


30804971

Chemical Information MINING

Facilitating Literature-Based Discovery

Edited by
DEBRA L. BANVILLE



CRC Press

Taylor & Francis Group

Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2009 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works
Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-1-4200-7649-3 (Hardcover)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data

Chemical information mining : facilitating literature-based discovery / Debra L. Banville, editor.

p. cm.

Includes bibliographical references and index.

ISBN 978-1-4200-7649-3 (alk. paper)

1. Chemical literature--Research. 2. Data mining. 3. Information storage and retrieval systems--Chemistry. I. Banville, Debra L.

QD8.5.C475 2009

025.06'54--dc22

2008030749

Visit the Taylor & Francis Web site at
<http://www.taylorandfrancis.com>

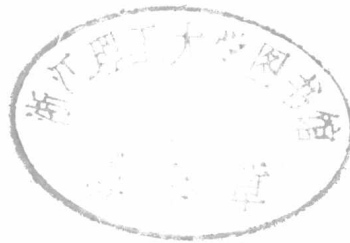
and the CRC Press Web site at
<http://www.crcpress.com>



30804971

Chemical Information MINING

Facilitating Literature-Based Discovery



Preface

Researchers in scientific endeavors such as the life sciences frequently face significant concerns when looking for the most relevant or “right” information from the literature. For example, what happens when we find too much information on a subject, cannot find any information, or cannot access the full text documents when interesting citations are found? These are concerns that most life science researchers face every day but rarely acknowledge. The magnitude of the problem is most commonly expressed as a growing interest in text extraction capabilities and our use of web search engines such as Google, and PubMed or PubChem to provide easy awareness of scientific life science information.

This book is about acknowledging concerns of information extraction, highlighting solutions available today, and underscoring the value these solutions bring to both academic and commercial scientists alike. A special focus is on chemical information extraction due to its importance in so many life science areas and to fill a gap in the literature that still exists at the time this book is being written. Chemical entity extraction is meant to complement the extensive literature on biological entity extraction. The ultimate goal, as described in this book, is to build relationships between chemical and biological entities—relationships that are at the heart of life science research.

The intent of this book is also holistic: to look at both the technological details, in this case the development of chemical structure extraction capabilities, and to provide a possible road map for how researchers can best think about these technologies in their daily work. On one hand, a road map is meant to underscore to developers that the ability to provide a great chemical text extraction capability is most valuable when the scientists needing this capability are factored into the process. On the other hand, we want to underscore to researchers that the capabilities of chemical text mining present new opportunities in how researchers think about and manage their information, and this requires openness to new techniques and capabilities. Ideally, those developing these new capabilities and the researchers needing those capabilities can collaborate on shaping the future of scientific information and knowledge management. This book is written with this vision in mind.

Acknowledgments

To all my colleagues at AstraZeneca who have supported me throughout the years I give thanks, with special thanks to Jim Rosamond and Jim Damewood, who I have worked very closely with in the area of chemical information mining. To the contributors of this book, thank you for your invaluable contributions! I could not have done this book without your generous gift of time and effort. We have something to be very proud of. A thanks also goes to the publishers for believing in this book. Finally, to my husband Don and daughter Janice, thank you for allowing me to take valuable personal time away from you to write and edit this book.

The Editor

Debra L. Banville is currently a scientific information analyst at AstraZeneca Pharmaceuticals within the R&D areas of drug discovery. Debra's focus is primarily on improving the management of both scientific information and knowledge about that information. In recognition of her work, Debra and her team were the recipients of an internal award for innovation. Debra has been an invited speaker at several conferences including most recently the Infonortics conference in Spain (October 2007) and the PharmaBioMed Conference in Portugal (2006), and an invited author for *Drug Discovery Today* (January 2006). Prior to her work in information science, Debra ran an active research program in the areas of drug binding to biological targets using multidimensional nuclear magnetic resonance techniques. Debra received her B.S. from Brandeis University, her Ph.D. from Emory University, and her post-doctorate at the University of California at San Francisco.

Contributors

Debra L. Banville

AstraZeneca Pharmaceuticals
Wilmington, Delaware
debra.banville@astrazeneca.com

Colin Batchelor

Informatics Department
Royal Society of Chemistry
Cambridge, United Kingdom
batchelorc@rsc.org

Roger Beckman

Chemistry Library
Indiana University
Bloomington, Indiana
beckmanr@indiana.edu

Martin Hofmann-Apitius

Fraunhofer Institute for
Algorithms and Scientific
Computing (SCAI)
Sankt Augustin, Germany
martin.hofmann-apitius@scai.fhg.de

A. Peter Johnson

School of Chemistry
University of Leeds
Leeds, United Kingdom
p.johnson@leeds.ac.uk

Richard Kidd

Informatics Department
Royal Society of Chemistry
Cambridge, United Kingdom
kiddr@rsc.org

Corinna Kolářik

Fraunhofer Institute for
Algorithms and Scientific
Computing (SCAI)
Sankt Augustin, Germany
corinna.kolarik@scai.fraunhofer.de

Bedřich Košata

Laboratory of Informatics and
Chemistry
Institute of Chemical Technology
Prague, Czech Republic
bedrich.kosata@vscht.cz

Miloslav Nic

Laboratory of Informatics and
Chemistry
Institute of Chemical Technology
Prague, Czech Republic
miloslav.nic@vscht.cz

Anikó T. Valkó

Keymodule Ltd.
Leeds, United Kingdom
aniko.valko@keymodule.co.uk

David J. Wild

School of Informatics
Indiana University
Bloomington, Indiana
djwild@indiana.edu

Antony J. Williams

ChemZoo Inc. and
ChemConnector Inc.
Wake Forest, North Carolina
antony.williams@chemspider.com

Andrey Yerin

Advanced Chemistry
Development Inc.
Moscow, Russian Federation
yerin@acd labs.ru

Contents

Preface.....	vii
Acknowledgments.....	ix
Editor	xi
Contributors	xiii

Part I Introduction to Information Mining for the Life Sciences

Chapter 1	Illustrating the Power of Information in Life Science Research.....	3
	<i>Debra L. Banville</i>	
Chapter 2	Chemical Information Mining: A New Paradigm	13
	<i>Debra L. Banville</i>	

Part II Chemical Semantics

Chapter 3	Automated Identification and Conversion of Chemical Names to Structure-Searchable Information.....	21
	<i>Antony J. Williams and Andrey Yerin</i>	
Chapter 4	Identification of Chemical Images and Conversion to Structure-Searchable Information	45
	<i>A. Peter Johnson and Anikó T. Valkó</i>	
Chapter 5	Chemical Entity Formatting.....	77
	<i>Bedřich Košata</i>	
Chapter 6	Chemical XML Formatting	99
	<i>Miloslav Nic</i>	

Part III Trends in Chemical Information Mining

- Chapter 7** Linking Chemical and Biological Information with Natural Language Processing..... 123
Corinna Kolářik and Martin Hofmann-Apitius

- Chapter 8** Semantic Web..... 151
Colin Batchelor and Richard Kidd

Part IV Involving the Researchers and Closing the Loop

- Chapter 9** The Future of Searching for Chemical Information 171
David J. Wild and Roger Beckman

- Chapter 10** Summary and Closing Statements 185
Debra L. Banville

- Index** 187

Part I

Introduction to Information Mining for the Life Sciences

1 Illustrating the Power of Information in Life Science Research

Debra L. Banville

CONTENTS

Introduction.....	3
Barriers to the Automation of These Discoveries	5
There Has Got to Be a Better Way to Do This — Shifting Paradigms	5
References.....	8

INTRODUCTION

The ironic proverbial saying that “a month in the lab can save you an hour in the library” is proving itself repeatedly and at a huge cost to both academic and commercial institutions alike. Missed information in the literature costs time, money, and quality. Both the quality of decisions made and the quality of subsequent research output is compromised when the available information is not realized. In monetary terms, incorrect decisions along the drug pipeline lifecycle in the pharmaceutical area can cost millions to billions of dollars (Adams and Brantner 2006; Banik and Westgren 2004; DiMasi 2002; DiMasi et al. 2003; Gaughan 2006; Leavitt 2003; Myers and Baker 2001).

Substantial costs have been experienced in academia as well and seen as missed funding opportunities due to a combination of access limitations to the information together with the inability to find and process the available information (Wilbanks and Boyle 2006). Access limitations are worse in academia than in industry. Lowering the barriers to access limitations has been the goal of individuals such as Paul Ginsparg, who in 1991 developed arXiv (Ginsparg 1991), the first free scientific online archive of non-peer reviewed physics articles that continues today (Ginsparg et al. 2004). Many groups have formed to increase the accessibility of academic information such as SPARC (Scholarly Publishing and Academic Resources Coalition), the Science Commons group (www.sciencecommons.org), and the World Wide Web Consortium (w3c.org).

The value of information mining the literature for knowledge has been illustrated repeatedly. In 1986, Donald R. Swanson, an information scientist, mathematician,

and professor emeritus at the University of Chicago, demonstrated the technique by using the literature to find a possible treatment for Raynaud's syndrome (Swanson 1986, 1987, 1988). Swanson went on to clinically prove the hypothesis suggested by the literature to use fish oils as a treatment for Raynaud's. This work set off a string of papers in an area coined as "literature-based discovery" or "literature-related discovery" (Smalheiser and Swanson 1994, 1996a,b, 1998; Swanson 1990, 1991; Swanson and Smalheiser 1999; Gordon and Lindsay 1996; Kostoff 2007; Weeber et al. 2001).

In fact, literature-based discovery and text mining of the literature are part of the same thing; they are both about extracting information from text to discover something new, novel, or not already known. Text mining and literature-based discovery go beyond the simple analysis of text. Ideally they led to the recognition of interesting patterns not explicitly stated. The most recent and prominent example of this, at the writing of this book, is a January 2008 article by a group from Peking University (Li et al. 2008). These researchers asked the question, Is there a common molecular pathway in addiction? They first identified ~1,000 relevant articles on the subject and manually extracted 2,343 items of evidence. They kept only well-established evidence and extensively annotated and then stored this evidence in a searchable database for further analysis. Based on their meticulous extraction and analysis, they identified five molecular pathways common to four different types of addictive drugs. This included discovering two new pathways and clues to the irreversible features of addiction. They did this without conducting a single experiment.

A rigorous description of literature-based discovery was published by Kostoff in an earlier paper (Kostoff 2007) and followed later by a series of eight papers that detailed the techniques used and demonstrated these techniques for a variety of life science areas including cataracts, Raynaud's, Parkinson's, and multiple sclerosis, and water purification (Kostoff 2008a,b; Kostoff et al. 2008a-f).

Other opportunities for knowledge discovery from the literature includes the area of *drug repurposing*, the development of novel uses for existing drugs. Most drug repurposing (also known as drug *reprofiling* or *repositioning*) discoveries were the result of researchers connecting key information to generate a valid hypothesis that could be tested in the clinic (Wilkinson 2002; Lipinski 2006; Oprea and Tropsha 2006; Ashburn and Thor 2004). Repurposed drugs frequently have the advantage of having been previously tested in the clinic for safety and are simply being reapplied to a novel area. This is not a new concept: in 2004, 84% of the 50 top-selling drugs had additional indications approved since their launch in the United States (Kregor 2007). For example, two drugs on the market for Parkinson's disease, Ropinirole/Requip (GSK) and Pramipexole/Mirapex (Boehringer Ingelheim), were later repurposed for restless leg syndrome. Repurposing involves additional expense for phase IV trials to support the new indication, application, and marketing fees, but this is nothing compared to the cost of running phase I, II, and III trials.

The bottom line is that in the hands of creative, experienced researchers, text mining of the literature or literature-based discovery can only serve to increase the opportunities within drug discovery and enhance life science research. Turning the ironic proverbial phrase around to read “an hour in the library saves a month in the lab” would be a more advisable approach.

BARRIERS TO THE AUTOMATION OF THESE DISCOVERIES

Finding relevant information involves finding relevant documents, accessing those documents, and finding relevant information within those documents. Numerous barriers exist along each step of the way (Banville 2006 and references within, 2008). For example, imagine that you are trying to find all the bicyclic compounds known to be selective for a specific target. What are some of the issues you would encounter?

- Too many sources to search
- No structure searching capability available within most of these sources
- Limited accessibility to all the necessary sources due to licensing costs
- Limited rights to download and manage the citations and documents found due to licensing restrictions

Getting a citation from PubMed, for example, does not mean that the scientist has access to the full text document cited or the right to use a computer to mine a large set of full text documents. Controversies over the announcement that researchers supported by the National Institute of Health (NIH) will be required to submit all peer-reviewed articles to the NIH for public access within 12 months of publication has predictably drawn positive reviews from most researchers and negative reviews from most publishers (Morrissey 2008). Even if full text access is available, the logistics of downloading all 100 or 1,000 “must read” full text articles are tedious, to say the least. For example, the group from Peking University engaged many students over two years to read, extract, and annotate information from 1,000 documents relevant to drug addiction (Li et al. 2008 and correspondence with L. Wei).

THERE HAS GOT TO BE A BETTER WAY TO DO THIS — SHIFTING PARADIGMS

A variety of technological advances including the advent of the Semantic Web and social networking are driving a cultural change in how information is found and presented back to the user (e.g., Murray-Rust et al. 1997; Rzepa 1998; Berners-Lee et al. 2001; Luo 2007; Chang 2007; Dong et al. 2007). It is no longer about publishing information in print form; it is about *ePublishing* with the ability for communities of readers to comment on this information. This effectively captures

knowledge about information, a new paradigm in information sharing. It is also about automatic linking to related information as a form of knowledge sharing and knowledge building.

Publishers like the Royal Society of Chemistry have initiated *Project Prospect* to enhance articles prior to publication with chemical and biological concepts (see rsc.org and *The Alchemist Newsletter* 2007 for details). Chapter 8 of this volume has a detailed discussion on publishing. The ability to find chemical structural information and its associated data is becoming much easier as the result of these endeavors and their many contributors (such as Rupp et al. 2007; Wilkinson 2002; Corbett et al. 2007; Batchelor and Corbett 2007; Corbett and Murray-Rust 2006; Nic et al. 2002; Murray-Rust and Rzepa 1999; Zimmermann and Hofmann 2007 and references within; Zimmerman et al. 2005; Williams 2005; Williams and Yerrin 1999; Rouse and Beckman 1998; Ibison et al. 1992, 1993a,b; Simon and Johnson 1997 and reference within).

Project Prospect endeavors to use and build acceptance of standards for chemical information by using the International Chemical Identifiers (InChIs) created by the International Union of Pure and Applied Chemistry (IUPAC) as a way to provide a nonproprietary way to make chemical information more machine-readable. To illustrate the potential of this in the simplest way, an InChI for benzene (i.e., **InChI=1/C6H6/c1-2-4-6-5-3-1/h1-6H**) was pasted into a Google search bar (www.google.com), this resulted in 37 hits in the fall of 2007 and over 1,000 hits 6 months later in the spring of 2008. The top hits were directed at the IUPAC Gold Book as shown in Figure 1.1.

Similar searches on common drugs resulted in many highly relevant hits. In the case of aspirin, shown in Figure 1.1, the links were made to several open-access chemical databases such as The Carcinogenic Potency Project database (<http://potency.berkeley.edu/chempages/ASPIRIN.html>), PubChem (<http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=2244>), Drug Bank (<http://www.drugbank.ca/cgi-bin/getCard.cgi?CARD=DB00945.txt>), and ChemSpider (<http://www.chemspider.com/RecordView.aspx?id=2157>). While this search does not provide a definitive capability and does not ensure a high degree of accuracy in the results found for these drugs, it does demonstrate the current ability we all have to perform a chemical structure search against a large body of information, the Internet, and retrieve highly relevant results.

Integration of select Internet resources, such as the public chemical databases mentioned above, provides a very practical approach to structure searching the Internet and internal resources (Dong et al. 2007). Chapter 8 elaborates on this concept. As summarized in Chapter 2, another facet of chemical structure mining involves finding information within full text documents that do not traditionally contain identifiers like InChI or SMILE strings. Chapter 5 contains an in-depth discussion of these identifiers.

