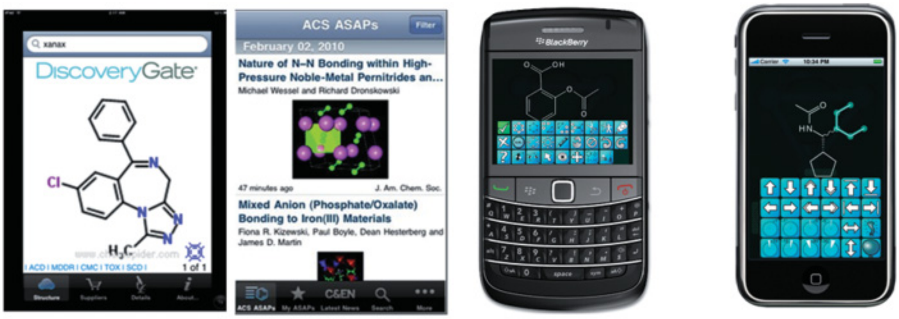
Mobile version | Full version | Search again

ChemSpider
Building community for chemists

Properties | Predicted | Identifiers | Google

Property	Value
ChemSpider ID	37171
Empirical Formula	C17H16F6N2O
Molecular Weight	378.3122
Nominal Mass	378 Da
Average Mass	378.3122 Da
Monoisotopic Mass	378.116682 Da
Systematic Name	(S)-[2,8-bis(trifluoromethyl)quinolin-4-yl][(2R)-piperidin-2-yl]methanol
InChIKey	XEEQGYMUWCZPDN-DOMZBBRYBA

Figure 28.2 Example of how a website can be adjusted for mobile applications on smart phones using the Royal Society of Chemistry's ChemSpider as an example. The left side of the figure shows screenshots of the full Internet version while the right represents screenshots of the mobile version.



xanax
DiscoveryGate
Cl
N
H₃C
www.chemspider.com
ACD | MDDR | CMC | TOX | SCD
1 of 1
Structure
Suppliers
Details
About
ACS ASAPs
Filter
February 02, 2010
Nature of N–N Bonding within High-Pressure Noble-Metal Pernitrides an...
Michael Wessel and Richard Dronskowski
47 minutes ago
J. Am. Chem. Soc.
Mixed Anion (Phosphate/Oxalate) Bonding to Iron(III) Materials
Fiona R. Kizewski, Paul Boyle, Dean Hesterberg and James D. Martin
ACS ASAPs
My ASAPs
Latest News
Search
More
C&EN
BlackBerry
O
OH
Carrier
10:34 PM
HN

Figure 28.3 Screenshots of various mobile chemistry applications.

cause applications to become optimized to these platforms in the future. One obvious route of value will probably be to provide online computational chemistry or biology algorithms that one would previously access on a PC, for example, by using APIs for prediction. Online chemistry and biology databases will also be popularized and increasingly accessed through services via mobile computing hardware [76].

What are the challenges ahead for mobile devices? Considering further the changing interfaces for an increasing abundance of devices, should we imagine a time without them? This may be hard to consider for molecules, but in reality molecules can already be Tweet'ed as SMILES [77] or InChIs [78]. But what about biology? What other uses can we put such devices to that might be collaborative in nature? Time will tell, but ultimately, as mobile devices increase in computing power, they may become our primary computers and means to access scientific research resources, analytical machines, and data in real time.

The science of biomedical R&D may change in the years ahead and will more likely involve more crowdsourcing, precompetitive collaborations, and aggregation of data from diverse sources. Working on technologies for mobile chemistry and biology applications may be a fruitful outlet for developers, especially if there is a collaborative component that can leverage network effects. There certainly needs to be more consideration of how these small powerful computers can be used, and that in itself is a challenge and may come about as much by accident as by design.

28.8 CROWDSOURCING TAIL FOR COLLABORATIVE DATABASES

Large-scale Internet systems (such as Twitter, Digg, Wikipedia, Amazon, Netflix) show a long tail in which a few people participate a lot and a lot of people participate a little [79]. Chapter 6 describes how major company efforts at crowdsourcing ideas [similar to Innocentive (www.innocentive.com)] likewise follow a power law [80] which is most often the signature of a system with positive feedback [81]. We have investigated whether different types of crowdsourcing environments requiring data contributions behave in a similar manner. For example, we have examined the data regarding the number of uploads of data in CDD (www.collaborativedrug.com) for each user and the number of depositions and curations for ChemSpider (www.chemspider.com) users. The CDD data suggest a power law with a considerable downward tail (Fig. 28.4), which is a signature of "saturation" of the audience; that is, in a fixed universe of users of these software, a majority of possible people are becoming active contributors (e.g., uploading data or adding content). In addition, the slope of the apparent power law (solid line) is quite different from the narrow range typical of much larger scale corporate intranet challenges (dashed line and Chapter 6) which are usually from 2.7 to 3. The ChemSpider data also show power law relationships with even more different slopes (Fig. 28.5).

The slope of a power law is a measure of contribution by whom. ChemSpider content is very strongly driven by a small but very active minority (one to five persons) of the total audience, while in contrast, corporate intranet challenges depend most on a large number of occasional contributors, with CDD in between. Because these statistical signatures are robust and affect our strategy to engage present and future contributors, this is an important area for inves-

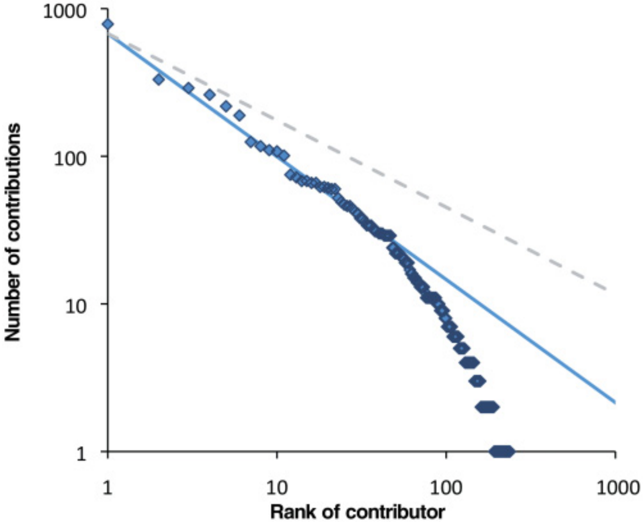


Figure 28.4 Rank–frequency plot of data contributions to CDD. Solid line: power law with α = 2.2; dashed line, α = 2.7.

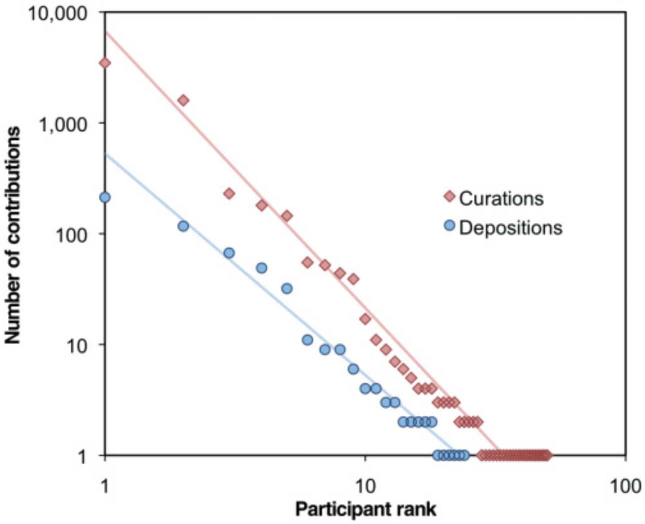


Figure 28.5 Rank–frequency plot of data contributions to ChemSpider. Diamonds: curations, line is power law with α = 1.4. Circles: depositions, line with α = 1.5.

tigation. There is evidence that in some cases the exponent of the power law reflects the difficulty of the task, for example, Twitter (easy) versus writing an original article for Wikipedia (hard) [82], but it does not appear to us that chemistry curation should be more or less complex than contributing to a business challenge. Instead we are drawn to the hypothesis that we are seeing signatures of task consistency, scale, duration, and human motivation: Nowak [83] presents multiple rules showing that sustained cooperation in a large community requires the contributors to perceive more benefit than in a small community. Only in a small community, such as ChemSpider today, are the intangible benefits (reputation, friendship, reciprocity) strong enough to sustain cooperation over time, and it is only over time and with a consistently applicable skill set (organic chemistry) that a small number of committed individuals can come to dominate the statistics. In contrast, the large challenges of Chapter 6 are diverse (and so will tend not to reengage the same expertise) and time limited (typically two to three weeks).

There is more to explore, such as to what extent crowdsourcing signatures reflect task difficulty, constrained access (anonymous versus password-restricted subscription), scientific discipline, or monetary reward (where analysis of Innocentive's statistics would be useful). In any case, we certainly agree with Chris Anderson [79] that wise players exploit both the head (the contributors you already know and access in traditional teams and workshops) *and* the tail (the huge number you do not know and now can access electronically).

28.9 ROLE OF "OPENNESS"—HOW FAR CAN COLLABORATION GO?

Collaboration in drug discovery and development has traditionally been a regulated exchange between known players, often based on formal contractual arrangements. At some level this is at odds with effectively exploiting the ability of densely connected networks, such as the Web, to provide opportunities for unexpected innovation and contributions. Success on the Web can be measured to a large degree by the extent to which your content is linked to, commented on, copied, and reused. Innovation on the Web is driven by connections made between people, between ideas, and often between the unexpected juxtaposition of, in many cases illegally copied, content. Where the free flow of information and content is supported, both technically through effective online tools for sharing and legally through open licensing, the Web effectively supports innovation around that content and information.

The potential of the Web lies in connecting people and insights through the discovery of shared common interests. In drug discovery the identity of those common interests is often proprietary or patentable secrets. While the collaborative frameworks described in this volume offer opportunities to bring costs down, bringing a drug to market will remain expensive. It is tempting to simply declare that the necessity to protect future return on investment in new treatments is simply incompatible with the free movement of information on the Web. It is challenging to enter into a nondisclosure agreement with people you cannot identify without disclosing the key information.

The distance, however, may not be as great as it seems. The traditional process of publication through peer-reviewed journals and patents is one of disclosure. The trade-off is between disclosure and the obtained protected interest, either a "moral right" in recognition and citation from a peer-reviewed article or in the right to exploit granted by a patent. The key decisions made here are when to disclose based on the legal framework, potential losses through not disclosing (e.g., by being "scooped" by a competitor gaining first publication), and the potential for identifying important additional information or collaborators through publication.

These are precisely the same issues associated with publication to the wider Internet. The key differences are that much greater granularity of "publication" is possible, with choices ranging from publication to a group, an organization, a community, or the world, and that the potential benefits in publication, unexpected insights, or connections to arise are much greater. This is, as it has always been, a risk assessment. What are the likely benefits and likely risks of different forms of publication. And, increasingly, what are the likely risks of *not* publishing in some form?

A risk management approach and the added range of options provided by the Internet mean that different solutions are likely to be most effective in different spaces. An explicit aim of SAGE Bionetworks (described earlier) is to make disease biology a precompetitive space by aggregating large quantities of valuable data in the public domain. The capital costs, and therefore the financial risks, of obtaining large volumes of high-quality disease or biology data are now so high that they pose a threat even to large pharmaceutical companies. The potential benefits in pooling these data to identify new potential treatment targets are huge. There is therefore likely to be a significant move toward more open publishing of disease biology data—and the growth of a range of business and academic opportunities around the curation and critical analysis of that data. At the other extreme, at the core of most drug patents, are specific compounds and their formulation. It is unlikely that such specific information on final stage development products will be published transparently prior to patent disclosure in the immediate future. Toxicology data on failed compounds fall perhaps somewhere in the middle.

What is clear is that to fully exploit the potential of the Web to support unexpected innovation and adventitious discovery requires a greater degree of open publication than is traditional in drug discovery. The potential for more effective and efficient flow of information is enormous. In an area where there is significant money to be made, this implies there is a significant market opportunity for the development of tools and approaches that take advantage of that potential. These tools will provide the technical and legal infrastructure that provides confidence about the level of sharing that is going on and confidence about the rights of parties to use or reuse information. High granular sharing (and easy publication) settings on collaborative author environments and widely used and legally respected patent licenses and material transfer agreements that provide clear rights and where appropriate limitations to those rights are likely to play a role here (http://sciencecommons.org/projects/patent-licenses/).

Fully exploiting the potential of the Web for collaboration, and particularly the effective creation and exploitation of unexpected opportunities, requires new approaches, new tools, and new thinking to be applied. This is particularly the case in traditionally secretive and highly controlled areas such as drug discovery. In the long term, however, these are just business risks and opportunities like any other. What is required is a much more sophisticated understanding of what those risks and opportunities represent and how they differ across different parts of an organization and for the different players in the drug discovery process. The change in the information and innovation landscape is such that the major players will need to reconfigure to stay competitive and there will be many opportunities for new and smaller players, including nonprofit organizations, to take advantage of the inevitable inertia of big pharma. Making choices about where and how to be more open and support wide-ranging collaboration opportunities will play a large part in the future health and survival of all of these interested parties.

28.10 CONCLUSIONS

There are a number of areas that we believe are important to monitor in terms of collaborative approaches in the biomedical sciences. For example, credit sharing and rewards for cooperation when various groups or individuals provide data as well as reliable methods for assigning attribution of data to such laboratories will be important as will measures of trust and quality assurance in collaborations.

While ideas are free, very often data are not, and one way to liberate these data is more connectivity between scientists across disciplines. The automation of scientific measurement so that the data are stored and accessible to those with permission may, in some ways, enable ready collaboration. A future research area could be empowering computers with the ability to identify collaborators (whether human or machine-based collaborators) and automatically push data to them. While a competitive advantage for a large or small company will be in the analysis and mining of data and not the data themselves, there needs to be a good understanding of data quality and automated recognition of these and any associated issues with the data will be required. Depending on the laboratory there may be very different accepted measurement standards and approaches and it will be important to understand the interlaboratory differences in data [63, 64].

Another issue is that if some of the precompetitive data in the pharmaceutical industry are to be freed up to the public domain, then they have to be accessible to those charged with the task of collating and providing the data for access. Big pharma and its associated legacy infrastructure has been notoriously bad at knowledge management. In recent years there has been recognition regarding the challenges of leveraging the decades of research-related information [65, 66]. Later bottlenecks will include researchers not being able to leverage public data because they either cannot locate or cannot access the data or the data may not be of sufficiently high quality. To date there are relatively few examples of using public domain data for research while literature data are more commonly used for the development and testing of computational ligand-based models [60, 61]. We believe that there need to be even greater efforts to demonstrate what can be achieved in regards to using literature data for model development.

There remain many questions to answer. Even when there is sufficient confidence in public domain data, how will researchers mesh public domain and precompetitive data into internal systems? Also, what are the legal challenges of protecting data such that they minimize hurdles for data that could be released to the public?

Getting people to collaborate and share data is a significant challenge. It is hard enough to coordinate such efforts inside an organization, so what is the payoff or cultural shift that needs to occur in order to have people participate in open collaborations across organizations? Is it human nature that most will focus on one's own research interests? We should be actively training the next generation of scientists to be more collaborative and use the various software systems that are available to facilitate this. The ultimate goal of collaborative software and tools is to speed up decision-making processes and enhance connectivity between scientists. Such tools may be useful for those involved in alliance management or needed for managing collaborator or researcher networks.

Many other issues could have been covered in this volume, but we are restricted by lack of space as well as the pace in this area of collaboration and affiliated technologies as it is very much a moving target. In the future there will be new and improved ways of getting biomedical researchers to collaborate, as yet not identified. Our present efforts are sure to represent only collaborative software for biomedical research version 1.0. There are, of course, many opportunities to address for version 2.0. These are exciting times.

ACKNOWLEDGMENTS

Stephan Schürer acknowledges funding from the National Institutes of Health for the BAO project described in "Ontologies and Collaborations" (RC2 HG005668).

REFERENCES

1. Williams AJ. Crowdsourcing, collaborations and text mining in a world of open chemistry. 2008. Available: http://www.slideshare.net/AntonyWilliams/crowdsourcing-collaborations-and-text-mining-in-a-world-of-open-chemistry-presentation.
2. Williams AJ. Internet-based tools for communication and collaboration in chemistry. *Drug Discov Today* 2008;13:502–506.

3. Hohman M, Gregory K, Chibale K, Smith PJ, Ekins S, Bunin B. Novel web-based tools combining chemistry informatics, biology and social networks for drug discovery. *Drug Discov Today* 2009;14:261–270.
4. Bailey DS, Zanders ED. Drug discovery in the era of Facebook—New tools for scientific networking. *Drug Discov Today* 2008;13:863–868.
5. Ekins S, Williams AJ. Precompetitive preclinical ADME/Tox data: Set it free on the web to facilitate computational model building to assist drug development. *Lab on a Chip* 2010;10:13–22.
6. Gupta RR, Gifford EM, Liston T, Waller CL, Bunin B, Ekins S. *Using open source computational tools for predicting human metabolic stability and additional ADME/TOX properties*. *Drug Metab Disposition*, doi: 10.1124/dmd.110.034918.
7. Kirschner NM, Pauker SG, Stubbs JW. How can cost-effectiveness information help control unsustainable growth in U.S. health care spending? *Ann Intern Med* 2009;150:57.
8. Arnold RJ, Gabrail N, Raut M, Kim R, Sung JC, Zhou Y. Clinical implications of chemotherapy-induced diarrhea in patients with cancer. *J Support Oncol* 2005;3:227–232.
9. Walker A, McMurray J, Stewart S, Berger W, McMahon AD, Dargie H. Economic evaluation of the impact of nicorandil in angina (IONA) trial. *Heart* 2006;92: 619–624.
10. Beinart SC, Kolm P, Veledar E, Zhang Z, Mahoney EM, Bouin O. Long-term cost effectiveness of early and sustained dual oral antiplatelet therapy with clopidogrel given for up to one year after percutaneous coronary intervention results: From the clopidogrel for the reduction of events during observation (CREDO) trial. *J Am Coll Cardiol* 2005;46:761–769.
11. Elliott RA, Hooper L, Payne K, Brown TJ, Roberts C, Symmons D. Preventing non-steroidal anti-inflammatory drug-induced gastrointestinal toxicity: Are older strategies more cost-effective in the general population? *Rheumatol (Oxford)* 2006;45:606–613.
12. Gaziano TA, Opie LH, Weinstein MC. Cardiovascular disease prevention with a multidrug regimen in the developing world: A cost-effectiveness analysis. *Lancet* 2006;368:679–686.
13. Parashar U, Bresee J, Widdowson M, Gentsch J. New breath for rotavirus vaccines. *Drug Discov Today Ther Strategies* 2006;3:159–165.
14. Simon J, Gray A, Duley L. Cost-effectiveness of prophylactic magnesium sulphate for 9996 women with pre-eclampsia from 33 countries: Economic evaluation of the Magpie Trial. *Bjog* 2006;113:144–151.
15. Sullivan SD, Lee TA, Blough DK, Finkelstein JA, Lozano P, Inui TS. A multisite randomized trial of the effects of physician education and organizational change in chronic asthma care: Cost-effectiveness analysis of the Pediatric Asthma Care Patient Outcomes Research Team II (PAC-PORT II). *Arch Pediatr Adolesc Med* 2005;159:428–434.
16. Arnold RJ. Cost-effectiveness analysis: Should it be required for drug registration and beyond? *Drug Discov Today* 2007;12:960–965.
17. Hakim Z. Using decision-analytic models wisely. *J Manag Care Pharm* 2003;9: 449–450.

18. Weinstein MC, O'Brien B, Hornberger J, Jackson J, Johannesson M, McCabe C. Principles of good practice for decision analytic modeling in health-care evaluation: Report of the ISPOR Task Force on Good Research Practices—Modeling studies. *Value Health* 2003;6:9–17.
19. Arnold RJ, Ekins S. Time for cooperation in health economics among the modelling community. *PharmacoEconomics* 2010;28:609–613.
20. Vandenbroucke JP, Psaty BM. Benefits and risks of drug treatments: How to combine the best evidence on benefits with the best data about adverse effects. *JAMA* 2008;300:2417–2419.
21. NIH. Clinical Research Policy Analysis and Coordination. Available: http://crpac.od.nih.gov/issue_adverse.asp.
22. Papanikolaou PN, Ioannidis JP. Availability of large-scale evidence on specific harms from systematic reviews of randomized trials. *Am J Med* 2004;117:582–589.
23. Scharf O, Colevas AD. Adverse event reporting in publications compared with sponsor database for cancer clinical trials. *J Clin Oncol* 2006;24:3933–3938.
24. (CBER) USDoHaHSFaDACfDEaRCaCfBEaR. Guidance for industry, development and use of risk minimization action plans. Available: http://www.fda.gov/cder/guidance/index.htm.
25. (CBER) USDoHaHSFaDACfDEaRCaCfBEaR. Guidance for industry, good pharmacovigilance practices and pharmacoepidemiologic assessment. Available: http://www.fda.gov/cder/guidance/index.htm.
26. (CBER) USDoHaHSFaDACfDEaRCaCfBEaR. Guidance for industry, premarketing risk assessment. Available: http://www.fda.gov/cder/guidance/index.htm.
27. Fleming TR. Identifying and addressing safety signals in clinical trials. *N Engl J Med* 2008;359:1400–1402.
28. Avorn J. Drug warnings that can cause fits—Communicating risks in a data-poor environment. *N Engl J Med* 2008;359:991–994.
29. Williams DP, Park BK. Idiosyncratic toxicity: The role of toxicophores and bioactivation. *Drug Discov Today* 2003;8:1044–1050.
30. Uetrecht J. Screening for the potential of a drug candidate to cause idiosyncratic drug reactions. *Drug Discov Today* 2003;8:832–837.
31. Li AP. A review of the common properties of drugs with idiosyncratic hepatotoxicity and the "multiple determinant hypothesis" for the manifestation of idiosyncratic drug toxicity. *Chem Biol Interact* 2002;142:7–23.
32. Kaplowitz N. Idiosyncratic drug hepatotoxicity. *Nat Rev Drug Discov* 2005;4:489–499.
33. Lee WM. Drug-induced hepatotoxicity. *N Engl J Med* 2003;349:474–485.
34. Ito K, Chiba K, Horikawa M, Ishigami M, Mizuno N, Aoki J. Which concentration of the inhibitor should be used to predict in vivo drug interactions from in vitro data? *AAPS PharmSci* 2002;4:E25.
35. Boelsterli UA, Ho HK, Zhou S, Leow KY. Bioactivation and hepatotoxicity of nitroaromatic drugs. *Curr Drug Metab* 2006;7:715–727.
36. Kassahun K, Pearson PG, Tang W, McIntosh I, Leung K, Elmore C. Studies on the metabolism of troglitazone to reactive intermediates in vitro and in vivo. Evidence for novel biotransformation pathways involving quinone methide formation and thiazolidinedione ring scission. *Chem Res Toxicol* 2001;14:62–70.

37. Walgren JL, Mitchell MD, Thompson DC. Role of metabolism in drug-induced idiosyncratic hepatotoxicity. *Crit Rev Toxicol* 2005;35:325–361.
38. Park BK, Kitteringham NR, Maggs JL, Pirmohamed M, Williams DP. The role of metabolic activation in drug-induced hepatotoxicity. *Annu Rev Pharmacol Toxicol* 2005;45:177–202.
39. Schuster D, Laggner C, Langer T. Why drugs fail—A study on side effects in new chemical entities. *Curr Pharm Des* 2005;11:3545–3559.
40. The PubChem Project. Available: http://pubchem.ncbi.nlm.nih.gov/.
41. Wang Y, Bolton E, Dracheva S, Karapetyan K, Shoemaker BA, Suzek TO. An overview of the PubChem BioAssay resource. *Nucleic Acids Res* 2010;38: D255–266.
42. ChEMBL Database. Available: http://www.ebi.ac.uk/chembldb/index.php.
43. Jensen NH, Roth BL. Massively parallel screening of the receptorome. *Comb Chem High Throughput Screen* 2008;11:420–426.
44. PDSP Ki Database. Available: http://pdsp.med.unc.edu/kidb.php.
45. ChemBank. Available: http://chembank.broad.harvard.edu/.
46. Seiler KP, et al. ChemBank: A small-molecule screening and cheminformatics resource database. *Nucleic Acids Res* 2008;36:D351–359.
47. Collaborative Drug Discovery. Available: http://www.collaborativedrug.com/.
48. Guarino N, Oberle D, Staab S. What is an ontology? In Staab S, Studer R, Eds. *Handbook on Ontologies*. Berlin Heidelberg: Springer, 2009.
49. Antoniou G, van Harmelen F. *A Semantic Web Primer*. Cambridge, MA, and London: MIT Press/Cambridge, 2004.
50. Inglese J, Shamu CE, Guy RK. Reporting data from high-throughput screening of small-molecule libraries. *Nat Chem Biol* 2007;3:438–441.
51. Bard JB, Rhee SY. Ontologies in biology: Design, applications and future challenges. *Nat Rev Genet* 2004;5:213–222.
52. Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H. Advancing translational research with the Semantic Web. *BMC Bioinform* 2007;8 (Suppl 3):S2.
53. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007;25:1251–1255.
54. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;25:25–29.
55. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 2004;32: D258–261.
56. Soldatova LN, King RD. Are the current ontologies in biology good ontologies? *Nat Biotechnol* 2005;23:1095–1098.
57. Zhou Y, Zhou B, Chen K, Yan SF, King FJ, Jiang S. Large-scale annotation of small-molecule libraries using public databases. *J Chem Inf Model* 2007;47:1386–1394.
58. Choi J, Davis MJ, Newman AF, Ragan MA. A semantic web ontology for small molecules and their biological targets. *J Chem Inf Model* 2010;50:732–741.
59. BioAssay Ontology. Available: http://www.bioassayontology.org/.
60. Staab S, Studer R (Eds.) *Handbook on Ontologies*. Berlin: Springer, 2009.

61. Sarntivijai S, Ade AS, Athey BD, States DJ. A bioinformatics analysis of the cell line nomenclature. *Bioinformatics* 2008;24:2760–2766.
62. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T. PID: The Pathway Interaction Database. *Nucleic Acids Res* 2009;37:D674–679.
63. BioPAX—Biological Pathway Exchange. Available: http://www.biopax.org/.
64. Jena—A Semantic Web framework for Java. Available: http://jena.sourceforge.net/index.html.
65. VIVO. Available: http://vivoweb.org/.
66. The Protégé Ontology Editor and Knowledge Acquisition System. Available: http://protege.stanford.edu/.
67. OWL 2 Web Ontology Language. Available: http://www.w3.org/TR/owl2-syntax/.
68. Barabasi A-L, Oltvai ZN. Network biology: Understanding the cell's functional organization. *Nat Rev Genet* 2004;5:101–113.
69. Ekins S, Giroux C. Computers and systems biology for pharmaceutical research and development. In Ekins S, Ed. *Computer Applications in Pharmaceutical Research and Development*. Hoboken, NJ: Wiley, 2006, pp. 139–165.
70. Dogrusoz U, Cetintas A, Demir E, Babur O. Algorithms for effective querying of compound graph-based pathway databases. *BMC Bioinform* 2009;10:376.
71. Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C. WikiPathways: Pathway editing for the people. *PLoS Biol* 2008;6:e184.
72. Kelder T, Pico AR, Hanspers K, van Iersel MP, Evelo C, Conklin BR. Mining biological pathways using WikiPathways web services. *PloS One* 2009;4:e6447.
73. Wright J, Wagner A. The Systems Biology Research Tool: Evolvable open-source software. *BMC Syst Biol* 2008;2:55.
74. Das S, Girard L, Green T, Weitzman L, Lewis-Bowen A, Clark T. Building biomedical web communities using a semantically aware content management system. *Brief Bioinform* 2009;10:129–138.
75. Martinez AW, Phillips ST, Carrilho E, Thomas SW, 3rd, Sindi H, Whitesides GM. Simple telemedicine for developing regions: Camera phones and paper-based microfluidic devices for real-time, off-site diagnosis. *Anal Chem* 2008;80: 3699–3707.
76. Williams AJ. Mobile chemistry—Chemistry in your hands and in your face. *Chem World* 2010;May. Available: http://www.rsc.org/chemistryworld/Issues/2010/May/MobileChemistryChemistryHandsFace.asp.
77. Weininger D. SMILES 1. Introduction and encoding rules. *J Chem Inf Comput Sci* 1988;28:31.
78. Prasanna MD, Vondrasek J, Wlodawer A, Bhat TN. Application of InChI to curate, index, and query 3-D structures. *Proteins* 2005;60:1–4.
79. Anderson C. *The Long Tail*. New York: Hyperion, 2006.
80. Spencer RW, Woods TJ. The long tail of idea generation. *Int J Innovation Sci* 2010;1:1–11.
81. Newman MEJ. Power laws, Pareto distributions and Zipf's law. Available: http://arxiv.org/abs/cond-mat/0412004v3, 2004.
82. Wilkinson D. Strong regularities in online peer production. In *Proceedings of the 2008 ACM Conference on E-Commerce*. Chicago, IL, 2008. New York: ACM.
83. Nowak MA. Five rules for the evolution of cooperation. *Science* 2006;314: 1560–1563.

HTTP
FEATURES
COMMERCIAL
DISPLAY
DESCRIBED
BIOMARKERS
COMMON
CHANGES
EFFECTIVE
FORM
BUSINESS
HOSTED
ALGORITHMS
DECISION
SUBJECT
EVOLVED
INTERNAL
DISTRIBUTED
CONFIDENTIALITY
STUDIES
WIKI
CHALLENGES
PUBLIC
DEVELOPMENT
STRUCTURE
ALLOWS
ASSOCIATED
UNDERSTANDING
DOMAIN
CHEMSPIDER
MECHANISM
MEDICINE
ESTABLISHED
DOCUMENTS
POINT
TRIALS
IDENTIFY
DISCOVERY
SUCCESSFUL
BIOLOGY
PLACE
ANALYTICAL
PRODUCTS
ENHANCE
MOLECULES
SEMANTIC
OPPORTUNITIES
SOCIAL
CONTENT
HELP
TEAM
TERMS
BENEFIT
PRECOMPETITIVE
ENVIRONMENT
BIOLOGICAL
RECORD
AVAILABLE
INNOVATION
DISEASE
CONCEPT
SERVICES
CONTENT
SCALE
RANGE
JOURNALS
MULTIPLE
COMPOUNDS
REVIEW
HUMAN
CASE
OPEN
REQUIRED
COLLECTION
REPOSITORIES
GROUP
RESEARCH
DIFFERENT
PRIMARY
REFERENCE
FIELD
INFORMATION
HEALTH
STORED
NETWORK
KNOWLEDGE
SHARING
VISUALIZATION
PROJECT
TECHNOLOGIES
INTERNET
LINK
FUNDING
CONTROL
VARIOUS
UNIVERSITY
BIOMEDICAL
DISCUSSIONS
LABORATORY
SUPPORT
IMAGES
REAL
LICENSE
TRANSLATIONAL
PARTNERS
ASSERTIONS
MATERIALS
CONNECTED
PROPOSED
BIOINFORMATICS
RESPONSIBLE
REPRESENT
SELECTED
FREE
STANDARDS
PROGRAM
DRUG
SMALL
SCIENTIFIC
ANNOTATED
APPLICATIONS
COMMUNITY
LEVEL
RIGHTS
PHARMACEUTICAL
PROBLEM
INCREASE
PROCESS
ELECTRONIC
WEB
DATABASE
BASED
YEARS
INTERFACE
IDEAS
CONTRIBUTIONS
ARTICLES
TYPES
INDUSTRY
NATIONAL
ANALYZE
MODEL
ANALYSIS
COST
COMPETITIVE
FRAMEWORK
VALIDATION
COMBINATION
MULTI
COLLABORATION
COMMENTS
STUDENTS
PRESENT
PUBLISHED
PARTICIPANTS
RELEVANT
EFFICIENT
CLINICAL
SOFTWARE
SECURITY
SCIENCE
RESOURCES
TOOLS
ANNOTATION
MEMBERS
QUESTIONS
INSTITUTE
EXPERIMENTAL
MOLECULAR
INFRASTRUCTURE
CHEMICAL
ACADEMIC
NEW
REPORT
GOOGLE
POLICIES
POSSIBLE
FOLLOWING
GOVERNMENT
ISSUES
ADVANTAGE
INCLUDING
FLEXIBILITY
CABIG
CLASS
CAPABILITIES
DATA
COMPANIES
BLOG
APPLIED
INITIATIVES
SET
GENOMIC
GENE
COMPUTING
ACCESS
CREATE
ONLINE
TRADITIONAL
NOTEBOOK
CHEMISTRY
SCREENING
CONTAINS
WIDE
INDIVIDUAL
CURATION
PERSONAL
INTERACTION
PATIENT
PARTICULAR
SEARCH
IMPLEMENTATION
IMPORTANT
PATENT
SPACE
EXPERIMENTS
ENABLE
WORLD
PEOPLE
FORMAT
CURRENT
DEFINED
CONSORTIUM
DESIGN
AUTHOR
VIRTUAL
ENGINE
SITES
MANAGEMENT
SOURCE
PROBLEM
INTEGRATION
PLATFORM
ENSURE
EXISTING
ACTIVITIES
APPROACH
INTEROPERABILITY
CREATIVE
CANCER
FEDERAL
SIMPLE
GOAL
FIGURE
ORGANIZATIONS
SCIENTISTS
TRUST
PORTAL
MAJOR
SIMPLY
NUMBER
SPECIFIC
FACILITATE
PRACTICE
OPERATIONS
PROTECTION

INDEX

Collaborative Computational Technologies for Biomedical Research, First Edition. Edited by Sean Ekins, Maggie A. Z. Hupcey, Antony J. Williams.