METHODS IN LITERATURE-BASED DRUG DISCOVERY

Nancy C. Baker

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the School of Information and Library Science

Chapel Hill
2010

Approved by:

Bradley M. Hemminger

Stephanie W. Haas

Javed Mostafa

Diane Pozefsky

Alexander Tropsha

ABSTRACT

NANCY C. BAKER: Methods in Literature-based Drug Discovery

(Under the direction of Bradley M. Hemminger)

This dissertation work implemented two literature-based methods for predicting new therapeutic uses for drugs, or drug reprofiling (also known as drug repositioning or drug repurposing). Both methods used data stored in ChemoText, a repository of MeSH terms extracted from Medline records and created and designed to support drug discovery algorithms.

The first method was an implementation of Swanson's ABC paradigm that used explicit connections between disease, protein, and chemical annotations to find implicit connections between drugs and disease that could be potential new therapeutic drug treatments. The validation approach implemented in the ABC study divided the corpus into two segments based on a year cutoff. The data in the earlier or baseline period was used to create the hypotheses, and the later period data was used to validate the hypotheses. Ranking approaches were used to put the likeliest drug reprofiling candidates near the top of the hypothesis set. The approaches were successful at reproducing Swanson's link between magnesium and migraine and at identifying other significant reprofiled drugs.

The second literature-based discovery method used the patterns in side effect annotations to predict drug molecular activity, specifically 5-HT6 binding and dopamine antagonism. Following a study design adopted from QSAR experiments, side effect information for chemicals with known activity was input as binary vectors into classification algorithms. Models were trained on this data to predict the molecular activity. When the best validated models were applied to a large set of chemicals in a virtual screening step, they successfully identified known 5-HT6 binders and dopamine antagonists based solely on side effect profiles.

Both studies addressed research areas relevant to current drug discovery, and both studies incorporated rigorous validation steps. For these reasons, the text mining methods presented here, in addition to the ChemoText repository, have the potential to be adopted in the computational drug discovery laboratory and integrated into existing toolsets.

ACKNOWLEDGEMENTS

I wish to thank my family for their steadfast support throughout this long research and writing process. My husband, Joshua, has provided support of every kind and helped to keep me balanced. My children Madeline, Natasha, Lily, Ella, and Verity have stood by me, even though they did not understand exactly what in the world I was doing. To my brothers and sisters Mende, Margaret, Melissa, Curry, Jim, and Tyler, I send a thank-you as well. They always asked the right questions and never mentioned the birthday cards I forgot to send over the last year. To all my friends – you know who you are – thank you for being there.

I want to thank my mother-in-law, Thelma S. Baker, Ph.D. She got her degree at age 54 and has been a great example for me. Her other children, my sisters-in-law (and their respective spouses), have been a great source of support and encouragement.

I could not have completed or even started this work without the support of my committee members: Brad Hemminger, Stephanie Haas, Javed Mostafa, Diane Pozefsky, and Alex Tropsha. Brad Hemminger and Stephanie Haas provided both leadership and structure and the essential encouragement along the way. Diane Pozefsky provided a rigorous review of the work and posed insightful questions. Thanks also to the faculty and staff of SILS, including Lara Bailey, who managed to get me registered each semester on time. I would like to thank my fellow SILS students, who over the years have provided kindness, help, encouragement, and great conversations.

I would like to give a special thanks to Alex Tropsha for his enthusiasm and support from the very beginning. And thanks as well to the members of Dr. Tropsha's lab. The exciting work and collegial environment in the lab has been an inspiration.

I would also like to recognize the sources of financial support that made this work possible. The one-year fellowship from the UNC Program in Bioinformatics and Computational Biology got me started. The NIH grant P20-HG003898 (through Alex Tropsha) has been a major source of support since then. The UNC School of Information and Library Science has also provided funding through research positions and the Lester Asheim Scholarship.

Thank you, finally, to my mother and father. My mother taught me how to work hard, but she was (and still is) a great example of how to enjoy life. Because of my father, a pharmacist, I sent many hours dusting shelves full of ointments, elixirs, and pills, wondering how such things could have been discovered.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

5-HT6	5-hydroxytryptamine or serotonin
ACS	American Chemical Society
CAS	Chemical Abstracting Service
CCR	Correct classification rate
CF	Cystic fibrosis
CFTR	Cystic fibrosis transmembrane conductance regulator protein
CLiDE	Chemical Literature Data Extraction
CML	Chemical Markup Language
DA	Dopamine antagonist
EPS	Extrapyramidal symptoms
FDA	Food and Drug Administration
IE	Information extraction
IR	Information retrieval
ISI	Institute for Scientific Information
IUPAC	International Union of Pure and Applied Chemistry
JACS	Journal of the Americal Chemical Society
LBD	Literature-based discovery
MeSH	Medical Subject Heading
NCBI	National Center for Biotechnology Information
NCE	New chemical entity
NCI	National Cancer Institute
NER	Named entity recognition
NLM	National Library of Medicine

NLP	Natural language processing
PDSP	Psychoactive Drug Screening Program
QSAR	Quantitative structure-activity relationship
SVG	Scalable vector graphics
UMLS	Unified Medical Language System
UNC	University of North Carolina

1. RESEARCH GOALS AND BACKGROUND

1.1 Research questions and their significance

The biomedical literature is a rich source of information about the activity of drugs in biological systems. This information, once extracted and stored in a usable format, could potentially guide researchers in their search for new safe and effective drug therapies. It is therefore no surprise that text mining techniques are increasingly applied to the chemical literature to extract this important information. But information extraction is only the first step. For literature to be useful in drug discovery, terms pulled from the literature must be used as input to some drug discovery algorithm. This dissertation investigates this second step in the process: what to do with the extracted information.

The broad research question motivating this work is:

How can information extracted from the biomedical literature be used in drug discovery?

This work will approach the broad question by concentrating on two specific methodologies. The research questions at the center of this dissertation are:

1. Can an extended and improved implementation of Swanson's ABC paradigm be used to predict new uses for existing drugs?

2. Can patterns in side effect annotations be used to predict a drug's molecular activity?

These are significant questions because, if they can be answered in the affirmative, literature-based discovery may acquire an accepted place alongside the traditional methods employed in the computational drug discovery laboratory. Currently few implementations of literature-based discovery are seen in day to day practice in the laboratory, despite the increasing interest in literature mining seen in recent years.

Robust validation is key to acceptance. It has been suggested that inadequate validation is one reason why Swanson's ABC approach, introduced to great excitement over 20 years ago, has received little notice outside the information science community (Bekhuis, 2006; Torvik, Renear, Smalheiser, & Marshall, 2009). In this research, therefore, validation will play a vital role, and one that should help foster greater acceptance from the drug research community.

1.1.1 Motivation

The discovery and development of new medicines is an expensive and high-risk endeavor. It was recently estimated that for drugs that reached clinical trials between 1989 and 2002, the average cost per drug was over $800 million (Adams & Brantner, 2006). Even when a drug has been approved for marketing, there is no guarantee it will be a success. Many drugs are pulled from the market because of adverse side effects (Giacomini et al., 2007).

To address these challenges, researchers are increasingly making use of data and computational methods to learn as much as they can about a drug *before* it undergoes expensive laboratory or clinical testing. This means analyzing data and looking for patterns that would allow prediction of chemical characteristics, both therapeutic and adverse.

Quantitative structure-activity relationship (QSAR) studies, for instance, are used to predict receptor binding, cellular transport, penetration of blood-brain barrier, and many types of toxicity. Fortunately, the repositories of chemical data needed for these quantitative experiments are growing in number and in size. The Molecular Libraries Initiative (NIH, 2007), with PubChem as its central repository, has spurred extensive testing of compounds and the results are all publicly available.

Increasingly, too, researchers are examining existing drugs to see if they can be reprofiled for a different indication. The reprofiling of drugs (also called repositioning or repurposing) can offer lowered costs and risks to the drug developer (Bradley, 2005). The safety profile of existing drugs is often well known, and expensive early stage animal studies may have already been performed, saving the expense of the studies and accelerating the development timeline.

Repositories of laboratory-based data for drugs may be growing in size, but most of the information about drugs remains locked up in the chemical and biomedical literature. For several hundred years, results from experiments with chemicals, drugs, and disease were reported only in the literature. Drug researchers are beginning to understand that this information could contribute greatly to their understanding of drugs, not just by finding relevant articles or facts and reading them, but by turning the literature into data and using it as input into computational experiments. In a manner similar to the methods used in the lab now, these experiments can *predict* activity or characteristics of drugs. A prediction of drug activity or effect is often the first step in drug reprofiling.

Only existing drugs have a literature record. This means that literature cannot be used to uncover a new chemical entity and predict its uses. Literature can, however, be used to predict new things about existing drugs, including how they might be used therapeutically in a disease where they have not been tested, i.e., drug reprofiling.

This dissertation research presents two literature-based drug discovery methods. Both methods use entities and relationships from the literature to predict new therapeutic uses for drugs. Validation is a central component of the study designs. The goal is to develop methods that can be integrated into the toolset already in use in the drug discovery laboratory.

1.1.2 Pilot Study

The Information Hierarchy or Information Pyramid is an important representation of learning and understanding in information science (Chaffey & Wood, 2005; Rowley, 2007). In this representation, data is depicted at the bottom, information in the middle, and knowledge at the top. The depiction illustrates, among other things, how humans learn. First we accumulate data, or the raw facts and observations about something. Next we organize it so that any patterns found can provide information about the data collection. Next we infer and reason from the information and conceptualize some tenets or generalizations that we can carry forward: this is knowledge.

This dissertation work concentrates on the top level of the pyramid: knowledge discovery. The essential prerequisite work in extracting the data and organizing it into information – the other two levels in the pyramid - were performed in a pilot study. In that work, a repository or knowledgebase was constructed from MeSH ()()()()annotations extracted from chemical and biomedical articles in PubMed (National Library of Medicine, 2010). The construction of this knowledgebase, called ChemoText, is described in Chapter 2. The pilot study also included an implementation of Swanson's ABC drug discovery methods; Chapter 2 also contains the results from this study.

1.2 Background

In this section we will look at how researchers are using literature data to make predictions. Before we examine methods to predict new things from the literature, we will look at the characteristics of the literature itself, including its historical development. Then we will review how other researchers have processed the literature to change it from language into data.

Drugs are chemicals. For that reason we will concentrate on processing chemical literature, starting with a look at the history and characteristics of chemical literature that make it a unique challenge to process. Drugs are chemicals that affect biological systems and the field of drug discovery sits at the intersection of biology and chemistry. So while we will focus on small molecule chemicals important to drug discovery, as a part of our methods overview we will often find it illustrative to describe implementations of important literature mining methods in biology, particularly at the molecular level.

The field of literature mining encompasses the steps, tools, and techniques to process the literature and find the relevant documents (Information Retrieval or IR), extract relevant facts (Information Extraction or IE), and learn new things from these facts (Text Mining/Knowledge Discovery). These three subfields are interdependent. Information extraction is often a first step in information retrieval. Both information retrieval and information extraction may be involved in finding and extracting the appropriate text and placing it into a data structure such as a database for later text mining.

At this point, a word about terminology may be helpful to prevent confusion. *Literature mining*, *text mining*, *knowledge discovery in text,* and *text data mining* are all terms which have been used more or less synonymously. In this dissertation we will adopt the terminology of Jensen et al. in which *literature mining* is used to describe the broad field which includes information extraction, information retrieval, and text mining (Jensen, Saric, & Bork, 2006). *Text mining* will be used interchangeably with *literature-based discovery*. They both refer to discovering new things from terms extracted from the literature.

1.2.1 Chemical and biomedical literature

The need for chemists to communicate their work and to learn about the research of others has existed since the dawn of chemistry. The early communication of chemical research in the 17th century took place primarily in private letters, pamphlets, and books. The 18th century saw the rise of scientific journals and periodicals, and much of the reporting of chemistry moved to these venues. In France, Lavoisier started *Annales de Chimie* in 1789 and in Germany in 1778, the *Chemishes Journal* was founded by Crell. With the advances in science and technology in the 18th and early 19th century, more outlets for communication were needed. Chemistry articles were included in the journals of the academies and learned societies such as the *Philosophical Transactions of the Royal Society* in Britain and *Memoires de l'Academie des Sciences* in France. There were also a number of journals run by commercial publishing companies, but these did not experience the longevity and influence of the journals produced by the more stable societies, with a few titles such as *Nature* being the exception. Later in the 19th century, societies devoted to chemistry began to form and to start publishing their own journals.   The Chemical Society in Great Britain was the earliest such society, formed in 1841, and was followed by societies in other European countries, among them the Societe Chimique de France in 1857, the Deutsche Chemische Gesellschaft in Germany in 1867, and the Russian Chemistry Society founded by Dmitrii Mendeleyev in 1868. (Cooke, 2004; Skolnik, 1982)

The journals published a variety of literature. Early publications were often proceedings of the organization's meetings. These proceedings included full text of some papers and abstracts of others. It soon became apparent, however, that as the volume of publications grew worldwide, a way to summarize the publications in other journals home and abroad was of great value and interest, and, as a result, collections of abstracts soon appeared, first as sections in the regular periodicals, and later as separate volumes. *Chemishes Zentralblatt*, founded in 1830 in Germany, was one of the early publications dedicated to abstracts, primarily of German research (Cooke, 2004).

In the United States, the American Chemical Society (ACS) was founded in 1876 and issued its first publication of meeting proceedings that year. The publication, which eventually became the *Journal of the American Chemical Society (JACS)*, included abstracts by 1897. In 1907 a separate publication dedicated to abstracts, *Chemical Abstracts*, was started. *JACS* has grown steadily since and has become one of the premier chemistry journals. *Chemical Abstracts* grew quickly as well. ACS started a division devoted to producing Chemical Abstracts that was eventually called Chemical Abstracting Service or CAS. They expanded their scope of coverage to books, dissertations, patents, government reports and extended their reach to most of the countries doing important chemical research. The types of information gathered on a research article included bibliographic data (e.g., author, journal, publication date, company) and a brief summary of the main findings of the article with an emphasis on chemicals, reactions, procedures and techniques (Skolnik, 1982).

CAS developed indexing schemes that proved immensely influential. The first was a subject index. In 1911 they started a patent index, and in 1920 came out with the first formula index. CAS developed their own nomenclature system that allowed them to index chemicals for efficient retrieval. In the 1960's they started to use computers and developed innovative computational methods to assist the indexing. With the creation of the Registry System, they began to store the structure of a chemical in computer files and assign unique numbers to each. This monumental effort took years, but as a result the CAS registry number became the most used chemical identifier worldwide. (Flaxbart, 2007; Weisgerber, 1997)

Other competing and complementary services emerged over the years. The Institute for Scientific Information (ISI), for instance, under the leadership of Eugene Garfield, developed the *Current Contents* and *Index Chemicus* (Garfield, 2001). ISI had a slightly different focus from CAS. They covered fewer journals over a broader scientific area and had a faster delivery time for their publications. They also captured citations in articles. Citations proved to be important to chemists who wanted to try a particular reaction method, for instance, because they could search the literature using the "primordial reference" to find all papers that used that method, and trace the modifications and improvements over time (Garfield, 1985).

The literature of medicine is also important background for this research. For medicine we will focus on the development of the United States National Library of

Medicine. In 1818 Joseph Lovell became the eighth Surgeon General in the U.S. Army medical department. Lovell collected books, both for his own use and the use of his staff of medical personnel. When he died in 1836, his books remained in the office and became the core of the official library of the Surgeon General. The library grew, and by 1840 the collection was large enough that someone felt the need to list the 134 titles in a small notebook, the first catalog. The Civil War brought rapid expansion to the Surgeon General's office and to the collection. In 1864 the new Surgeon General, William A. Hammond, oversaw the production of the first printed catalog. It listed 2,100 volumes. The real growth in the library, however, came when Surgeon General Joseph Barnes made John Shaw Billings his assistant in charge of the library, which they agreed should become a "National Medical Library". Billings energetically started collecting books and pamphlets, old and new, contacting physicians all over the country to send past copies of journals. By 1875 the library was the largest medical library in the country. (Blake, 1980; Blake, 1986)

Billings was no less energetic in organizing and cataloging the collection. Here he had examples to follow. Following the example of abstracting journals in Europe and particularly the bibliographies of J.D. Reuss and W. G. Ploucquet, Billings eventually created an index called *Index-Catalogue* that indexed books by title and author, journals by title, and journal articles by subject. Because his library was the most comprehensive collection of medical literature in the country, the *Index-Catalogue* became the most extensive guide to medical literature available. (Blake, 1980; Blake, 1986)

Keeping current was still a problem. With years between the publication of each volume, a physician in need of current information had to refer to the European abstracting publications, the best of which were in German. To fill this need, Billings worked with New

York publisher F. Leypoldt to produce a monthly subject guide to medical books and journals, which they called *Index Medicus*. Though very successful, the *Index Medicus* struggled financially and for a time merged with a similar publication of the American Medical Association. After a number of years of slow growth in the early part of the 20th century, the library grew rapidly during World War II and began to modernize its cataloging operations. Microfilm, mechanization, and finally computerization have brought the library and the catalog efforts into the modern age.

The computerization of a catalog yields a database. Today the National Library of Medicine's collection of citations, reaching back to the Civil War, is publicly available as the Medline database and can be freely searched through the PubMed web site (National Library of Medicine, 2008). Medline covers medical and biomedical literature, primarily journals, including drug research, and, importantly, it is free; these qualities make it the most commonly used corpus for biomedical text mining.

While the focus of PubMed remains bibliographic, CAS has broadened its functions. The CAS registry number has become such an important identifier for chemicals that the database has become a point of entry and control to the world of chemicals, as well as a bibliographic resource. The centrality of CAS when discussing information in chemistry is hard to overestimate. CAS is like a planet with an immense gravitational pull. One is either going with the pull, or fighting it, but ignoring it is impossible. Its gravitational pull affects this literature review in the following way. Early and very substantial work in named entity recognition, information extraction, and information retrieval in chemistry was dominated by scientists at CAS (as well as ISI). Later, as the field of bioinformatics developed, the preponderance of literature mining work was concentrated in molecular biology and on large biological molecules - genes and proteins. In more recent years, literature mining in chemistry has gained interest as scientists look to extract their own chemical information from the literature, in part to build their own repositories separate from CAS. The recent work in literature mining draws on both the previous work in both chemistry and biology, and therefore the discussion of methods and applications in this review will include techniques and methods in both those fields.

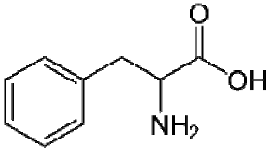
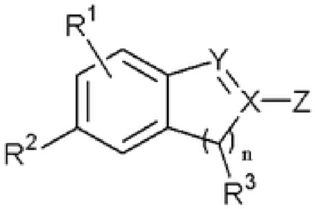
A very key difference in the literature of chemistry and the literature of biology is the role played by the structure of a molecule (Fugmann, 1985). In the chemistry of small molecules such as drugs, the structure is central; in contrast, the biology of large chemical molecules such as proteins and DNA-encoding genes does not pivot on exact molecular structure. The chemical structure of a DNA strand for Gene A, for instance, may vary between individuals or undergo mutations. It is, however, still Gene A. (Location as well as chemical makeup is important for genes.) By contrast, if a small molecule chemical B undergoes a structural change, it is no longer chemical B; it is now a different chemical, with a different name, and with perhaps dramatically different properties. Because the precise structure is so vital, communication of that structure plays a role in information extraction and information retrieval, and adds new wrinkles to recognizing chemical entities in text, a necessary prelude to extracting them.

The task of finding the entities of interest in the text is called named entity recognition (NER). Before we can learn how computers recognize chemicals in the text, we must first discuss how scientists represent chemicals in their published work.

Representation of chemicals in text

A chemical is a collection of atoms bonded together and taking up three dimensional space. The structure of a chemical makes it unique and gives it its physical and biological characteristics. In written communication chemists portray this structure in a variety of ways. Representative samples of the most common structures are listed in Table 1.1.

Table 1.1 Structural representation of chemicals

	Structure Representation Example	**Communication characteristics**
1.	$HO_2CCH(NH_2)CH_2C_6H_5$C9H8O4	Chemical formula. Specifies type and number of atoms but no information on 3D structure. Computer can read but cannot translate to structure accurately. Humans cannot get complete structural information.
2.		Chemical structure diagram. Very understandable to humans. Preferred mode of human – human written communication, however cannot be used to reference the molecule in a line of text or in the spoken word. Computer can generate but not understand easily.
3		Markush structure. This structure indicates a family of molecules. The letters can be replaced by a variety of substructures. Used in patents to gain coverage on a variety of molecules with a similar core structure.

A publication reporting the synthesis of a new compound or a chemical reaction will likely contain a chemical structure diagram like the one in row 2 of Table 1.1. When the chemical is referred to in the text, however, a name must be used.

Every chemical has a unique, standardized name that can be used in text. The standard nomenclature system for chemicals is the IUPAC (International Union of Pure and Applied Chemistry) standard (IUPAC, 2009). In this name, called the *systematic name*, each component of the chemical structure has a corresponding syllable in the nomenclature. The use of the systematic name results in an unambiguous translation of the structure into words (Gasteiger & Engel, 2003).

The IUPAC name is long and cumbersome however, and most chemists, though they may use it to introduce a molecule, will often refer to the chemical by its common name. These names, also called *trivial* or *generic* names, have their origins in history or in custom and are shorter and easier to read, write, and remember than IUPAC names. In contrast to systematic names, they give little to no information about the structure of the chemical. Because of their widespread use, a place for trivial names has been included in the IUPAC standards. A semi-systematic or semi-trivial name has elements of both, often a parent structure which is trivial, modified by a systematic prefix (Cooke-Fox, Kirby, & Rayner, 1989a; Cooke-Fox, Kirby, & Rayner, 1989b; Cooke-Fox, Kirby, & Rayner, 1989c).

Other commonly used chemical names are trade names. These include the names of marketed pharmaceuticals, and, as a number of companies may market the same chemical under different trade names, the names for a chemical can mount up. For instance, one chemical database contains 174 different names for aspirin (Williams, 2008).

The author may not want to identify a chemical in a way that indicates its structure. This is often the case when researchers in the pharmaceutical industry are publishing findings but not ready to reveal the structure of a potential new drug. In this case company codes are often used (Banville, 2006). Chemicals are often referenced by their identifier or reference number in a repository or library. CAS Registry Number and National Cancer Institute (NCI) numbers are common examples. Table 1.2 contains examples of commonly used names.

Table 1.2 Examples of names used for chemicals

Type	Examples
Systematic chemical names	2-amino-3-phenylpropanoic acid, 2-(acetyloxy)benzoic acid
Trivial, common, generic names	Phenylalanine, aspirin, methylphenidate, water
Trade Names	Ritalin, Concerta
Organization/Company codes	NCI455, BMS 181339-01, NSC125973
Abbreviations	AZT, DMS

Computer-readable representations of chemicals

Many software programs have been written that help scientists study molecules. These programs take a chemical as input or deliver chemical information as output. A variety of ways have been developed to format a chemical structure so that it can be used by software. A few representative ones are listed in Table 1.3. While these are formats designed for computer use, some, such as SMILES, can be composed and understood by humans, although they are rarely the preferred format for human – human information exchange.

Table 1.3 Representative computer readable structures

Type	Comments
SMILES	Line notation. A variation of the original SMILES creates unique structures.
Molfile	Connection table. Originated by Molecular Design Limited (MDL).
SDfile	Connection table; used for exchanging multiple chemicals.
InChI	Line notation. International Chemical Identifier.
InChI key	Binary form of InChI indentifier.
PDB	Protein Data Bank 3D conformation.

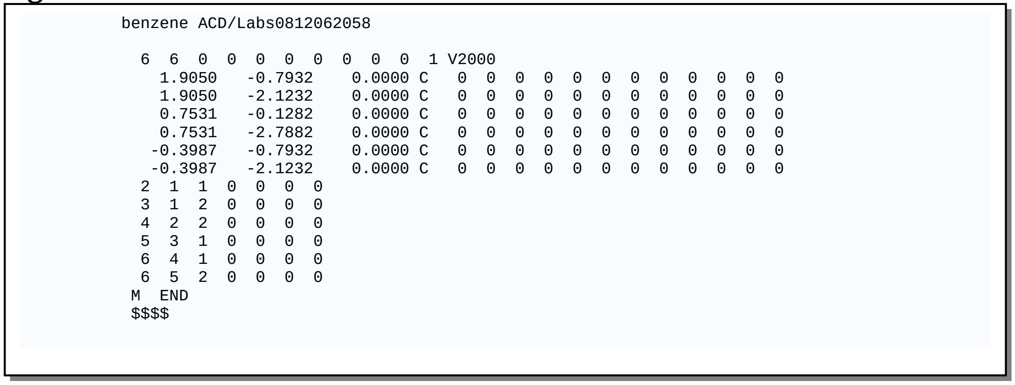
SMILES strings and InChI identifiers are both line notations, compact forms of the chemical structure that can be stored in a line of text. The InChI key is a fixed length, hashed representation of the InChI identifier, designed with the goal of making web searches faster than they were with the InChI string representation (Gasteiger & Engel, 2003). Because they are digital, they are not human-readable. Table 1.4 shows the SMILES and InChI representations for phenylalanine.

Table 1.4. Representative line notations for phenylalanine

SMILES string	O=C(O)C(N)Cc1ccccc1
InChI string	InChI=1/C9H11NO2/c10-8(9(11)12)6-7-4-2-1-3-5-7/h1-5,8H,6,10H2,(H,11,12)
InChI key	COLNVLDHVKWLRT-UHFFFAOYAL

Another general type of representation is connection tables. Connection tables store the atoms and bonds of the molecule in tabular format. Figure 1.1 contains an example.

Figure 1.1 Molfile connection table for benzene.



```
benzene ACD/Labs0812062058

  6  6  0  0  0  0  0  0  0  0  1 V2000
    1.9050   -0.7932    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.9050   -2.1232    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.7531   -0.1282    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.7531   -2.7882    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3987   -0.7932    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3987   -2.1232    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  2  1  1  0  0  0  0
  3  1  2  0  0  0  0
  4  2  2  0  0  0  0
  5  3  1  0  0  0  0
  6  4  1  0  0  0  0
  6  5  2  0  0  0  0
 M  END
 $$$$
```

All of the different forms of chemical representation have their own purpose, advantages, and disadvantages, and all have many flavors as they are extended and improved (Gasteiger & Engel, 2003).

These representations of chemicals will rarely be seen in the text of an article. There is still compelling reason to include them in this background literature review. A chemical name or identifier pulled from the text must generally be converted to one of the computer readable formats in order to use it as input to any software that performs computational routines on the molecules. In addition, it is the hope of many chemists that in the future, computer readable structures will be imbedded in the literature so that one can search the literature by structure.

1.2.2 Information Extraction

Information extraction (IE) concerns itself with finding the desired information in the text, extracting it, and (often) storing it in some kind of data structure for later use, either as input to text mining or as permanent storage, a way to make it available to others. In this regard, it can be an important component in the construction of public repositories.

Natural Language Processing

Natural language processing (NLP) techniques play an important role in many IE applications. Natural language processing is a set of computational tools employed to manipulate the text so that meaning can be extracted.

Often NLP approaches begin with preprocessing steps to reduce the volume and dimensionality of the data. A common first step is to tokenize the text, which means to break it into units called *tokens*, commonly words or punctuation. Stop words, a set of words deemed beforehand to be without semantic significance (e.g., *the, a, an, be, for,* etc.) are generally eliminated (Manning & Schuetze, 1999).

Stemming, another common technique to reduce volume and dimensionality, eliminates suffixes to create the stem form of each word. Porter’s stemming algorithm is one of the earliest and the most commonly used (van Rijsbergen, Robertson, & Porter, 1980). After stemming, the words *act*, *acted*, *acting* would be reduced to the semantic essential: *act*. Through stemming, the meaning is to a great extent retained while data dimensionality is reduced.

NLP methods can be used to parse the sentence or analyze it to determine its grammatical structure. Parsing can be performed at several levels (Shatkay & Feldman, 2003). Shallow parsing analyzes the sentence to find important parts such as the noun phrases and pull them out for further processing. Deep parsing can yield more information about the meaning of the sentence but is more computationally expensive. It turns out that significant sense can be extracted from text without parsing. The bag-of-words approach treats each word the same and instead of drawing meaning from word order and sentence structure, infers meaning from associations of words.

Named Entity Recognition

A critical component of information extraction is entity recognition or named entity recognition (NER) (Jensen et al., 2006). This task involves identifying the entities (genes, proteins, chemicals, etc.) of interest. Once an entity is identified, it is tagged with a unique, standardized identifier in a step called normalization.

Identification is fraught with difficulties because of the complex ways humans employ language to refer to things and people (Manning & Schuetze, 1999). We saw in our earlier discussion of chemical names that chemistry is no exception. Chemicals, particularly drugs, can accrue many synonyms. Polysemy, where one word can have many meanings, is a problem as well. Short forms such as abbreviations and acronyms can often be interpreted in many ways. All these wrinkles in word usage present challenges to computer algorithms.

The techniques for entity recognition can be divided into those that use external sources, such as dictionaries and lexicons, and those techniques that use only clues available in the text. The clues in the text that lead to entity identification are actually very rich and include the appearance of the word (morphology, upper case, lower case, patterns of letters, numbers, and symbols), syntax (part of speech), and the context of the word. Dictionary-based methods can be very effective, but face the challenge of needing continual updates to stay current (Jensen et al., 2006). Often combinations of dictionary and text-based techniques are used to achieve the best results.

In 1989, Hodge et al. were one of the first groups to recognize chemical names embedded in text (Hodge, 1989). Their goal was to extract the name, decipher it, and assign the correct CAS number to it. They tokenized the text, eliminated stopwords and punctuation. Nonchemical words were flagged and subsequently ignored. All remaining words were matched against a lexicon of chemical names. The maximal matching string decided the match. The CAS number stored with the matched chemical in the lexicon was indexed to the article.

Chowdhury and Lynch extended NLP techniques to patents (Chowdhury, 1992a; Chowdhury, 1992b). They analyzed the patent sublanguage and found generic terms are often used in order to gain coverage on a family of chemicals, not just a specific chemical. They tokenized the text and processed the tokens using both morphological and dictionary approaches.

While the aforementioned approaches rely on hand-crafted rules, many groups have implemented machine learning approaches. While exact implementations vary, these methods involve composing a vector for each term. The positions in the vector contain numeric values that indicate features of the term such as word length, number of digits, number of dashes, whether the term has a Greek symbol, etc. A training corpus is used as input to a classifier, such as Naïve Bayes or support vector machine (Chang, Schutze, & Altman, 2004). The advantage to machine learning is that the algorithms are not subject-area specific and therefore can be implemented in various fields. Machine learning approaches have similarly been applied to disambiguating genes, proteins, and mRNA (e.g., (Hatzivassiloglou, Duboue, & Rzhetsky, 2001)) and to deciphering abbreviations in biomedical text (e.g., (Yu, Kim, Hatzivassiloglou, & Wilbur, 2007)).

In 1999 Wilbur and colleagues from the National Library of Medicine (NLM) and National Center for Biotechnology Information (NCBI) compared three methods for identifying chemical names in text (Wilbur et al., 1999), with the goal of improving tools offered by the NLM such as MetaMap, which had historically showed weaker performance in chemistry than in other biomedical fields. One was lexically-based and the other two were flavors of Bayesian methods. The lexical method started with a list of chemical morphemes or name segments. Words from the test corpus were analyzed to find segments that matched the chemical morphemes. The algorithms matched the longest left most segment and moved across the word from left to right checking each segment. This routine was designed to handle IUPAC nomenclature. Trade names and generic names have no such regular construction and required handling by construction of their own morpheme dictionary and by lookup in NLM's Medical Subject Heading (MeSH) database. Regular expressions were also used to match patterns common in semi-systematic names. For instance, 3'5'-dichloromethotrexate could be recognized by pattern matching to the 3'5' component and then lookup in MeSH would identify the remainder of the term. All three methods produced satisfactory results, but one of the statistical methods slightly outperformed the others. Acronyms and abbreviations were a weak point for the lexical method.

Zimmermann, et al. modified their ProMiner literature mining system to work on chemicals (Zimmermann et al., 2005). ProMiner was originally designed to identify genes and proteins. Because the system was dictionary-based, they customized it for the chemical literature by developing a specialized dictionary of chemical terms drawn from MeSH and ChEBI (Degtyarenko et al., 2008). The system performed well on trivial and generic names, but the long, complex IUPAC names with their braces and parentheses proved a challenge to their tokenizing algorithms.

Translation of extracted entities

A key step in entity extraction in chemistry is to translate the chemical name into structure or the tructure into name, and either into a unique identifier such as a CAS number or SMILES string.

Early progress in automation of the translation process came in the 1960's with the work of Eugene Garfield (Garfield, 1964). He formulated a methodology to translate a systematic chemical name in the literature to its corresponding molecular formula. Garfield built a dictionary of morphemes or name segments used in systematic names. When given a word, his algorithm would search the dictionary for the morphemes in the name, and then decide whether the morphemes were indicating a structure formation or a structural modification. This algorithm was put to use when Garfield produced the *Index Chemicus*.

In contrast to Garfield's dictionary-based methods, Cooke-Fox et al. employed grammar-based techniques (Cooke-Fox, Kirby, & Rayner, 1989a; Cooke-Fox, Kirby, & Rayner, 1989b; Cooke-Fox, Kirby, & Rayner, 1989c). They created a formal grammar for the IUPAC nomenclature that allowed them to build structure diagrams from the names. They added routines to handle semi-systematic names and specialist nomenclature.

In the 1970’s as a part of a comprehensive name editing system, Vander Stouw et al. developed parallel techniques for translating CAS nomenclature into structures in the form of atom-bond connection tables, the format used as input to the CAS registry system (Vander Stouw, Naznitsky, & Rush, 1967). CAS nomenclature differs somewhat from IUPAC, and for a number of years linguistic methods applied to IUPAC were paralleled by researchers working in or closely with CAS.

A number of projects have addressed the translation of the graphical representation of a chemical structure printed in a journal article into a computer readable format. The CLiDE (Chemical Literature Data Extraction) project is the most extensive (Ibison et al., 1993). Started in 1990 at the University of Leeds under A. Peter Johnson, this project looked broadly at scientific articles and developed a methodology to understand the structure of the whole article and then to break it into pieces in three main steps. First, they analyzed the article and identified its physical layout. The program then processed and recognized each of the primitives or basic components. From this information, the program was able to determine the logical layout, what component was what: introduction, body, structural image, etc. Logical objects were associated with certain characteristics that were signals as to their type: font (size, type, style such as bold), alignment (justified, centered, flush left or right), position, and relative alignment. Once the software understands the document, the chemical structures are recognized and decomposed in a manner similar to the way the document was decomposed. The pieces of the structural depiction are analyzed to find lines, wedges, and chemical name strings. CLiDE produces a connection table which can then be used as input to a chemical drawing program. CLiDE is now maintained and distributed by SimBioSys, Inc.

Kekule, a software package developed in the early 1990’s by McDaniel and Balmuth, has similar goals, but does not as broad a broad scope and focuses on structural images alone (McDaniel & Balmuth, 1992). Kekule takes a scanned image and applies optical character recognition and rule-based logic to create connection tables that are then entered into a database.

Gkoutos et al. shifted the focus from scanned journal articles to the web and argued the need for structures to be embedded in HTML as vector images (Gkoutos, Rzepa, Clark, Adjei, & Johal, 2003). This format allowed attachment of metadata that could be read and interpreted by a computer. They tested two already known programs for converting raster images to SVG (scalable vector graphics) and got promising results with simple chemicals.

In a recent project Hattori and colleagues describe an application that mines patent applications to predict the key compounds (Hattori, Wakabayashi, & Tamaki, 2008). A patent may list an extensive number of compounds that are structurally similar but often only one or two are key, or the most important to the patent seeker. Medicinal chemists often have the job in industry to read the patents and discern the key compounds. Hattori’s theory was that the listed compounds will cluster around the one or two key compounds. They extracted the compound names from the patent text, converted them to structures, and measured and plotted the chemical similarity between them. The plots showed definite clusters. They achieved significant recall of key compounds by identifying the central point in the cluster and mapping it back to the molecule name.

Beyond chemical entity: properties

The chemical entity, whether extracted as a name and translated into a structure, or vice-versa, is the desired outcome of many extraction projects. Other researchers, however, see it as only the beginning of the extraction process. Many researchers aim to extract reactions, chemical or physical properties, biological activity, or patent claims along with the chemical.

Zamora and Blower developed a methodology to extract chemical reactions from the full text of ACS journal articles (Zamora & Blower, 1984a; Zamora & Blower, 1984b). They closely analyzed paragraphs describing synthesis reactions from the *Journal of Organic Chemistry* and determined there was a very predictable pattern in the way reactions were reported. Their routines examined the structure of the paragraph as well as the structure of each sentence to look for keywords and syntactic clues. Their goal was to extract reactants, reagents, quantities, and conditions, including solvents, temperature, equipment used, time, etc. and to populate a data structure with the results.

In their ChemXtreme application, Karthikeyan et al. mined the World Wide Web for very specific physical properties (Karthikeyan, Krishnan, Pandey, & Bender, 2006). The process started by feeding a list of chemicals to the Google search API. This Google routine retrieved all the URL addresses indexed to the selection terms and passed them to a client process that downloaded the pages and combed them for information fitting a set of templates or regular expressions. Text matching the patterns was extracted and placed in a database. A few of the physiochemical properties they extracted were LC50, LD50, melting point, freezing point, and density.

Murray-Rust and colleagues at Cambridge have created OSCAR (Townsend et al., 2004), an extraction program with a variety of capabilities. It not only recognized chemical names, but also found and extracted results from a wide variety of laboratory tests such as mass spectroscopy and NMR. Their methods are lexical, but also include extensive use of pattern matching routines that take advantage of the highly structured reporting of lab results.

Beyond chemical entity: relationships

Another important goal of information extraction is to find *relationships* between entities: between genes to understand expression patterns, between proteins to build protein interaction networks, and in the realm of drug research, between genes and drugs, and drugs and disease.

Two main processing approaches have been used to extract relationships from biomedical text: co-occurrence and NLP. Co-occurrence methods look for entities that appear together in sentences, titles, abstracts, or Medline records. The underlying premise is that if two things are mentioned in proximity then they are likely related. While generally a robust technique, co-occurrence based approaches suffer from two main weaknesses. First, entities that are not related can indeed be co-mentioned. Additionally, even if the entities are related, we gain no information on the nature of the relationship (Jensen et al., 2006).

NLP techniques can examine syntax and semantics and can both establish relationships with higher accuracy, and determine in many cases what kind of relationship exists. To do the latter, they look for specific verbs such as *inhibit, phosphorylate, activate* (e.g., (Blaschke, Andrade, Ouzounis, & Valencia, 1999)), or identify patterns in the entity-verb occurrences (e.g., (Rindflesch, Tanabe, Weinstein, & Hunter, 2000)). NLP methods have their disadvantages as well. They are generally tailored to specific applications and therefore do not generalize well to other biomedical areas. Because they depend on sentence structure, they do not perform well when finding relationships between sentences. Co-occurrence methods can find relationships beyond the sentence boundary and are often general enough to translate between specialties (Jensen et al., 2006).

Rindflesch et al. use NLP techniques to extract very specific information about drugs from Medline abstracts: the interaction of drugs and genes in cancer cells (Rindflesch et al., 2000). They parsed the text and tagged parts of speech. The identified noun phrases were matched against the UMLS Metathesaurus (Bodenreider, 2004) to find drug names. The program identified cells and genes using knowledgebases in addition to contextual information. The output of the application is a first order calculus statement expressing the drug/gene entities and their relationship. The example below shows how the software captures the relationship between the cells (HAG/src3-1), the drug (CDDP) and the gene(v-src).

> *Original sentence:* "Compared with parental or mock-tranfected HAG-1 cells, v-src-transfected HAG/src3-1 cells showed a 3.5-fold resistance to cisdiamminedichloroplatinum (CDDP)."
>
> *Extracted relationship:* I_resistant(v-src,HAG/src3-1,CDDP)

Future Directions

The open science movement reflects a changing attitude toward the dissemination of information by scientists in many domains. Led by a few far-sighted individuals, chemistry, too has started to embrace the tenets of open science, although the field still lags behind biology and bioinformatics. Peter Murray-Rust, Henry Rzepa, and others have promoted a vision of a Chemical Semantic Web (Murray-Rust, Rzepa, Tyrrell, & Zhang, 2004; Murray-Rust, Rzepa, Stewart, & Zhang, 2005). In this vision, the primary communication of chemical information would be journal articles published on the web with CML (Chemical Markup Language) (Gkoutos, Murray-Rust, Rzepa, & Wright, 2001; Murray-Rust & Rzepa, 2001; Murray-Rust & Rzepa, 2003; Murray-Rust, Rzepa, Williamson, & Willighagen, 2004). The rigorous use of CML would make the articles machine understandable. The authors use the term "datuments" to illustrate the combination of documents and data. In these datuments, each mentioned chemical would be accompanied by a machine-understandable depiction of the structure (InChI string or connection table). If this vision were realized, the sophisticated named entity recognition routines of the past would no longer be necessary. Chemical property data would be equally transparent. The CML schema would ensure that each reported data element follow a particular structure and be expressed in a standard vocabulary. Data types, data values and the associated limits can be checked and validated by the restriction expressed in the schema. The data could be accompanied by metadata indicating quality, provenance, or key words for later retrieval.

This vision would require the concerted effort and support of many chemists and the cooperation of far-sighted publishers. While these forces are coalescing, Murray-Rust et al. argue that the most important intermediate step is that chemists make their data available at the time of publication. Data, they point out, is not copyrightable, and for the most part publishers are not interested in publishing the complete data associated with an article, so they have nothing to lose. Murray-Rust et al. recommend that authors submit their data to a public or institutional repository under the Open Access protocol. This is not an outlandish request. In the bioinformatics field, authors have for years submitted protein and nucleic acid sequences to public repositories such as GenBank as a requirement of publication.

While the techniques and technology have changed over the years, the motivation behind information retrieval and extraction in chemistry has fundamentally not changed: the need to answer questions about chemicals.

1.2.3 Text Mining

Text mining, another important subtask in literature mining, finds new knowledge in the literature. It is often preceded by information retrieval and information extraction. Often the extracted information is put into some sort of data structure to facilitate the mining activity.

Text mining can enable the practitioner to take a bird's eye view of the literature. This perspective allows connections to be made between facts in one document and facts in another. The documents may have been written in different decades by people in different scientific disciplines, but through text mining the connections can be brought to light where they can be examined and evaluated. This computer-assisted observation can reveal relationships that would have been difficult or prohibitively time consuming to find manually. Text mining can also find patterns in large sets of data – in this regard text mining is closely akin to data mining. The bird's eye view can pick out correlations, associations, and trends not possible to see when examining documents individually.

These two characteristics of literature – its rich connections and its patterns – have been used to discover new things, and, specifically, to find new therapeutic uses for drugs. Don Swanson pioneered the understanding of literature connections and their potential in uncovering new knowledge (Swanson, 1990). His literature-based discovery work and the work of the researchers who followed in his footsteps will be discussed in depth. Before that discussion, however, we will look at the smaller body of work that uses patterns in side effects to predict new uses for drugs.

Text mining and adverse events

A drug can have both targeted, desired effects on an organism, and undesired effects, called side effects or adverse events. Several research groups have shown that the array of side effects attributed to a drug can indicate what molecular interactions it has, particularly what receptors it binds. Fliri et al. converted the side effects available through the CEREP Bioprint database (Krejsa et al., 2003) to create binary descriptor sets or side effect spectra (Fliri, Loging, Thadeio, & Volkmann, 2005). They clustered the spectra and found that drugs with similar known molecular mechanisms had similar side effects. They point out that understanding this relationship between molecular mechanisms and side effects may help drug developers avoid drug candidates with high risk for undesired effects.

In a more recent study, Campillos et al. used side effect information to infer off-target binding (Campillos, Kuhn, Gavin, Jensen, & Bork, 2008). They retrieved package insert text files from a variety of sources such as the FDA and manufacturers' websites. The section of the package inserts listing side effects was extracted and parsed. Terms were matched against a dictionary they had assembled from the UMLS (National Library of Medicine, 2006) and COSTART (Food and Drug Administration, 1989). Presence or absence of each side effect was coded in a binary fashion. They developed a side effect similarity measure and used it to make pairwise comparisons of each drug in their reference set to every other drug. The measures were adjusted to account for very common side effects, very rare side effects, and side effects with a high correlation (nausea and vomiting, for instance). In addition to the side effect similarity measure, they calculated the structural similarity of each pair of drugs using the Tanimoto (Willett, Barnard, & Downs, 1998) method. The known protein targets of each drug were downloaded from online databases including Matador (Gunther et al., 2008), DrugBank (Wishart et al., 2006), and PDSP K_i (Psychoactive Drug Screening Program database) (Roth, Lopez, Patel, & Kroeze, 2000). They clustered the drugs by side effect similarity and structural similarity and looked for pairs which had a high side effect similarity but did not show significant structural similarity. They wanted to reduce the weighting of pairs with structural similarity, a known predictor of similar biological activity. They also eliminated pairs found to bind to the same proteins. What remained were pairs of drugs with similar side effect profiles, but no other known indicators of similar molecular activity. For instance, the Alzheimer's treatment donepezil was found to have a very similar side effect profile to the antidepressant venlafaxine, but structurally they are diverse, and donepezil has not been known to bind to proteins associated with depression. A protein binding assay performed by the authors showed donepezil to have affinity for the 5HTT receptor, a key receptor in depression treatment. In total they identified 261 drugs with possible novel targets. They tested twenty drugs and found 13 of them active in *in vitro* binding assays. The activity of nine of these was confirmed in cell assays, and the study resulted in two new patent applications.

Literature-based discovery and Swanson

Swanson, a researcher in information science, developed a methodology for literature-based discovery based on his observations of scientific literature (Swanson, 1990). He noted that the increasing specialization of scientists was paralleled by an increasing specialization in scientific journals. He described a situation where scientific domains no longer interacted through the reading and publishing of their literatures: researchers reading and publishing in one set of journals were not aware of articles in other journals. The literatures become islands and, in Swanson's terms, *non-interactive*.   This situation, according to Swanson, creates the potential for knowledge to go unconnected, relationships not recognized and inferences not made, a situation he termed *undiscovered public knowledge*. Swanson demonstrated that these connections might be made using through literature mining. Using his ABC literature-based methodology he made several discoveries, among them a connection between Reynaud's disease and fish oil (Swanson, 1986) and the potential of magnesium to treat migraines (Swanson, 1988). Swanson emphasized that literature-based methods only assisted with hypothesis generation or hypothesis support, and that any hypothesis derived from the literature, must, like any other, be substantiated by experimental science.

Swanson's ABC methodology starts with identifying a disease or condition of interest. As an example we will consider migraine. The term *migraine* becomes the C term. In the next step, the literature is searched for terms that co-occur with *migraine*. These are the intermediary B terms and include, in the case of migraine, terms such as *spreading cortical depression*, *vasoconstriction,* and *vasodilation.* The B terms can be seen as terms for physiological conditions or states or processes that underlie the disease state. In the next step potential treatments – the A terms – are identified by finding drug or chemicals associated with any of the B terms. Next the C – A connection is tested and the only potential treatments retained for further examination are those that have not yet been explicitly linked to migraine.

The best hunting ground for finding this undiscovered knowledge is in what Swanson termed *complementary but disjoint* literatures. Complementary but disjoint literatures have common areas or subjects that can provide rich opportunities for linkages. The literature describing diseases for instance, can contain many descriptions of molecular or physiological phenomena that accompany the disease. Drug researchers may quite independently write about molecular or physiological phenomena that are modulated by a particular drug. No one may have thought to search the literature exhaustively for a link between the disease and drug. A link is implied, however, if there is common ground, and a novel hypothesis could be in the making. Finding an implicit connection between two things based on an examination of the explicit connections is the fundamental notion behind ABC.

The ABC paradigm has two approaches, termed by Weeber et al. as *open* and *closed* (Weeber, Klein, de Jong-van den Berg, & Vos, 2001). The open approach starts with a concept of interest such as a disease and proceeds through the steps described above. The *closed* approach starts with a hypothesis (e.g., drug A treats disease C) and looks for B terms connected to both A and C that may support or explain the link from A to C.

Although literature-based drug discovery has generally followed Swanson’s footsteps, the ABC method has been adapted and implemented in a variety of ways. Swanson himself in collaboration with Smalheiser extended and automated his methods in an application called Arrowsmith (Smalheiser & Swanson, 1998) and continued to find novel connections (Smalheiser & Swanson, 1996a; Smalheiser & Swanson, 1996b). The subsequent implementations of the ABC method retain the essential technique of using explicit connections to find implicit connections, but creative and increasingly rigorous enhancements have emerged. The next section of this literature review will discuss the major themes in the adaptation of Swanson's groundbreaking methodology.

Paradigms

Often the adaptations of Swanson's ABC recast the paradigm in terms of other analytical models in order to take advantage of the properties and methods associated with those models. The A, B, and C terms, for instance, may be depicted as nodes in a mathematical graph model and the relationships between them considered the edges. Both Wren et al. (Wren, Bekeredjian, Stewart, Shohet, & Garner, 2004) and Narayanasamy et al. (Narayanasamy, Mukhopadhyay, Palakal, & Potter, 2004) employ this terminology. In the development of their Transminer application, Narayanasamy and colleagues take advantage of graph terminology, properties, and visualization techniques. Concepts extracted from the literature become nodes and known associations between concepts are identified by co-occurrence in the literature and depicted as edges. Moving along the edges from one node to another is termed traversing the graph. Possible new associations are identified through transitivity, a property of graphs that maintains if A is related to B and B is related to C, then A is related to C. Stated in this way, it is clear how effectively graph terminology not only describes Swanson's ABC, but also extends it, as graphing can include many more than three nodes and transitive closure can posit an implicit relationship after transversal of many nodes.

Similar to graph models, network models are useful in literature-based discovery. Seki and Mostafa employed a formal information retrieval model called the *inference network* (Seki & Mostafa, 2007). The network they depict has nodes and edges, but has more inherent structure than the graph model of Narayanasamy. The network's nodes are typed and arranged in layers according to type. In the information retrieval context, top and bottom level nodes would represent the user query and the documents in the collection, respectively. Intermediate nodes represent key words in the documents. When they apply this model to searching for genes related to diseases, disease and genes take the outside positions and gene functions and disease phenotypes are represented by the intermediate nodes. This depiction again is more extensive than the ABC paradigm of Swanson, but the principles of relating concepts and entities are the same.

Corpora

Researchers in literature-based discovery in biomedical science generally choose some part of Medline (National Library of Medicine, 2008) as a corpus. Medline is the most comprehensive bibliographic source of biomedical literature. It is also free. Medline is compiled by the U.S. National Library of Medicine and includes articles from over 5,000 journals. As of this writing, it contains records for more than19.5 million articles. Medline can be downloaded from the NLM and loaded into a local database for access or it can be accessed through the PubMed Entrez browser (Wheeler et al., 2008).

Medline contains language structured in two distinct ways. The title and abstract are in natural language, usually English. The Medline record also contains the structured MeSH annotations attached to each record by indexers at the NLM. These annotations are selected from a controlled vocabulary.

Researchers who select title and abstract as their corpus often employ natural language processing (NLP) methods to turn the language into data. Ahlers et al. use NLP to extract the semantic relationships from abstract text (Ahlers, Hristovski, Kilicoglu, &

Rindflesch, 2007) . Lindsay and Gordon used the word tokens to create bigrams (two word combinations) and trigrams to use as their units of analysis (Lindsay & Gordon, 1999). They based this choice on the observation that many medical concepts comprise more than one word. In a similar vein, Weeber et al. mapped the tokens of the title and abstract to concepts in the Unified Medical Language System (UMLS), a thesaurus of medical terms provided by the NLM (Weeber et al., 2001). Using the UMLS has another advantage: terms that map to its entries have medical significance. Terms that do not map to the UMLS are more likely outside the medical domain and less likely to be of interest and therefore can be eliminated.

MeSH terms are another corpus selected by many researchers in literature-based biomedical discovery. The MeSH vocabulary has its own hierarchical ontology in the Trees database, but the MeSH terms are also a component of the UMLS. This gives the researcher using MeSH the ability to sort and filter the MeSH terms. Srinivasan (Srinivasan, 2004) bases her system on MeSH terms and uses their relationship to UMLS to help rank them. Yetisgen-Yildiz and Pratt similarly extract MeSH terms and then use the UMLS to filter them (Yetisgen-Yildiz & Pratt, 2006). Hristovski et al. use MeSH and restrict their extraction to only MeSH headings that the annotators flagged as major headings (Hristovski, Stare, Peterlin, & Dzeroski, 2001).

Data reduction and focus: relevance

Once the data or units of analysis are gathered, a number of methodologies are employed for defining a relationship between data elements. Co-occurrence is behind them all.

The sheer volume of articles in Medline means that whether natural language or MeSH is selected as a corpus, the combinatorics of connecting one concept to another will mount up and the volume of data will be large. Many techniques are employed by researchers with the aim of finding those connections that are both interesting and significant.

The task of finding what is interesting starts with the user. In every implementation of literature-based discovery, the user specifies a starting point such as a disease. Often the user controls other decisions beyond the starting direction. In the work of Lindsay and Gordon (Lindsay & Gordon, 1999) and Weeber et al. (Weeber et al., 2003) the user plays a large role in making decisions about which intermediary terms will be investigated further. In (Weeber et al., 2003), the central role of the user-expert is demonstrated as the authors investigate novel therapeutic uses for thalidomide. Their decisions to pursue one set of linkages over another based on prior knowledge is considered essential to the utility of the application. In a recent paper by Petrič, et al. the researchers limit terms to the rarest ones, based on the idea that rarity may indicate novel and innovative information, and then they use subject area experts to select the intermediate terms linked to the rare terms (Petrič, Urbančič, Cestnik, & Macedoni-Lukšič, 2008).

The UMLS concept types or concept groups are used to designate the domain and direction of the exploration in (R. N. Kostoff, Briggs, Solka, & Rushenberg, 2008; Srinivasan, 2004; Weeber et al., 2001; Yetisgen-Yildiz & Pratt, 2006). In the LitLinker system of Yetisgen-Yildiz and Pratt, for instance, the user controls the domain and the direction of discovery by specifying the UMLS concept groups for the starting, linking, and target terms. (In Swanson's paradigm these are the C terms, B terms, and A terms.) Through the software's user interface, the user can designate a starting concept such as a disease, then select the category such as physiological conditions as the linking or intermediary terms, and finally specify genes as the category for the target concepts. Similarly in (Srinivasan, 2004) the user specifies what profiles are to be constructed and analyzed.   Wren et al. (2004) (Wren et al., 2004)start with the construction of a dictionary that contains only those terms they are interested in. They pull diseases from OMIM (Hamosh, Scott, Amberger, Bocchini, & McKusick, 2005), genes from Locuslink (Pruitt, Katz, Sicotte, & Maglott, 2000) and chemical names from MeSH. Terms outside their dictionary are ignored by their algorithms.

By allowing the user to concentrate the literature extraction to terms that are interesting and relevant the volume of data is reduced considerably. However, the resulting connections may still number in the thousands, and some mechanism to rank the results is a crucial part of most literature-based discovery implementations. Through ranking the output, researchers attempt to put the most promising connections at the top. Estimating the importance or significance of a connection is challenging and it has been approached in various ways.

Yetisgen-Yildiz and Pratt (2006) rank the target (or C) terms in order of the number of linking (B) terms that connect the C term to the A term. Then they apply a threshold level to eliminate low scoring terms. Hristovski et al. (2001) have a pre-calculated set of association rules that establish the significance of a co-occurrence of two terms. Each association has a support and confidence level that can be used both as a screening metric and a ranking metric for the final output. Lindsay and Gordon (1999) use frequencies of terms. They found relative frequencies perform best in ranking C terms. Their frequency calculations rely on metrics commonly used in information retrieval such as tf*idf (token frequency * inverse document frequency). Srinivasan (2004) computes weights for the

MeSH term profiles in the intermediate steps and the final list is ranked by combining these weights.

Wren et al. (2004) calculate what they call the strength of the relationship between entities. They rank the relationships they find against a random network of relationships to estimate the significance of the relationship. Input requires the co-occurrence count.

All literature-based discovery applications aim to produce hypotheses. There is a wide variation in the extent to which the final list of hypotheses has been influenced more by user input or statistics. In all implementations, the user selects the hypotheses deserving of further study.

Validation and Evaluation

Validation is a challenge for discovery systems because, if the system works, it is by definition finding something unknown (Yetisgen-Yildiz & Pratt, 2006). The most common approach to validation has been to treat Swanson's discoveries as the gold standard and reproduce them. This approach is taken by (Lindsay & Gordon, 1999; Srinivasan, 2004; Weeber et al., 2001). A key requirement to using a previous discovery as a gold standard is to limit the input data to a timeframe before the discovery was explicitly known and written about.

A variation of this approach is to divide the corpus into two groups based on a pre-selected date. Hypothesis sets can be produced on the earlier baseline period and tested against the later period to see if the implicit connections derived from the earlier data are explicitly present in the second period. Yetisgen-Yildiz and Pratt (2006) use this approach to test LitLinker. They used the cutoff date January 1, 2004 and concentrated on finding implicit relationships in three disease areas: Alzheimer's disease, migraine, and schizophrenia. Then they examined the literature between January 1, 2004 and September 30, 2005 to ascertain how many of the identified implicit relationships became explicitly stated in the literature. They measured their results using precision and recall and were able to track changes in precision and recall over time. In a similar vein Hristovski et al. (2001) picked a baseline and test time frame and tracked connections in terms associated with ten different diseases. They found they were quite good at finding future connections, but their hypothesis sets were too large to be useful, so they tested various thresholds to lower the number of hypotheses.

In an experimental approach to validation, Wren et al. (2004) take advantage of their expertise as laboratory scientists and test their hypothesis that chlorpromazine can treat cardiac hypertrophy by conducting experiments on mouse models of the disease. Narayanasamy et al. both reproduce the magnesium-migraine connection, and, for their other cancer gene hypotheses, rely on the verification by experts in the field (Narayanasamy et al., 2004).

Because disparate methods have been used by authors to evaluate their literature-based discovery systems there has been to date no way to compare the efficacy of applications. In a very recent paper, Yetisgen-Yildiz and Pratt describe promising methodologies to remedy this situation (Yetisgen-Yildiz & Pratt, 2009). They base their recommendation on four principles. First, 1) the quality of all target terms or hypotheses should be evaluated, not just those that replicate the gold standard. 2) The evaluation of a system should be based on multiple experiments, not just one. 3) Evaluation should be independent of prior knowledge in order to avoid bias. Many literature-based discovery systems require a human expert to decide on which the intermediate or linking terms should be selected, a step open to bias if the expert knows the desired outcome of the experiment. Last 4), an evaluation method should allow valid comparison of different systems.

Guided by these principles, Yetisgen-Yildiz and Pratt describe performance metrics that can be adopted by any researcher whose application produces a set of hypotheses upon which recall and precision can be calculated. In essence these metrics go beyond measuring precision and recall for the complete set. They recommend measuring precision and recall at increments on a ranked set to evaluate how effectively the ranking algorithms place the most important and relevant terms at the top.

Future Directions

In her recent review, Bekhuis discussed the progress in literature-based discovery since Swanson's early work (Bekhuis, 2006). Her comments are a good starting point to assess the progress in the field and important future directions. She cites system appraisal as a problem for developers. There are few choices for evaluation of systems because the yardsticks are few. She implied that more known discoveries to use as gold standards would be an asset for researchers to validate their systems. With the lack of agreed upon yardsticks, division of data into time periods is a good alternative, especially since the technique can be applied to any area of science. Certainly the recommendations of Yetisgen-Yildiz and Pratt (2009), if implemented by future researchers, will go a long way toward satisfying Bekhuis' concerns.

Bekhuis also encourages developers of literature-based discovery systems to participate with research teams and work on substantive problems rather than methodological problems. This will help garner credibility to the field and gain the attention of the wider biomedical research community. Bekhuis speculated about what why the research in literature-based discovery was so little known outside the field of information science. Biology has a solid foundation on experimental, empirical science. The notion that experiments can be conducted on data alone, even when the data was collected by other researchers, is a difficult paradigm shift for many scientists.

Concern for this hurdle has been discussed by others. While there are still many scientists who are skeptical about experimenting on data, there are those advocating it and proposing new names for it. Bray describes the shift between biology being a data-collection science to hypothesis-driven science where the hypotheses may be the result of reasoning from pre-existing data and those who make and test the hypotheses may not be those who wielded the pipette in the lab (Bray, 2001). Blagosklonny and Pardee (Blagosklonny & Pardee, 2002) agree with Bray and emphasize that computational biology or conceptual biology, as they term it, is not a distinct type of science, but just has a different source for its data: information in databases.

1.3 Conclusion

Complete and accurate information is as critical to chemists as it is to practitioners in any other scientific field. The landscape of chemical information is undergoing rapid and fundamental changes. Central to this change is the move to publicly accessible information on the web. Here the number of chemical entities is growing at a rapid rate, and the biological effects and activity resources are expanding to new areas. This comprehensive and interconnected chemical information, founded as it is on rich data, should ensure that the rate of acquiring new knowledge will increase as well.

2. PILOT STUDY

2.1 Introduction

This dissertation research was preceded by a pilot study with two goals. The first goal was to build a repository or knowledgebase of terms extracted from the literature that represent the bioactivity and effect of the chemicals, particularly drugs. It was hypothesized that this repository, called ChemoText, could be used in drug research to predict new uses for drugs. The second goal of the pilot study was to test this hypothesis by implementing a version of Swanson's ABC methodology. This implementation would use the data in the ChemoText repository to find implicit links between entities and generate predictions for new uses for drugs - drug reprofiling.

This dissertation work builds on the fundamental research conducted in the pilot study. ChemoText is the source of data for the studies under both aims of this dissertation, and the first aim will extend the ABC study conducted in the pilot. Section 2.2 outlines the steps taken to design and build ChemoText. Section 2.3 presents the pilot implementation of the ABC methodology.

2.2 Construction of ChemoText

2.2.1 Corpus and Theory

Text extraction requires a corpus. The corpus selected for this research was the annotation section of Medline records. Medline (National Library of Medicine, 2008) is the database of bibliographic information created and maintained by the National Library of Medicine (NLM).

Medical Subject Headings (MeSH) are keywords added to Medline records by trained annotators at the National Library of Medicine in order to facilitate search and retrieval. The annotators choose the terms that reflect the main points of the article from a controlled vocabulary. The headings can be accompanied by subheadings or qualifiers. These terms, also selected from a controlled vocabulary, reflect what aspect of the heading is under study. For example, an article that discusses the origins of Huntington Disease might be annotated with *Huntington Disease/etiology*. A heading may be accompanied by several subheadings or none at all.

When an article discusses chemicals, a Registry Number (RN) entry is included in Medline. Although not strictly MeSH annotations, these lines are also extracted in the course of this project. For brevity, both the RN and MeSH terms will be referred to collectively in this work as MeSH annotations.

MeSH terms have been written about extensively, both with regard to their function in search and retrieval, as well as their usefulness in other database and computational applications (Bodenreider, 2008). Funk and Reid looked at the quality of MeSH annotations using inter-annotator agreement as a measure of quality (Funk & Reid, 1983). Kostoff et al. in (R. N. Kostoff, Block, Stump, & Pfeil, 2004) evaluated the information contents of MeSH and title to see if they approximated the information of the abstract.

Advances in computational linguistics in tandem with the steep increase in journal articles have spurred research on replacing manual indexing with automatic methods, work spearheaded by the National Library of Medicine (Aronson et al., 2000; Neveol, Zeng, & Bodenreider, 2006). The goal of the work is to build software that can assign MeSH headings that result in retrieval performance equal (or better than) the current manual indexing.

MeSH has been evaluated in the context of statistically-based information retrieval applications. Rubin et al. compared the efficacy of several feature sets in computationally retrieving articles about pharmacogenomics and found that MeSH terms compared favorably in their discriminative power to terms extracted from the natural language of the abstract and title (Rubin, Thorn, Klein, & Altman, 2005). This finding is similar to that of Chen et al., who compared disease-drug relationships extracted from full text of articles and clinical narratives to MeSH and UMLS annotations. They concluded that the two sources produced consistent and complementary results (Chen, Hripcsak, Xu, Markatou, & Friedman, 2008).

Of more relevance to this project, MeSH terms have been extracted by a number of developers to create a knowledgebase for biomedical applications. Cimino and colleagues studied MeSH extensively and observed patterns they were then able to capitalize on in constructing an evidence-based medicine knowledgebase (Cimino & Barnett, 1993; Mendonca & Cimino, 2000). His group’s observations of the relationship between MeSH headings and subheadings in a Medline record were built on and extended by researchers in literature-based discovery, e.g., (Hristovski, Friedman, Rindflesch, & Peterlin, 2006; Srinivasan, 2004).

Cimino’s tactic was to look very closely at the headings and subheading co-occurrence patterns in a limited domain, in this case clinical medicine, formalize them, and then attach meaning to them. Cimino and colleagues were able to do this semantic analysis and rule-building because they restricted their domain. If they had attempted to observe patterns in the whole of Medline, important patterns may have been obscured.

The more one restricts the domain, the more one can say about it. This is the essence of the theory of sublanguage, the theory that explains why close study of a restricted domain yields patterns that can be exploited in computational linguistic methods (Harris, 2002). The rationale behind the theory is that people who work in a specialized area develop language patterns to help them communicate effectively (Haas, 1997). In the case of Cimino and colleagues, as in this research, the theory is being extended from natural language to annotations of natural language.

The pilot project restricted the domain to articles (or annotations of articles) about chemicals. The terms targeted for extraction were restricted as well, to annotations indicating chemical activity and effect. It was hoped that this narrow focus would yield strong signals useful in drug research.

2.2.2 Analysis and Design

The analysis and design stages of development started with observing and recording the patterns in a small subset of articles. Once the terms indicating chemical effect and activity were identified, algorithms were developed to extract them. The algorithms were tested on the initial small test set and then implemented on the entire Medline corpus.

The sublanguage observations in the pilot study were based on a sampling of 125 randomly chosen articles about the chemical genistein. Genistein is a chemical found in soybeans that has been studied for its connection to a number of diseases, particularly its potential to treat cancer. Just one chemical was chosen in order to get a well rounded view of the types of research a chemical undergoes. A number of articles reported the results of *in vitro* experiments such as protein-binding assays and cell assays where the molecular activity of the drug is studied. Studies on whole organisms were also present, both on animal models such as rabbits and in human clinical trials.

The 125 sample Medline records were printed, read, and the MeSH terms were manually extracted, tabulated and compared to the contents of the abstract and title. This dataset was termed the *PMID125Set.*

The MeSH terms indicating biological activity became quickly apparent when the PMID125Set was examined. They included protein annotations, disease annotations, and the group of biological effects identified by the *drug effects* annotations.

On the molecular level, protein annotations stood out. A protein is a large molecule constructed of amino acids. The proteins in the human body are ubiquitous, and in addition to being vital structural elements, play many active roles in metabolism, signaling, growth and development. Proteins are the targets of most drugs. The goal of a drug is to bind to and modulate the activity of a protein, in order to suppress or enhance its activity. A large body of research concentrates on studying the relationship between drugs and proteins.

The PMID125Set contained 304 instances of protein annotations. This represents 180 unique names, many of which are protein family names (e.g., *Kinases*) rather than the name of individual proteins. Ninety-five of the 125 articles had at least one protein annotation. The most commonly occurring entry was *Protein Tyrosine Kinase* with 15 appearances, followed by *Receptors, Estrogen* with 12 occurrences. Several specific names like *NF-kappa*

B are included in the list, but so is the extremely general term *Proteins*. Table 2.1 shows the most commonly annotated proteins.

Table 2.1 Top most common protein annotations in the PMID125Set

Protein Name
Protein-Tyrosine Kinases
Receptors, Estrogen
Cyclin-Dependent Kinase Inhibitor p21
NF-kappa B
Proliferating Cell Nuclear Antigen
Tumor Necrosis Factor-alpha
Caspase 3
Caspases
CF Transmembrane Conductance Regulator
DNA Binding proteins
Receptor, Epidermal Growth Factor

Table 2.2 Counts of top 5 disease/condition annotations in PMID125Set

Disease/Condition	Count
Breast Neoplasms	18
Prostatic Neoplasms	7
Body Weight	6
Adenocarcinoma	4

Disease annotations were a significant indicator of drug activity. Disease annotations were found in 69 of 125 articles (53.6%). A total of 111 disease annotations were extracted, representing 57 unique diseases. The most common disease annotation in the PMID125Set was *Breast Neoplasms*, one of the many forms of neoplasms mentioned in the articles. Table 2.2 lists the top four most frequently occurring diseases in the PMID125Set.

The diseases were identified by looking up the headings in the MeSH Tree file. This data source is a hierarchical ontology available from the NLM. The category C contains diseases and conditions, signs, and symptoms such as *Body Weight*. For brevity we will refer to this collection of terms as *diseases*.

The articles with disease annotations fall into somewhat distinct categories. Many of the articles state in their introductory remarks that genistein is known to have action against a particular disease (breast cancer, for instance) and, given that, the research of the paper endeavors to understand either how or why (mechanisms) and when (under what conditions).

Other articles start by discussing a molecular level activity genistein is known to have and then test the drug against a new disease for which this activity might prove fruitful. In one article, for example, the researchers note that genistein has been shown in previous studies to have anti-inflammatory activity and then test whether this activity might extend to beneficial results in treating *alopecia areata* (hair loss) in the mouse model.

In most cases the subject drug was under study as a treatment for the annotated disease. In some cases however, the article reported that genistein caused a disease or had particular adverse effects. The patterns in the annotations indicate to a great extent whether the drug treats or causes the disease. For instance, when the subject drug was annotated with either *adverse effects* or *toxicity*, it was reported to cause the disease. When the subject drug was annotated with *therapeutic use* or the disease was annotated with *prevention & control*, the drug is generally discussed as a treatment for the disease. The combination of the drug annotation *toxicity* with the disease annotation *chemically induced* was a strong contextual marker for indicating the paper described the drug as causing the disease. Other researchers have noted these patterns in pairs of annotations, e.g., (Mendonca & Cimino, 2000). As an illustration, consider PMID 12132873. In this study the authors fed mice special diets with varying amounts of genistein and daidzein. They found that the incidence of vulvar carcinomas was associated with the amount of the drugs in the diet. The relevant annotations for genistein were *Genistein/*toxicity* and *Vulvar Neoplasms/*chemically induced/pathology*. Patterns in the annotations were used to categorize and tag the disease terms into *treat* or *cause* categories. Of the 111 disease annotations, 16 were tagged as *cause*, among them several forms of neoplasms.

The next area of the Medline record containing evidence of drug activity is the qualifier *drug effects.* Drug effects annotations were found in 90 articles out of 125, with an average of 2.7 per article. Two hundred forty-five separate headings associated with the effects were extracted, representing 152 unique annotations. Table 2.3 lists the most commonly occurring headings paired with drug effects. *Cell Division* tops the list with 21 occurrences followed by *Apoptosis* with twelve.

Table 2.3 Most common headings co-occurring with drug effects annotations in PMID125Set

MeSH Descriptor	Count	Pct
Cell Division	21	8.6%
Apoptosis	12	4.9%
Endothelium, Vascular	5	2.0%
Uterus	5	2.0%
Gene Expression Reg., Neoplastic	4	1.6%
Cell Cycle	4	1.6%
Phosphorylation	4	1.6%

Table 2.4 Top occurring drug effects categories in PMID125Set

MeSH Descriptor	Count	Pct
Biological/Cell Phys. Phenomena, Immunity	64	16%
Physiological Processes	50	13%
Genetic Processes	39	10%
Cells	38	10%
Biochem.Phen., Metabolism, Nutrition	22	6%
Urogenital System	19	5%
Tissues	17	4%
Amino Acids, Peptides, and Proteins	16	4%

The records were examined for false positives, records that code for drug effect but the article reports that the drug has no effect, and three such instances were found. PMID 16557470 is an example. Genistein was investigated to see if it had an effect on cell proliferation and on mammary glands. The study confirmed the latter but found no effect of genistein on cell proliferation. Automated routines cannot discern these negative results at this time and will include these incorrect drug effects. It is likely that so few false positives were found because negative results are not published at the same rate as positive results, and, particularly with the comparative studies, the heading linked to *drug effect* is often general, indicating the direction of the research presented in the paper.

To determine whether the drug effects describe drug activities from a wide spectrum of physiological levels, each drug effect annotation was looked up in the National Library of Medicine's MeSH Tree file. This file contains all the MeSH annotations arranged in a tree structure that allows one to travel from a given annotation to a higher node in the tree that represents a family or category to which the annotation belongs. The effect *Apoptosis* (programmed cell death) for instance can be mapped to the more general term *Cell Physiological Phenomena*. Table 2.4 contains the categories and the number and percentage of annotations falling into each, and shows that the entries are distributed among a number of physiological levels.

The Medline record can list more than one chemical. One or more of them may be the subject of the research, while other chemicals are peripheral, perhaps discussed or used in the experimental procedure, but not the central object of study. In order to reduce the volume of data to remove incidental chemical annotations it was important to identify the chemicals that were the subjects of the article and then associate the activity terms *only* with the subject chemical(s). A heuristic algorithm was developed that evaluated the MeSH subheadings or qualifiers occurring with the chemical annotations and identified the chemicals most likely to be the subjects. The heuristic followed a rule-based stepwise procedure, a procedure developed based on the detailed analysis of the PMID125Set. In this process, the annotations from each Medline record were examined to see if more than one chemical was annotated and identified as a major topic. If only one chemical was found and major, it was tagged as the subject chemical. If more than one chemical was identified as major, then the subheadings or qualifiers of each were examined. If the subheadings were the same for each of the chemicals, then they were all tagged as subjects.

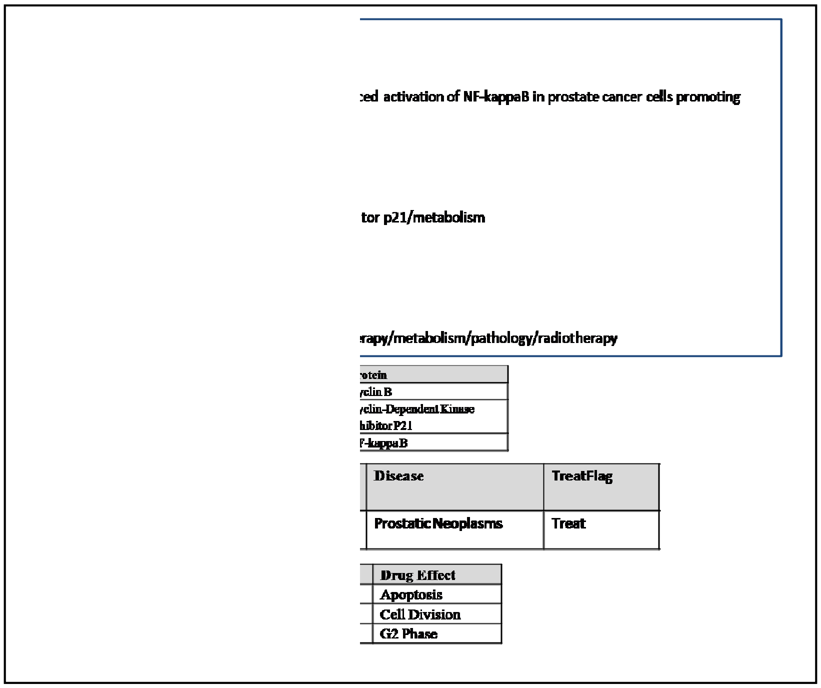
Table 2.5 Hierarchy of MeSH subheadings (qualifiers) when establishing subject chemicals

Level	MeSH subheadings
1	*Pharmacology* OR *Adverse Effects* OR *Therapeutic Use* OR *Administration & Dosage* OR *Toxicity* OR *Pharmacokinetics*
2	Any subheadings except *Biosynthesis*, *Metabolism*, *Chemistry*
3	*Biosynthesis* OR *Metabolism* OR *Chemistry*

Preliminary analysis of the PMID125Set showed that certain subheadings were more commonly associated with subjects then other headings. *Pharmacology*, *therapeutic use*, and *administration & dosage*, for instance, are subheadings commonly annotated to the subject chemical, while the subheadings *metabolism* and *biosynthesis* are less common annotations for subject chemicals. A hierarchy of subheadings was assembled, starting with those most commonly associated with subjects to those rarely seen associated with subjects. (See Table 2.5.) This hierarchy was used to compare the chemicals in the remainder of the records and tag those most likely to be subjects. Only chemicals flagged as major in at least one of their subheadings are used as input to the algorithm. If a subheading from level one was found, the associated chemical(s) were designated subjects. Only if no chemical had a subheading from the first group did the algorithm look at subheadings from the second group. If no chemicals have been identified annotated with subheadings from the first two groups, then chemicals tagged with a subheading from level 3 were tagged as subjects.

Medline records with more than one subject are common. Forty percent have more than one subject chemical, and the average number of subject chemicals per Medline record is 1.65. In the next step of the processing each of the subject chemicals was associated with the previously extracted activity and effects terms. Figure 2.1 below shows the MeSH annotations for one sample Medline record and the ChemoText database records produced from it.

Figure 2.1 Sample Medline record with MeSH annotations and the resulting database records in ChemoText.



:ed activation of NF-kappaB in prostate cancer cells promoting

tor p21/metabolism

rapy/metabolism/pathology/radiotherapy

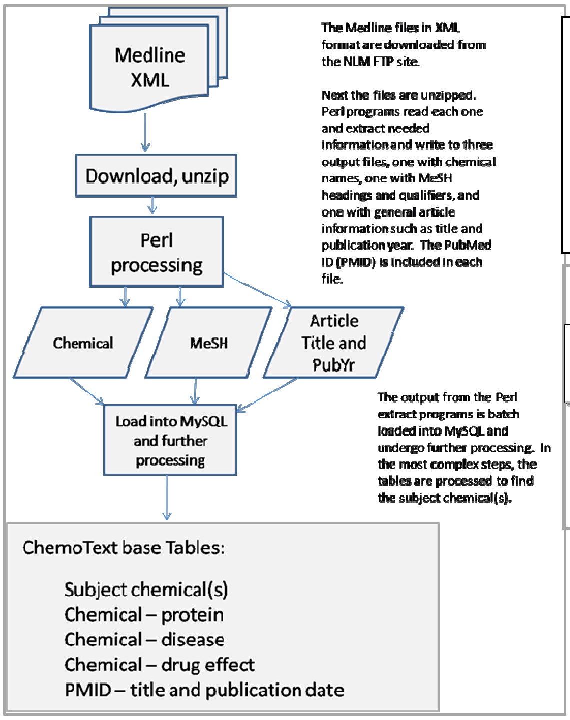
rotein
yclin B
yclin-Dependent Kinase hibitor P21
F-kappaB

Disease	TreatFlag
Prostatic Neoplasms	Treat

Drug Effect
Apoptosis
Cell Division
G2 Phase

2.2.3 Construction

The 2008 baseline version of Medline was downloaded from the National Library of Medicine web site and used as the corpus for extraction routines. The baseline files consist of over 500 zipped XML files. Once the files were downloaded and expanded, the extract routines were run on each. The extraction routines were written in Perl. The data was loaded into a MySQL database and subsequent processing was performed in SQL. The processing steps are illustrated in Figure 2.2, and the completed database depicted as a network is shown in Figure 2.3. The diagram shows the number of unique entities in each category as well as the number of relationships between entities stored in the database, which was named ChemoText. The baseline file contained 16,880,015 records; 6,635,344 records had identified subject chemicals and were included in ChemoText.



The Medline files in XML format are downloaded from the NLM FTP site.

Next the files are unzipped. Perl programs read each one and extract needed information and write to three output files, one with chemical names, one with MeSH headings and qualifiers, and one with general article information such as title and publication year. The PubMed ID (PMID) is included in each file.

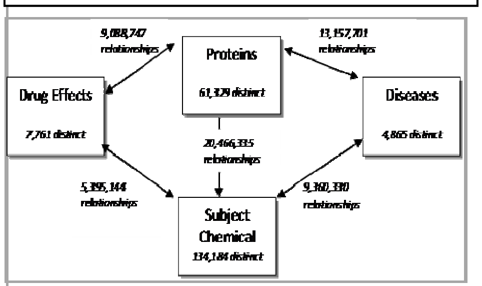
The output from the Perl extract programs is batch loaded into MySQL and undergo further processing. In the most complex steps, the tables are processed to find the subject chemical(s).

ChemoText base Tables:

Subject chemical(s)
Chemical – protein
Chemical – disease
Chemical – drug effect
PMID – title and publication date

/iew

xt.



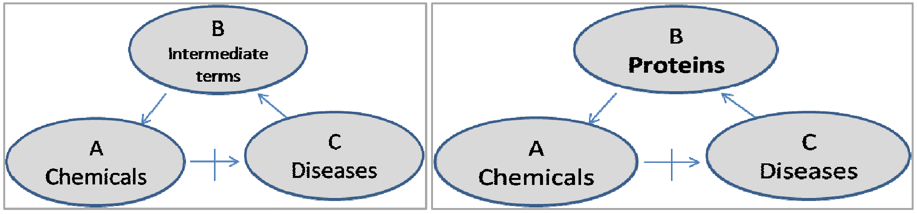
2.3 Drug Discovery Application

The potential of using ChemoText for drug discovery was explored in the next phase of the pilot study. The goal was to generate a list of chemicals linked implicitly but not explicitly to a particular disease through the literature. Such a list or hypothesis set may contain chemicals important to drug research either as new treatments or as key chemicals in the physiology of the disease. To generate the hypotheses, the ABC methodology (described in Chapter 1) of Swanson (Swanson, 1988) was adopted.

2.3.1 Methods

The implementation of Swanson's ABC paradigm using ChemoText incorporated several features that differentiate it from other implementations. A critical design decision made at the onset was to limit the B terms (also called linking or intermediate terms) to protein annotations. See Figure 2.4 below. This limitation was applied not only to reduce the volume of data, but also because proteins are the agents behind most physiological processes and are therefore studied both by scientists researching disease and by scientists looking at drugs. Because these very different groups of scientists may not be aware of each others' work, there is a strong potential for finding undiscovered implicit relationships between drugs (A terms) and diseases (C terms) via proteins (B terms).

Figure 2.4 On the left, Swanson's ABC paradigm. On the right the design for this study: protein annotations only were used as the B terms.



In order to facilitate validation of the results, the common literature-based methodology of identifying a cutoff date and dividing the data into a pre-cutoff set and a post-cutoff set was adopted. This segmentation meant that a hypothesis set could be constructed from the earlier set and then validated by looking at the results in the second, later set. Because the study used migraine as the disease and 1985 as the cutoff year, the study was additionally able to attempt a reproduction of Swanson's link between migraine and magnesium.

The first article to directly connect magnesium to migraines was published in 1985. The routines were limited to evidence before that year for the baseline data. The ChemoText database was queried for all articles published before 1985 in which *migraine disorders*, *migraine with aura*, or *migraine without aura* were included in the MeSH annotations. These were the C terms. In the next step each protein annotation included in any of these articles was extracted. This was the pool of proteins associated with migraine (B terms). This pool contained 131 proteins and included names for specific proteins as well as protein families (e.g., *Receptors, Adrenergic*). The next step extracted any chemical that was identified as the subject of a study in which any of the migraine pool proteins was annotated. Chemical family names such as *Amines* or *Lactones* were programmatically eliminated to reduce the data volume and because this study seeks new uses for specific chemicals, not chemical families. The resulting set of terms were the A terms. The number of migraine pool proteins associated with each chemical was counted. Any chemical from this list which already had a direct link to migraine was eliminated.

The entire ChemoText database was examined to determine which chemicals predicted to have a link to migraine based on the evidence of the baseline period did indeed have literature evidence of a link by the test period. The most common MeSH subheadings appearing with these chemicals when they were annotated with migraine were also extracted to help elucidate what kind of link emerged.

2.3.2 Results

The experiment produced a list of 4,725 chemicals potentially connected with migraine. (See Table 2.6 Part A.) We term this list the hypothesis set. When the set was ranked by protein count (*Prot Ct*), magnesium appeared near the top of the list at position 3. This closely reproduces Swanson's discovery.

Table 2.6 Comparison of baseline and test periods. Ranked by protein count the top 12 chemicals out of 4,725 that are predicted to have a connection to migraine based on their associations with migraine proteins before 1985. Part A contains information available in ChemoText during the baseline period before 1985. Part B contains data extracted from ChemoText in the test period.

A. Baseline Data: 1984 and before			**B. Test Data: After 1984**			
Rank	**Chemical Name**	**Prot Ct**	**First Yr**	**Article Ct**	**Disease Qualifier**	**Chemical Qualifier**
1	Sodium	104	2006	1	blood	cerebrospinal fluid
2	Zinc	93	0	0		
3	Magnesium	91	1985	39	blood	blood
4	Copper	88	1986	1	etiology	adverse effects
5	Corticosterone	86	0	0		
6	Prednisolone	84	2007	1	complications	therapeutic use
7	Cysteine	81	1994	3	radionuclide imaging	analogs & derivatives
8	Edetic Acid	80	1989	1	physiopathology	admin & dosage
9	Lead	79	0	0		
10	Colchicine	77	0	0		
11	Cyclic GMP	76	1995	4	physiopathology	physiology
12	Nicotine	75	1999	3	drug therapy	adverse effects

Many researchers have reproduced Swanson's magnesium – migraine discovery; thus the results are not novel, but can be viewed as a method validation. However, the design of ChemoText enabled an extension of this analysis in a novel direction. For each chemical in the hypothesis set, the ChemoText database was searched for any link between the chemical and migraine after 1984. These results were summarized and combined with the results from the baseline period. Table 2.5 Part B contains these new columns: *First Year* (abbreviated *First Yr*, the first year an article appeared directly associating the chemical to migraine), *Article Count* (abbreviated *Article Ct*, the count of articles with this direct association) and the most common qualifiers or subheadings (based on occurrence counts) appearing in the annotations of the disease and the chemical with migraine (*Disease Qualifier* and *Chemical*

Qualifier). Magnesium was first connected to migraine in 1985 and has had 39 articles since connecting it to migraine. Both the most common disease qualifier and the most common chemical qualifier occurring in records in which migraine and magnesium occur together were *blood*, indicating the blood levels of magnesium are important in migraine.

The set was examined to see what general observations could be made. The set contains many types of chemicals. Sodium, zinc, copper and magnesium are elements. Cysteine is an amino acid and cyclic GMP is a nucleotide. Pharmaceuticals become more common as one scans down the list. The disease and chemical qualifiers indicate that the connections between the chemicals and migraine were varied. A number of chemicals were annotated indicating they treat migraine. Some chemicals like copper apparently cause migraine, and some appear to be involved in the physiological mechanisms of migraine (e.g., cyclic GMP).

The total set contained 154 chemicals that had no connection to migraine in the baseline period but developed a connection by 2007. Among the top 12 chemicals, eight (66%) have developed links to migraine since 1984. The *Article Count* element was adopted as a rough indicator of the significance of a chemical's connection to migraine. Magnesium has had 39 articles linking it to migraine since 1985 while copper has only one since its first connection in 1986. Sodium has only one article linking it directly to migraine, but the article is recent therefore the connection is newly established and its significance as of today is understandably low.

Table 2.7 Baseline and test period results for valproic acid and nitric oxide. Ranked by protein count. Sections of the output set containing valproic acid and nitric oxide, two chemicals with high article counts in the test period. Part A contains information available in Medline during the baseline period before 1985. Part B contains data extracted from Medline records in the test period.

A. Baseline data: 1984 and before			**B. Test Data: After 1984**			
Rank	**Chemical Name**	**Prot Ct**	**First Yr**	**Article Ct**	**Disease Qualifier**	**Chemical Qualifier**
103	Mannitol	44	0	0		
104	Penicillin G	43	0	0		
105	**Valproic Acid**	**43**	**1988**	**83**	**drug therapy**	**therapeutic use**
106	Deuterium	43	0	0		
107	Aluminum	42	0	0		
108	Orotic Acid	42	0	0		
	…		0	0		
598	Quartz	11	0	0		
599	**Nitric Oxide**	**11**	**1991**	**40**	**physiopathology**	**physiology**
600	Orciprenaline	11	0	0		
601	Methaqualone	11	0	0		

Based on the article count metric, two chemicals, valproic acid and nitric oxide, warrant further discussion. (See Table 2.7) Valproic acid, found in position 105, has only 43 migraine-related proteins. The first article discussing its therapeutic use in migraine appeared in 1988 and by 2007, 83 articles linked valproic acid to migraine, twice as many as magnesium. Valproic acid is an example of drug re-profiling. It was used for many years as an anti-epileptic drug before being tried in migraine prophylaxis (Sorensen, 1988). Valproic acid developed the strongest link to migraine based on the article count metric, yet it did not appear as high as magnesium in the hypothesis set based on baseline protein count.

Nitric oxide appears relatively low in the list as well at position 599, linked to only 11 proteins in common with the pool of migraine-linked proteins, but by 2007 it had 40 articles linking it to migraine, one more than magnesium. Nitric oxide is an important signaling molecule in the body, and the qualifiers in the last two columns indicate that this chemical plays a role in the physiopathology of the disease.

Precision and Recall

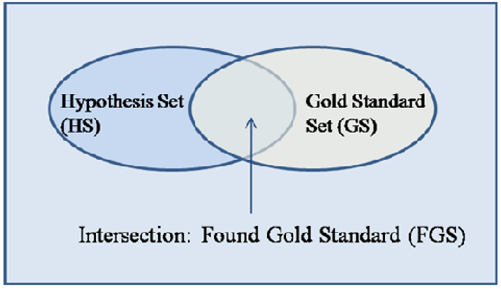
Precision and recall were calculated using the following formulas.

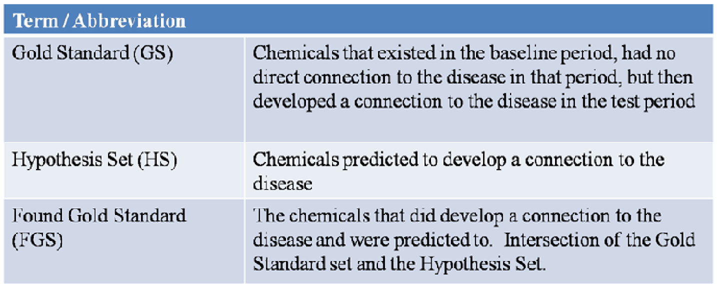
Chemical Precision= $(HS \cap GS) / HS$ and

Chemical Recall: $(HS \cap GS) / GS$ (1)

HS is the number of entries in the hypothesis set and GS stands for the number of *g*old *s*tandard chemicals, the chemicals that the experiment ideally should have predicted. GS chemicals are those that existed in the baseline period, and had no direct link to migraine during that period, but by the end of the 1985-2007 test period had developed a direct link to migraine. There were 177 total GS chemicals; our routines found 154 of them. The 23 chemicals were missed because they did not have proteins linked to them from the migraine protein pool. In other words, the B – C connection did not pick up these chemicals. The intersection of the hypothesis set and the GS chemicals gives the number of GS chemicals found by our experiments. The variables used in the prediction of precision and recall are summarized in Figure 2.5.

Figure 2.5 Explanation of chemical sets and term definitions





Term / Abbreviation	
Gold Standard (GS)	Chemicals that existed in the baseline period, had no direct connection to the disease in that period, but then developed a connection to the disease in the test period
Hypothesis Set (HS)	Chemicals predicted to develop a connection to the disease
Found Gold Standard (FGS)	The chemicals that did develop a connection to the disease and were predicted to. Intersection of the Gold Standard set and the Hypothesis Set.

The results for recall and precision are as follows.

Chemical Precision = $\frac{154}{4725}$ = 0.033 = 3.3% Chemical Recall = $\frac{154}{177}$ = 0.870 = 87.0%

The recall results are high. Selecting migraine drugs based on proteins identified 87% of the future chemicals connected to migraine. Our precision results, however, are weak. Only 3.3% of the chemicals in the hypothesis set developed a connection to migraine after 1984.

One likely reason for the low precision is that the 131 proteins connected to migraine include many protein families. These annotations can be very general and therefore have the likelihood of being annotated with many chemicals. For instance, *Adenosine Triphosphatases* and *Peptide Hydrolases* are two protein annotations from the migraine protein pool. While these families certainly have a connection to migraine, they are so broad that they will have connections to many other diseases and chemicals. As a result they will likely significantly increase the size of the hypothesis set with chemicals of little potential connection to migraine. Not all protein families can be discounted, however. *Receptors, Serotonin* is also a protein family, but it has a well-known importance to the physiology of migraine and should not be undervalued. In future work we hope to develop other metrics that attribute a weight to the protein annotations that will reflect their importance to the disease being investigated.

Increasing Precision

The relationship between protein count and the strength of the connection of a chemical to migraine was investigated. To reflect the importance of the connection between a chemical and migraine the article count metric was used. This metric acts as a weighted count, giving chemicals a weight equal to the number of publications connecting them with migraine. Counting co-occurrences to estimate relationship strength is a common technique in text mining (e.g., (Stapley & Benoit, 2000)). Using article count, however, does have limitations. It is a direct measure of publication activity, and publications may not always accurately reflect significance of a chemical as a potential treatment for a disease. Publication rates may increase, for instance, if a certain drug is suspected of having dangerous side effects. Additionally, a chemical that has ten articles connecting it to migraine cannot be said to be ten times more important than a chemical with only one article. Despite these limitations the article count metric will be used as a rough indicator for the importance of a connection between a chemical and migraine.

For a graphic understanding of these relationships between protein count (the number of proteins from the protein pool associated with the chemical in the baseline period), the hypothesis set chemicals and the gold standard chemicals, a bar chart was generated that grouped the hypothesis set by protein count ranges. (See Figure 2.6.) For each protein count range, the following percentages were depicted as bars: the percentage of the hypothesis set, percentage of gold standard (GS) chemicals, and percentage of gold standard articles. The graph shows that over 80% of the hypothesis set chemicals have fewer than 10 proteins linking them to migraine. This large group has around 50% of the future linked chemicals. However, this group only has around 30% of the articles linking chemicals to migraine. Because so many chemicals in the hypothesis set had fewer than 10 proteins, a separate bar chart (Figure 2.7) was created to look at the 0-10 range in detail. This graph shows that over 40% of the chemicals in the hypothesis set had only one protein from the migraine protein pool. This large group contained only 10% of the true migraine chemicals and less than 5%

of the migraine articles. Eliminating this group of chemicals could improve precision without significantly degrading recall.

To test this idea, precision and recall were recalculated as the chemicals with the lowest protein counts were consecutively eliminated. The results are contained in Table 2.8.

Figure 2.6 Bar chart showing percentages by protein count. HS – count of hypothesis set chemicals. GS is count of gold standard chemicals. Art Ct is article count.

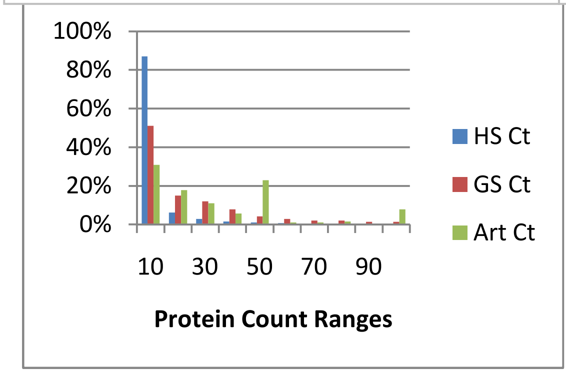
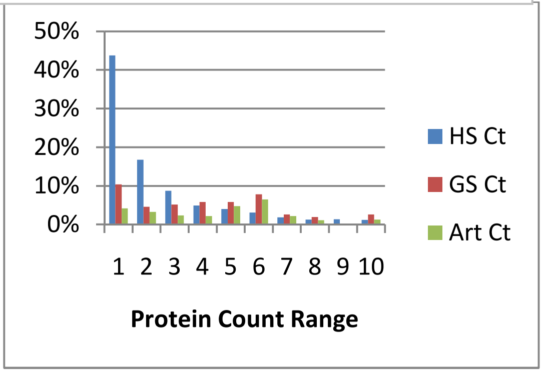


Figure 2.7 Bar chart showing percentages for chemicals with 10 or fewer associated proteins.



This table includes a new element: A*rticle Recall.* To calculate this we used the following formula.

Article recall = *(Found GS Articles) / (All GS Articles)* (2)

We will illustrate this formula using the results from the entire hypothesis set.

Article recall = 552/(552 + 55) = .909 = 90.9%

The numerator in this equation is the number of articles associated with the 154 chemicals from our hypothesis set that did indeed develop a future link to migraine and are in the gold standard set. The denominator is the number of articles for the gold standard chemicals in our hypothesis in addition to the 55 articles associated with the 23 chemicals that the routines did not find. Article recall overall was 90.9%. Article recall is higher than chemical recall because the chemicals we did find on average had more articles associated with them then the chemicals we did not find.

Table 2.8 Precision and recall results as thresholds are applied

Threshold Applied	Hypothesis Set Count	Found GS Chemicals	Found GS Articles	Precision	Recall	Article Recall
none	4725	154	552	0.03	0.870	0.909
protct > 1	2658	138	529	0.05	0.780	0.871
protct > 2	1867	131	511	0.07	0.740	0.842
protct > 3	1454	123	498	0.08	0.695	0.820
protct > 4	1223	114	486	0.09	0.644	0.801
protct > 5	1034	105	460	0.10	0.593	0.758
protct > 6	888	93	424	0.10	0.525	0.699
protct > 7	801	89	412	0.11	0.503	0.679
protct > 8	739	86	406	0.12	0.486	0.669
protct > 9	674	86	406	0.13	0.486	0.669
protct > 10	617	82	399	0.13	0.463	0.657

Table 2.8 records the change in precision and recall as protein count thresholds were applied to the hypothesis set. The elimination of each group of chemicals caused an increase in precision and a decrease in recall. By eliminating all chemicals with 10 or fewer proteins, the hypothesis set contains 617 chemicals. Of these 82 or 13% are future linked. While the chemical recall was decreased to 46.3%, the article recall only decreased to 65.7%, showing that the chemicals remaining had a more significant connection to migraine as measured by article count. The three chemicals that eventually developed the strongest link to migraine (magnesium, nitric oxide, and valproic acid) are all included in the set of 617, although nitric oxide, with only 11 chemicals from the protein pool, was close to the cutoff. Our results on

the whole compare favorably to other similar studies (Hristovski et al., 2001; Yetisgen-Yildiz & Pratt, 2006).

2.3.2 Evaluation of pilot study and next steps

The pilot study was successful in revealing both strengths and weaknesses of both ChemoText and the drug discovery application. The ABC implementation using ChemoText was able to reproduce Swanson's link between magnesium and migraine.

The strategy of using proteins as the intermediate B terms was effective in creating a hypothesis set with high recall. The reason for this likely lies in the central role proteins play in both disease and drug research. The study of disease increasingly focuses on the physiology of the disease state at the molecular level, a level in which observations of proteins and their interaction with other molecules is central. Drug research focuses on proteins as well, searching for drugs that modulate the behavior of proteins involved in the disease pathway.

While recall was high, precision was low. The technique of applying cutoffs to the protein counts improved precision, but still left large hypothesis sets. Metrics other than protein count may be more effective in ranking the hypothesis set and putting the best candidates near the top. There are many examples in the literature of rankings based on weighted counts of connecting terms that could yield better results. This dissertation research will investigate other ranking approaches.

When other metrics are explored in ranking the hypothesis set, there must be a way to evaluate the results of each ranking so that they can be rigorously compared to find the best. The methods outlined by Yetisgen-Yildiz and Pratt in a recent paper form the basis for such a line of evaluation (Yetisgen-Yildiz & Pratt, 2009). The methods involve calculating metrics that measure how well the ranking approach puts the relevant (i.e., future-linked or gold standard) entries toward the top of the ranked hypothesis set, where they are more likely to come to the notice of researchers. The metrics are Precision@K, MAP, and 11-point average precision. These metrics have been adopted from the field of information retrieval and are used to evaluate the performance of IR applications such as search engines.

The goal of this dissertation is to produce text mining applications that could be adopted as tools in the computational drug research laboratory. That will only happen if that application can be rigorously validated and the results comprehensively evaluated. The new implementation of this ABC study will concentrate on developing these validation and evaluation components.

3. EXTENDED IMPLEMENTATION OF SWANSON'S ABC METHODS

3.1 Introduction

In this study the explicit connections between entities in the biomedical literature were used to identify implicit connections between biomedical entities. These implied connections are potential new discoveries. Specifically, the co-occurring annotations between diseases, proteins, and chemicals were examined to find implied connections between chemicals and disease, and therefore to predict new uses for existing drugs or drug reprofiling.

This work extended the pilot study. The pilot study implemented Swanson's ABC paradigm using the MeSH annotations extracted from Medline records and stored in ChemoText. In the pilot the most significant design strategy introduced was to limit the B intermediary terms to protein annotations. This strategy was very effective and was retained for this research. The reason for the success in using proteins as intermediary linking terms likely lies in the central role proteins play in both disease and drug research. The study of disease increasingly focuses on the physiology of the disease state at the molecular level, a level in which observations of proteins and their interactions with other molecules are central. Drug research focuses on proteins as well, searching for drugs that will modulate the behavior of proteins involved in the disease pathway.

The validation approach used in the pilot study was also retained. In that approach the corpus was divided into two sets by a cutoff year. The data from the early time period was used to create the discovery hypotheses and data from the later time period was used to validate the hypotheses.

This study went beyond the pilot work in its scope. Three diseases were included and three year cutoffs were applied to each. New approaches were used to rank the hypothesis set and the rankings were evaluated using techniques adopted from the information retrieval field, techniques that evaluate how well the ranking places the most important or relevant chemicals at the top of the returned list.

3.2 Overall Design

The diseases chosen for this study were cystic fibrosis, psoriasis, and migraine. Migraine was chosen in order to reproduce and extend the pilot study. Cystic fibrosis was selected because it is a very serious rare disease with few successful treatments. Psoriasis provides a contrast to cystic fibrosis; it is common, not life-threatening, and there are many treatments, although no cures. It was thought this group of diseases would provide an interesting diversity in the results.

Three cutoff points were selected: 1984-1985, 1989-1990, and 1994-1995. The 1984-85 cutoff was chosen to reproduce the pilot study. The 1989-1990 and 1994-1995 cutoffs were selected to see how the chemicals and treatments changed over time. Each year cutoff partitioned the data into two sets. The baseline set contained the data from any relevant article published in the baseline period, which is defined as any article in ChemoText with a publication year up to and including the first cutoff year (e.g., 1984). The test set contains any article from the test period. The test period includes all relevant articles published after the baseline period (e.g., 1985 and after) through 2008. Table 3.1 below contains details about each baseline and test period.

Table 3.1 Description of baseline and test period construction. In each case the baseline period starts with the earliest relevant article pulled from ChemoText before the year cutoff.

Cut-off	**Baseline period ends with and includes year**	**Test period starts with (and includes) year**	**Test period ends**
1984-85	1984	1985	2008
1989-90	1989	1990	2008
1994-95	1994	1995	2008

The combination of a disease and time period will be called a *test run*. Each test run produced a hypothesis set, or a list of chemicals found to have an implicit connection to the disease in question. The names for each test run and the datasets produced are listed in Table 3.2.

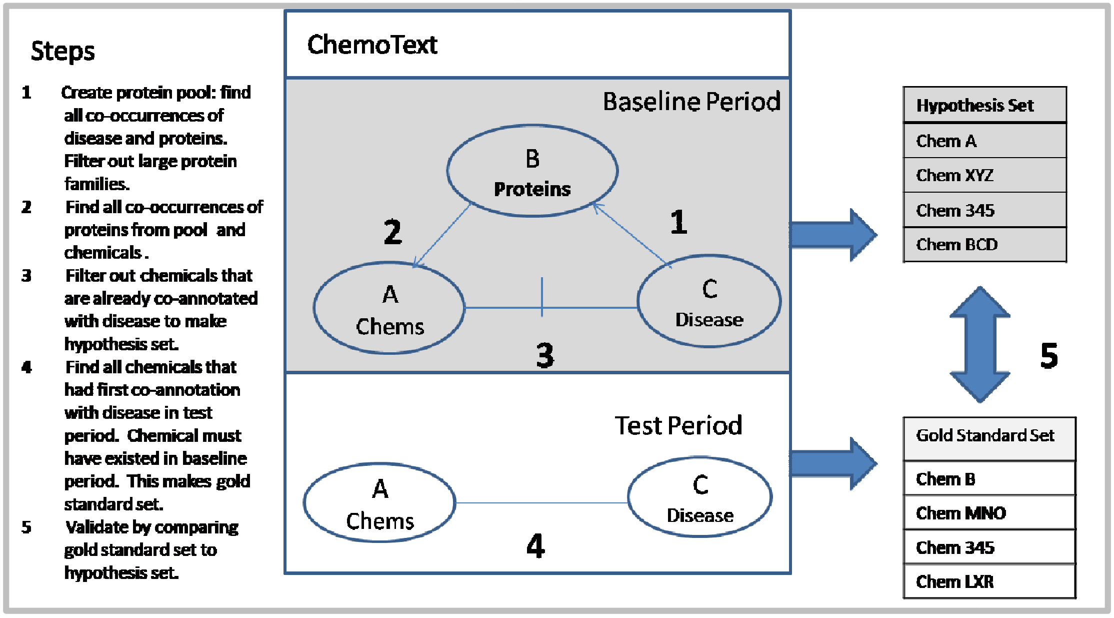
Table 3.2 Description of each test run and name of resulting hypothesis sets

Disease	Year cut-off	Test run name	Hypothesis set name
Cystic Fibrosis	1984-1985	CF 1984-85 test run	CF 1984-85 Set
Cystic Fibrosis	1989-1990	CF 1989-89 test run	CF 1989-90 Set
Cystic Fibrosis	1994-1995	CF 1994-95 test run	CF 1994-95 Set
Psoriasis	1984-1985	Psoriasis 1984-85 test run	Psoriasis 1984-85 Set
Psoriasis	1989-1990	Psoriasis 1989-89 test run	Psoriasis 1989-90 Set
Psoriasis	1994-1995	Psoriasis 1994-95 test run	Psoriasis 1994-95 Set
Migraine	1984-1985	Migraine 1984-85 test run	Migraine 1984-85 Set
Migraine	1989-1990	Migraine 1989-89 test run	Migraine 1989-90 Set
Migraine	1994-1995	Migraine 1994-95 test run	Migraine 1994-95 Set

3.3 Methods

A graphic representation of the method is presented in Figure 3.1. For each test run (disease and year cutoff), the following steps were performed. The ChemoText database was queried for any occurrence of the disease annotation with a protein annotation in any article published in the baseline period. The disease annotation for cystic fibrosis was *Cystic Fibrosis* and for psoriasis was *Psoriasis*. Three annotations were used in the case of migraine: *Migraine Disorders, Migraine with Aura*, and *Migraine without Aura*. The resulting set of proteins was then cleaned by removing protein annotations identified beforehand as being too broad to be useful. They represent large families of proteins that likely have members that play a role in most physiological processes and therefore most diseases. They would therefore provide little specific information about a disease. These annotations include terms such as *Proteins* and *Amino Acids*. The complete list of eliminated proteins is included in Appendix 1. The same list was used for each test run.

Figure 3.1 Flowchart of method. Note that the ChemoText Knowledgebase is logically divided into Baseline period and Test period.


Steps
1 Create protein pool: find all co-occurrences of disease and proteins. Filter out large protein families.
2 Find all co-occurrences of proteins from pool and chemicals.
3 Filter out chemicals that are already co-annotated with disease to make hypothesis set.
4 Find all chemicals that had first co-annotation with disease in test period. Chemical must have existed in baseline period. This makes gold standard set.
5 Validate by comparing gold standard set to hypothesis set.
ChemoText
Baseline Period
B
Proteins
2
1
A
Chems
C
Disease
3
Test Period
A
Chems
C
Disease
4
Hypothesis Set
Chem A
Chem XYZ
Chem 345
Chem BCD
5
Gold Standard Set
Chem B
Chem MNO
Chem 345
Chem LXR

The resulting list of proteins was termed the *protein pool*. For each protein in the protein pool, ChemoText was again queried for co-occurrence between the protein and a chemical annotation in an article published in the baseline period. The resulting dataset was summarized by adding up the number of proteins from the pool linked to each chemical and storing the total in a variable called Protein Count (ProtCt). To reduce the number of entries and to try to find only the significant co-occurrences of protein and chemical, only those chemicals were chosen that were subject chemicals of the articles in question. (The identification of the subject chemical was described in Chapter 2.) Because this study targets specific drugs to reprofile, chemical families were eliminated from the results. Examples of chemical families are *Acids*, *Benzoflavones,* and *Hydrazines*.

It is important to note that this study is designed to focus on the classic drug type: a small organic molecule. Protein-based therapies and solutions and mixtures are excluded from the hypothesis sets.

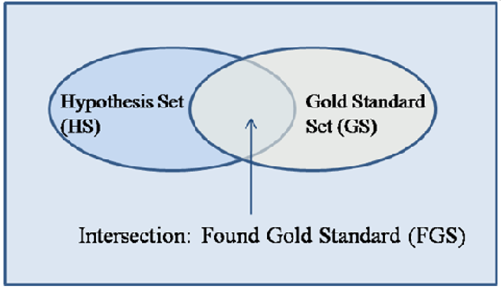
The resulting set represented the list of chemicals connected through intermediary protein annotations to the disease. In the next step those chemicals that already had in the baseline period an *explicit* or known relationship to the disease in the baseline period were eliminated and what remained was a set of chemicals with only an *implicit* connection to the disease. To find the set of known connections, the baseline period was queried for co-annotations of the chemical and the disease in the same article. Again, because of the way ChemoText was constructed around subject chemicals, this step only looked for and identified articles in which the chemical was the subject of the article and co-annotated with the disease. Chemicals found to have this connection were eliminated from the list. The resulting set of chemicals was the *hypothesis set(HS)*. These chemicals were predicted to have a connection to the disease, either as a potential treatment, an endogenous chemical playing a role in the disease mechanism, or as a causative agent.

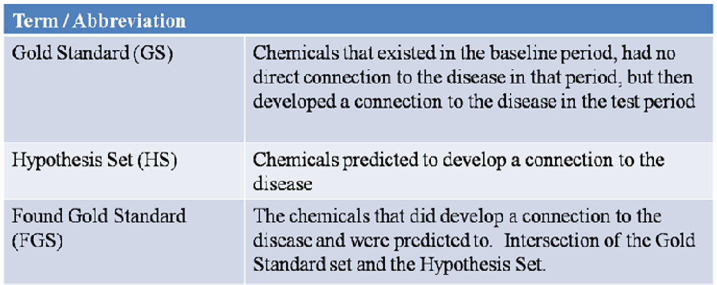
Next, ChemoText was queried for all the chemicals that represent those chemicals that *should* have been included in the hypothesis set. This set includes any chemical that existed in the baseline period, had no direct connection to the disease (that is, was never a subject chemical in an article in which the disease was annotated), but did develop a direct connection in the test period (again, as a subject chemical in an article where the disease was annotated). This set of chemicals was termed the *gold standard (GS)* set.

The chemicals in the gold standard set were further described by adding columns that helped to illuminate the link between the chemical and the disease that developed. The number of proteins linking it to the disease in the baseline period was added to the set (ProtCt). The number of articles (Article Ct or ArtCt) linking the chemical to the disease in the test period was included as well. (See Table 2.7 for an example.) Article count is a rough measure of how important the link was that eventually developed. In addition, the most common disease subheadings or qualifiers and the most common chemical subheadings annotated with the drug and disease were also collected and appended to the chemical records.

In the next step the hypothesis set was validated by checking to see which entries in the hypothesis set were also in the gold standard set. This group of chemicals represents the true positive predictions and will be termed the *found gold standard (FGS)* chemicals. The following figure depicts the hypothesis set, the gold standard set, and the intersection of the two.

Figure 3.2 Depiction of chemical sets and term definitions. The same sets and definitions were used in the pilot study.





Term / Abbreviation	
Gold Standard (GS)	Chemicals that existed in the baseline period, had no direct connection to the disease in that period, but then developed a connection to the disease in the test period
Hypothesis Set (HS)	Chemicals predicted to develop a connection to the disease
Found Gold Standard (FGS)	The chemicals that did develop a connection to the disease and were predicted to. Intersection of the Gold Standard set and the Hypothesis Set.

3.3.1 Calculation of precision and recall

Precision and recall were calculated using the following formulas.

Chemical Precision= $(HS \cap GS) / HS$ and

Chemical Recall: $(HS \cap GS) / GS$ (1)

HS is the number of entries in the hypothesis set. GS stands for gold standard, the number of chemicals which developed a link to the disease. Gold standard chemicals are those that existed in the baseline period, and had no direct link to the disease during that period, but by the end of the test period had developed a direct link to the disease.

3.3.2 Calculation of ranking variables

Each hypothesis set was initially ranked separately on three variables calculated with data elements retrieved in the baseline period. The first variable was *protein count* (*ProtCt*). This is the total number of proteins from the protein pool that are co-annotated with the chemical in the baseline period. If two chemicals have the same protein count, the value WtCOS (described below) was used as a secondary ranking value.

The next ranking approach, called *WtCOS*, was devised to rank high the chemicals with a protein profile similar to the disease protein profile, where protein profile is defined as the specific proteins and the relative number of articles associated with each. To calculate WtCOS, the relationships between the disease and its proteins and the chemical and its proteins were represented as weighted vectors. Each position in both the disease and chemical vector represented a protein. To weight positions in the disease vector the number of articles linking the protein to the disease in question was totaled into a variable called LCF or local co-occurrence frequency. The number of articles linking the protein to *any disease* was totaled into a variable called GCF or global co-occurrence frequency. The LCF was divided by the GCF in a variable called DisLCFIGCF. This number represented the proportion of articles linking the protein to the disease.

The chemical vectors are weighted in a similar way. The number of articles which link the protein to the chemical (LCF) is divided by the number of articles which link the protein to all chemicals (GCF) and placed in a variable called ChemLCFIGCF. To compute WtCOS, the cosine of the two vectors is calculated by the following equation (Manning & Schuetze, 1999):

$$\mathrm{WtCOS} = \cos(x, y) = \frac{x \ast y}{|x||y|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\,\sqrt{\sum_{i=1}^{n} y_i^2}} \qquad (2)$$

where x = DisLCFIGCF and y = ChemLCFIGCF. The chemicals with the vectors most similar to the disease vector will have the smallest value for WtCOS and will be ranked first.

The WtProp metric looks only at the proteins annotated with each chemical. It calculates the percentage equal to the number of disease proteins annotated with the chemical divided by all the proteins annotated with the chemical. The protein count (number of proteins from the protein pool) was divided by the total number of proteins annotated with that chemical in the baseline period. Because a simple proportion gives chemicals with few proteins the advantage, the proportion was multiplied again by the protein count. For instance chemicals with only one protein annotation that happened to come from the protein pool would always have the WtProp = 1 and appear at the top of the list. To avoid this, the proportion was multiplied by the Prot Count again to weigh chemicals with more proteins. If for instance a chemical is annotated with 50 proteins in the literature until 1985, for instance, and 20 of those have been annotated with migraine (migraine protein pool) then WtProp will be equal to 20/50 = .4 *20 = 8.0.

$$\text{WtProp} = \frac{Prot\ Count}{Protein\ Total} * Prot\ Count \tag{3}$$

WtProp is designed to identify chemicals that may not have many proteins annotated with them, but have proteins significant to the disease in question.

The resulting rankings from each of the three ranking strategies were averaged. Each hypothesis set was then ranked based on the average. This rank was called *Average Rank (AvgRank)*. A random ranking (*RandomRank*) was also calculated in order to see whether the rankings performed better than chance. Each entry in the hypothesis set was assigned a random number drawn from the set of numbers between 1 and n, where n is the number of entries in the set. The set was then ranked on this random value. The ranking approaches are summarized in Table 3.3.

Table 3.3 Summary of ranking approaches

Ranking Approach	**Description**
ProtCt	Count of protein pool members associated with chemical
WtCOS	Cosine similarity between the disease-protein vector and chemical-protein vector
WtProp	Proportion of proteins that are related to the disease
AvgRank	The three above rankings are averaged, then the set is ranked on the average
RandomRank	A random number is assigned to each chemical in HS, then ranked on that number

The five sets of ranking results were evaluated by three different methods that in different ways try to measure how well the ranking strategy puts the gold standard chemicals at the top of the list. The first of these methods is the *11-point average interpolated precision.* For each of eleven standard recall levels (0, .1, .2, .3, etc.), that will be denoted as *i,* a variable called the interpolated precision is set to the maximum precision obtained for any recall level greater or equal to *i*.

Precision at K measures performance by calculating precision at specified points in the hypothesis set. If the K threshold values are 10, 20, 30, 40, 50 then precision will be calculated for the top 10 ranked entries in the hypothesis set, the top 20 ranked entries, the top 30 ranked entries, etc. Precison@K is probably the most intuitive measure. It answers the straightforward question, how many found gold standard chemicals were found in the top 10, 20, 30, etc. entries of the list.

MAP or *mean average precision* takes the precision value at each found gold standard chemical. The precision values are averaged when the number of gold standard terms equals k, where k is 10, 20, 30, etc.

3.4 Results

Record counts and overall precision and recall for each hypothesis set are recorded in Table 3.4. In every one of the three diseases the number of proteins in the protein pool increased over each of the three cutoff points. The hypothesis set counts increased similarly. Conversely, and not surprisingly, the number of gold standard chemicals decreased. This trend was expected because the number of years from the cutoff into the future diminished with each time period. The potential discoveries identified in 1984 have over 20 years to be realized, while those after 1994 have only 10 years.

Table 3.4 Summary of precision and recall results from cystic fibrosis(CF), psoriasis, and migraine

Disease	Year Cutoff	Prot Pool Count	Hypothesis Set Count (HS)	Found GS Chems	Total Gold Standard (GS)	Overall Precision (%)	Overall Recall (%)
CF	84-85	346	5,555	215	243	3.9	88.5
CF	89-90	482	9,292	204	219	2.2	93.2
CF	94-95	698	14,143	157	158	1.1	99.4
Psoriasis	84-85	370	5,532	173	220	3.1	78.6
Psoriasis	89-90	537	9,192	134	158	1.5	84.8
Psoriasis	94-95	739	13,393	115	125	0.9	92.0
Migraine	84-85	110	4,006	147	169	3.7	87.0
Migraine	89-90	149	7,122	140	158	2.0	88.6
Migraine	94-95	189	10,467	120	134	1.1	89.6

The changes in precision over time reflect the strong growth in the number of entries in the hypothesis set and the simultaneous reduction of the gold standard chemicals, and consequently the gold standard chemicals that the routines were able to identify. Precision declined by roughly a percentage point in all diseases from one time period to another.

Psoriasis recall in the 1984-85 test run was at 78.6%, the lowest of any test run for any disease. The algorithm missed 47 chemicals. They did not appear in the hypothesis set at all. These chemicals were not found because they had no proteins co-annotated with them from the protein pool. Although many of the missed chemicals had only a few articles linking them to psoriasis, one chemical *1 alpha,24-dihydroxyvitamin D3* had 46 articles linking it to psoriasis, making a significant omission. This chemical is an analog of vitamin D. In the 1989-90 period the recall was improved, with only 24 chemicals missed because they had no proteins annotated with them in common with the protein pool. The most significant of them was ethyl fumarate with 14 articles. By the 1994-95 test run the recall was at 92%. Only 10 chemicals were missed; the most significant was cyclopamine with four articles.

Recall, however, improved over time, particularly in the cases of psoriasis and cystic fibrosis. Although recall did improve with migraine, it was less dramatic. Why recall should improve is not entirely clear. One can speculate that research has increasingly put focus on proteins, both the study of proteins in the etiology and physiology of disease as well as proteins as drug targets. If this is true, then using proteins as the intermediary has become even more effective over time.

Overall recall for migraine was on average lower that for psoriasis and cystic fibrosis. This may be because some drugs are tried on migraine by virtue of their primary indication, not because any basic research has led a researcher to investigate the proteins implicated in the drug’s activity. Anti-convulsant drugs, for instance, are tried on migraine because a number of anti-convulsant drugs have already shown some efficacy against migraine.

The number of proteins in the migraine pool is considerably smaller than the number in the pools for the two other diseases in each of the test period cutoffs. One can speculate that much of the focus in migraine has been on the specific receptors such as 5-HT1, which in the 1990's were discovered to be key players in migraine. The focus on 5-HT1 receptors may have worked to limit for a time basic research on other proteins involved in migraine.

Ranking Evaluation

The hypothesis sets are very large and the number of gold standard chemicals is very small. This needle-in-a-haystack condition is most dramatic in the 1994-95 cystic fibrosis test run. Only 157 chemicals out of 14,143 turned out to be gold standard. Unless the ranking approaches perform very well at putting the gold standard chemicals near the top, there is little chance that this methodology will attract the attention of drug researchers.

Table 3.5 contains the evaluation results of each of the ranking approaches applied to the cystic fibrosis hypothesis sets. In each time period the rankings performed significantly better than random ranking. The metrics ProtCt and AvgRank had the strongest results consistently over all three test runs while WtCOS performed the worst. As with all the diseases studied, results were strongest in the 1984-85 runs and grew successively weaker, reflecting the shrinking window of time in the test period.

The 11-point average precision approach divides the found gold standard chemicals into ten groups called recall levels. The highest precision value within each recall level is reported. Both AvgRank and the ProtCt rankings put gold standard or gold standard chemicals at the first position, so the value is the first column of each is 100%. MAP@K averages precision over the gold standard chemicals. The precision of the first ten GS chemicals resulting from the AvgRank was the highest, followed by ProtCt.

Precision@K gives the results that are the most intuitively easy to understand. The first 7 out of 10 chemicals (70%) presented by the AvgRank approach were gold standard. Three and four of the first ten ranked by WtProp and ProtCt, respectively, made it to the top ten while none of the top ranked chemicals in the WtCOS approach were gold standard.

Table 3.6 contains the ranking evaluation for psoriasis. Each of the ranking methods showed strong performance in the 1984-85 psoriasis test runs and in all cases showed significantly better performance than random ranking. The ProtCt and WtProp showed similar performance to those measures for cystic fibrosis, while surprisingly WtCOS performed considerably better for psoriasis than it did with CF in 1984-85 time period. In later test runs, WtCOS was weaker. As expected, performance deteriorated over the three time periods for psoriasis, but not as strongly for cystic fibrosis. The WtProp and ProtCt ranking approaches showed a weaker performance in 1989-90 compared to 1984-85, but improved for the 1994-95 period, while WtCOS showed further decline in performance in the same period. This likely indicates that proteins have become more central to disease and drug research through the study period.

An evaluation of each ranking approach for migraine test runs are presented in Table 3.7. All ranking approaches performed well for migraine in the 1984-85 test runs. The 1989-90 runs WtCOS was strong while WtProp and ProtCt weakened, while in the 1989-90 test runs WtCOS decreased significantly. The ranking approaches performed significantly better than random rankings in all periods.

Table 3.5 Ranking evaluation results for Cystic Fibrosis. Highest ranks in each range are bolded.

	1984 – 1985					**1989 – 1990**					**1994 - 1995**				
Evaluation method : 11 Point Average Precision (%) at 10%, 20%, 30%, 40%, 50% recall															
Ranking Approach	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%
WtCOS	17.1	16.4	14.9	14.4	11.2	8.6	9.0	9.4	8.9	7.9	8.9	7.1	5.4	5.5	5.6
WtProp	50.0	37.9	28.3	**24.5**	20.3	40.0	31.2	**28.4**	**21.8**	15.8	30.0	**27.4**	17.7	14.0	12.4
ProtCt	**100.0**	**48.1**	**31.4**	23.8	**20.4**	**100.0**	33.0	26.7	21.7	**16.9**	37.5	25.4	**19.4**	**14.1**	**12.7**
AvgRank	**100.0**	37.5	30.7	24.2	18.6	**100.0**	**35.7**	25.5	20.2	15.5	**66.7**	26.1	18.3	12.9	10.8
RandomRank	5.7	4.8	4.4	4.3	4.4	3.3	2.3	2.4	2.3	2.4	0.9	1.1	1.2	1.2	1.2
Evaluation method: MAP@K (%) where K = 10, 20, 30, 40, 50 gold standard terms found from top of ranking															
Ranking Approach	K= 10	20	30	40	50	10	20	30	40	50	10	20	30	40	50
WtCOS	13.3	14.0	14.2	14.5	14.5	5.7	7.0	7.4	7.7	7.9	6.6	6.5	6.4	6.1	5.8
WtProp	35.9	39.3	38.3	36.9	35.2	33.2	32.5	31.5	30.6	29.8	24.4	24.4	23.5	21.8	20.5
ProtCt	47.9	50.2	48.8	45.8	42.8	**53.7**	**48.4**	**42.8**	**40.0**	**37.2**	32.1	28.9	27.1	24.6	22.6
AvgRank	**67.3**	**56.4**	**49.8**	**46.0**	**43.1**	46.2	41.7	39.1	36.6	34.4	**37.2**	**32.2**	**28.0**	**25.2**	**22.8**
RandomRank	4.4	4.2	4.3	4.3	4.3	2.5	2.4	2.3	2.3	2.3	0.6	0.8	0.9	0.9	0.9
Evaluation method: Precision@K (%) where K = 10, 20, 30, 40, 50 top ranked entries on hypothesis set															
Ranking Approach	K= 10	20	30	40	50	10	20	30	40	50	10	20	30	40	50
WtCOS	0.0	10.0	13.3	12.5	14.0	0.0	5.0	3.3	2.5	2.0	0.0	0.0	3.3	2.5	2.0
WtProp	30.0	25.0	46.7	40.0	40.0	20.0	35.0	36.7	32.5	30.0	20.0	30.0	26.7	20.0	24.0
ProtCt	40.0	50.0	**53.3**	**50.0**	**48.0**	**50.0**	**50.0**	40.0	**42.5**	**38.0**	**30.0**	30.0	30.0	25.0	**26.0**
AvgRank	**70.0**	**55.0**	43.3	42.5	40.0	40.0	40.0	**43.3**	35.0	36.0	20.0	**35.0**	**33.3**	**30.0**	**26.0**
RandomRank	0.0	0.0	0.0	2.5	2.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0

Table 3.6 Ranking evaluation results for Psoriasis. Highest ranks in each range are bolded.

	1984 – 1985					**1989 – 1990**					**1994 - 1995**				
Evaluation method : 11 Point Average Precision (%) at 10%, 20%, 30%, 40%, 50% recall															
Ranking Approach	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%
WtCOS	50.0	11.8	11.1	9.3	8.6	42.9	6.9	4.7	4.3	4.1	33.3	4.2	3.8	3.1	2.4
WtProp	50.0	24.5	20.4	15.0	13.2	22.0	16.7	13.0	9.1	6.8	**50.0**	**13.3**	**9.0**	**7.5**	5.7
ProtCt	50.0	**26.6**	**20.9**	16.0	**13.7**	**50.0**	**18.1**	**13.6**	**9.6**	**7.0**	**50.0**	**13.3**	7.9	7.0	**5.8**
AvgRank	**100.0**	19.5	18.0	**16.8**	13.0	45.5	13.8	11.0	7.6	6.5	**50.0**	9.8	7.0	6.0	5.6
RandomRank	6.7	4.5	3.7	3.6	3.5	9.1	1.8	1.8	1.6	1.6	1.2	0.9	0.9	0.9	1.0
Evaluation method: MAP@K (%) where K = 10, 20, 30, 40, 50 gold standard terms found from top of ranking															
Ranking Approach	K= 10	20	30	40	50	10	20	30	40	50	10	20	30	40	50
WtCOS	38.9	27.5	22.0	19.4	17.7	16.0	11.4	9.6	8.3	7.5	17.1	10.6	8.3	7.0	6.1
WtProp	45.3	35.4	31.2	28.5	**26.7**	19.1	18.1	16.9	15.4	14.1	**27.7**	**19.9**	**16.1**	**14.0**	**12.4**
ProtCt	44.2	34.6	30.9	28.4	**26.7**	24.1	21.1	**18.9**	**17.1**	**15.5**	26.6	19.6	15.7	13.6	12.1
AvgRank	**50.2**	**39.1**	**32.2**	**28.8**	26.6	**27.8**	**21.6**	18.1	16.3	14.6	17.9	13.2	11.0	9.7	8.9
RandomRank	2.6	3.1	3.5	3.6	3.6	3.0	2.2	2.0	1.9	1.9	0.8	0.8	0.8	0.8	0.9
Evaluation method: Precision@K (%) where K = 10, 20, 30, 40, 50 top ranked entries on hypothesis set															
Ranking Approach	K= 10	20	30	40	50	10	20	30	40	50	10	20	30	40	50
WtCOS	40.0	30.0	30.0	25.0	22.0	30.0	15.0	10.0	10.0	8.0	10.0	20.0	16.7	12.5	12.0
WtProp	40.0	**45.0**	**33.3**	30.0	26.0	20.0	20.0	13.3	20.0	20.0	**30.0**	25.0	**26.7**	**20.0**	**18.0**
ProtCt	40.0	40.0	30.0	27.5	26.0	20.0	20.0	20.0	15.0	16.0	20.0	**30.0**	20.0	15.0	14.0
AvgRank	**50.0**	35.0	**33.3**	**32.5**	**28.0**	**40.0**	**25.0**	**23.3**	**25.0**	**24.0**	10.0	10.0	16.7	12.5	12.0
RandomRank	0.0	5.0	3.3	2.5	2.0	0.0	5.0	3.3	5.0	4.0	0.0	0.0	0.0	0.0	0.0

Table 3.7 Ranking evaluation results for Migraine. Highest ranks in each range are bolded.

	1984 – 1985					**1989 – 1990**					**1994 - 1995**				
Evaluation method : 11 Point Average Precision (%) at 10%, 20%, 30%, 40%, 50% recall															
Ranking Approach	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%
WtCOS	100.0	19.4	14.3	14.2	11.8	50.0	11.0	8.6	8.3	6.6	11.5	5.3	5.3	5.3	4.8
WtProp	37.2	32.7	30.6	20.3	18.0	42.9	25.7	21.3	18.8	13.5	40.0	20.3	15.3	10.3	7.9
ProtCt	100.0	22.0	18.7	16.7	13.4	100.0	18.8	16.4	13.9	9.4	100.0	18.7	11.6	9.7	7.6
AvgRank	50.0	27.5	25.6	20.1	13.4	100.0	24.3	19.4	12.3	9.8	100.0	12.0	10.6	9.5	7.2
RandomRank	3.9	4.3	4.4	4.1	3.8	7.1	2.1	1.9	2.0	2.1	8.3	1.4	1.2	1.2	1.2
Evaluation method: MAP@K (%) where K = 10, 20, 30, 40, 50 gold standard terms found from top of ranking															
Ranking Approach	K= 10	20	30	40	50	10	20	30	40	50	10	20	30	40	50
WtCOS	36.8	27.9	24.3	21.7	20.1	20.1	15.2	12.9	11.7	11.0	6.3	5.5	5.3	5.2	5.2
ProtCt	66.3	45.8	37.3	32.5	29.5	42.3	29.5	25.3	22.7	21.0	36.9	27.2	21.9	18.9	16.8
AvgRank	35.5	30.4	28.9	27.8	26.5	43.9	33.2	29.0	26.3	23.9	31.1	21.6	17.8	15.8	14.4
RandomRank	3.0	3.3	3.5	3.7	3.8	4.0	3.1	2.7	2.5	2.4	2.2	1.7	1.5	1.4	1.4
Evaluation method: Precision@K (%) where K = 10, 20, 30, 40, 50 top ranked entries on hypothesis set															
Ranking Approach	K = 10	20	30	40	50	10	20	30	40	50	10	20	30	40	50
WtCOS	30.0	25.0	16.7	20.0	18.0	30.0	20.0	13.3	12.5	14.0	0.0	5.0	10.0	7.5	6.0
WtProp	30.0	35.0	33.3	35.0	32.0	40.0	40.0	33.3	30.0	28.0	40.0	30.0	23.3	20.0	22.0
ProtCt	50.0	40.0	33.3	32.5	26.0	40.0	30.0	23.3	22.5	20.0	40.0	25.0	26.7	22.5	22.0
AvgRank	30.0	25.0	26.7	25.0	26.0	40.0	25.0	23.3	22.5	20.0	20.0	10.0	10.0	15.0	16.0
RandomRank	0.0	0.0	0.0	0.0	0.0	0.0	5.0	6.7	5.0	4.0	0.0	5.0	3.3	2.5	2.0

To get a picture of how the ranking strategies worked overall, the results were averaged over all three diseases and each of the three cutoff periods. The averages are presented in Table 3.8 and Figure 3.3, 3.4, and 3.4 show the results graphically. The WtProp ranking approach had the highest average results for recall levels over 10. The ProtCt approach returned the highest average results measured by MAP@K, although WtProp and AvgRank were close behind. The Precision@K results were also close with WtProp and ProtCt achieving the top results.

Table 3.8 Average evaluation scores for each ranking approach. Scores are averaged over all three diseases and the three cutoffs.

Evaluation method: 11-Point Average Precision (%) at 10%, 20%, 30%, 40%, 50% recall

Ranking Approach	10%	20%	30%	40%	50%
WtCOS	35.8	10.1	8.6	8.1	7.0
WtProp	40.2	25.5	20.4	15.7	12.6
ProtCt	76.4	24.9	18.5	14.7	11.9
AvgRank	79.1	22.9	18.5	14.4	11.2
RandomRank	5.1	2.6	2.4	2.4	2.4

Evaluation method: MAP@K where K=10, 20, 30, 40, 50 gold standard **terms found from top of ranking**

Ranking Approach	K= 10	20	30	40	50
WtCOS	17.9	14.0	12.3	11.3	10.6
WtProp	31.6	28.8	26.7	24.9	23.4
ProtCt	41.6	33.9	29.9	27.1	24.9
AvgRank	39.7	32.2	28.2	25.8	23.9
RandomRank	2.6	2.4	2.4	2.4	2.4

Evaluation method: Precision@K (%) where K = 10, 20, 30, 40, 50 top ranked entries in hypothesis set

Ranking Approach	K= 10	20	30	40	50
WtCOS	15.6	14.4	13.0	11.7	10.9
WtProp	30.0	31.7	30.4	27.5	26.7
ProtCt	36.7	35.0	30.7	28.1	26.2
AvgRank	35.6	28.9	28.1	26.7	25.3
RandomRank	0.0	2.2	1.8	1.9	1.8

Figure 3.3 Graph of average values for 11-Point Average Precision

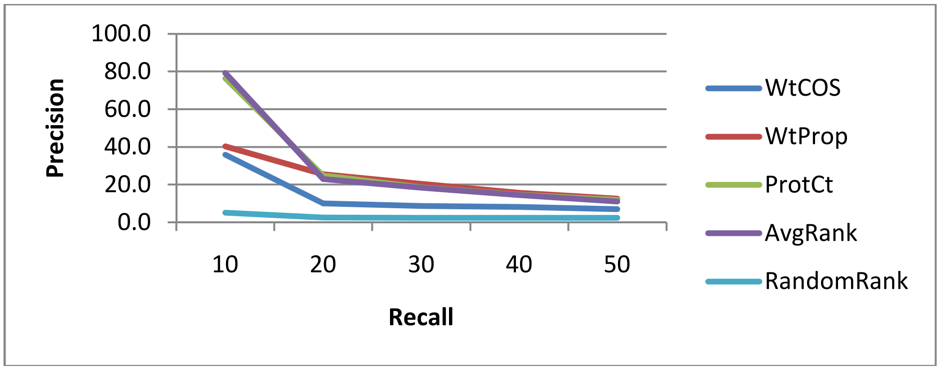


Figure 3.4 Graph of average values for MAP@K

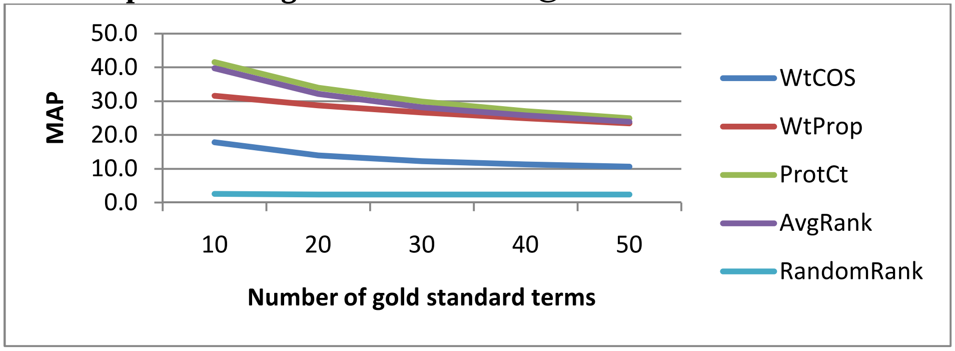
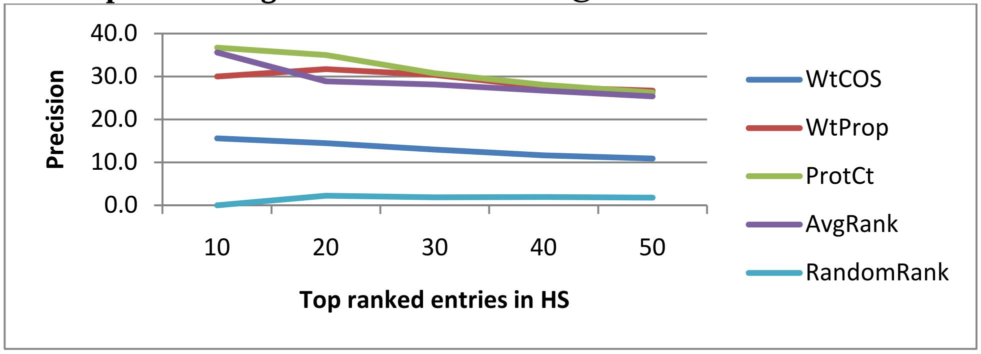


Figure 3.5 Graph of average values for Precision@K



3.5 Discussion

Before we move on to a discussion of each disease individually, we will look at the hypothesis sets in some detail and note characteristics shared by each of the sets. See

Appendix 2, 3, and 4 for the first twenty records returned by each ranking in each test run. In these tables, the found gold standard chemicals can be identified by the columns on the right with the white background. The elements ArtCt (Article Count), FirstYr (first year of a direct connection between chemical and disease), and the subheadings are pulled from the test period. The chemical and disease subheadings or qualifiers are the most commonly occurring ones when the disease and chemical are annotated together. The columns in gray (chemical name and protein count) represent data from the baseline period; the white columns contain pulled from or calculated from data pulled from the test period.

The hypothesis sets have some striking similarities. First, most of the entries in the hypothesis set were not found in the gold standard set, meaning the routines did not find a direct link between the chemical and the disease in the test period, as it was predicted to. This is not surprising given the large hypothesis sets and the low number of gold standard chemicals in each.

The entries in each set are a mixture of all kinds of chemicals. They include potential drugs (exogenous) but also endogenous chemicals or those naturally found in the body. Endogenous chemicals include elements such as magnesium, zinc, and calcium. These elements are important signaling chemicals. Nucleic acids (e.g., Cyclic GMP) and steroids (e.g., estrone) are also apparent.

The hypothesis sets are also diverse in the *type* of connections that evolve between the chemicals and the disease. There are drugs which appear to have been tried in disease treatment. This is evident through the disease and chemical qualifiers such as *drug therapy* and *administration & dosage*. Other chemicals appear to play a role in the physiology or etiology of the disease. This is evidenced by the *blood, physiopathology,* and *etiology* qualifiers. Endogenous molecules can often be recognized by the *metabolism* or *biosynthesis* subheadings. The *chemically induced* qualifier indicates that a chemical appears to cause the disease.

Our goal in this study is to find drugs that can be reprofiled for new therapeutic uses. We cannot evaluate reprofiling potential from just the ranking results, because the ranking results reflect the diverse ways a chemical can be connected to a disease.

To evaluate reprofiling specifically, we will use two methods. First, review articles will be identified and studied to find any examples of drug reprofiling.   The examples of reprofiling we will include in this discussion will be limited to those that could have picked up by this study: drugs that existed before the cutoff, had no connection to the disease, and then developed a connection to the disease in one of the test periods. We will then see if the reprofiled drugs are in the relevant hypothesis set and how highly they are ranked.

Next we will use the article count metric to rank the found gold standard chemicals in each hypothesis set. The article count is a rough indicator of how much publication attention a drug received and we will use it to find the most promising reprofiled drugs and then look to see how high the ranking approaches placed these drugs.

Before we look at the details of each disease and its respective reprofiled therapies, background on the disease itself will be presented along with a description of the therapeutic strategies used to treat the disease.

3.5.1 Cystic Fibrosis

Overview

Cystic fibrosis (CF) is the most lethal genetic disease among Caucasians. CF is caused by a mutation in the gene that encodes the cystic fibrosis transmembrane conductance regulator (CFTR) protein. This protein play a number of important roles in the body and therefore a defective protein can adversely affect several organs, including lungs, pancreas, liver, and the reproductive organs. The CF mutation in the CFTR causes a thickness in mucus, making normal clearing of the mucus difficult. The buildup of mucus in turn impairs the function of the affected organ. The lung manifestations are the most life-threatening. 80% of deaths from CF result from pulmonary insufficiency (O'Sullivan & Freedman, 2009). Because the mucus is a host for bacteria, many CF patients develop chronic respiratory infections, exacerbating the already reduced pulmonary capacity. Diabetes mellitus is a growing complication of cystic fibrosis.

Drug therapies for CF target the manifestations of the CFTR deficiency in specific organs. Therapies directed at the respiratory system try to improve the viscosity of the mucus to enable better clearing. Antibiotics treat the chronic infections in the lungs. Because CF complications in the liver and pancreas impede the normal metabolism of food, diet therapy is critical in CF patients, including supplementing the diet with nutrients that are poorly absorbed (e.g., Vitamins K and D). Complications such as diabetes must also be treated. Newer therapies target the CFTR protein itself by attempting to rectify incorrect transcription or by activating the protein’s activity. Gene therapy has received some attention, but clinical application of the therapies has so far been unsuccessful (O'Sullivan & Freedman, 2009).

Cystic fibrosis: reprofiled drugs

We will approach our evaluation of reprofiling in two ways. First we will examine two recent reviews of cystic fibrosis for current or potential therapies that represent re-profiling of drugs and see whether the drugs reprofiled in practice have shown a presence in any of our three time period analyses. Next we will look at the found gold standard chemicals to find reprofiled drugs that met with some success, or at least received some attention, as measured by the number of articles linking them in the test periods to cystic fibrosis. This second step will allow us to give attention to drugs that may not be mentioned in the reviews but did at some time in the recent past receive attention from researchers in the form of publications.

We must limit our examination to the chemicals which we *could* have predicted: chemicals that have a literature record in the baseline period, but no connection to CF, but then did develop a connection in the test period. This means new chemicals entities (NCE) are generally outside our scope. An NCE is a compound that has not yet been approved for any therapeutic indication therefore likely has little if any literature history. Besides new chemical entities, as discussed previously, there are other drug therapies that by design do not make it into these results. Protein therapies and solutions are two examples. It is important to note these omissions in the case of cystic fibrosis. Two important therapies for CF noted in both reviews are dornase alfa, a recombinant deoxyribonuclease (protein), and hypertonic saline solution. Even if they were examples of re-profiling, they would not appear in the results reported here. Endogenous chemicals and elements appear frequently on the hypothesis lists. Although these substances may be of interest to some researchers, but because the goal of this study is re-profiling of small molecule pharmaceuticals, we will not focus on endogenous molecules.

In their review of cystic fibrosis, O’Sullivan and Freedman (O'Sullivan & Freedman, 2009) describe the current treatment recommendations from the US Cystic Fibrosis Foundation for chronic pulmonary disease. Two of these may be considered examples of re-profiling. Azithromycin belongs to the macrolide antibiotic family. It appears to not only kill bacteria, but also stimulate anti-inflammatory activity. In ChemoText it appeared first (as a subject drug) in 1987 and was first linked directly to CF in 1995. In the 1989-90 hypothesis sets Azithromycin was not ranked high. The WtProp ranking put it highest at position 2895 out of 9,292 entries in the hypothesis set. In the 1994-95 sets, it had moved up to position 710 out of 14,143. While this is a large jump, this position may not have brought the drug to the attention of a researcher.

Ibuprofen is a nonsteroidal anti-inflammatory drug that in long term studies slows down the deterioration of lung function (O'Sullivan & Freedman, 2009). The first appearance of ibuprofen in ChemoText was 1968 and its first link to CF was in an article published in 1990. In the 1984-85 study ibuprofen was ranked 357 out of 5,555 members of the hypothesis set and by 1989-90 it was ranked at 229 out of 9,292. Again, it may not have been ranked high enough ever to garner a researcher’s attention.

O’Sullivan and Freedman also reviewed the emerging therapies for cystic fibrosis. Genistein, a chemical found in soybeans, was being studied for its ability to modify CFTR activity. Genistein’s first appearance as a subject drug in ChemoText was 1981. In the 1984-85 period it did not have any proteins in common with CF and did not make the hypothesis set. In the 1989-90 period it made the hypothesis set, but its highest ranking was 1,783 out of 9,292. By the 1994-95 period, genistein had moved all the way up to position 66 on the AvgRank list out of 14,143 chemicals in the list. Although genistein was not directly connected to the disease CF through disease and subject chemical annotations, genistein was explicitly studied for its affects on the CFTR using *in vitro* and animal models. Likely the researchers had the disease in mind and the potential of genistein to treat CF cannot really be regarded as a novel connection.

In the second review, Frerichs and Smyth list mannitol as a promising treatment in Phase III trials (Frerichs & Smyth, 2009). Mannitol is a diuretic that has appeared in the literature for many years, described primarily as a diagnostic aid to test renal function. Its first appearance in ChemoText as a subject drug is 1949 and its first direct connection to cystic fibrosis appeared in 1993. In this article however, an oral form of mannitol was used to help assess pancreatic dysfunction of children (Green, Austin, & Weaver, 1993). The first pilot study appeared in 1999 (Robinson et al., 1999) testing the inhaled mannitol on cystic fibrosis patients. In the lungs, mannitol helps move water across the lung surface and reduces mucus viscosity (Storey & Wald, 2008) . An inhaled dosage form is now in Phase III trials for CF. In the 1984-85 hypothesis set, mannitol was placed in position 107 by the WtProp ranking and in position 103 by the ProtCt ranking, and by the 1989-90 period mannitol had moved up to positions 95 and 98, respectively, where the drug might have been noticed by a drug researcher.

Two other drugs being investigated for use in CF deserve mention: curcumin and miglustat. Curcurmin, an extract of turmeric, has been proposed as a corrector of the protein misfolding that often accompanies the CFTR mutation (Frerichs & Smyth, 2009). It was first associated with CF in 2004. Although tests on proteins showed some success, clinical Phase I trials have so far been negative. In the 1994-95 test run, curcumin only had 24 proteins connecting it to CF. It did not rank high by any measure, with the highest rank at position 1000. Miglustat first appears in PubMed in 1994 and only garners four proteins from the protein pool. It is ranked very low. These potential reprofiled drugs come a little too late to be picked up by our studies. It would be interesting to see how high they would appear in later cutoff dates.

Next, we will look for significant drugs by examining the gold standard output set for each test run presented in Appendices 5A, 5B, and 5C. These tables are sorted by article count and should provide us with reprofiled drugs that, because of timing and other reasons, were not mentioned in the reviews. The chemicals are listed in descending order of the number of articles that link each to cystic fibrosis in an attempt to put the most important gold standard chemicals at the top. Because the lists are lengthy, only those chemicals with four or more articles are included. The number of proteins, most common disease qualifier (DisQual) and chemical qualifier (ChemQual) are shown next. At the right hand side are the four rankings produced by the study: WtCOS, ProtCt, WtProp, and AvgRank. Selected chemicals from this list will be discussed.

Several of the drugs already mentioned are evident (e.g., ibuprofen and mannitol). Although we will concentrate on drugs with the potential to be reprofiled, it will be noted briefly that many of the top ranked chemicals are endogenous substances such as nitric oxide, hydrogen peroxide, and uridine triphosphate. The ranking routines were very good at ranking nitric oxide high in the 1994-95 period (at position 25 by the ProtCt approach) and putting hydrogen peroxide near the top in 1989-90 and 1994-95 (position 1 by the AvgRank approach and position 4 by ProtCt, respectively). The ranking routines also successfully put the nutrients taurine and carnitine near the top of several hypothesis sets. The AvgRank for taurine in 1984-85 was position 47 while carnitine appeared at position 68. It should also be noted that although taurine first appears in 1985 directly connected to CF, a derivative of taurine called taurocholic acid was directly connected to CF in 1982.

Nitric oxide is high on the tables in Appendix 5 with 64 articles linking it to cystic fibrosis. Nitric oxide was named Molecule of the Year in 1992, and the years preceding 1992 and the years since have seen a dramatic increase in the research on nitric oxide (Gibaldi, 1993; Koshland, 1992). This small but highly reactive endogenous molecule plays a signaling role in many physiological processes. Drugs are being developed that can therapeutically modulate the activity of nitric oxide. The first article directly linking nitric oxide to cystic fibrosis was published in 1995 and a total of 64 articles link the two by the end of the test period. In the ABC analysis in 1984-85 (see Appendix 2A) nitric oxide was ranked best by ProtCt at position 905. By 1989-90 it had risen to position 288 and by 1994-95 it was ranked at position 25 by ProtCt. The amount of basic research on the molecule caused the number of proteins from the CF protein pool associated with it to climb dramatically from 16 to 182, resulting in its jump in the rankings. A similar increase in protein counts and in higher rankings will be seen with psoriasis and migraine.

The top reprofiled drug on the 1984-85 list is Ciprofloxacin. This antibiotic came onto the scene in 1983 and had only one protein linking it to CF in the 1984-85 period and therefore it ranked very low. Its first connection to CF came in 1985. It is likely that research physicians readily try new antibiotics on cystic fibrosis patients as the bacteria grow resistant to older forms. Rifampin, another antibiotic, was ranked more highly by all of the ranking approaches. Rifampin is used in CF patients, but not as widely as Ciprofloxacin. Lithium, which ranked high on all approaches except for WtCOS, was tested on CF patients and found to have a detrimental effect, reducing the key measures of lung function and signaling researchers that CF patients with manic-depressive disease should not be treated with lithium or if they do take the drug, they should be monitored closely (Anbar et al., 1990).

Like mannitol, furosemide is a diuretic, promoting excretion of urine by the kidneys. Its connection with CF started when a furosemide-treated mouse was proposed as an animal model for the disease (Szeifert, Varga, Damjanovich, & Gomba, 1987). In later studies it was examined for its ability to help CF patients improve kidney function. Its diuretic and anti-inflammatory effects have also been thought to improve lung function in patients (Prandota, 2001).

Forskolin is a plant extract with a number of properties. It has been used to study the molecular level activity of the CFTR for a number of years and does seem to affect the chloride conductance by CFTR channels, although it does not yet seem to have been proposed as a CF treatment (Kerem, 2006). It eventually has nine articles linking it to cystic fibrosis. It was predicted at position 152 in the 1984-85 table, but had moved up to position 69 in 1989-90.

Ranitidine is a blocker of gastric H2 receptors. It evidently improves the fat absorption in patients with cystic fibrosis (DiMagno, 2001). Caffeine was ranked highly in three of the four approaches in the 1984-85 period. Hepatic enzymes are often affected by

CF and administration of caffeine was shown to be useful diagnostic tool in measuring liver function in CF patients, specifically breakdown removal of caffeine by the liver.

So far we have looked at the drugs the routines should have identified and put high on the ranked lists. Next we will look at what the routines did rank high. The first observation is one that has been mentioned before: a high percentage of endogenous chemicals including elements appear at the top portion of each list. We will ignore these and focus on potential reprofiled drugs.

In the 1984-85 test run, edetic acid appears in position 1 and 2 of the ProtCt and WtProp rankings, respectively. Edetic acid is a chelating agent used in manufacturing of pharmaceuticals and in the preservation of food. In 1985 edetic acid in combination with antimicrobials was tested in CF patients as a therapy for chronic lung infection but showed no signs of efficacy (Brown, Mellis, & Wood, 1985).In later studies edetic acid was used as a probe molecule to test intestinal permeability in CF patients (Escobar et al., 1992) Dimethyl sulfoxide (high on all lists) and warfarin are other compounds used in testing cellular permeability and protein function. Chloroquine was suggested as a treatment for lung inflammation seen in CF 2003 (Derleth, 2003). In 2006 a cell based assay found that chloroquine, because it is a permeable weak base, was able to show some effect on TGF-beta, anther protein involved in CF.

An overview of the reprofiled chemicals discussed in this section is presented in Table 3.9 below.

Table 3.9 Cystic Fibrosis – selected reprofiled chemicals. Best rank is the highest rank from any test run. HS is hypothesis set. ArtCt is the number of articles connecting the drug to the disease.

Chemical	Best rank / HS count	Previous use / activity	Status	Art Ct	Reprofiling type
Azithromycin	710 / 14,143	Antibiotic	Recommended for chronic pulmonary disease	40	Functional
Ibuprofen	229 / 9,292	Anti-inflammatory	Slows deterioration of lung function	27	Functional
Genistein	66 / 14,143	Anticancer; CFTR activity	Phase II showed efficacy	10	Molecular Functional
Mannitol	95 / 9,292	Diuretic	Ongoing clinical trials (2010)	8	Functional
Curcumin	1000 / 14,143	Spice; CFTR activity	Phase I clinical trials negative	13	Molecular Functional
Ciprofloxacin	3,484 / 5,555	Antibiotic	In use	109	Functional
Rifampin	49 / 5,555	Antibiotic	Combination therapy effective in small trial	6	Functional
Lithium	9 / 9,292	Ion transport; psychosis	In trials exacerbated CF	4	Molecular Functional
Furosemide	20 / 5,555	Diuretic and anti-inflammatory	Seems to improve kidney function	6	Functional
Forskolin	69 / 9,292	CFTR activity	Still basic research	9	Molecular Functional
Ranitidine	29 / 9,292	Anti-ulcer; reduces acid	Improves fat absorption and gastric emptying	7	Functional
Caffeine	31 / 5,555	Stimulant	Diagnostic	4	Functional
Edetic acid	1 / 5,555	Chelating agent	No effect in trials; diagnostic for intestinal permeability	3	Functional?

Cystic fibrosis summary

Before leaving this examination of cystic fibrosis, it may be beneficial to step back and summarize what has been observed. The most striking characteristic of the collection of drugs that develop a connection to CF is the wide variety of *ways* in which they are connected to the disease. Although we did not encounter drugs that cause cystic fibrosis (as we likely will with migraine) we did find lithium exacerbated respiratory symptoms. We did of course find many drugs that have been reprofiled to treat CF, but here, too, variety is a striking characteristic. Drugs treat the myriad of manifestations of the broken CFTR protein in a variety of organ systems, while some target the protein itself, and still others target the DNA mutation that causes the CFTR problem.

Functional reprofiling is seen most commonly in cystic fibrosis. Researchers know what function a drug has on a tissue or organ and reason that the function would be beneficial in cystic fibrosis. Mannitol, for instance, is a diuretic; it promotes fluid removal from tissues and was used extensively to increase kidney output. Applied to lung tissue, mannitol has a parallel effect, moving fluid from the lungs to the mucus layer where it hydrates the mucus for easier clearance.

We also saw cases of reprofiling based on knowledge of what the drug does at the molecular level and what parallel molecular mechanisms are at work in the disease state. This kind of reprofiling we will call *molecular functional reprofiling*. In the case of cystic fibrosis, genistein, curcumin, and forskolin have been studied *in vitro* for their effects on the CFTR protein in hopes they can correct the protein malfunction.

Other chemicals were reprofiled not to treat CF, but to probe, test, or measure physiological functions important to CF. Warfarin has been used to test plasma clearance in CF patients compared to control to see if CF has affected the patient's metabolism. Similarly caffeine has been used to test hepatic function in CF patients. Tests like this can be used as a diagnostic. Caffeine levels too high or low can indicate that the organ (e.g., liver) has become affected by the disease. Edetic acid is used as a probe to test intestinal permeability in CF patients. Other chemicals create an *in vitro* or *in vivo* environment where therapies can be tested. An example of this is furosemide: a study suggested giving furosemide to mice makes them a valid animal model for CF. A number of other chemicals create the needed environments (e.g., acidic, basic) to test other chemicals that may be useful in the treatment of CF.

How might the landscape of chemicals associated with cystic fibrosis be different from that of other diseases? Cystic fibrosis is a serious disease. CF patients are chronically sick and experience deterioration of organ function over many years. There are no truly successful therapies for CF and certainly no cures. Drug re-profiling in CF may be different from other diseases. We did not see, for instance, a case of observational or chance reprofiling, where a drug is noticed by chance to have an effect on a disease, and this observation is picked up and acted on by researchers. This sort of serendipitous event is perhaps less likely in a chronic disease like CF than it would be in a disease like psoriasis, where any change in the disease state is readily visible. As we have seen, functional reprofiling, taking a drug with known function and safety profile, and applying it to cystic fibrosis, is the most common approach.

3.5.2 Psoriasis

Overview

Psoriasis is a common skin disease that is characterized by red, scaly patches called plaques. The plaques are discrete areas of inflammation and excessive skin production. Although the etiology of psoriasis is unclear, it is thought to have origins in the immune system. (Levine & Gottlieb, 2009)

The severity of psoriasis can range anywhere from mild to severe, depending on the location and coverage of the plaques. Psoriasis has several forms as well, including plaque psoriasis, (the most common), pustular, and guttate psoriasis. Guttate psoriasis is associated with a streptococcal throat infection.

The choice of treatment depends on the location and severity of the patches. The first line of treatment generally is limited to topical applications such as corticosteroids, vitamin D derivatives, vitamin A derivatives, tar preparations, and anthralin, or combinations of these. These topicals work on several ways in the psoriatic skin. Corticosteroids, for instance, reduce the inflammation, and vitamin D analogs work by suppressing the skin proliferation. Non-pharmaceutical products are also used; creams and emollients help to moisturize the skin and reduce the itching (Levine & Gottlieb, 2009; Naldi & Gambini, 2007).

When topical remedies are ineffective or the disease is too widespread, systemic therapies are used. Recent research in psoriasis has revealed that the immune system plays a major role in the disease pathway, so many of the systemic medications are directed at the immune system. (Sabat, Sterry, Philipp, & Wolk, 2007) These treatments include small molecule drugs as well as the new protein-based biologicals. Light therapy, often in combination with other therapies, is common. Because there is no cure for psoriasis, patients often rotate through many therapies.

Because psoriasis is so common and its manifestations are visible – and unpleasant - the disease has a long history of motivated and imaginative patients taking charge of their own treatment. The National Psoriasis Foundation (*National Psoriasis Foundation,*2009) even hosts a web page called *It Works for Me* where patients can tell others of their personal treatment successes. In addition to testimonials for prescription therapies, patients recount their success with a variety of over-the-counter and home remedies such as Listerine, salt baths, olive oil, lime juice, and banana peels.

Just as patients have re-directed household substances to gain relief from psoriasis, researchers have actively sought to reprofile drugs for use in the disease. As more is learned about the physiology and etiology of the disease, the opportunities for reprofiling expand. For instance, since researchers learned that psoriasis involves the immune system, a number of immunomodulatory drugs have been studied in clinical trials.

Psoriasis and reprofiled drugs

In this section we will look beyond the quantitative measures and evaluate the results qualitatively to answer the question: how *useful* were the results. The purpose is to see whether these results – had they been available early in the test periods – could have helped to accelerate the development of important treatment options for psoriasis. Similarly to the evaluation of the CF results, we will first look at a recent review article and see if any of the reprofiled drugs discussed are in the hypothesis sets and where they are ranked. Then we will look at the gold standard drugs that have significant numbers of articles linking them to psoriasis and see where the rankings put these drugs.

A 2008 review by Halverstam and Lebwohl described nonstandard and off-label therapies for psoriasis (Halverstam & Lebwohl, 2008), including a number of reprofiled therapies. We will limit our discussion to those drugs that could have been identified by the algorithms in this research: small molecule drugs that existed in the baseline period with no direct link to psoriasis, but which did develop a link in the test period. Three drugs reviewed met these criteria and made it into our hypothesis sets: mycophenolate mofetil, sulfasalazine, and paclitaxel. The first two were examples of functional reprofiling, the third was an instance of observational reprofiling.

Mycophenolate mofetil is an immunosuppressive drug that has been used to prevent organ rejection in transplant patients. The drug is a form of mycophenolic acid, a drug that was tried on psoriasis patients but discontinued because of adverse events. Mycophenolate mofetil demonstrated anti-inflammatory effects in addition to its immune system effects and had been used in other skin diseases. In 1997 it was used successfully to treat a man with psoriasis. This case study was followed by more trials with larger patient populations, and by the 1994-95 test period there were 20 articles linking this drug to psoriasis. While the ranking algorithms did not rank it in the top 100, the WtCOS approach did put mycophenolate mofetil at position 543 out of 13,393 entries in the hypothesis set.

The review also discusses sulfasalazine, a drug used to treat Crohn's disease and ulcerative colitis. While this drug's mechanism of action is not entirely clear, it is thought to have anti-inflammatory activity through its interference of folate metabolism. In double-blinded randomized trial conducted in the early 1990's, sulfasalzine was reported to improve psoriasis in a majority of patients (Halverstam & Lebwohl, 2008). The WtProp ranking approach in the 1984-85 test runs put sulfasalazine at position 171 out of 5,532 entries in the hypothesis set.

The review also included a discussion of paclitaxel in the treatment of psoriasis. Paclitaxel is a chemotherapeutic drug used in treating breast and ovarian cancer. It had been observed in an early study of paclitaxel that patients on the drug experienced improvement of their psoriasis symptoms (Halverstam & Lebwohl, 2008). On that basis, a small clinical trial was conducted (Ehrlich et al., 2004). All of the patients showed improvement and the drug was well tolerated by most of the patients. The WtCOS ranking algorithm in 1989-90 ranked paclitaxel at position 66 out of 9,192 where it would likely have been noticed. The authors note that for patients who suffer from both breast cancer and psoriasis, paclitaxel is a treatment to be considered.

Next we will look at the gold standard chemicals, ranking them by article count and see what reprofiled chemicals the ABC algorithms were able to find. Appendices 6A, 6B, and 6C list the most important gold standard chemicals by virtue of their article counts for each of the three cutoff year test runs. Once again it is interesting to note that the lists contain endogenous molecules and elements as well as drugs, although there appear to be fewer endogenous substances and more drugs in these lists than in the same lists created for cystic fibrosis.

The two top entries in Appendix 6A are analogs of vitamin D. Calcitriol is the physiologically active form of vitamin D and cholecalciferol is a vitamin D analog. Vitamin D fits somewhere in between endogenous and drug. For many years Vitamin D and its various forms or analogs have been important treatments for psoriasis and are thought to suppress cell proliferation. These two forms of vitamin D have received a lot of attention from researchers (353 articles for calcitriol) and even though they were also ranked high on the hypothesis set lists, they cannot be considered novel connections because the association between psoriasis and vitamin D is a longstanding one.

In the 1984 time period the drug propylthiouracil appears high on each of the rankings, particularly AvgRank, where it appeared at position six. Because propylthiouracil was used for many years as a treatment for hyperthyroidism before being tested in psoriasis, it represents a good example of drug reprofiling. In 1993 researchers reasoned that because the drug had immunomodulatory and free radical scavenging effects, they would try it as a psoriasis treatment in a small clinical trial. It is an oral systemic with lower toxicity than other treatments of psoriasis and did show some benefit (Elias, Goodman, Liem, & Barr, 1993). Methimazole is a drug from the same family as propylthiouracil and is thought to have a similar mechanism of action. Methimazole has also received attention for its potential to treat psoriasis. Although the ABC ranking mechanism did not put it as high as propylthiouracil, it did achieve an average rank of 58 in 1984-85.

Capsaicin appears high on the tables in Appendix 6 with 11 articles. The highest rank it acquired from the ABC analysis was 149 out of 5,532 in the 1984-85 test run. Capsaicin is the active chemical in chili peppers and although known for its burning and irritant effects, has also been used as an anti-itch treatment (antipruritic). It is thought that one of the mechanisms of capsaisin action is that it inhibits vasodilation. With this knowledge, researchers reasoned that it might have useful activity in the cutaneous vascular changes caused by psoriasis (Bernstein, Parish, Rapaport, Rosenbaum, & Roenigk, 1986). At least one double-blind controlled study demonstrated the efficacy of capsaicin, particularly in reducing the itch associated with the disease (Ellis et al., 1993).

Ranitidine and psoriasis have an interesting history that can be traced by reviewing the seven articles linking it to psoriasis. A 1991 article (Andersen, 1991) reports the worsening of a case of psoriasis for a patient taking ranitidine, a histamine H2 blocker used to treat gastrointestinal ulcers, while another article published the same year speculates there is reason to think ranitidine might treat psoriasis. The reasoning is based on the knowledge that histamine released from mast cells plays a role in psoriasis, and therefore blocking the histamine could improve the disease symptoms (Nielsen, Nielsen, & Georgsen, 1991). An open, prospective study of twenty patients had promising results (Kristensen et al., 1995). Most of the patients showed long term improvement. In 1997 a larger study, blinded and placebo-controlled, produced contrary results, showing no significant difference between the control and treatment groups (Zonneveld et al., 1997). Whether or not ranitidine is ever determined to have an effect on psoriasis, it was predicted in this study, and in 1989-90 ranked at position 47 by the AvgRank method.

The drug pentoxifylline has five articles connecting it to psoriasis in the 1994-95 period and it was identified by the ABC algorithms and ranked very high, at position 20 on the 1994-95 test run WtProp ranking. Pentoxifylline affects blood flow, platelet aggregation, and cell proliferation and has been investigated as a treatment for a wide variety of conditions. In 1996 it was suggested as a potential treatment for psoriasis. *In vitro* and *in vivo* studies demonstrated that it did inhibit skin cell proliferation (Omulecki, Broniarczyk-Dyla, Zak-Prelich, & Choczaj-Kukula, 1996). In 2006 the drug was tested in a placebo-controlled clinical trial and, although it produced few side effects, it also showed little efficacy (Magela Magalhaes et al., 2006).

Two antibiotics, rifampin and erythromycin, are listed in Appendix 6A and both were ranked in the top 100 by at least one ranking approach. Rifampin was ranked high by every ranking approach, appearing at position one in the average rank. Rifampin has been used to treat tuberculosis since the 1960's and has also been used to treat other bacterial infections such as meningitis and leprosy. In 1986 a preliminary report was published describing a study in which rifampin was used in combination therapy with either penicillin or erythromycin in psoriasis associated with streptococcal carriage (Rosenberg et al., 1986). The rate of streptococcal carriage was reduced and the psoriasis markedly improved. The subsequent studies of rifampin in monotherapy for psoriasis produced somewhat conflicting results, partly because researchers designed the studies around streptococcal-related psoriasis. Further studies indicated that the antibiotic activity of rifampin was not the reason for its effects. Instead, rifampin was shown to have immunomodulatory effects on the innate immune system (Tsankov & Grozdev, 2009). The articles about rifampin and psoriasis continue up to 2009. Although rifampin does not seem to have become a standard therapy for psoriasis, the research on its use in psoriasis continues.

Erythromycin also appears on the 1984-85 list in Appendix 6A and it also received fairly high rankings from the algorithms, appearing at position 38 on the WtProp list. The first article directly connecting erythromycin was the article noted above that described a study combining rifampin with either erythromycin or penicillin in guttate psoriasis, the kind of psoriasis that appears commonly when the patient has a streptococcal infection such as strep throat (Rosenberg et al., 1986). Research in the ensuing years indicated that macrolide antibiotics such as erythromycin have anti-inflammatory effects. In a 2007 study (Polat et al., 2007) showed a statistically significant improvement for patients taking erythromycin in addition to topical corticosteroids as compared to the group of patients using topical corticosteroids alone. Curiously the patients in this study had psoriasis vulgaris, not guttate psoriasis. A 2008 study indicated that erythromycin showed no significant efficacy in using erythromycin against guttate psoriasis (Dogan, Karabudak, & Harmanyeri, 2008). The connection between erythromycin and psoriasis, similar to the rifampin and psoriasis, is still not clear but is receiving continued attention from the research community.

Like paclitaxel discussed earlier, tamoxifen is a treatment for breast cancer. Tamoxifen works by blocking estrogen. Evidence for tamoxifen’s use in psoriasis started in a manner similar to paclitaxel: a woman treated for breast cancer with the drug experienced a clearance of psoriasis (Ferrari & Jirillo, 1996). While several case studies have supported this claim, large scale clinical trials have not been carried out. Tamoxifen ranked high at position 7 on 1994-95 WtProp ranking (Appendix 3C).

A summary of the drugs reprofiled for psoriasis and discussed here is presented in Table 3.10.

Table 3.10 Psoriasis – selected reprofiled chemicals. Best rank is the highest rank from any test run. HS is hypothesis set. ArtCt is the number of articles connecting the drug to the disease in the test period.

Chemical	Best rank/ HS count	Previous Use / Activity	Status	Art Ct	Reprofiling type
Mycophenolate mofetil	543 / 13,393	Immunosuppressive; transplant	In use; recent clinical trials	20	Functional
Sulfasalazine	171 / 5,532	Crohn’s, Ulcerative Colitis\ anti-inflammatory	Good results in trials	13	Functional
Paclitaxel	66 / 9,192	Breast cancer	Effective in small trial		Observational
Calcitriol	2 / 5,532	Vitamin	In use	353	Class-based
Cholecalciferol	9 / 5,532	Vitamin	In use	41	Class-based
Propylthiouracil	6 / 5,532	Antithyroid, antiproliferative, Immunomodulatory	Good results in small trials	16	Functional
Methimazole	58 / 5532	Antithyroid, antiproliferative	Good results in small trials	7	Functional
Capsaicin	149 / 5,532	Antipruritic, flavoring	Reduced itch in trials	11	Functional
Ranitidine	47 / 9,192	H2 Antagonist/anti-ulcer	No improvement	7	Molecular Functional
Pentoxifylline	20 / 13,393	Antiproliferative, blood flow	Showed no efficacy in trial	5	Functional
Rifampin	1 / 5,532	Antibiotic	Unclear, still under study	6	Functional
Erythromycin	38 / 5,532	Antibiotic	No effect in 2008 trial	4	Functional
Tamoxifen	7 / 13,393	Breast cancer	Effective in case study	3	Observational

3.5.3 Migraine

Overview

Migraine is a chronic neurological disorder affecting nearly 12% of the adult population. It is characterized by often debilitating headache, photophobia, nausea, and phonophobia. Some migraines are accompanied or preceded by an aura. The physiology of migraines is not completely understood, although in recent years enormous progess has been made in understanding the underlying mechanics of the disorder. During a migraine attack, events in the neurological system trigger dilation of the meningeal blood vessels, which in turn causes pain and further disturbances of the nervous system. Because the neural system affects the vascular system, migraine is often considered a neurovascular disorder (Bigal & Krymchantowski, 2006).

Migraine therapies can be divided into two groups: those that prevent an attack and those that treat a migraine once it has begun, a strategy called acute therapy. Acute therapies can further be categorized by whether they are migraine-specific or not. Pain relief medications (aspirin, acetaminophen, opiates, etc.) are non-specific. The acute therapies specific to migraine include ergotamine, dihydroergotamine, and the triptan drugs. The triptan drugs, beginning with the launch of sumatriptan in 1991, represent the most significant introduction to the arsenal of drugs to treat migraine. These drugs are 5-HT1B and 5-HT1D agonists, meaning that they bind and enhance the activity of these 5-HT1 postsynaptic receptors, ultimately causing vasoconstriction. Although highly effective in some patients, binding to the 5-HT1 receptors can also have negative cardiovascular effects. Triptans, for that reason, cannot be prescribed for anyone at risk for cardiac problems. In addition, triptans do not work for everyone (Bigal & Krymchantowski, 2006).

Preventing migraines has proven more challenging than treating migraine attacks. The causes behind an onset of a migraine attack are multifactorial and vary from person to person. Several classes of drugs have commonly been reprofiled in migraine prevention: anti-convulsants, beta-blockers, serotonin antagonists, anti-depressants, and calcium-channel blockers. Given the side effect profiles of the drugs used in prevention, they are not recommended unless the patient has severely debilitating attacks (Bigal & Krymchantowski, 2006). New preventive strategies are sought.

Migraine and reprofiled drugs

In a 2006 review article discussing the emerging drugs for migraine, Bigal and colleagues included a number of potential new treatments. Most of the treatments represent new chemical entities, but there are a few examples of potential drug reprofiling, of which only two could have been found by this ABC study. One of those is the anticonvulsant zonisamide. Like many anticonvulsants, zonisamide was identified as a possible treatment for the prevention of migraines. It has been studied in two clinical trials with favorable results (Bigal & Krymchantowski, 2006). Zonisamide appeared in the hypothesis sets for 1989-90 and 1994-95 and had its first direct link to migraine in 2004. It appeared very low in the 1989-90 set (position 2397 out of 7,122 entries) but by 1994 had risen to position 627 out of 10,467 entries (Appendix 7).

Because zonisamide is in a class of drugs commonly reprofiled for migraine, it would have likely received attention on that basis alone. This type of reprofiling will be termed *class-based reprofiling*.

Another more unexpected example of reprofiling is capsaicin, the pepper extract that also saw reprofiling activity for psoriasis. Capsaicin is known to activate the vanilloid receptors that reside on neurons. Activation of vanilloid receptors is thought to desensitize the nerve fibers. For this reason an intranasal form of capsaicin called civamide has been tested for efficacy against acute migraine in a small clinical trial. Despite nasal burning and lacrimation, many of the patients experienced relief (Bigal & Krymchantowski, 2006). Capsaicin was predicted quite high on each test run. The highest were position 34 in 1984-85, 24 in 1989-90, and 21 in 1994-95.

The 2006 review by Bigal et al. also mentioned a class of drugs under development that target nitric oxide synthase, the protein that produces endogenous nitric oxide. Nitric oxide, in addition to its many other roles, is thought to be behind migraine etiology in some patients. Physicians were alerted to this possibility when patients taking nitroglycerine for heart attacks experienced the onset of migraines. Drugs that inhibit nitric oxide synthase are being investigated. Most of these drugs are new chemical entities and therefore not included on any hypothesis set. The molecule nitric oxide, however, is on the 1984-85 set and ranked by WtCOS at position 19 (Appendix 7A). As mentioned previously, the explosion of investigations into nitric oxide leading up to and following its designation as molecule of the year likely plays a role in its ranking.

In a 1999 review of nutritional and botanical approaches to migraine prevention, two endogenous substances are discussed which may be deficient in migraine patients: magnesium and melatonin (Sinclair, 1999). Studies have shown that supplementing these substances can help reduce the severity and number of migraines. Magnesium concentration in the body has an effect on several important proteins implicated in migraine pathogenesis, including the serotonin receptor (also known as 5-HT receptor) and nitric oxide synthase. Magnesium has also been linked to reduction in vasospasm and platelet aggregation. Magnesium supplements as preventative treatment of migraine were studied in a number of clinical trials. In a randomized, double-blind, placebo-controlled study of 81 patients, magnesium was shown to reduce the attack frequency by 41.6% as compared to the 15.8% in the control group (Peikert, Wilimzig, & Kohne-Volland, 1996). Magnesium sulfate has been shown to be effective as an intravenous treatment for acute attacks (Bigal & Krymchantowski, 2006). Magnesium and magnesium sulfate combined have had over 40 articles connecting it to migraine. All the 1984-85 ABC rankings placed magnesium high, with ProtCt at position 2 and AvgRank at position 11.

Some migraine sufferers have imbalances in their endogenous melatonin levels. Although no large scale blinded and randomized trials have been conducted to study melatonin, a small open-label study was conducted on 22 children with a history of migraine. The subjects took 3 mg of melatonin before bed for three months. Fourteen of the subjects reported significant reduction in migraine attacks and four reported no headaches at all during the study period (Miano et al., 2008). The first year melatonin was directly connected to migraine in ChemoText was 1986. In the Appendix 7A table, we can see that melatonin was ranked at position 34 out of 4,006. The AvgRank and ProtCt rankings were also high.

Next we will examine briefly the tables found in Appendices 7A, 7B, and 7C. These tables list the drugs and endogenous molecules that over time accrued the most articles written about them and give visibility to reprofiled drugs not mentioned in the reviews.

Valproic acid has the highest article count in the 1984-85 table presented in Appendix 7A. Valproic acid is an example of class-based reprofiling. It is an anticonvulsant, and like many in that class before it, was reprofiled for migraine. Valproic acid has been a very successful reprofiling example. Since 1988 when it was first tried in migraine prevention, it has accrued 88 articles connecting it to migraine. The WtProp ranking approach put it at position 72 in the 1984-85 set, where it may have come to the notice of researchers, but it is likely that because it is an anticonvulsant it would have been suggested as a migraine treatment as a matter of course and would not have been studied any earlier had these results been available in 1984. (Valproic acid appeared at position 105 in the pilot study hypothesis set ranked by protein count.)

Similarly, many of the compounds found in Appendices 7A, 7B, and 7C are examples of class-based reprofiling. Acetazolamide and lamotrigine are anticonvulsants; fluoxetine, moclobemide, and sertraline are antidepressants; butorphanol, ketorolac, and dipirone are analgesics. Vomiting is common during migraines; droperidol and ondansetron are antiemetics.

A summary of the drugs reprofiled for migraine and discussed here is presented in Table 3.11.

Table 3.11 Migraine – selected reprofiled chemicals. Best rank is the highest rank from any test run. HS is hypothesis set. ArtCt is the number of articles connecting the drug to the disease.

Chemical	**Best rank/ HS count**	**Previous Use / Activity**	**Status**	**Art Ct**	**Reprofiling type**
Zonisamide	627 / 10,467	Anticonvulsant	Trial successful	4	Class-based
Capsaicin	21 / 10,467	Antipruritic, flavoring, activates vanilloid receptor	Trial successful	12	Molecular Functional
Nitric Oxide	19 / 4,006	Endogenous; NO synthase is target	Inhibitors under development	41	Molecular Functional
Magnesium	2 / 4,006	Endogenous	Used in prevention and acute treatment	40	Molecular Functional
Melatonin	34 / 4,006	Endogenous	Trial showed efficacy	15	Molecular Functional
Valproic acid	72 / 4,006	Anticonvulsant	In use	88	Class-based

3.6 Conclusion

In this chapter an implementation of Swanson's ABC paradigm has been described and evaluated. The evaluation was based on dividing the corpus into two parts by a cutoff year, running the experiment on the earlier data, and validating the results on the data drawn from the latter time period. The goal was to use protein connections between drugs and diseases to predict new uses for existing drugs.

The most important difference between this study and the pilot study was the addition of new ranking approaches and the evaluation of the rankings through the use of metrics devised to evaluate information retrieval applications. Finding a ranking approach (or several approaches) that puts the most significant, relevant, true positives, gold standard entries near the top is critical, particularly in a list of returned entries that is numbered in the thousands.

Each of the ranking approaches was able to put found gold-standard chemicals nearer the top of the list than a randomly ranked list. In many cases the improvement over random was 20-fold. WtProp and ProtCt often had very similar results, but they each had instances when they performed better than the other. WtCOS performed worst overall except for several striking instances – the 1984-85 psoriasis, where it put the drugs with the highest number of articles in the top 20 and 1984-85 migraine where nitric oxide was placed at position 19. There is no obvious need to add another ranking approach. Because they returned different sets of chemicals, one future strategy could be to merge the top drugs from each ranking approach.

This study, like the pilot study, was able to put magnesium in a high position on the 1984-1985 set using the ProtCt ranking approach. This closely reproduces Swanson's findings.

The basis for establishing the implicit connections between drugs and disease was proteins.   The proteins in common between the drug and disease were the basis for putting a drug in the hypothesis set, and some aspect of the protein-disease relationship (e.g., articles in common) was used as input into the ranking mechanisms. The strategy of putting proteins in this central position worked well. There were drugs that did not make it into the hypothesis sets because they had no proteins in common with the protein pools, but they were few, and with a few exceptions, not very significant. Many of the drugs missed by the analyses did in time develop links to the disease through protein annotations. Had the analyses been done at more time intervals, these drugs would have likely been included in the hypothesis sets.

The role of time in this study warrants further discussion. The data upon which this study depended was pulled from the Medline 2009 baseline file. Many articles, hundreds of thousands in fact, have been published since the baseline file was loaded into ChemoText. It is highly likely that more of the chemicals on the hypothesis set have now been associated with cystic fibrosis, psoriasis, or migraine. It would certainly be interesting to rerun the experiment with new data on a regular basis.

ChemoText has proved to be an effective data repository for storing and allowing the programmatic extraction of literature data for these experiments. There are some caveats that must be declared when using ChemoText. Every researcher who uses co-occurrence as a way to find explicit relationships between entities defines what they mean by co-occurrence. It can mean co-mention in an abstract, title, sentence, MeSH annotations, or something else entirely. In this application, co-occurrence is based on the relationship between chemicals and disease and proteins when the chemical is identified in ChemoText as the *subject* chemical. In most cases this design method worked well to reduce noise of incidental and insubstantial connections, although because it is a heuristic algorithm, it was not always correct. But this feature was designed with drugs in mind and does not work as well to depict the relationships between endogenous molecules (including elements) and a disease. Endogenous substances can be annotated many times with a disease before they receive the focus and are deemed the subject chemical by the ChemoText algorithm. The relationship of the element sodium to migraine is a good example. Annotations of sodium appeared in many articles before the article published in 2006 that focuses on the sodium levels in the cerebrospinal fluid. For that reason caution should be exercised before calling a connection between a chemical and a disease a novel one. Connections such as these can also cause rankings to receive high evaluations by MAP, Precision@K, etc. For this reason these evaluations will be used only to compare runs within this implementation and not to the ABC implementations of other researchers. Whatever the definition of a co-occurrence, every researcher must conduct extensive research in many sources before a literature connection is claimed to be a discovery.

An unexpected result from this study has been the insight it has offered into drug reprofiling. We have seen that there are several ways that a drug can be selected for its reprofiling potential. Table 3.12 summarizes the reprofiling approaches we have seen in this study.

Table 3.12 Summary of reprofiling approaches observed in this study

Reprofiling approach	**Description**	**Example**
Functional	Known physiological function of a drug targeted to a different disease	Mannitol known to have diuretic function on kidneys. Made into inhaled form to be used in CF patients to move water to lung surface.
Molecular functional	Molecular function of chemical known – matched to known or hypothesized disease mechanism	Histamine thought to be involved in psoriasis. Histamine antagonist (ranitidine) tried.
Class-based	Certain classes of drugs regularly reprofiled in different indication because previously drugs in that class worked	Anticonvulsants used in migraine prevention.
Observational	Researcher or patient notices improvement in one disease or condition when being treated by the drug for another condition	Breast cancer patients showed improvement in psoriasis when being treated with paclitaxel.

Functional reprofiling seems the most common approach. Functional profiling starts with knowing what activity a drug has in one disease setting (anti-inflammatory, for instance) and translating that function to another disease. Judging from the number of cases we have encountered in this study, functional reprofiling is applied often.

We have seen cases of *molecular functional reprofiling*. This takes place when researchers establish the molecular activity of a drug and they also know the molecular mechanisms behind the disease. They put these two lines of evidence together and hypothesize that the drug may treat the disease.

We also saw examples of *class-based reprofiling,* where researchers reprofiled a drug because other drugs in the same class had previously been successfully reprofiled. This was a commonly seen reprofiling approach in migraine prevention.

Chance or *observational reprofiling* is less commonly seen than the other reprofiling approaches. In these instances a drug is studied or used for one indication and is by chance observed to treat another condition. Chance reprofiling receives the most press because of the famous example of sildenafil (Viagra) (Bradley, 2005). While this drug was in clinical trials for angina, male participants and the researchers noticed and appreciated the occurrence of penile erections shortly after taking the drug. Sildenafil was reprofiled for male erectile dysfunction and has become a blockbuster. In the studies described here we saw several (less famous) examples of observational reprofiling.

The three diseases selected for this study proved highly informative about the varying approaches to drug reprofiling. In many ways the diseases are very different. Cystic fibrosis is a genetic disease that slowly causes loss of lung function and eventually - generally before the age of forty - the disease is fatal. Although it is generally long-term and has a genetic component, psoriasis is irritating and uncomfortable, but rarely fatal. Migraine is episodic, but when it strikes, it can be debilitating. Both treatment of the migraine attack and prevention are important aspects of the therapy. The manifestations of CF develop slowly in internal organs; psoriasis shows itself on the skin.

Despite these differences, researchers in each of these diseases have used reprofiling as a method to find new therapies, alongside the development of new chemical entities. The examples of reprofiling we have seen were mostly functional reprofiling, based on knowledge of the disease and drug mechanism, and transferring that function from one disease to another. We did see a few examples of observational reprofiling with psoriasis: both tamoxifen and paclitaxel were observed to improve psoriasis symptoms when they were administered to cancer patients. Cystic fibrosis is likely less amenable to observational reprofiling because it works on the less visible parts of the body.

Functional reprofiling in CF was seen in the transfer of diuretic action from kidneys to lungs in the cases of mannitol and furosemide. Ranitidine was reprofiled to help improve fat absorption in patients whose cystic fibrosis had affected the function of their liver and pancreas. Warfarin, caffeine, and edetic acid were reprofiled to test and measure the effect of the disease on organ function. Although clinical research is always cautious, reprofiling in CF seemed more circumspect than in psoriasis, involving more preliminary *in vitro* studies to establish efficacy before clinical trials were undertaken.

Psoriasis has a long and colorful history of reprofiling, both by patients and by medical professionals. The knowledge that psoriasis is an immune system disorder spurred many experiments in reprofiling drugs with known immunomodulatory activity. These included mycophenolate mofetil, propylthiouracil, and methimazole. Capsaisin was reprofiled to target itching. On the molecular level, researchers suspected that histamine might play a part in the etiology of psoriasis and tried ranitidine, a histamine blocker. Some attempts to reprofile do not follow a direct path. Rifampin was tried on guttate psoriasis patients with the reasoning that it would reduce the bacterial load, but after positive results were obtained, more studies were done that determined the antibacterial action of rifampin played no role in its efficacy, leading researchers to suspect the drug had immunomodulatory effects. Even when functional reprofiling fails, researchers can learn from their experiments.

Migraine, too, has a solid history of reprofiling, particularly for preventative therapeutics, where class-based reprofiling is particularly common. The same classes of drugs (e.g., anticonvulsants, beta-blockers, antidepressants) are routinely tried in migraine prevention.

While we did see reprofiling for acute migraine in the case of capsaicin, reprofiling in general is not as important in the acute treatment of migraine as it was in psoriasis or CF. The success of the triptan drugs has been followed by intense research into the protein receptors involved in migraine and new chemical entities are being developed that target these receptors in different ways. A number of new chemical entities were in development for their activity against nitric oxide synthase; although these compounds are too new to be picked up by this ABC study, nitric oxide was identified.

Although the term reprofiling is not generally used in the context of vitamin and mineral supplements, we did see novel application of supplements. Given the high ranking of both magnesium and melatonin in these results, it is possible that that literature connections can indicate what endogenous molecules should be examined in a disease context to see if their levels play a role in the disease onset.

Future Directions

The question of how high on a ranked hypothesis set a drug should be in order to be noticed by researchers is a question with no absolute answer. The answer depends heavily on the context these studies are performed in. If the output of these analyses is to be examined manually by a researcher, then it is likely that only the top couple of hundred may ever be considered.

The purpose of this research, however, is to determine whether this literature-based tool could fruitfully be used in a computational drug discovery laboratory as an additional tool to help understand the working of drugs and to find new therapies for disease. Such laboratories employ many computer-based techniques to analyze drugs and have many resources to draw on. In such a context, the limitations of manually analyzing the ABC output are less relevant.

In the computational drug research lab, the commonly applied methods center on chemical structure and the relationship of that structure to molecular and clinical activity. Like the ABC study described here, some of the methods produce large lists of chemicals hypothesized to have therapeutic use in a particular disease. The hypothesized drug candidates are tested in wet lab experiments such as protein binding assays. This step is expensive and generally an effort is made to send to the lab only those drugs with a high likelihood of testing positive.

It is desirable therefore to investigate other bodies of information that might strengthen or weaken the case of the compounds so that only the strongest candidates move to the wet lab. The ABC analysis can play this role. Results from the ABC analysis can be used to eliminate some candidates or to increase the confidence in others. Conversely, the ABC analysis could be used to generate hypothesis sets that are subsequently passed through screening routines using QSAR models for the second stage of hypothesis strengthening or elimination.

This combination of ABC results and the results of another validated hypothesis-generating tool may work synergistically to highlight the candidates most likely to succeed in the clinic. Indeed, as drug research becomes more expensive and high risk, every line of evidence that can be brought to bear to identify and prioritize potential therapies should be explored.

4. PREDICTING DRUG MOLECULAR ACTIVITY FROM SIDE EFFECTS

4.1 Introduction and Background

In the last chapter the connections between biomedical entities present in the literature were used to predict new therapies for disease. The goal of this study is to explore the possibility that patterns in the side effect profiles of drugs can predict their molecular activity.

Determining the molecular activity of a drug can be another way to initiate drug reprofiling. In the last chapter this type of reprofiling was termed *molecular functional reprofiling*. It starts with observing the molecular level activity of a molecule and then combines that knowledge with the diseases that might benefit therapeutically from such molecular activity. To take an example from the previous chapter, the triptan drugs so important in the acute treatment of migraine are all 5-HT1B/D agonists. This means that they bind and enhance the work of the 5-HT1B and 5-HT1D receptors. If a drug with previously untested activity at this receptor was found to bind to 5-HT1B and 5-HT1D in a laboratory experiment, then that drug might be a candidate for migraine therapy. Often the complete picture of the molecular mechanisms of the action of a drug is unknown even when it has been used successfully to treat a disease. The discovery that it binds to a protein related to a different disease may be a signal that it could be reprofiled. Binding to an unexpected target is called *off-target binding*.

One of the main endeavors in a computational drug research laboratory is to predict the molecular activity of drugs. Quantitative structure activity relationship (QSAR) techniques are commonly used to find elements in the chemical structure called descriptors that can be used in statistical algorithms in order to predict activity. The study described in this chapter has the same goal as a QSAR experiment – to predict the molecular activity of drugs, and the experiments have a similar design. Instead of chemical descriptors, however, these studies use side effect terms drawn from the literature as the basis for prediction.

4.1.1 Previous Work

Physicians and drug researchers have known for a long time that a relationship exists between the molecular activity of a drug and its clinical effects. Serotonin syndrome, for instance, is the name given to a set of physical symptoms associated with long term use of drugs that have an effect on the serotonin receptors.

One of the first *computational* studies to examine the relationship between side effects and molecular activity was conducted by Fliri, et al. (2005). They looked at the relationship from a global perspective by examining data from protein binding assays alongside side effect information. They found a strong correlation between binding patterns and side effect patterns.

Campillos et al. (2008) used the relationships illustrated by Fliri in order to predict off-target binding. They created side effect vectors by extracting adverse effect terms from drug package inserts and mapping the terms to a controlled vocabulary. They then calculated a normalized pairwise vector similarity between each pair of drug in their set. Because they were looking for off-target or unexpected binding, they eliminated pairs of drugs known to bind to the same targets. They also eliminated drugs that because of chemical structure similarity would have been likely to bind to the same targets. Of the resulting 121 drugs with the highest similarity score, twenty were tested in *in vitro* binding assays. Thirteen of these drugs bound to the predicted targets and subsequent cell assays were used to confirm nine drug-protein interactions. From these strong results they filed two new patent applications.

4.1.2 Data sources

Molecular Activity

There are two sources for molecular activity information used in this study. First, 5-HT6 binders and nonbinders will be extracted from the PDSP K_i database (Roth et al., 2000). This database is a resource supported by the National Institute of Mental Health Psychoactive Drug Screening Program. PDSP K_i contains receptor binding results for psychoactive drugs and receptors involved in pathways important to the nervous system. Some of the results stored in the database are established experimentally by the Roth lab and some are collected from the literature.

The other source of molecular activity is the MeSH pharmaceutical action codes. These codes are assigned to chemicals by the indexers at the National Library of Medicine and are available online or from a file that can be downloaded from the MeSH web site. Examples of the types of pharmaceutical actions available through this resource are listed in Table 4.1.

Table 4.1 Sample MeSH Pharmaceutical Action records

Pharmaceutical Action	Chemical Name
Adrenergic Agonists	adrafinil
Adrenergic Agonists	Albuterol
Adrenergic Agonists	amidephrine
Adrenergic Agonists	amitraz
Adrenergic Agonists	anisodamine
Adrenergic Agonists	Apraclonidine
Adrenergic alpha-Antagonists	Phenoxybenzamine
Adrenergic alpha-Antagonists	Phentolamine
Adrenergic alpha-Antagonists	phenylpiperazine
Adrenergic alpha-Antagonists	Piperoxan
Adrenergic alpha-Antagonists	Prazosin

The pharmaceutical action designations differ from the binding data stored in PDSP K_i. On one hand they are more informative. They describe what kind of activity the drug has because of its binding, whether the binding blocks the normal action of the protein (antagonists) or enhances it (agonists). On the other hand, the pharmaceutical action is less specific about which receptor is blocked or enhanced. The code may designate Dopamine Agonist or Histamine Antagonist, but not give any information on which of several dopamine receptors D1, D2, D3 are enhanced, or which of the histamine receptors H1, H2, or H3.

Side effect data

Side effects are clinical manifestations of a drug treatment that are unplanned for or unexpected and often adverse. Studies that infer molecular activity from side effect information are uncommon in drug research, likely because of the difficulty in establishing a corpus of side effect data. Until very recently there was no publicly available resource with clinical effects data structured for use in computational experiments. On the other hand, there are many sources of side effects recorded in textual format, including drug package inserts, web sites, and the biomedical literature.

Fliri et al. (2005) used a commercially available database called CEREP BioPrint to retrieve their side effect profiles. Campillos et al. (2008) used text mining techniques to extract terms from package insert pdf files downloaded from various web sites. Each package insert was put through a series of processing steps that extracted the side effects terms and mapped them to a standard vocabulary using the COSTART (Food and Drug Administration, 1989) data source. In January of 2010 this data was made available to the public and it is now the only public source of side effect data for marketed drugs (Kuhn, Campillos, Letunic, Jensen, & Bork, 2010).

Many articles published in the biomedical literature discuss the side effects of drugs. Some of these effects are included in the MeSH annotations and will therefore be extracted and stored in ChemoText. As a result, ChemoText is also a source of side effect information.

MeSH annotations of side effects or adverse effects can be differentiated from annotations of therapeutic effects by subheadings or qualifiers. The subheadings such as *adverse effects* indicate that the effect is unwanted and probably adverse, what we are calling a side effect. When these effects are identified and loaded into the ChemoText Disease Table, the field called TreatFlag is set to *Cause*. The process by which the ChemoText processing identifies side effects is described is detailed in Chapter 2.

For this study, a separate side effects table called CTSideEffects was created from the Disease Table. This table was built by pulling all the records in the Disease Table with the Treat Flag equal to *C* (cause). Two additional filters were applied to the records. First, side effects were limited to those occurring in an article with only one subject drug. In articles with more than one subject drug, such as comparative studies, it was impossible to tell which of the drugs caused the effects. For this reason these articles were omitted from this analysis. An additional filter was put on species to ensure that only studies performed on humans were included in CTSideEffects. *Drug effects* annotations occurring in articles with an adverse event disease annotation were also extracted.

The data in CTSideEffects was evaluated as a data source for this study in two ways. First, the side effects for specific drugs were examined and compared to the side effects described in that drug's package insert, the document that could be considered the gold standard. Second, counts of chemicals and their side effects were calculated in order to get an idea of the scope of CTSideEffects.

The side effects in the package insert were manually compared to the side effects in CTSideEffects for several drugs. The results for one of these drugs, risperidone, are shown in Table 4.2. The left side of the table contains the side effects extracted from the Warnings and Precautions and Adverse Reactions section of the package insert for risperidone. The right hand column contains the CTSideEffects annotation for risperidone which was thought to be the closest in meaning. The MeSH Browser was used to look up terms and their meanings and possible synonyms. The weakest correlations between terms from each source are indicated by italics. For instance, *Nausea* could not be found in the CTSideEffects terms. *Abdominal pain* was found in CTSideEffects and it may be related to nausea. The terms are not synonyms, however, and the weakness of this correlation is indicated by italics. In parentheses is a PubMed ID from one of the articles in which the annotation was found. Note that often the language varies between the two sources even though the meaning is the same.   The package insert term *Dysphagia* and the MeSH term *Deglutition Disorders* both mean having difficulty in swallowing, and the MeSH term *Sialorrhea* means *Saliva Increased*, the term seen in the package insert.

Table 4.2 Concordance of side effects reported in the package insert vs. CTSideEffects for drug risperidone. PMID is PubMed identifier for an example of an article annotated with that effect. Italics indicate a MeSH annotation more weakly linked to the package insert term.

Package Insert	**CTSideEffects Entry (PMID)**
Cerebrovascular Events, incl. stroke	Stroke (12451085)
Neuroleptic Malignant Syndrome	Neuroleptic Malignant Syndrome (15495506)
Tardive dyskinesias	Dyskinesia, Drug-Induced (15363485)
Hyperglycemia and diabetes mellitus	Hyperglycemia(16395845), Diabetes Mellitus (11526997) (Type 1 and 2)
Hyperprolactinemia	Hyperprolactinemia (17519641)
Orthostatic Hypotension	Hypotension, Orthostatic (9496415)
Potential for cognitive and motor impairment	*Parkinson Disease, secondary (8990067)*
Seizures	
Dysphagia	Deglutition Disorders (14571332)
Priapism	Priapism (12716256)
Suicide	
Somnolence, Fatigue	Disorders of Excessive Somnolence(16965213), Fatigue(11757991)
Appetite Increased	Appetite Regulation(17199131), Obesity(14961939), *Weight Gain(18759643)*
Rhinitis	*Respiration Disorders (15795553)*
Upper respiratory tract infection, cough	Cough(12717324), *Dyspnea (10756565)*
Vomiting, Nausea, Dyspepsia	*Abdominal pain(17984854)*
Urinary incontinence	Urinary incontinence(18387724)
Saliva increased	Sialorrhea(11351120)
Constipation	
Fever	Fever(17119106)
Parkinsonism	Parkinson Disease, secondary (10087680)
Dystonia	Dystonia(8862861)
Abdominal pain	Abdominal pain (17984854)
Anxiety	
Dry mouth	
Tremor	Tremor(10087680)
Rash	
Akathisia	Akathisia, Drug-Induced (16013909)

In general there was a high concordance between ChemoText side effects for risperidone and those in the package insert. There were, however, examples of side effects occurring in one source but not the other. Some package insert side effects (e.g., *Constipation*, *Rash*, and *Dry Mouth*) were not found in ChemoText. There were also annotations in ChemoText that were not found in the package insert. *Jaundice* for instance was found annotated in CTSideEffects, but was not seen in the package insert. While *Rash* (or the MeSH term *Exanthema*) was not found in CTSideEffects, several skin conditions were found: *Erythema Multiforme, Pruritus,* and *Pemphigoid, Bullous*. Similarly, *Rhinitis* was not found in the MeSH annotations for Risperidone, although several annotations indicating an adverse effect on respiration were found, including *Respiration Disorders* and *Respiration Insufficiency*. A search in PubMed (*risperidone[majr] AND rhinitis*) yielded several mentions in abstracts of risperidone causing rhinitis(e.g., PMID 15056514), but these connections between drug and disease did not make it into the annotations.

The comparison of the package inserts to CTSideEffects brought to light some other characteristics of each data source. The package insert will often contain information about the percentage of patients experiencing the side effect in both the test group and the control group. MeSH annotations do not indicate side effect prevalence. Some side effects are annotated many times with a drug, but it is difficult to know whether high occurrence rates indicate that the side effect occurs commonly or is a severe effect, both, or neither.

While there is much similarity in the language used in package inserts, there is no enforced controlled vocabulary. MeSH side effects are pulled from a controlled vocabulary. The MeSH vocabulary, however, often lacks the specificity of the package insert terms. While the package insert states *Appetite Increased*, the more general MeSH term states *Appetite, Appetite Regulation*, and *Hunger.* These terms do not indicate whether these conditions are increased or decreased. It is difficult to assess whether the lack of specificity poses a problem when analyzing the data. Fliri and colleagues mapped specific side effects to body systems, but they were still able to find a strong relationship between effects and binding.

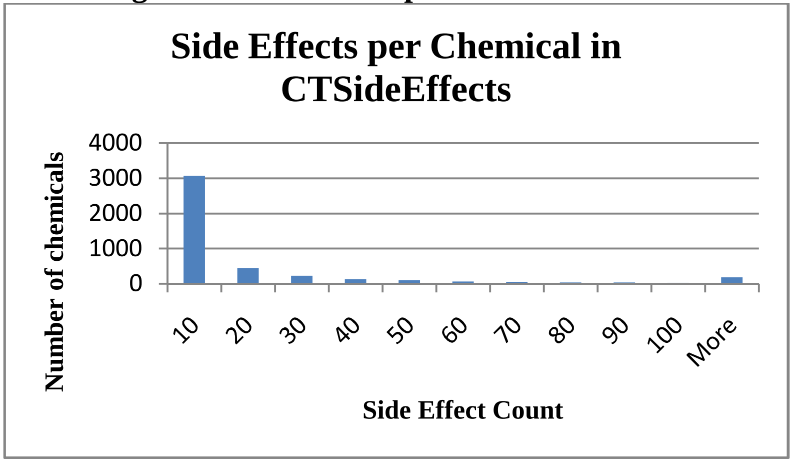
Because the CTSideEffects are drawn from literature annotations, they have some other inherent weaknesses. Negative results are not annotated in a way to differentiate them from positive results. The drug lisuride, for instance, was studied to see if it had the potential to cause cardiac myopathies. It was found not to bind to the receptor responsible for cardiac myopathies. Despite these negative findings, the annotations, and the resulting CTSideEffects entries, were the same as it would be if lisuride *did* cause cardiac myopathies. Negative findings such as this are not common, but they introduce an element of noise into the data.

The indexers apparently annotate the most important or most discussed side effects, but do not document every side effect mentioned in the study. The side effects therefore are not as exhaustive as side effects listed in the package insert. Therefore, there are fewer records in ChemoText for drugs with relatively few side effects or relatively mild side effects.

A global comparison of literature side effects to package inserts offers some interesting observations. The scope of the literature is broader than the scope of the package inserts. Any chemical that is the subject of an article will be included in PubMed annotations, whereas the package insert is a document prepared under a very specific set of circumstances - when a prescription drug is approved in the United States. Approved prescription drugs comprise a small subset of the chemical space and are a subset of the drug space as well. Investigational drugs, drugs pulled from the market, and drugs approved outside the U.S. may not have a package insert, but they will very likely have a literature record.

The CTSideEffects table has 4,393 chemicals with at least one side effect. The number of side effects per drug varies greatly. Most of the chemicals in CTSideEffects have only a handful of annotated side effects, while some have hundreds. Ethanol has the most with 655, followed by methotrexate with 573. A histogram of the side effect counts is seen below in Figure 4.1. Approximately 1,100 chemicals have 15 side effects or more.

Figure 4.1 Histogram of side effects per chemical in CTSideEffects



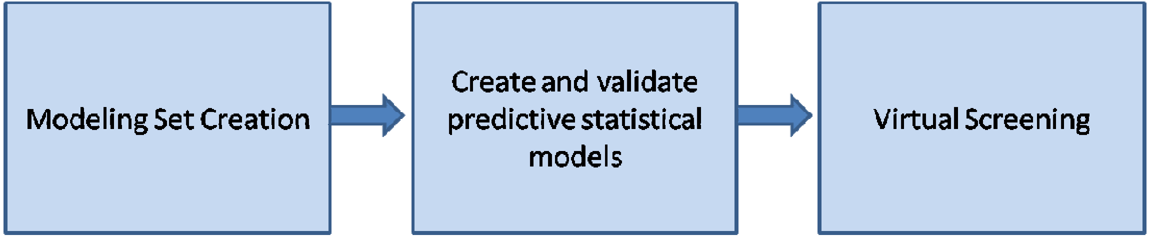
4.2 Overall design

The goal of this study is to investigate whether side effects are predictive of two different molecular activities: 5-HT6 receptor binding and dopamine antagonism. 5-HT6 is one of the many serotonin receptors. (5-HT or 5-hydroxytryptamine is a synonym for serotonin.) 5-HT6 binders are thought to have potential in enhancing cognition deficits related to Alzheimers (Geldenhuys & Van der Schyf, 2009; Mitchell & Neumaier, 2005). 5-HT6 binders were chosen because they were the subject of a recent QSAR study in the Molecular Modeling Laboratory at UNC (Hajjo, Fourches, Roth, & Tropsha, 2009).

Dopamine antagonists are typically used as anti-psychotics, anti-emetics, and antidepressants. Dopamine antagonists were chosen because there are a substantial number of dopamine antagonists identified in the MeSH Pharmaceutical Action file.

The overall design of each experiment is depicted in Figure 4.2 below. The terminology used in this chapter is defined in Table 4.4. The three major steps in the process are 1) create the modeling datasets, 2) build statistical models that predict the molecular activity, and, 3) perform virtual screening of a large set of chemicals (screening set) to identify potential chemicals with the desired activity (5-HT6 binders or dopamine antagonists).

Figure 4.2 Overall design of side effect prediction studies.



The modeling datasets consist of side effect vectors, one vector per drug. The side effect data is extracted from the CTSideEffects table. Each vector position corresponds to one side effect. A 1 in the position indicates the drug has been annotated with that side effect; a zero indicates the drug has no record for that side effect in the table. Each vector also contains the class variable. For the 5-HT6 study, this variable will indicate whether the drug is a 5-HT6 binder or nonbinder and for the dopamine antagonism study the variable will indicate whether or not the drug is a dopamine antagonist. A simplified illustration of the modeling set construction is pictured in Table 4.3.

Table 4.3 Illustration of side effect vectors in a modeling set. Each chemical is called an instance and each side effect is an attribute.

	Nausea	Vomiting	Dizziness	Sleep Disorders	Hyperprolactinemia	Hypotension	Deglutition Disorders	Stroke	Dystonia	**CLASS**
Chem 1	1	0	0	0	0	0	0	0	0	1
Chem 2	0	0	1	1	1	1	1	1	1	0
Chem 3	1	1	1	1	1	1	0	0	0	1
Chem 4	0	0	0	0	1	1	0	1	1	1

In the second step of the study the models will be built. This step is broken into smaller substeps. First, several classifiers and attribute selection algorithms are run against the modeling sets to find the combinations of classifiers and attribute selection methods that perform the best. To perform this testing, 80% of the modeling set (the training set) will be used to train the classifier and the resulting model will be tested on the remaining 20% of the modeling set. This procedure will be termed *80/20 validation*. The best performing combinations of classifier and attribute selection algorithm will be further validated by Y-randomization and any weak performers will be eliminated. The selected algorithms will be trained on the modeling set to produce the final models. The Weka machine learning tool will be used for classification (Hall et al., 2009).

The final models will be used in virtual screening. The purpose of the virtual screening is to *predict* the molecular activity in chemicals where it is so far unknown. If the models are robust this step may identify novel drug candidates. In this step the models are run against a screening dataset. This dataset contains side effect vectors of all the drugs from

CTSideEffects that were not included in a modeling set. Each model is applied to the screening set and each chemical is predicted to be a binder (antagonist) or a nonbinder (not an antagonist). The prediction is accompanied by a probability.

Table 4.4 Selected terms used in this study.

Term	Meaning
Class	What is being predicted. In this case either 5-HT6 binding or dopamine antagonism.
Modeling set	Set of chemicals (positive and negative instances) with known class variable.
Instances	The members of the modeling set. In this case, chemicals with known class.
Attributes	The characteristics of the instances that are being analyzed to see if they predict the class, i.e., side effects.
Training set	All or some of the instances in the modeling set that are used in to train the classifier.
Test set	Some of the instances from the modeling set which are not used in training. The model constructed from the training set is used against the test set to measure how well the model performs.
Screening set	Large pool of chemicals with unknown class for which the class will be predicted.
Model	A classifier algorithm and attribute selection algorithm trained on a dataset
CTSideEffects table	ChemoText table with MeSH annotations of disease extracted from articles where the TreatFlag=Cause.

4.3 Methods

4.3.1 Predicting 5-HT6 binding using side effects

Step 1. 5-HT6 - Preparation of Modeling Sets

The PDSP database (version kidb100108) was downloaded in January, 2010 and searched for all drugs that have been tested against the serotonin 5-HT6 receptor. In the cases where the PDSP chemical names did not match the MeSH names, a manual lookup step was necessary to map the names. For instance, the PDSP name *Acetylsalicylic Acid* was mapped to the MeSH term *Aspirin*. Many PDSP chemicals did not have entries in MeSH. Some are in early stages of drug development and do not have a literature record. Several filters were applied to the PDSP data. Only assays performed against human cloned proteins were included. The K_i values for all entries meeting the filtering criteria were averaged. Drugs with average K_i values less than 10,000 nm were considered binders. Drugs with K_i greater than or equal to 10,000 were considered nonbinders. Sumatriptan was omitted because of conflicting results.

In preliminary work, we found that setting a threshold for side effects improved the classification results. This is likely because few side effects create a very sparse dataset and therefore are weak predictors. All chemicals that had fewer than 15 side effects therefore were eliminated from the study. Campillos et al. (2008), likely for similar reasons, applied a similar threshold to the side effect count when creating their vectors. Twenty-nine 5-HT6 binders and twenty nonbinders met the inclusion criteria. The drugs are listed in Appendix 8.

This set of drugs has two weaknesses as a classification dataset. First, the number of binders is greater than the number of nonbinders. This imbalance in classes will reduce the accuracy of the predictive models. Because there are no more eligible instances of nonbinders in PDSP, random drugs were randomly drawn from CTSideEffects to augment the nonbinders. To reduce the chances that these drugs were 5-HT6 binders, drugs known to bind to any of the serotonin receptors were omitted. The second limitation of the dataset is that the PDSP drugs are biased toward psychoactive compounds and therefore not representative of the screening set. Randomly pulling drugs from CTSideEffects will not eliminate this bias, but it may weaken its effects. In three rounds, nine drugs were selected randomly and classed as nonbinders and added to the known binders and nonbinders. The resulting three datasets will be termed the 5HT6Set1, 5HT6Set2, and 5HT6Set3. The composition of binders and nonbinders for each set is presented in Table 4.5. Second, the PDSP drugs are likely biased toward psychoactive drugs and therefore not representative of the pool of drugs used in the screening step.

Table 4.5 5-HT6 Binding : Composition of modeling sets. Mapping refers to the step of mapping side effects to more general MeSH Tree node.

Modeling Set	Binder Count	True non-binder Count	Presumed nonbinder Count	Total	Side effect count before cleanup / mapping	Side effect count after cleanup / mapping
5HT6Set1	29	20	9	58	1408	368
5HT6Set2	29	20	9	58	1316	333
5HT6Set3	29	20	9	58	1385	351

The number of unique side effects in each modeling set was very large and would have yielded large, sparse vectors. To reduce the dimensionality of the vectors that were produced, a subset of the side effects was mapped to a more general effect using the MeSH Tree file. In addition, side effects only annotated with one or two of the drugs were removed because they would have little predictive power. The 15 side effect threshold was applied to the set before these mapping and cleanup steps.

The mapping to more general descriptors was carried out by programmatically looking up each side effect in the MeSH Tree file and mapping it to a higher (broader and more general) level in the tree. The MeSH Tree file contains the MeSH annotation hierarchy and allows one to find annotations higher and lower on the tree. If an annotation term was more specific than level 3 it was replaced by the descriptor at level 3. (Level 3 is the way we will refer to the number of nodes, where a node is three digits separated by period.) Table

4.6 illustrates how this summarization step changes the data using the example of the level 3 node *Bone Diseases, Infectious*. This table shows all the MeSH disease and condition annotations that were mapped to *Bone Diseases, Infectious*.

In preliminary work we tried grouping the side effects at various levels. We found that results were somewhat better if two categories of side effects, movement disorders and cardiovascular effects, were not mapped to a more general descriptor. In both of these studies, therefore, annotations in these two categories were left at their original level of specificity. These categories of side effects play a large role in the receptors studied and the specificity of the annotation was likely important. Column 6 of Table 4.5 shows the number of side effects that were included in the set before the steps were taken to reduce the dimensionality. The reduced number of side effects (and therefore the number of vector positions) for each modeling set is displayed in the last column.

The drug side effect vectors were created. In each position of the vector a 1 or a 0 was entered indicating that the drug was or was not annotated to this side effect (or category of side effect). Each vector also contained a class variable.

Table 4.6 Illustration of side effect summary using MeSH Tree file hierarchy. The annotations in column 2 were mapped to the higher level annotations in column 3 before creation of the side effect vector.

MeSH Tree Category	**Annotated side effect**	**Higher level**
C01.539.160.412	Osteitis	Bone Diseases, Infectious
C01.539.160.495	Osteomyelitis	Bone Diseases, Infectious
C01.539.160.595	Periostitis	Bone Diseases, Infectious
C01.539.160.762	Spondylitis	Bone Diseases, Infectious
C01.539.160.762.301	Discitis	Bone Diseases, Infectious

Step 2. 5-HT6 - Model Creation

The three modeling sets are very similar. They differ only in the nine randomly selected nonbinders. Because of these nonbinders, however, the predictive models created from them will perform differently on the virtual screening set. It is not possible to know which of the randomly selected nonbinders are the most representative and therefore provides the best training data. For that reason models were built on each of the three modeling sets for use in screening, and the prediction results were averaged. It is hoped that this step compensatedfor any bias inherent in any one of the sets.

The two major components of a model are the attribute selection algorithm and the classifier. The Weka machine learning tool implements many different attribute selection algorithms and classifiers. Two attribute selection algorithms and two classifiers showed strong performance in preliminary work and were evaluated on each modeling set. These algorithms are described in Table 4.7.

Table 4.7 Classifiers and attribute selection algorithms used in model building

Classifiers	
Short Name	**Description**
NB	Naïve Bayes
Bagging	Combines results from NB, Random Forest, and K-nearest neighbor(IBk)
Attribute selection algorithms	
Short Name	**Description**
Subset	CfsSubsetEval: Selects features or attributes that are correlated highly with the class, but are not highly correlated with each other
Chi-squared	Uses the chi-squared statistic to evaluate the importance of each attribute to the class.

The 12 models (combinations of attribute selection, classifier, and modeling set) were tested in 80/20 validation. In this step the modeling sets were segmented. Eighty percent or 4/5 of the modeling set was randomly selected to train the classifier and build a model. The model was used to predict the binding on the remaining 20 percent of the modeling set. The exercise was repeated 50 times. Sensitivity, specificity, and the correct classification rate (CCR), the average of sensitivity and specificity, and the standard deviation were calculated for each run. The results are presented in Table 4.8.

Sensitivity is calculated as follows:

True Positives / (True Positives + False Negatives)

Specificity is calculated as follows:

True Negatives / (True Negatives + False Positives)

CCR or correct classification rate is the average of specificity and sensitivity:

(Sensitivity + Specificity)/2

Six models (shown in bold) were selected from these 12 models from the first step. Many of the original models showed an imbalance of sensitivity and specificity. The two best models for each modeling set were selected based on a high CCR and a balance between specificity and sensitivity. Each of these was then validated further using Y-randomization. In this validation technique, a training set was built by extracting a random 80% of the modeling set and setting the class variable of these instances randomly to one or zero (representing bind and nobind). This scrambled set was used to train the classifier and then the model was tested against the corresponding test set. Sensitivity, specificity and CCR were calculated. Because a high CCR in Y-randomization indicates the model is weak, any model with a CCR greater than .60 was eliminated. There were none which fit these criteria. Results from Y-randomization are in Table 4.9.

Table 4.8 5-HT6 Binders : Results from 80/20 validation. Descriptions of classifiers and attribute selection methods are in Table 4.7. Models selected for use in virtual screening are in bold.

Modeling Set	Classifier	Attribute Selection	Sensitivity Avg	Specificity Avg	CCR Avg	CCR StdDev
5HT6Set1	Bagging	Chi-squared	0.78	0.76	0.77	0.13
5HT6Set1	Bagging	Subset	0.82	0.73	0.78	0.10
5HT6Set1	NB	Chi-squared	0.88	0.66	0.77	0.11
5HT6Set1	NB	Subset	0.78	0.74	0.76	0.10
5HT6Set2	Bagging	Chi-squared	0.83	0.74	0.79	0.11
5HT6Set2	Bagging	Subset	0.77	0.78	0.77	0.14
5HT6Set2	NB	Chi-squared	0.86	0.68	0.77	0.11
5HT6Set2	NB	Subset	0.69	0.79	0.74	0.12
5HT6Set3	Bagging	Chi-squared	0.87	0.77	0.82	0.11
5HT6Set3	Bagging	Subset	0.87	0.76	0.81	0.11
5HT6Set3	NB	Chi-squared	0.93	0.68	0.80	0.09
5HT6Set3	NB	Subset	0.83	0.78	0.80	0.13

Table 4.9 5-HT6 Binders : Results from Y-randomization. Descriptions of classifiers and attribute selection methods are in Table 4.7. Good models will have low sensitivity, specificity, and CCR.

Model	Modeling Set	Classifier	Attribute Selection	Sensivity Avg	Specificity Avg	CCR Avg
5HT6Model1	5HT6Set1	Bagging	Chi-squared	0.81	0.27	0.54
5HT6Model2	5HT6Set1	Bagging	Subset	0.44	0.60	0.52
5HT6Model3	5HT6Set2	Bagging	Chi-squared	0.25	0.32	0.28
5HT6Model4	5HT6Set2	Bagging	Subset	0.45	0.39	0.42
5HT6Model5	5HT6Set3	Bagging	Chi-squared	0.70	0.16	0.43
5HT6Model6	5HT6Set3	Bagging	Subset	0.71	0.44	0.58

Step 3. 5-HT6 - Virtual Screening

Each of the six selected models was retrained on the entire modeling set and saved. A screening set was constructed by extracting any chemical from CTSideEffects that was not in a modeling set and had greater than 14 side effects. Vectors were created for the screening set in a procedure similar to the modeling sets. The screening set had 1,089 chemicals.

The saved models were used to predict the binding of the chemicals in the screening set. For each chemical a prediction (bind or no bind) was produced in addition to a probability measure. Six sets of predictions were produced, one for each model. The results were merged and the probabilities were averaged.

4.3.2 Predicting dopamine antagonists using side effects

Step 1. Dopamine antagonists – Creation of modeling sets

The methods used to predict dopamine antagonism were similar to those above, except in the construction of the modeling sets. The known dopamine antagonists were identified by finding the MeSH chemicals with the pharmaceutical action *Dopamine Antagonists*. Twenty-six drugs were identified that were dopamine antagonists and also met the side effect cutoff. These drugs are listed in Appendix 9.

Six modeling sets were constructed. In each of the sets the 26 dopamine antagonists were used as the positive instances. The assembly of the negative instances varied. For three of the modeling sets the negative examples were pulled randomly from the pool of drugs in the CTSideEffects table. It is being assumed because the drug is not designated as a dopamine antagonist that the drug indeed is not a dopamine antagonist. Each of the first three sets had a different set of randomly selected instances assumed to be negative.

For the other three modeling sets, the negative instances were drawn from PDSP. Twenty-four to 26 drugs tested and determined to be nonbinders to any dopamine receptor were randomly chosen from the 34 drugs that were nonbinders and met the side effect count threshold of 15. These modeling sets have the advantage of containing tested negatives. If the drugs do not bind to dopamine they cannot be dopamine antagonists. However, these sets also have the disadvantage of being skewed toward psychoactive drugs because they are drawn from PDSP. It was hoped that having modeling sets with negatives instances drawn in various ways will give robust results when the predictions are combined in the virtual screening step.

Table 4.10 Dopamine antagonists : Construction of modeling sets (DA=dopamine antagonists). Mapping refers to the step of mapping side effects to more general MeSH Tree node.

Set Name	How were negative instances selected?	True DA Count	Negative Count (not DA)	Side effect count before clean up / mapping	Side effect count after clean up / mapping
DASet1	Randomly from CTSideEffects	24	25	1,093	258
DASet2	Randomly from CTSideEffects	24	24	944	223
DASet3	Randomly from CTSideEffects	24	26	1,039	250
DASet4	Randomly from PDSP dopamine non-binders	24	25	1,292	324
DASet5	Randomly from PDSP dopamine non-binders	24	24	1,293	324
DASet6	Randomly from PDSP dopamine non-binders	24	25	1,215	297

Step 2. Dopamine Antagonists – Creating models

Each of the six modeling sets was trained with the bagging and Naïve Bayes classifiers in combination with each of the attribute selection algorithms. Each model was tested in 50 iterations of 80/20 validation. The sensitivity, specificity, CCR, and the standard deviation of the CCR were calculated and averaged. The averages are recorded in Table 4.11. The models with the high CCR results and a good balance between sensitivity and specificity were selected. At least one model per modeling set was selected. The selected models are in bold.

Table 4.11 Dopamine Antagonists: Model performance in 80/20 validation. Selected models are in bold.

Model Components			Results			
Dataset	**Classifier**	**Attribute Selection**	**Average Sensitivity**	**Average Specificity**	**Average CCR**	**StdDev CCR**
DASet1	**Bagging**	**Chi-squared**	**0.88**	**0.88**	**0.88**	**0.11**
DASet1	Bagging	Subset	0.83	0.86	0.85	0.13
DASet1	NB	Chi-squared	0.88	0.82	0.85	0.14
DASet1	NB	Subset	0.82	0.87	0.84	0.12
DASet2	Bagging	Chi-squared	0.96	0.93	0.94	0.08
DASet2	Bagging	Subset	0.99	1.00	0.99	0.04
DASet2	NB	Chi-squared	0.81	0.93	0.87	0.14
DASet2	NB	Subset	0.99	1.00	0.99	0.04
DASet3	Bagging	Chi-squared	0.91	0.88	0.89	0.10
DASet3	Bagging	Subset	0.85	0.88	0.86	0.09
DASet3	NB	Chi-squared	0.91	0.73	0.82	0.13
DASet3	NB	Subset	0.84	0.89	0.87	0.10
DASet4	Bagging	Chi-squared	0.92	0.74	0.83	0.10
DASet4	**Bagging**	**Subset**	**0.90**	**0.80**	**0.85**	**0.10**
DASet4	NB	Chi-squared	0.92	0.47	0.70	0.12
DASet4	NB	Subset	0.90	0.78	0.84	0.09
DASet5	Bagging	Chi-squared	0.93	0.70	0.82	0.11
DASet5	**Bagging**	**Subset**	**0.92**	**0.75**	**0.83**	**0.12**
DASet5	NB	Chi-squared	0.92	0.48	0.70	0.12
DASet5	NB	Subset	0.93	0.73	0.83	0.11
DASet6	Bagging	Chi-squared	0.93	0.79	0.86	0.11
DASet6	**Bagging**	**Subset**	**0.94**	**0.84**	**0.89**	**0.09**
DASet6	NB	Chi-squared	0.91	0.56	0.74	0.11
DASet6	NB	Subset	0.95	0.83	0.89	0.09

The six selected models were validated further using Y-randomization. The results are displayed in Table 4.12 below. All models passed this validation step and were used in the virtual screening.

Table 4.12 Dopamine Antagonists: Y-randomization results on selected models.

Model	Dataset	Classifier	Attribute Selection	Sensitivity Avg	Specificity Avg	CCR Avg
DAModel1	DASet1	Bagging	Chi-squared	0.75	0.20	0.47
DAModel2	DASet2	NB	Subset	0.82	0.17	0.49
DAModel3	DASet3	Bagging	Chi-squared	0.37	0.40	0.39
DAModel4	DASet4	Bagging	Subset	0.77	0.30	0.54
DAModel5	DASet5	Bagging	Subset	0.18	0.58	0.38
DAModel6	DASet6	Bagging	Subset	0.77	0.14	0.45

Step 3. Dopamine Antagonists – Virtual Screening

A virtual screening set was created from chemicals drawn from CTSideEffects that were not in any of the modeling sets and passed the side effect count threshold. Each of the six selected models was run against the screening set. The prediction and score from each run were stored and the average score from the six runs was calculated.

4.4 Results

4.4.1 5-HT6 Binding

The 1089 chemicals in the 5-HT6 binder screening set were analyzed by each of the final models in order to predict whether the chemical was a 5-HT6 binder. Forty-five (45) chemicals were predicted by all models to be 5-HT6 binders. Five hundred and ninety-three (593) were predicted by all models to be nonbinders. Two hundred eighty-three (283) chemicals had an average score greater than 0.5 and therefore are predicted binders overall. The drugs with the highest probability score are listed in Table 4.13 below.

Table 4.13 5-HT6 Screening Results. Chemicals predicted to be 5-HT6 binders with highest average probability. Probability scores returned by each model are listed next to average.

Chem Name	Average	5-HT6 Model1	5-HT6 Model2	5-HT6 Model3	5-HT6 Model4	5-HT6 Model5	5-HT6 Model6
Mirtazapine	0.94	0.89	1.00	0.84	1.00	0.92	1.00
Phenelzine	0.89	0.71	1.00	0.78	1.00	0.86	1.00
Metoclopramide	0.86	0.81	0.94	0.75	1.00	0.75	0.93
Reboxetine	0.86	0.82	0.98	0.65	0.89	0.90	0.94
Bupropion	0.85	0.76	0.89	0.77	0.98	0.80	0.90
Tiapride	0.83	0.68	0.98	0.69	0.96	0.70	0.97
Sultopride	0.83	0.73	0.93	0.78	0.98	0.69	0.87
Triazolam	0.83	0.71	0.93	0.77	0.94	0.75	0.88
Clomipramine	0.83	0.81	0.71	0.77	0.96	0.82	0.91
Sodium_Oxybate	0.83	0.71	0.98	0.67	0.89	0.79	0.94
Sertraline	0.83	0.74	0.94	0.73	0.98	0.68	0.89
Fluvoxamine	0.82	0.71	0.90	0.83	1.00	0.80	0.69
Levodopa	0.82	0.77	0.97	0.79	0.98	0.65	0.76
Domperidone	0.82	0.59	0.95	0.73	1.00	0.65	0.97
Modafinil	0.81	0.68	0.98	0.62	0.89	0.76	0.94
Apomorphine	0.81	0.74	0.88	0.73	0.93	0.77	0.80
Citalopram	0.81	0.73	0.81	0.70	0.88	0.74	0.98
Disulfiram	0.80	0.66	0.91	0.65	0.97	0.72	0.92
Oxazepam	0.80	0.64	0.98	0.66	0.89	0.70	0.94

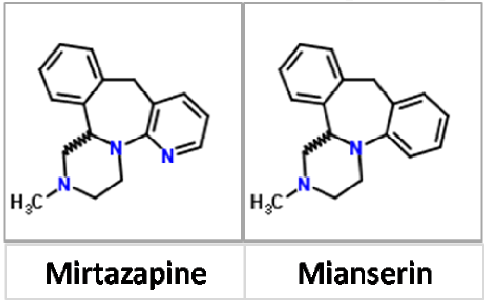
These drugs all have some known molecular activity. This established activity and its relationship to the predicted 5-HT6 binding activity is summarized in Table 4.14 and will be discussed briefly. The web resources DrugBank (Wishart et al., 2008) and the MeSH browser were used to gather this information.

Table 4.14 Known activities of high predicted potential 5-HT6 binders

Chemical Name	**Description**
Mirtazapine	Analog of mianserin, a known 5-HT6 binder
Phenelzine	Monoamine oxidase inhibitor (MAOI)
Metoclopramide	Serotonin (5-HT) antagonist and dopamine antagonist
Reboxetine	norepinephrine reuptake inhibitor (NRI)
Bupropion	Inhibits reuptake of norepinephrine, dopamine, and serotonin; Anti-cholinergic activity
Tiapride	Dopamine antagonist
Sultopride	Dopamine antagonist
Triazolam	GABA neurotransmitter enhancement
Clomipramine	Selective serotonin reuptake inhibitor(SSRI), norepinephrine reuptake inhibitor (NRI)

Mirtazapine appears at the top of the results list with an average probability of 0.94 of being a binder to the 5-HT6 receptor. Mirtazapine has not been tested against 5-HT6. It is, however, a close analog of the drug mianserin which is a known 5-HT6 binder (Figure 4.2). Chemicals that have a high structural similarity often have similar molecular activity. It is very likely therefore that the top predicted chemical is indeed a 5-HT6 binder.

Figure 4.3 Chemical structures of mirtazapine (left) and mianserin (right). Mianserin is a known 5-HT6 binder and mirtazapine is predicted to be one.



The next highest ranked drug on the screening results is the antidepressant phenelzine. It is known to be a monoamine oxidase inhibitor. Monoamine oxidase breaks down monoamines that are responsible for signaling. Serotonin is one of the monamines. By inhibiting the oxidase, the the breakdown of serotonin is blocked, resulting in increased levels of serotonin. While we do not know if the prediction that phenelzine is a 5-HT6 binder is correct, we do know that it has an effect on a serotonin pathway.

Similarly, metoclopramide and bupropion are also known to have effects on the serotonin pathway. Metoclopramide binds to and blocks at least one 5-HT (serotonin) receptor. Bupropion inhibits the reuptake of serotonin into the neuron.

Clomipramine has actually been tested in 5-HT6 binding assays that were completed after the build of the PDSP database used in this study. The drug was indeed found to bind to 5-HT6 with a nanomolar concentration of 112. Clomipramine was predicted by this side effect study correctly. Two other drugs that were tested positive as binders in later binding assays were also found by their average score in this study to be binders: nortriptyline and doxepin. In the same batch of tests, however, two drugs were found to be actual binders to 5-HT6 that were not predicted so by this study – raloxifene and tamoxifen. The average probability for these two drugs was under 0.50. A number of other drugs tested in this batch were not included in this study because they did not meet the side effect count threshold. Table 4.15 contains a summary of the results.

Table 4.15 5-HT6 Binding results not included in PDSP and predicted 5-HT6 binding from side effect profiles.

Chemical Name	**Binding Assay Data**			**Screening Prediction**
	% Inhibition	Ki(nM)	Binder?	Avg Probability
Clomipramine	98.6	112	Yes	0.83
Nortriptyline	99.1	214	Yes	0.71
Doxepin	98.1	105	Yes	0.72
Raloxifene	88.2	750	Yes	0.35
Tamoxifen	91.1	1,041	Yes	0.42

4.4.2 Dopamine antagonists

The 976 chemicals in the screening set were analyzed by each of the final models in order to predict whether the chemical was a dopamine antagonist. Thirty-six (36) chemicals were predicted by all models to be dopamine antagonists. Seven hundred and eight (708) were predicted by all models not to be dopamine antagonists. Seventy-five (75) chemicals had an average score greater than 0.5 and therefore are predicted overall to be dopamine antagonists. The top 14 (0.85 or greater) of the 36 chemicals predicted by all models to be dopamine antagonists are listed in table 4.16 below. These 14 chemicals received the highest average probability.

Table 4.16 Dopamine antagonist - predictions

		DA	DA	DA	DA	DA	DA
Chemical Name	**Avg**	**Model1**	**Model2**	**Model3**	**Model4**	**Model5**	**Model6**
Molindone	0.96	0.83	1.00	0.94	1.00	1.00	1.00
Tetrabenazine	0.95	0.81	1.00	0.90	1.00	1.00	1.00
Fluphenazine depot	0.95	0.76	1.00	0.93	1.00	1.00	1.00
Cetirizine	0.92	0.81	1.00	0.80	1.00	0.90	1.00
Trihexyphenidyl	0.90	0.68	1.00	0.79	0.96	1.00	1.00
Benztropine	0.89	0.63	0.99	0.74	1.00	0.96	0.99
Ziprasidone	0.88	0.84	1.00	0.80	0.80	1.00	0.86
Potassium Cyanide	0.88	0.60	0.92	0.82	1.00	0.96	0.97
Veralipride	0.87	0.79	1.00	0.85	0.80	0.95	0.84
Pemoline	0.86	0.66	0.99	0.75	0.99	0.81	0.95
Pirenzepine	0.85	0.74	1.00	0.75	1.00	0.63	0.97
Diphenhydramine	0.85	0.58	0.96	0.83	0.81	0.94	0.97
Bromazepam	0.85	0.52	0.92	0.72	1.00	0.96	0.97
Sertraline	0.85	0.67	1.00	0.66	0.95	0.84	0.96

According to DrugBank, molindone occupies dopamine receptor sites in the brain and decreases dopamine activity. Although the site does not use the term antagonists, the terms it does use describe antagonist activity. Molindone is a likely dopamine antagonist.

Tetrabenazine is used to treat movement disorders. DrugBank reports that it works as an inhibitor of monamine transport (dopamine is also a monamine) and as such promotes the early degradation of dopamine. This activity may have many of the same effects as a dopamine antagonist and it may be the reason this drug was predicted to be a dopamine antagonist with a fairly high probability.

Ziprasidone is a known dopamine antagonist that was inadvertently omitted from the modeling set. It was however in the screening set and identified correctly as a dopamine antagonist with a high probability. Fluphenazine depot is an analog of fluphenazine and veralipride is an analog of sulpiride. Both of these drugs are known dopamine antagonists. It is therefore likely that fluphenzine depot and veralipride are dopamine antagonists as well.

On the other hand, there seems to be no connection between cetirizine and dopamine antagonism. Cetirizine is a histamine H1 antagonist used in the treatment of rhinitis, urticaria, and asthma. Curiously, the poison potassium cyanide causes movement problems as the poisoning progresses, and these effects are likely the reason the chemical scored highly.

Both triheyxphenidyl and benztropine, while structurally dissimilar, are both M1 muscarinic acetylcholine receptor antagonists used to treat the extrapyramidal symptoms of parkinsons disorders. They are also both thought to increase the availability of dopamine. Their possible effect on the dopamine pathway in addition to their association with movement disorders may account for their relatively high average prediction scores. The information for these drugs is summarized in Table 4.17.

Table 4.17 Predicted dopamine antagonists. Information primarily taken from DrugBank and MeSH browser.

Chemical Name	Description of uses and known molecular activities
Molindone	Used to treat psychotic symptoms. Known to occupy dopamine receptor sites and decrease dopamine activity.
Tetrabenazine	Used to treat movement disorders. VMAT inhibitor which promotes early degradation of dopamine.
Fluphenazine depot	Analog of fluphenazine, a known dopamine antagonist
Cetirizine	Used in treatment of rhinitis and asthma. Histamine H1antagonist.
Trihexyphenidyl	Used to treat extrapyramidal symptoms of parkinsons. M1 muscarinic acetylcholine receptor antagonist. Also thought to increase availability of dopamine.
Benztropine	Similar to trihexyphenidyl. Used to treat extrapyramidal symptoms of parkinsons. M1 muscarinic acetylcholine receptor antagonist. Also thought to increase availability of dopamine.
Ziprasidone	Known dopamine antagonist.
Potassium cyanide	Poison. Can cause movement disorders.
Veralipride	Analog of sulpiride, a known dopamine antagonist

PDSP was examined to see if any of the top predicted dopamine antagonists (Table 4.17) had been tested in dopamine binding assays. Binding is a prerequisite to antagonism. Only molindone and ziprasidone had been tested. Molindone was found to bind to the dopamine D2, D3, and D4 receptor subtypes. Ziprasidone was found to bind to the dopamine D1, D2, D3, D4, and D5 receptor subtypes.

4.5 Discussion

The models for the dopamine antagonist study were strong. The average sensitivity and specificity were 0.92 and 0.86 for the models selected for virtual screening and the average CCR was 0.89. The dopamine antagonist datasets constructed with negative instances pulled from PDSP resulted in models with weaker sensitivity and specificity in the validation steps than the models created from datasets with negative instances randomly selected from the CTSideEffects pool of chemicals. This difference likely reflects the strong bias in the composition of PDSP toward drugs in specific psychoactive drug classes.

Dopamine antagonists are known for the movement impairments associated with their use. These side effects are termed extrapyramidal symptoms (EPS). The range of symptoms includes the inability to start movement, called *akinesia*, as well as the inability to refrain from moving (*akathesia* or *dyskinesia*). The EPS were reflected in the side effects chosen by the attribute selection algorithm in Weka to have the highest discriminatory power. Five of the top ten side effects identified by the chi-squared attribute selection process were some type of movement and muscular disorders. The right hand column of Table 4.18 contains an example taken from one of the selected dopamine antagonism models.

Table 4.18 Sample of most discriminative side effects for the dopamine antagonism study.
Dyskinesia Drug Induced
Dystonia
Movement Disorders
Brain Diseases
Muscle Rigidity
Akathisia, Drug-induced
Puerperal Disorders
Stomatitis
Gastroenteritis
Salivary Gland Diseases

In the 5-HT6 models, the accuracy varied as the negative instances were selected differently. Overall, however, the accuracy of the 5-HT6 binding models was considerably lower than the accuracy of the dopamine antagonist models. The average CCR of the final models was 0.79, as compared to 0.89, the average CCR for dopamine antagonist study. The models with the best CCR were unbalanced, showing high sensitivity and low specificity.

The specificity results were less than 0.80 for all the selected models. Low specificity indicates that the models were not strong in identifying negative instances.

In the validation process there were 5-HT6 binders that were consistently misclassified. Ketanserin was one of these drugs. Ketanserin is highly promiscuous, binding to many receptors including several in the serotonin (5-HT) family, histamine H1, and the alpha-1 adrenergic receptor. This promiscuity may be the cause of side effects that are unrelated to 5-HT6 and consequently may have weakened the modeling set. In general, serotonin binding is known to be promiscuous (Roth et al., 2000). The training set may have contained a number of other 5-HT6 binders that likely fall into this category and contributed to the weak performance of the classifier.

Another likely contributor to the low prediction rate of the 5-HT6 models is that *binding* was predicted and not what happens *after* binding. Binding can result in promoting the activity of the receptor or blocking the activity of the receptor. These two actions can result in very different sets of downstream effects. The modeling set for 5-HT6 may contain some agonists and some antagonists and the divergent side effect profiles may not contain enough common ground to produce good models for binding.

The topmost ranking chemicals in Tables 4.13 do have a high likelihood of being predicted correctly as 5-HT6 binders. We have seen that mirtazapine is a close chemical analog of a drug known to be a 5-HT6 binder and this relationship increases the chances that mirtazpine will be a binder. Beyond the first few drugs, however, there may be other biological reasons for their high scores. Each of these drugs has some known molecular functions that would influence the classification process. The drug phenelzine, for instance, is a known monoamine oxidase inhibitor. This activity has a net effect of increasing serotonin levels. While it may also be a 5-HT6 binder, its already known role in the serotonin pathway may be responsible for some of its side effects.

Drugs that modulate serotonin receptors or serotonin levels can also affect dopamine levels (Di Giovanni, Di Matteo, Pierucci, & Esposito, 2008). This pathway interaction or crosstalk between pathways may account for the overlap in side effects identified as significant by the attribute selection routines. Movement disorders were significant side effects for both dopamine antagonists and 5-HT6 binders, although they were less significant for 5-HT6 binders. Movement disorders represented two of the top ten side effects with the highest discriminatory power in one of the 5-HT6 models with attributes determined by chi-squared (Table 4.19). Movement disorders represent half of the top ten side effects in one representative dopamine antagonism run (Table 4.18). Having movement disorders in common may be the reason that two known dopamine antagonists, tiapride and sultopride, were predicted with high probability to be 5-HT6 binders. These drugs may indeed be 5-HT6 binders, or the side effects arising from their dopamine antagonism may make them look like 5-HT6 binders.

Table 4.19 Sample of most discriminative side effects for the 5-HT5 binding study.
Behavioral Symptoms
Gastrointestinal Hemorrhage
Dystonia
Dyskinesia, Drug-Induced
Peptic Ulcer
Skin Diseases, Vascular
Hypersensitivity, Intermediate
Sexual Dysfunction, Physiological
Puerperal Disorders
Arrhythmias, Cardiac

The drugs that were tested in a 5-HT6 binding assay after the download of PDSP (Table 4.15) provide an opportunity to check the screening results for these drugs. All of the five drugs were true binders, but only three were identified as binders in the screening process. Only one (clomipramine) was predicted with a high probability to be a binder. Tamoxifen and raloxifene were incorrectly predicted by this study to be nonbinders. It is interesting that these two drugs showed the lowest affinity for the receptor and the lowest percent inhibition. While this is an interesting observation, more cases need to be studied to see if binding affinity has any consistent relationship to side effects.

ChemoText has been a robust source of side effect information for this study. This repository has several advantages over a data source constructed from processing the text of package insert. First, the coverage of the chemical space is significantly broader than package inserts. Second, the MeSH side effects are publicly available in electronic format, making them easy to gather and access. The collection of drugs and side effects will be updated automatically during the yearly update of ChemoText.

Future Work

The feasibility and benefit of combining the side effect annotations stored in ChemoText with side effects drawn from package inserts should be investigated. It is possible that the side effects from package inserts will augment the ChemoText records. With better side effect coverage, more drugs may meet the side effect count threshold, making the modeling sets larger and the models potentially more robust. Fortunately, a structured source of package insert side effects called SIDER (Kuhn et al., 2010) became available in early 2010. This resource could facilitate combining side effects from the two sources.

The annotations that fall under the category of drug effects include many types of effects that are not related to adverse events. Many studies, for instance, report on the cellular level effects of drugs (e.g., apoptosis, mitosis). These effects could be used in addition to adverse effects to give the classifiers more attributes to choose from in the attribute selection process.

Animal side effects can be explored as well. Drugs undergo extensive animal testing before human trials and the side effects are reported in the literature. The data on animal trials in ChemoText is extensive, but it is fragmented among various species. It would have to be determined whether the data for each species should be considered separately or could be combined.

Other sources of molecular activity data should be investigated. There are many other public and commercial sources of binding and activity information that could potentially be used. PubChem, for instance, as the central repository for the Molecular Libraries Roadmap Initiative, is a growing resource for many kinds of chemical assays.

Other prediction methods may yield better results. Campillos et al. (2008) used a similarity search approach in their study. This approach may be better suited to the complex polypharmacology of psychoactive drugs in particular (Keiser et al., 2009). Visualization tools and other machine learning software may provide additional insight into the side effect data.

In several cases (e.g., mirtazapine) the methods predicted binding activity in chemicals that are structural analogs of known 5-HT6 binders. We can be fairly sure in these cases that the predicted chemical is indeed a binder. While this is a welcome validation of these side effect based methods, these predictions are not useful in practical terms. Structure-based QSAR methods would have been able to identify these chemicals as binders. We would like these new methods to predict binders in drugs whose structure is *dissimilar* to known binders and therefore the structure-based methods would be inadequate. If the side effect methods can identify such drugs, then we have found a way to complement and enhance the QSAR methods in use in the lab.

Campillos et al. (2008) had a similar goal and eliminated structurally similar chemicals from their prediction set using a structural similarity measure called the Tanimoto coefficient. What remained were chemicals unexpectedly linked to binding through their side effects alone. We could employ a similar technique in our future work. The Tanimoto coefficient could be calculated between each predicted binder and each known binder in the modeling set. Drugs with high similarity could be flagged and omitted from the results. The remaining drugs would be those that *only* side effect data predict as binders.

4.6 Conclusion

The goal of this study was to develop a literature-based methodology to hypothesize new uses for drugs by predicting their molecular activity. The molecular activity of a drug indicates how it might be reprofiled. Dopamine antagonists are used as antipsychotics, anti-emetics, and antidepressants. 5-HT6 binders are thought to have potential in treating Alzheimers.

This study is the first of its kind. No other researcher has constructed predictive models for receptor binding and antagonism from side effect annotations extracted from the biomedical literature. It has necessarily been exploratory in nature.

The models constructed to predict dopamine antagonism performed better than the 5-HT6 binding models in validation runs performed in Weka. Although more experiments are needed to generalize from these results, it does make sense that side effect profiles would be more indicative of antagonism than simple binding. Binding can result in two very different sets of effects, depending on whether the receptor activity is blocked or enhanced.

Dopamine antagonists are well-known for their extrapyramidal side effects. These prevalent and serious side effects likely helped the performance of the classifiers. We did not directly test whether dopamine agonists could be reliably discriminated from dopamine antagonists. This is a study planned for future work.

The 5-HT6 prediction models produced results well above random in validation procedures and the drugs returned by the virtual screening step with the highest probabilities look like they may indeed be 5-HT6 binders. Clomipramine, a drug tested after the publication of the version of the PDSP database used, was indeed found to be a binder with moderate affinity. On the other hand, tamoxifen and raloxifene, also confirmed binders, were predicted to be nonbinders.

The methods described here show promise in identifying drugs with specific molecular activity which could be the basis for reprofiling the drug for a new therapeutic indication. In addition, the literature-based discovery methods introduced here have the potential of bringing new insight into the complexity of chemical and biological interactions in the human body.

5. CONCLUSION

This dissertation research investigated two different literature-based discovery methodologies to determine their potential in identifying new uses for drugs, or drug reprofiling. Both studies used data in the ChemoText knowledgebase and both included validation steps.

The first method, referred to as ABC, took advantage of the rich literature connections between disease, proteins, and drugs to predict new uses for existing drugs. The strategy of using protein annotations as the intermediary B terms was very effective in finding chemicals that developed links to the diseases under study. The recall was very high. The reason for this likely lies in the central role proteins play in both disease and drug research. The study of disease increasingly focuses on the physiology of the disease state at the protein level. Drug research focuses on proteins as well, searching for drugs that will modulate the behavior of proteins involved in the disease pathway. Although proteins may be in common between the two fields, the literatures may not always interact and the authors may not be totally aware of each others' work, giving rise to potential undiscovered implicit relationships between chemicals and disease.

The validation method used in the ABC study was based on dividing the corpus into two segments based on a year cutoff. The earlier or baseline period was used to create the hypotheses and the later period was used to validate the hypotheses. The large hypothesis sets and the small number of gold standard chemicals meant that although recall was high, overall precision was very low. Ranking the hypothesis sets is a way to compensate for low precision. Rankings that effectively put the gold standard chemicals toward the top allow the practitioner to choose cutoffs that are likely to give the desired levels of precision and recall. The rankings in this study, particularly ProtCt and WtProp, turned out to be very robust. The average precision for the top 50 chemicals ranked by the WtProp or ProtCt approach was over 26% (Table 3.8). This represents more than a ten-fold improvement over the 2% precision of the random ranking.

In practice, the acceptable levels of precision and recall (and sensitivity and specificity) are decided by the user based on what is to be done with the results. If, for example, an expensive laboratory test were to be run on the top ten chemicals in a hypothesis set, then precision may be more important than recall; with high precision, the lab tests are more likely to return positive results. The goal of this dissertation work, however, is to develop methods that can be used in coordination with the other computational methods in place in the drug discovery lab, methods like QSAR. These other methods produce prediction sets as well. The predictions from various lines of evidence can be combined or compared to arrive at a consensus prediction and the weakest candidates can be removed. Low precision ceases to be a significant problem when computational techniques such as these can be applied to reduce and strengthen the hypothesis set.

While the ranking results were good, they did not provide specific information about reprofiling. In order to evaluate the performance in identifying reprofiled drugs, actual examples of reprofiling were gleaned from review articles and compared to the results. We were able to confirm that many drugs reprofiled in practice were ranked highly by at least one of the ranking approaches. This step demonstrated a link between these results and actual discovery. Had the results been available in the baseline period, they may have indeed have accelerated the drug discovery process.

The design of the study allowed the focus to move back and forth in time. In the later test period the significance of an emergent link between a drug and a disease was measured by the article count, the number of articles in which the drug, as a subject chemical, was co-annotated with the disease. Article count proved a useful tool to measure the significance of a connection between the drug and the disease.

This study was able to reproduce Swanson's link between magnesium and the prevention of migraines. In the 1984-85 time period magnesium was placed at position two in the ranking based on protein count. Forty (40) articles were found in the test period to link magnesium to migraine. Two other chemicals identified in the same time period developed an even stronger connection to migraine: nitric oxide and the anticonvulsant valproic acid. They were both ranked highly by at least one of the ranking approaches. Despite all the literature-based discovery projects endeavoring to reproduce Swanson's migraine-magnesium connection, no one has identified the strong link between these chemicals and migraine. (Swanson himself, however, noted the connection between epilepsy and migraine. (Swanson, 1988))

An unexpected result of this ABC study was the light it shed on the practice of drug reprofiling. Discussion of reprofiling in the pharmaceutical literature is generally limited to a few well-known cases, such as sildenafil (Viagra) for erectile dysfunction and bupropion for smoking cessation. In practice, at least for the diseases studied here, reprofiling was a common approach to finding new drug therapies.

There are many ways this methodology could be extended and enhanced. The methods should be applied to a variety of other diseases in order to establish whether the methods can be extended successfully or if there are diseases where different strategies should be explored. The role played by time in these studies is worthy of more attention. We saw definite trends in the growth of the protein pool, hypothesis sets, and gold standard terms over time. Treating time as a variable and performing the same analyses with varying temporal cutoffs would help further address the robustness of the models and evaluation techniques, as well as provide fruitful insight into the role that time plays in the evolution of discoveries.

The second study in this dissertation research used patterns in the side effect annotations of drugs to predict molecular activity. This study was novel in several ways. Whereas other studies have used side effects from package inserts, this study uses side effects annotations pulled from Medline records and stored in ChemoText. This study also focused on a particular molecular activity and trained and validated classifier models to predict that activity. The validated models were used in virtual screening to predict 5-HT6 binding and dopamine antagonism in a large library of chemicals where those activities were previously unknown.

The side effect study was challenged by biological complexity of neurotransmitter pathways. Dopamine and serotonin pathways intersect and interact with each other and therefore a drug working on one pathway may affect the other pathway. The side effects may be the downstream effect of either one of the pathways. Drug promiscuity also added a challenging complexity to the data. Psychoactive drugs notoriously act on many receptors. Untangling the clinical effects from each receptor would likely require more sophisticated techniques and significantly more data, including nontextual data such as chemical structure. Despite the challenges, the validation results were strong, particularly for the dopamine antagonist models, and the studies were able to identify examples of 5-HT6 binders and dopamine antagonists, respectively.

Validation is an indispensible component of the research methods in the drug discovery laboratory. For that reason, validation has been placed in a central position in the design of these studies. The ABC study started with the validation and evaluation guideline set down by previous researchers (Yetisgen-Yildiz & Pratt, 2009) but also included a comparison to random ranking, as well as the evaluation of reprofiling through manual examination of review articles. The design of the side effect study followed the design of QSAR experiments, and therefore adopted and adapted the stringent validation steps implemented in those studies.

Historically, validation has not been a strong component of literature-based discovery methodologies. This is unfortunate, because validation is essential. Literature-based discovery is a tool, and with any tool, it is vital to know where to apply it: where it works and where it does not work. Without the measuring stick provided by validation, researchers cannot be sure they have learned something from their experiments. Any field of study needs these measures to move forward, and the lack of them may be the reason that the field of literature-based discovery has progressed more slowly than it should have. The studies presented here demonstrate that literature-based methods can be validated just like methods based on laboratory data.

Through its distillation of a large body of chemical and disease research, ChemoText has proved itself to be a rich source of information for drug discovery.   There is no other repository that contains MeSH terms structured in a way to be useful in drug discovery algorithms. ChemoText adds value to MeSH annotations with its routines that identify the subject chemical, in addition to the way it organizes and links the annotations. The complexity and dynamic nature of the literature means that improving these routines will likely continue to be an ongoing activity. In addition to maintenance and enhancements, there are also plans to make ChemoText publicly available.

Future work should go beyond data improvements and methods development. The end goal of this work is to discover new therapeutic uses for drugs. To see that goal realized, these literature-based methods must be adopted in the computational drug discovery laboratory and put to use on real, substantive problems. The question of how to integrate these methods with the toolset already in use in the lab remains the next significant challenge.

APPENDICES

Appendix 1. Proteins excluded from all protein pools

(MeSH category D12- amino acids, peptides, and proteins)

Protein Name
Amino Acids
Aminopeptidases
Antibodies
Antibodies, Monoclonal
Antibodies, Viral
Antilymphocyte Serum
Autoantibodies
Bacterial Proteins
Caerulein
Captopril
Carrier Proteins
Cytokines
Dietary Proteins
Enzyme Precursors
Enzymes
Fenclonine
gamma-Globulins
Gelatin
Globulins
Glycoproteins
Hydrolases
Immune Sera
Immunoglobulins
Isoantibodies
Isoenzymes
Lipoproteins
Macroglobulins
Mucoproteins
Neoplasm Proteins
Nerve Tissue Proteins
Oligopeptides
Papain
Peptide Fragments
Peptides
Pituitary Hormones
Placental Hormones
Plant Proteins
Pregnancy Proteins
Protein Kinases

Protein Precursors
Protein Subunits
Proteins
Proteoglycans
Proteolipids
Proteome
Receptors, Cell Surface
Receptors, Drug
Receptors, Peptide
Receptors, Virus
Recombinant Proteins
Recombinases
Ribonucleases
Serum Albumin, Bovine
Transcription Factors
Vasopressins
Vegetable Proteins
Viral Proteins
Xenopus Proteins

Appendix 2. Cystic Fibrosis: Top 20 chemicals returned by each ranking

The columns with white background represent data from the Baseline Period. The gray columns are drawn from the Test Period. ProtCt is the count of proteins from the protein pool the chemical has annotated with it. FirstYr is the first year the chemical appears as the subject chemical in an article that also has an annotation of the disease. DisQual and ChemQual are the most common disease qualifiers (or subheadings) and chemical qualifiers (subheadings) appearing in the annotations when the chemical is annotated with the disease.

Appendix 2A. Cystic Fibrosis 1984-1985					
Ranked by ProtCt					
ChemName	**Protct**	**FirstYr**	**ArtCt**	**DisQual**	**ChemQual**
Edetic Acid	173	1985	3	complications	pharmacokinetics
Cortisone	164	0	0		
Chlorpromazine	163	0	0		
Mercury	152	0	0		
Cycloheximide	148	0	0		
Lead	147	0	0		
Propranolol	145	1995	1		pharmacology
Phenobarbital	144	1993	1	complications	therapeutic use
Cyclophosphamide	139	0	0		
Morphine	134	1986	3	complications	administration & dosage
Puromycin	132	0	0		
Lithium	131	1990	4	drug therapy	therapeutic use
Diethylstilbestrol	131	0	0		
Chloroquine	131	2003	2	blood	pharmacology
Cadmium	130	1994	1	genetics	toxicity
Indomethacin	129	0	0		
Dimethyl Sulfoxide	128	0	0		
Folic Acid	126	2006	1	drug therapy	pharmacology
Choline	124	2007	1	blood	therapeutic use
Tetradecanoylphorbol Acetate	122	1991	2	genetics	pharmacology
Ranked by WtProp					
Cortisone	164	0	0		
Edetic Acid	173	1985	3	complications	pharmacokinetics
Chlorpromazine	163	0	0		
Propranolol	145	1995	1		pharmacology
Lead	147	0	0		
Mercury	152	0	0		

Cyclophosphamide	139	0	0		
Puromycin	132	0	0		
Chloroquine	131	2003	2	blood	pharmacology
Phenytoin	122	0	0		
Indomethacin	129	0	0		
Vinblastine	112	0	0		
Cycloheximide	148	0	0		
Diethylstilbestrol	131	0	0		
Lithium	131	1990	4	drug therapy	therapeutic use
Gold	110	0	0		
Dimethyl Sulfoxide	128	0	0		
Formaldehyde	121	0	0		
Mercaptoethanol	109	1999	2	physiopathology	
Isoflurophate	99	0	0		
Ranked by WtCOS					
Clomiphene	38	0	0		
20-alpha-Dihydroprogesterone	14	0	0		
ATP gamma-p-azidoanilide	2	0	0		
Procainamide	51	0	0		
Idoxuridine	28	0	0		
Bromocriptine	67	0	0		
Ethyl Biscoumacetate	14	0	0		
Dicumarol	57	0	0		
Congo Red	25	0	0		
Echothiophate Iodide	13	0	0		
testosterone enanthate	6	0	0		
Warfarin	60	1993	2	metabolism	pharmacokinetics
Dihydrotachysterol	20	0	0		
Apomorphine	43	0	0		
Haloperidol	65	0	0		
cholesteryl linoleyl ether	5	0	0		
Molybdenum	53	2001	1	urine	
Metyrapone	50	0	0		
Carbimazole	15	0	0		
sodium thiocyanate	8	0	0		
Ranked by AvgRank					
Adenosine	119	1992	5	metabolism	pharmacology
Cortisone	164	0	0		
Hydrogen Peroxide	115	1998	10	metabolism	metabolism
Choline	124	2007	1	blood	therapeutic use
Dimethyl Sulfoxide	128	0	0		
Bromodeoxyuridine	80	0	0		
Silver	73	2007	1	drug therapy	adverse effects
Dopamine	113	1988	1	blood	blood
Folic Acid	126	2006	1	drug therapy	pharmacology
Tetradecanoylphorbol Acetate	122	1991	2	genetics	pharmacology

Estrone	88	0	0		
Ethinyl Estradiol	109	1987	1	blood	blood
Nandrolone	71	0	0		
Niacin	73	0	0		
Lead	147	0	0		
Bromocriptine	67	0	0		
Lidocaine	72	2001	1	metabolism	analogs & derivatives
Pyridoxine	102	1996	1	metabolism	analysis
Clofibrate	98	0	0		
Furosemide	82	1987	6	metabolism	toxicity

Appendix 2B. Cystic Fibrosis 1989-1990					
Ranked by ProtCt					
ChemName	**Protct**	**First Yr**	**ArtCt**	**DisQual**	**ChemQual**
Tetrad.Acetate	236	1991	2	genetics	pharmacology
Chlorpromazine	208	0	0		
Indomethacin	193	0	0		
Propranolol	189	1995	1		pharmacology
Cycloheximide	187	0	0		
Cortisone	186	0	0		
Chloroquine	182	2003	2	blood	pharmacology
Phenobarbital	180	1993	1	complications	therapeutic use
Lithium	180	1990	4	drug therapy	therapeutic use
Lead	179	0	0		
Cyclophosphamide	179	0	0		
Cadmium	179	1994	1	genetics	toxicity
Mercury	178	0	0		
Dimethyl Sulfoxide	176	0	0		
Tretinoin	176	0	0		
Hydrogen Peroxide	167	1998	10	metabolism	metabolism
Adenosine	166	1992	5	metabolism	pharmacology
Diethylstilbestrol	164	0	0		
Methotrexate	163	2003	1	drug therapy	therapeutic use
Choline	160	2007	1	blood	therapeutic use
Ranked by WtProp					
Cortisone	186	0	0		
Chlorpromazine	208	0	0		
Indomethacin	193	0	0		
Chloroquine	182	2003	2	blood	pharmacology
Propranolol	189	1995	1		pharmacology
Gold	155	0	0		
Lead	179	0	0		
Cyclophosphamide	179	0	0		
Tretinoin	176	0	0		
Dimethyl Sulfoxide	176	0	0		
Mercury	178	0	0		
Lithium	180	1990	4	drug therapy	therapeutic use
Tetra. Acetate	236	1991	2	genetics	pharmacology
Cycloheximide	187	0	0		
Vinblastine	135	0	0		
Diethylstilbestrol	164	0	0		
Cadmium	179	1994	1	genetics	toxicity
Phenytoin	153	0	0		
Choline	160	2007	1	blood	therapeutic use
Methotrexate	163	2003	1	drug therapy	therapeutic use
Ranked by WtCOS					
4-hydroxytamoxifen	14	0	0		
Tamoxifen	93	0	0		
N-Methylscopolamine	9	0	0		

Metribolone	13	0	0		
triperiden	2	0	0		
Congo Red	34	0	0		
Bromocriptine	93	0	0		
20-alpha-Dihydroprogesterone	15	0	0		
otenzepad	5	0	0		
Capsaicin	55	0	0		
Clomiphene	50	0	0		
Apomorphine	60	0	0		
Spiperone	15	0	0		
Quinuclidinyl Benzilate	11	0	0		
Dizocilpine Maleate	7	0	0		
ATP gamma-p-azidoanilide	3	0	0		
Procainamide	65	0	0		
Haloperidol	86	0	0		
Idoxuridine	35	0	0		
Warfarin	71	1993	2	metabolism	pharmacokinetics
Ranked by AvgRank					
Hydrogen Peroxide	167	1998	10	metabolism	metabolism
Bromocriptine	93	0	0		
Tamoxifen	93	0	0		
Estrone	110	0	0		
Adenosine	166	1992	5	metabolism	pharmacology
Niacin	90	0	0		
Dimethyl Sulfoxide	176	0	0		
Lidocaine	95	2001	1	metabolism	analogs & derivatives
Clomiphene	50	0	0		
Haloperidol	86	0	0		
Folic Acid	141	2006	1	drug therapy	pharmacology
Guanosine Triphosphate	113	0	0		
Tetradecanoylphorbol Acetate	236	1991	2	genetics	pharmacology
Clonidine	82	0	0		
Dehydroepiandrosterone	85	0	0		
Pyridoxine	120	1996	1	metabolism	analysis
Deferoxamine	53	0	0		
Calcium, Dietary	53	2004	1	metabolism	pharmacokinetics
Procainamide	65	0	0		
Silver	94	2007	1	drug therapy	adverse effects

Appendix 2C. Cystic Fibrosis 1994-1995					
Ranked by ProtCt					
ChemName	**Protct**	**First Yr**	**ArtCt**	**DisQual**	**ChemQual**
Tretinoin	295	0	0		
Cycloheximide	258	0	0		
Indomethacin	255	0	0		
Hydrogen Peroxide	249	1998	10	metabolism	metabolism
Chlorpromazine	249	0	0		
Dimethyl Sulfoxide	245	0	0		
Lead	243	0	0		
Methotrexate	242	2003	1	drug therapy	therapeutic use
Cyclophosphamide	241	0	0		
Propranolol	240	1995	1		pharmacology
Mercury	237	0	0		
Doxorubicin	232	2001	2	genetics	pharmacology
Cisplatin	230	0	0		
Chloroquine	226	2003	2	blood	pharmacology
Diethylstilbestrol	216	0	0		
Platelet Activating Factor	208	1999	1	blood	administration & dosage
Cortisone	207	0	0		
Nicotine	198	0	0		
Nickel	195	0	0		
Formaldehyde	195	0	0		
Ranked by WtProp					
Chlorpromazine	249	0	0		
Indomethacin	255	0	0		
Lead	243	0	0		
Cyclophosphamide	241	0	0		
Propranolol	240	1995	1		pharmacology
Cortisone	207	0	0		
Mercury	237	0	0		
Dimethyl Sulfoxide	245	0	0		
Cycloheximide	258	0	0		
Chloroquine	226	2003	2	blood	pharmacology
Methotrexate	242	2003	1	drug therapy	therapeutic use
Diethylstilbestrol	216	0	0		
Vinblastine	176	0	0		
Tretinoin	295	0	0		
Gold	181	0	0		
Hydrogen Peroxide	249	1998	10	metabolism	metabolism
Cisplatin	230	0	0		
Phenytoin	193	0	0		
Doxorubicin	232	2001	2	genetics	pharmacology
Choline	195	2007	1	blood	therapeutic use
Ranked by WtCOS					
Spiperone	25	0	0		
otenzepad	6	0	0		

Tamoxifen	167	0	0		
Clomiphene	57	0	0		
Nafoxidine	22	0	0		
Congo Red	41	0	0		
Idazoxan	27	0	0		
CP 96345	24	0	0		
3-(2-carboxypiperazin-4-yl)propyl-1-phosphonic acid	8	0	0		
5-(N-methyl-N-isobutyl)amiloride	3	0	0		
Citalopram	14	0	0		
Pentostatin	32	0	0		
Dizocilpine Maleate	57	0	0		
Capsaicin	115	0	0		
tamoxifen aziridine	6	0	0		
N(6)-cyclohexyladenosine	29	0	0		
Bromocriptine	120	0	0		
Ketanserin	45	0	0		
chrysarobin	7	0	0		
tricalcium phosphate	6	0	0		
Ranked by AvgRank					
Tamoxifen	167	0	0		
Pyridoxine	158	1996	1	metabolism	analysis
Chloroquine	226	2003	2	blood	pharmacology
Bromocriptine	120	0	0		
Dimethyl Sulfoxide	245	0	0		
Haloperidol	126	0	0		
Kainic Acid	139	0	0		
Capsaicin	115	0	0		
Lead	243	0	0		
Clonidine	115	0	0		
Hydroxyurea	106	0	0		
Molybdenum	112	2001	1	urine	
Vanadium	148	0	0		
Silver	116	2007	1	drug therapy	adverse effects
Dipyridamole	106	0	0		
Guanosine Triphosphate	163	0	0		
Uridine	119	2002	2	drug therapy	analogs & derivatives
Cadmium Chloride	106	0	0		
Naloxone	141	1995	1		pharmacology
Lidocaine	133	2001	1	metabolism	analogs & derivatives

Appendix 3. Psoriasis: Top 20 chemicals returned by each ranking

The columns with white background represent data from the Baseline Period. The gray columns are drawn from the Test Period. ProtCt is the count of proteins from the protein pool the chemical has annotated with it. FirstYr is the first year the chemical appears as the subject chemical in an article that also has an annotation of the disease. DisQual and ChemQual are the most common disease qualifiers (or subheadings) and chemical qualifiers (subheadings) appearing in the annotations when the chemical is annotated with the disease.

Appendix 3A. Psoriasis 1984-1985					
Ranked by ProtCt					
ChemName	**Protct**	**FirstYr**	**ArtCt**	**DisQual**	**ChemQual**
Estradiol	232	0	0		
Phenobarbital	160	1994	1	complications	adverse effects
Lead	147	0	0		
Tetra. Acetate	144	1989	1	blood	pharmacology
Cadmium	138	0	0		
Vitamin E	135	1988	5	blood	blood
Puromycin	134	0	0		
Glycerol	129	0	0		
Hydrogen Peroxide	127	1989	2	blood	pharmacology
Morphine	126	0	0		
Adenine	124	1999	1	complications	
Phenytoin	123	1985	1	complications	adverse effects
Formaldehyde	122	0	0		
Heme	119	0	0		
Mercaptoethanol	118	0	0		
Clofibrate	115	1991	1	drug therapy	therapeutic use
Ethinyl Estradiol	114	0	0		
Rifampin	110	1986	6	drug therapy	therapeutic use
Halothane	110	0	0		
Methylcholanthrene	109	0	0		
Ranked by WtProp					
Estradiol	232	0	0		
Lead	147	0	0		
Phenobarbital	160	1994	1	complications	adverse effects
Vitamin E	135	1988	5	blood	blood
Puromycin	134	0	0		
Tetradecanoylphorbol Acetate	144	1989	1	blood	pharmacology
Mercaptoethanol	118	0	0		

Phenytoin	123	1985	1	complications	adverse effects
Cadmium	138	0	0		
Bromodeoxyuridine	104	0	0		
Rifampin	110	1986	6	drug therapy	therapeutic use
Ozone	94	2000	1	therapy	adverse effects
Carbon Tetrachloride	107	0	0		
Formaldehyde	122	0	0		
Halothane	110	0	0		
Hydrogen Peroxide	127	1989	2	blood	pharmacology
Adenine	124	1999	1	complications	
Glycerol	129	0	0		
Periodic Acid	95	0	0		
Clofibrate	115	1991	1	drug therapy	therapeutic use
Ranked by WtCOS					
Congo Red	26	0	0		
Calcitriol	48	1985	353	drug therapy	analogs & derivatives
Carbazilquinone	4	0	0		
Warfarin	67	1992	2	drug therapy	therapeutic use
Selenious Acid	20	0	0		
Metiamide	13	0	0		
Succinylcholine	41	2007	1	complications	therapeutic use
Metoclopramide	18	0	0		
Cholecalciferol	66	1986	41	drug therapy	therapeutic use
Danazol	46	0	0		
oxmetidine	5	0	0		
Yohimbine	13	1988	1	blood	therapeutic use
Acenocoumarol	19	0	0		
Phenindione	25	0	0		
Dextromoramide	2	0	0		
Carbimazole	16	0	0		
Glyburide	40	1987	1	pathology	adverse effects
Dimethadione	6	0	0		
Pregnenolone	44	0	0		
Famotidine	3	0	0		
Ranked by AvgRank					
Rifampin	110	1986	6	drug therapy	therapeutic use
Lead	147	0	0		
Hydrochloric Acid	85	0	0		
Ethinyl Estradiol	114	0	0		
Vitamin E	135	1988	5	blood	blood
Propylthiouracil	85	1993	16	drug therapy	therapeutic use
Phenobarbital	160	1994	1	complications	adverse effects
Bromodeoxyuridine	104	0	0		
Cholecalciferol	66	1986	41	drug therapy	therapeutic use
Cisplatin	76	0	0		
Warfarin	67	1992	2	drug therapy	therapeutic use
Formaldehyde	122	0	0		
Sodium Dodecyl Sulfate	88	0	0		

Methylcholanthrene	109	0	0		
Puromycin	134	0	0		
Estriol	89	0	0		
Glycerol	129	0	0		
Adenine	124	1999	1	complications	
Ouabain	104	0	0		
Thiourea	90	0	0		

Appendix 3B. Psoriasis 1989-1990					
Ranked by ProtCt					
ChemName	**Protct**	**FirstYr**	**ArtCt**	**DisQual**	**ChemQual**
Estradiol	337	0	0		
Phenobarbital	202	1994	1	complications	adverse effects
Cadmium	197	0	0		
Lead	187	0	0		
Morphine	175	0	0		
Doxorubicin	167	2004	1	complications	administration & dosage
Formaldehyde	160	0	0		
Glycerol	155	0	0		
Puromycin	155	0	0		
Calcimycin	151	0	0		
Ethinyl Estradiol	150	0	0		
Adenine	150	1999	1	complications	
Mercaptoethanol	149	0	0		
Heme	143	0	0		
Aluminum	142	0	0		
Halothane	142	0	0		
Carbon Tetrachloride	141	0	0		
Cisplatin	140	0	0		
Putrescine	140	0	0		
Nicotine	139	2006	1	drug therapy	pharmacology
Ranked by WtProp					
Estradiol	337	0	0		
Lead	187	0	0		
Cadmium	197	0	0		
Mercaptoethanol	149	0	0		
Phenobarbital	202	1994	1	complications	adverse effects
Puromycin	155	0	0		
Formaldehyde	160	0	0		
Carbon Tetrachloride	141	0	0		
Asbestos	109	0	0		
Doxorubicin	167	2004	1	complications	administration & dosage
Ethinyl Estradiol	150	0	0		
Calcimycin	151	0	0		
Aluminum	142	0	0		
Halothane	142	0	0		
Glycerol	155	0	0		
Periodic Acid	114	0	0		
Morphine	175	0	0		
Ozone	121	2000	1	therapy	adverse effects
Deuterium	123	0	0		
Adenine	150	1999	1	complications	
Ranked by WtCOS					
Congo Red	32	0	0		

Clomiphene	48	0	0		
Pregnenolone	53	0	0		
Succinylcholine	46	2007	1	complications	therapeutic use
Clorgyline	20	0	0		
Warfarin	78	1992	2	drug therapy	therapeutic use
Omeprazole	28	1993	1	complications	therapeutic use
Tolazamide	6	0	0		
Selegiline	15	0	0		
Ouabain	130	0	0		
Metiamide	14	0	0		
1-Methyl-4-phenyl-1,2,3,6-tetrahydropyridine	36	0	0		
Vitamin K 1	36	0	0		
15-Hydroxy-11 alpha,9 alpha-(epoxymethano)prosta-5,13-dienoic Acid	18	0	0		
Promegestone	15	0	0		
SQ 29548	9	0	0		
lipid-associated sialic acid	3	0	0		
Hydrochloric Acid	95	0	0		
Carbazilquinone	6	0	0		
Mesterolone	8	0	0		
Ranked by AvgRank					
Lead	187	0	0		
Ouabain	130	0	0		
Cisplatin	140	0	0		
Cadmium	197	0	0		
Hydrochloric Acid	95	0	0		
Ethinyl Estradiol	150	0	0		
Phenobarbital	202	1994	1	complications	adverse effects
Adenine	150	1999	1	complications	
Propylthiouracil	108	1993	16	drug therapy	therapeutic use
Warfarin	78	1992	2	drug therapy	therapeutic use
Nicotine	139	2006	1	drug therapy	pharmacology
Silver	100	0	0		
Glycerol	155	0	0		
Danazol	72	0	0		
Estriol	107	0	0		
Carbon Tetrachloride	141	0	0		
Vincristine	109	0	0		
Methylcholanthrene	132	0	0		
Bromodeoxyuridine	120	0	0		
Carbachol	100	0	0		

Appendix 3C. Psoriasis 1994-95					
Ranked by ProtCt					
ChemName	**Prot Ct**	**First Yr**	**ArtCt**	**DisQual**	**ChemQual**
Estradiol	435	0	0		
Doxorubicin	259	2004	1	complications	administration & dosage
Cadmium	257	0	0		
Cisplatin	245	0	0		
Morphine	240	0	0		
Lead	236	0	0		
Calcimycin	232	0	0		
Formaldehyde	202	0	0		
Nitric Oxide	201	1997	15	metabolism	biosynthesis
Aluminum	198	0	0		
Nicotine	197	2006	1	drug therapy	pharmacology
Tamoxifen	193	1996	3	drug therapy	therapeutic use
Adenine	191	1999	1	complications	
Glycerol	185	0	0		
Butyric Acid	185	0	0		
Halothane	184	0	0		
Puromycin	183	0	0		
Carbon Tetrachloride	179	0	0		
Ozone	179	2000	1	therapy	adverse effects
Putrescine	179	0	0		
Ranked by WtProp					
Estradiol	435	0	0		
Doxorubicin	259	2004	1	complications	administration & dosage
Calcimycin	232	0	0		
Lead	236	0	0		
Cisplatin	245	0	0		
Cadmium	257	0	0		
Tamoxifen	193	1996	3	drug therapy	therapeutic use
Ozone	179	2000	1	therapy	adverse effects
Carbon Tetrachloride	179	0	0		
Morphine	240	0	0		
Aluminum	198	0	0		
Formaldehyde	202	0	0		
Mercaptoethanol	169	0	0		
Asbestos	137	0	0		
Puromycin	183	0	0		
Ethinyl Estradiol	177	0	0		
Halothane	184	0	0		
Nicotine	197	2006	1	drug therapy	pharmacology
Suramin	160	0	0		
Pentoxifylline	135	1996	5	drug therapy	therapeutic use

Ranked by WtCOS					
Congo Red	37	0	0		
Cromakalim	27	0	0		
Losartan	25	2008	1	drug therapy	adverse effects
Clorgyline	23	0	0		
DPI 201-106	7	0	0		
Amiloride	107	0	0		
PD 123177	4	0	0		
Veratridine	37	0	0		
Tetraethylammonium	18	0	0		
Tetrodotoxin	77	0	0		
Succinylcholine	50	2007	1	complications	therapeutic use
Paclitaxel	73	2004	1	drug therapy	administration & dosage
Pregnenolone	75	0	0		
L 365260	16	0	0		
1-Methyl-4-phenyl-1,2,3,6-tetrahydropyridine	56	0	0		
SQ 29548	11	0	0		
vapiprost	6	0	0		
L 158809	3	0	0		
Tolazamide	6	0	0		
Sodium, Dietary	66	2004	2	drug therapy	administration & dosage
Ranked by AvgRank					
Cadmium	257	0	0		
Nitric Oxide	201	1997	15	metabolism	biosynthesis
Ouabain	156	0	0		
Lead	236	0	0		
Amiloride	107	0	0		
Carbon Tetrachloride	179	0	0		
Silver	120	0	0		
Morphine	240	0	0		
Cisplatin	245	0	0		
Hydrochloric Acid	102	0	0		
Cadmium Chloride	114	0	0		
Naloxone	134	0	0		
Ethinyl Estradiol	177	0	0		
Penicillin G	134	0	0		
Estriol	111	0	0		
Glycerol	185	0	0		
Dimethylnitrosamine	99	0	0		
Phosphorylcholine	83	2006	1	complications	analogs & derivatives
Kainic Acid	134	0	0		
Danazol	98	0	0		

Appendix 4. Migraine: Top 20 chemicals returned by each ranking

The columns with white background represent data from the Baseline Period. The gray columns are drawn from the Test Period. ProtCt is the count of proteins from the protein pool the chemical has annotated with it. FirstYr is the first year the chemical appears as the subject chemical in an article that also has an annotation of the disease. DisQual and ChemQual are the most common disease qualifiers (or subheadings) and chemical qualifiers (subheadings) appearing in the annotations when the chemical is annotated with the disease.

Appendix 4A. Migraine 1984-85					
Ranked by ProtCt					
ChemName	**Protct**	**FirstYr**	**ArtCt**	**DisQual**	**ChemQual**
Sodium	81	2006	1	blood	cerebrospinal fluid
Magnesium	74	1985	40	blood	blood
Zinc	74	0	0		
Copper	69	1986	1	etiology	adverse effects
Corticosterone	67	0	0		
Prednisolone	67	2007	1	complications	therapeutic use
Edetic Acid	66	1989	1	physiopathology	administration & dosage
Colchicine	65	0	0		
Lead	64	0	0		
Atropine	61	0	0		
Nicotine	61	1999	3	drug therapy	adverse effects
Bucladesine	60	0	0		
Cycloheximide	60	0	0		
Cyclic GMP	60	1995	4	physiopathology	blood
Manganese	59	0	0		
Iodine	55	1990	1	diagnosis	administration & dosage
Isoflurophate	55	0	0		
Nitrogen	55	0	0		
Mercury	54	0	0		
Halothane	54	0	0		
Ranked by WtProp					
Phenoxybenzamine	51	0	0		
Phentolamine	47	0	0		
Nicotine	61	1999	3	drug therapy	adverse effects
Atropine	61	0	0		
Isoflurophate	55	0	0		

Guanethidine	36	0	0		
Prednisolone	67	2007	1	complications	therapeutic use
Desipramine	36	0	0		
Corticosterone	67	0	0		
Sodium	81	2006	1	blood	cerebrospinal fluid
Pilocarpine	38	0	0		
Thiopental	38	0	0		
Halothane	54	0	0		
Carbachol	44	0	0		
Lead	64	0	0		
Methylprednisolone	49	2000	3	therapy	therapeutic use
Apomorphine	37	1990	6	physiopathology	pharmacology
Ketamine	35	1995	2	drug therapy	administration & dosage
Baclofen	26	1990	3	drug therapy	therapeutic use
Mazindol	17	0	0		
Ranked by WtCOS					
Vitamin D	36	1994	1	drug therapy	therapeutic use
Ouabain	44	0	0		
Parathion	23	0	0		
Clomiphene	21	1992	2	chemically induced	adverse effects
Iodine	55	1990	1	diagnosis	administration & dosage
Succinylcholine	20	0	0		
Nitromifene	8	0	0		
Carbimazole	7	0	0		
Dihydrotestosterone	35	0	0		
Phenformin	26	0	0		
Oxotremorine	16	0	0		
Propylthiouracil	42	0	0		
Mitoguazone	7	0	0		
Creatinine	43	0	0		
Carbon Monoxide	20	0	0		
Medroxyprogesterone 17-Acetate	15	1997	1	drug therapy	administration & dosage
Quinuclidinyl Benzilate	10	0	0		
Ethambutol	5	0	0		
Nitric Oxide	7	1991	41	physiopathology	blood
Silver	25	0	0		
Ranked by AvgRank					
Corticosterone	67	0	0		
Sodium	81	2006	1	blood	cerebrospinal fluid
Atropine	61	0	0		
Iodine	55	1990	1	diagnosis	administration & dosage
Creatinine	43	0	0		
Prednisolone	67	2007	1	complications	therapeutic use

Isoflurophate	55	0	0		
Propylthiouracil	42	0	0		
Phentolamine	47	0	0		
Ouabain	44	0	0		
Magnesium	74	1985	40	blood	blood
Apomorphine	37	1990	6	physiopathology	pharmacology
Zinc	74	0	0		
Pilocarpine	38	0	0		
Bilirubin	45	0	0		
Carbachol	44	0	0		
DDT	42	0	0		
Puromycin	49	0	0		
Calcimycin	45	0	0		
Cysteamine	38	0	0		

Appendix 4B. Migraine 1989-1990					
Ranked by ProtCt					
ChemName	**Protct**	**FirstYr**	**ArtCt**	**DisQual**	**ChemQual**
Sodium	109	2006	1	blood	cerebrospinal fluid
Zinc	102	0	0		
Tetradecanoylphorbol Acetate	87	0	0		
Colchicine	87	0	0		
Prednisolone	85	2007	1	complications	therapeutic use
Nicotine	84	1999	3	drug therapy	adverse effects
Cyclic GMP	83	1995	4	physiopathology	blood
Corticosterone	83	0	0		
Bucladesine	83	0	0		
Atropine	82	0	0		
Lead	80	0	0		
Cycloheximide	79	0	0		
Manganese	77	0	0		
Cyclophosphamide	70	2001	1	etiology	administration & dosage
Iodine	69	1990	1	diagnosis	administration & dosage
Nitrogen	69	0	0		
Halothane	68	0	0		
Vitamin A	67	0	0		
Calcimycin	67	0	0		
Cadmium	67	0	0		
Ranked by WtProp					
Phenoxybenzamine	60	0	0		
Atropine	82	0	0		
Phentolamine	59	0	0		
Nicotine	84	1999	3	drug therapy	adverse effects
Guanethidine	45	0	0		
Sodium	109	2006	1	blood	cerebrospinal fluid
Prednisolone	85	2007	1	complications	therapeutic use
Isoflurophate	62	0	0		
Pilocarpine	51	0	0		
Cyclic GMP	83	1995	4	physiopathology	blood
Thiopental	47	0	0		
Colchicine	87	0	0		
Pentylenetetrazole	47	0	0		
Methylprednisolone	65	2000	3	therapy	therapeutic use
Ketamine	47	1995	2	drug therapy	administration & dosage
Carbachol	63	0	0		
Baclofen	38	1990	3	drug therapy	therapeutic use
Desoxycorticosterone	66	0	0		
Apomorphine	49	1990	6	physiopathology	pharmacology

Lead	80	0	0		
Ranked by WtCOS			ArtCt	DisQual	ChemQual
Parathion	28	0	0		
Vitamin D	48	1994	1	drug therapy	therapeutic use
Quinuclidinyl Benzilate	12	0	0		
ethylcholine aziridinium	5	0	0		
Succinylcholine	24	0	0		
Oxotremorine	18	0	0		
Clomiphene	29	1992	2	chemically induced	adverse effects
Dizocilpine Maleate	10	0	0		
Calcitriol	50	0	0		
Medroxyprogesterone 17-Acetate	24	1997	1	drug therapy	administration & dosage
Ouabain	64	0	0		
Heme	45	0	0		
1,4-dihydropyridine	18	0	0		
W 7	15	0	0		
Iodine	69	1990	1	diagnosis	administration & dosage
Phenformin	32	0	0		
Gallamine Triethiodide	13	0	0		
BE 2254	6	0	0		
Dihydrotestosterone	52	0	0		
Methylcholanthrene	39	0	0		
Ranked by AvgRank					
Sodium	109	2006	1	blood	cerebrospinal fluid
Ouabain	64	0	0		
Iodine	69	1990	1	diagnosis	administration & dosage
Cyclic GMP	83	1995	4	physiopathology	blood
Atropine	82	0	0		
Creatinine	52	0	0		
Isoflurophate	62	0	0		
Zinc	102	0	0		
Apomorphine	49	1990	6	physiopathology	pharmacology
Aluminum	61	0	0		
Corticosterone	83	0	0		
Calcimycin	67	0	0		
Cysteamine	54	0	0		
Carbachol	63	0	0		
Vitamin D	48	1994	1	drug therapy	therapeutic use
Pilocarpine	51	0	0		
Dihydrotestosterone	52	0	0		
Phentolamine	59	0	0		
Hydrochloric Acid	49	0	0		
Thiourea	52	0	0		

Appendix 4C. Migraine 1994-1995					
Ranked by ProtCt					
ChemName	**Prot Ct**	**FirstYr**	**ArtCt**	**DisQual**	**ChemQual**
Sodium	139	2006	1	blood	cerebrospinal fluid
Zinc	132	0	0		
Tetradecanoylphorbol Acetate	126	0	0		
Colchicine	114	0	0		
Bucladesine	112	0	0		
Nicotine	110	1999	3	drug therapy	adverse effects
Corticosterone	109	0	0		
Prednisolone	109	2007	1	complications	therapeutic use
Cyclic GMP	108	1995	4	physiopathology	blood
Cycloheximide	105	0	0		
Lead	105	0	0		
Cadmium	102	0	0		
Atropine	99	0	0		
Hydrogen Peroxide	96	0	0		
Calcimycin	95	0	0		
Manganese	95	0	0		
Halothane	94	0	0		
Cyclophosphamide	93	2001	1	etiology	administration & dosage
Tretinoin	91	0	0		
Forskolin	89	0	0		
Ranked by WtProp					
Atropine	99	0	0		
Phentolamine	73	0	0		
Phenoxybenzamine	65	0	0		
Thiopental	66	0	0		
Ketamine	72	1995	2	drug therapy	administration & dosage
Nicotine	110	1999	3	drug therapy	adverse effects
Guanethidine	54	0	0		
Colchicine	114	0	0		
Prednisolone	109	2007	1	complications	therapeutic use
Sodium	139	2006	1	blood	cerebrospinal fluid
Pentylenetetrazole	63	0	0		
Halothane	94	0	0		
Pilocarpine	66	0	0		
Isoflurophate	76	0	0		
Cyclic GMP	108	1995	4	physiopathology	blood
Methylprednisolone	87	2000	3	therapy	therapeutic use
Ouabain	84	0	0		
Lead	105	0	0		
Corticosterone	109	0	0		

Potassium Chloride	85	0	0		
Ranked by WtCOS					
Quinuclidinyl Benzilate	15	0	0		
ethylcholine aziridinium	14	0	0		
1,4-dihydropyridine	26	0	0		
Parathion	33	0	0		
beta-Naphthoflavone	16	0	0		
Oxotremorine	28	0	0		
Hydrochlorothiazide	45	0	0		
(4-(m-Chlorophenylcarbamoyloxy)-2-butynyl)trimethylammonium Chloride	10	0	0		
N(6)-cyclohexyladenosine	19	0	0		
Succinylcholine	28	0	0		
Promegestone	13	0	0		
Tolbutamide	53	0	0		
Ouabain	84	0	0		
CGP 12177	5	0	0		
W 7	26	0	0		
Sodium, Dietary	49	0	0		
Losartan	18	1995	1	chemically induced	
Prostaglandins H	19	0	0		
BE 2254	10	0	0		
N(6)-cyclopentyladenosine	9	0	0		
Ranked by AvgRank					
Sodium	139	2006	1	blood	cerebrospinal fluid
Cyclic GMP	108	1995	4	physiopathology	blood
Ouabain	84	0	0		
Atropine	99	0	0		
Carbachol	81	0	0		
Calcimycin	95	0	0		
Isoflurophate	76	0	0		
Zinc	132	0	0		
Creatinine	62	0	0		
Forskolin	89	0	0		
Pilocarpine	66	0	0		
Aluminum	84	0	0		
Corticosterone	109	0	0		
Tolbutamide	53	0	0		
Kainic Acid	79	0	0		
Yohimbine	51	0	0		
Sodium, Dietary	49	0	0		
Hydrochloric Acid	55	0	0		
Amiloride	56	0	0		
Cadmium	102	0	0		

Appendix 5. Cystic Fibrosis: Gold standard chemicals by highest article count

This table shows what the ABC routines should have found and ranked high. Number 1 is the highest rank. ArtCt is the number of articles that connect the chemical to the disease in the Test Period. FirstYr is the first year the chemical (as subject chemical) is annotated with the disease. ProtCt is the number of proteins from the disease protein pool that the chemical has annotated with it in the Baseline Period. The four ranking methodologies are described in the text of Chapter 3. The data in the columns shaded in gray are data elements derived from ChemoText in the Baseline period. The columns with the white background are pulled from the Test Period.

Appendix 5A. Cystic Fibrosis 1984-1985								
					Rankings (out of 5,555 chems in HS)			
ArtCt	First Yr	DisQual	Prot Ct	ChemName	Avg Rank	Wt COS	Wt Prop	Prot Ct
109	1985	complications	1	Ciprofloxacin	4184	4160	3649	4290
64	1995	metabolism	16	Nitric Oxide	905	1218	1175	645
27	1990	drug therapy	27	Ibuprofen	602	1427	357	396
22	1985	metabolism	91	Taurine	48	308	66	52
21	1985	complications	7	Aztreonam	1405	1975	1034	1260
13	1985	microbiology	10	Imipenem	851	1073	872	936
11	1991	metabolism	6	Uridine Triphosphate	1646	1117	4030	1353
11	1999	microbiology	14	4-Butyrolactone	1154	1956	991	731
10	1991	drug therapy	10	Omeprazole	1015	1547	795	954
10	1998	metabolism	115	Hydrogen Peroxide	3	70	43	24
9	1996	blood	2	beta Carotene	3175	2808	4266	3009
9	1992	metabolism	39	Forskolin	152	296	234	268
8	1985	drug therapy	2	Cisapride	3144	3099	2600	3166
8	1995	complications	3	Budesonide	1794	1612	1514	2106
8	1993	drug therapy	71	Mannitol	390	1320	107	103
7	1988	microbiology	4	Pyocyanine	1318	821	1699	1713
7	1990	metabolism	30	Ranitidine	162	202	282	355
6	1998	drug therapy	4	pamidronate	1149	468	1706	1689
6	1985	metabolism	35	Lactic Acid	544	1308	412	301
6	1989	blood	76	Carnitine	68	276	108	87
6	1987	metabolism	82	Furosemide	20	156	57	74

6	1989	microbiology	93	Rifampin	126	597	58	49
5	1999	complications	10	Megestrol Acetate	908	1211	873	940
5	1987	blood	24	Malondialdehyde	597	1174	541	439
5	1992	metabolism	119	Adenosine	1	50	48	23
4	1985	metabolism	2	Cilastatin	2470	1915	2618	2679
4	1997	therapy	8	Polyethyleneimine	1279	1136	1902	1106
4	1986	complications	24	Talc	166	106	331	422
4	2001	physiopathology	46	Glyburide	87	216	122	208
4	1995	complications	54	Amphotericin B	102	234	209	158
4	1988	metabolism	106	Caffeine	96	499	47	31
4	1990	drug therapy	131	Lithium	32	293	15	12

Appendix 5B. Cystic Fibrosis 1989-1990

					Rankings (out of 9,292 chems in HS)			
Art Ct	First Yr	DisQual	Prot Ct	ChemName	Avg Rank	Wt COS	Wt Prop	Prot Ct
64	1995	metabolism	40	Nitric Oxide	278	567	615	366
40	1995	drug therapy	3	Azithromycin	4267	4910	2895	3867
27	1990	drug therapy	50	Ibuprofen	295	1336	229	30
17	1991	complications	1	Itraconazole	9288	6566	9287	6566
14	1995	drug therapy	9	meropenem	1696	2648	1251	1525
13	2004	drug therapy	5	Curcumin	2148	2363	1997	2334
11	1991	metabolism	13	Uridine Triphosphate	1102	719	2521	1099
11	1999	microbiology	22	4-Butyrolactone	917	1125	907	697
10	1998	drug therapy	7	Genistein	1540	1032	2586	1783
10	1991	drug therapy	34	Omeprazole	179	309	430	431
10	1998	metabolism	167	Hydrogen Peroxide	1	94	23	50
9	1996	blood	3	beta Carotene	4073	3572	6728	3541
9	1992	metabolism	105	Forskolin	139	821	76	103
8	1992	drug therapy	5	1,3-dipropyl-8-cyclopentylxanthine	2746	3385	2257	2436
8	1995	complications	11	Budesonide	1019	726	1039	1278
8	1993	drug therapy	92	Mannitol	610	1620	95	132
7	1990	metabolism	64	Ranitidine	33	126	134	208
6	2000	drug therapy	3	Clarithromycin	2186	699	3745	3135
6	1998	drug therapy	9	pamidronate	349	803	1073	1466
5	1992	drug therapy	2	benzamil	5207	4628	7105	4961
5	1999	complications	15	Megestrol Acetate	699	301	805	981
5	1992	metabolism	166	Adenosine	5	199	40	51
4	1992	genetics	10	8-((4-chlorophenyl)thio)cyclic-3',5'-AMP	1675	2409	1575	1413
4	1997	therapy	16	Polyethyleneimine	784	218	1206	930
4	2001	physiopathology	67	Glyburide	71	339	125	199
4	1995	complications	84	Amphotericin B	266	1225	131	156
4	1990	drug therapy	180	Lithium	54	495	12	43

Appendix 5C. Cystic Fibrosis 1994-1995

					Rankings (out of 14,143 entries in HS)			
ArtCt	First Yr	DisQual	Protct	ChemName	Avg Rank	Wt COS	Wt Prop	Prot Ct
64	1995	metabolism	182	Nitric Oxide	30	382	73	25
40	1995	drug therapy	23	Azithromycin	1770	4308	710	1038
14	1995	drug therapy	12	meropenem	1549	2264	1470	1760
13	2004	drug therapy	24	Curcumin	1134	2360	1016	1000
11	1999	microbiology	38	4-Butyrolactone	985	2477	852	586
10	1998	drug therapy	66	Genistein	66	101	375	274
10	1998	metabolism	249	Hydrogen Peroxide	21	365	16	4
9	1997	genetics	2	4-phenylbutyric acid	5347	3884	5448	6573
9	1996	blood	8	beta Carotene	2075	780	4432	2394
8	2000	microbiology	2	homoserine lactone	8518	8414	7668	8454
8	1995	complications	30	Budesonide	714	1446	767	764
7	1997	surgery	127	Tacrolimus	197	1166	86	81
6	1997	drug therapy	11	fluticasone	2443	3922	2103	1949
6	2000	drug therapy	18	Clarithromycin	2516	5587	1190	1288
6	1998	drug therapy	21	pamidronate	661	725	1006	1091
5	1999	complications	22	Megestrol Acetate	786	1387	792	1061
4	2001	blood	1	25-hydroxyvitamin D	9754	10918	9505	10977
4	1997	drug therapy	9	salmeterol	2620	4223	1805	2298
4	1997	therapy	27	Polyethyleneimine	700	1016	1065	862
4	2001	physiopathology	105	Glyburide	402	1803	94	130
4	1995	complications	119	Amphotericin B	212	1188	90	97

Appendix 6. Psoriasis: Gold standard chemicals by highest article count

This table shows what the ABC routines should have found and ranked high. Number 1 is the highest rank. ArtCt is the number of articles that connect the chemical to the disease in the Test Period. FirstYr is the first year the chemical (as subject chemical) is annotated with the disease. ProtCt is the number of proteins from the disease protein pool that the chemical has annotated with it in the Baseline Period. The four ranking methodologies are described in the text of Chapter 3. The data in the columns shaded in gray are data elements derived from ChemoText in the Baseline period. The columns with the white background are pulled from the Test Period.

Appendix 6A. Psoriasis 1984-1985								
					Rankings (out of 5,532 entries in HS)			
ArtCt	First Yr	DisQual	Prot Ct	ChemName	Avg Rank	Wt COS	Wt Prop	Prot Ct
353	1985	drug therapy	48	Calcitriol	30	2	173	141
41	1986	drug therapy	66	Cholecalciferol	9	9	64	80
16	1993	drug therapy	85	Propylthiouracil	6	44	40	47
15	1997	metabolism	23	Nitric Oxide	233	255	551	379
13	1987	drug therapy	36	Sulfasalazine	224	746	171	225
12	1997	drug therapy	2	zinc pyrithione	2753	2391	2628	2777
11	1986	drug therapy	16	Capsaicin	386	149	1014	525
8	1987	drug therapy	1	Zidovudine	5102	5149	5092	5086
7	1986	drug therapy	7	Trimethoprim-Sulfamethoxazole Combination	970	839	1444	1084
7	1991	drug therapy	24	Ranitidine	164	178	408	363
7	1993	drug therapy	49	Methimazole	58	260	106	138
6	1985	drug therapy	9	1-hydroxycholecalciferol	373	78	666	878
6	1985	blood	30	Malondialdehyde	340	932	278	291
6	1986	drug therapy	110	Rifampin	1	24	11	18
5	1996	drug therapy	26	Pentoxifylline	226	489	323	339
5	1985	drug therapy	49	Thalidomide	40	133	104	137
5	1988	blood	135	Vitamin E	5	97	4	6
4	1988	chemically induced	1	Terfenadine	5302	5320	5301	5469

4	1994	drug therapy	3	fludarabine	2273	1893	2765	2054
4	1986	metabolism	8	Urocanic Acid	796	914	843	981
4	1989	diagnosis	8	Amoxicillin	982	1043	1264	988
4	1997	drug therapy	10	Minocycline	979	1634	726	854
4	1985	drug therapy	17	Flurbiprofen	726	1453	545	521
4	1994	drug therapy	22	Vidarabine	172	29	554	395
4	1986	drug therapy	39	Sulfamethoxazole	152	516	167	202
4	1993	drug therapy	46	Nifedipine	155	588	155	156
4	1986	drug therapy	80	Erythromycin	82	515	38	58

Appendix 6B. Psoriasis 1989-1990

					Rankings (out of 9,192 entries in HS)			
ArtCt	First Yr	DisQual	Prot Ct	ChemName	Avg Rank	Wt COS	Wt Prop	Prot Ct
34	1990	drug therapy	1	dimethyl fumarate	8448	7819	8432	7893
16	1993	drug therapy	108	Propylthiouracil	9	109	37	50
15	1997	metabolism	51	Nitric Oxide	123	359	292	202
12	1997	drug therapy	3	zinc pyrithione	4040	4089	3954	3618
9	1993	chemically induced	6	terbinafine	2191	2561	2232	1986
7	1990	drug therapy	2	maxacalcitol	5451	5414	5040	5423
7	1991	drug therapy	56	Ranitidine	47	119	180	172
7	1993	drug therapy	72	Methimazole	90	426	91	111
5	1991	drug therapy	1	bimolane	6251	6611	5933	6611
5	2004	chemically induced	2	imiquimod	5114	4873	4482	5050
5	1996	drug therapy	54	Pentoxifylline	271	1113	157	182
4	1994	drug therapy	5	fludarabine	3045	3376	3645	2373
4	1997	drug therapy	20	Minocycline	640	1433	570	688
4	1994	drug therapy	36	Vidarabine	117	57	427	330
4	1993	drug therapy	106	Nifedipine	744	2868	47	54

Appendix 6C. Psoriasis 1994-1995

					Rankings (out of 13,393 entries in HS)			
ArtCt	FirstYr	DisQual	Protct	ChemName	Avg Rank	Wt COS	Wt Prop	Prot Ct
20	1997	drug therapy	14	mycophenolate mofetil	635	543	1156	1389
15	1997	metabolism	201	Nitric Oxide	2	29	28	9
12	1997	drug therapy	3	zinc pyrithione	5830	5708	5990	5287
11	1995	drug therapy	13	fluticasone	2487	4978	1422	1576
6	1996	drug therapy	5	citraconic acid	5331	7612	3337	3667
5	1995	drug therapy	4	liarozole	3285	3035	3156	3926
5	2003	drug therapy	5	pioglitazone	4372	5240	4011	3543
5	2004	chemically induced	10	imiquimod	982	470	1796	1857
5	2002	drug therapy	15	leflunomide	1338	2512	1209	1347
5	1996	drug therapy	135	Pentoxifylline	346	1940	20	52

4	1997	drug therapy	42	Minocycline	621	2291	309	434

Appendix 7. Migraine: Gold standard chemicals by highest article count

This table shows what the ABC routines should have found and ranked high. Number 1 is the highest rank. ArtCt is the number of articles that connect the chemical to the disease in the Test Period. FirstYr is the first year the chemical (as subject chemical) is annotated with the disease. ProtCt is the number of proteins from the disease protein pool that the chemical has annotated with it in the Baseline Period. The four ranking methodologies are described in the text of Chapter 3. The data in the columns shaded in gray are data elements derived from ChemoText in the Baseline period. The columns with the white background are pulled from the Test Period.

7A. Migraine – Highest gold standard chemicals 1984-1985 order by descending Article Count (ArtCt).								
					Rankings (out of 4,006 chems in HS)			
Art Ct	First Yr	DisQual	Prot Ct	ChemName	Avg Rank	Wt COS	Wt Prop	Prot Ct
88	1988	drug therapy	32	Valproic Acid	129	369	72	111
41	1991	physiopathology	7	Nitric Oxide	610	19	2231	638
40	1985	blood	74	Magnesium	11	41	61	2
19	1992	drug therapy	13	Fluoxetine	671	1701	121	395
15	1986	drug therapy	37	Melatonin	48	193	34	76
13	1992	drug therapy	25	Acetazolamide	148	257	195	169
12	1995	drug therapy	20	Capsaicin	83	31	158	229
11	1991	drug therapy	9	Butorphanol	676	1443	218	578
10	1988	chemically induced	6	1-(3-chlorophenyl)piperazine	861	1588	314	818
10	1989	drug therapy	33	Meperidine	130	424	26	105
10	2001	drug therapy	8	Dipyrone	435	577	364	605
9	1991	drug therapy	18	Magnesium Sulfate	144	85	272	256
8	1989	drug therapy	6	Nicardipine	900	1253	792	799
8	1997	drug therapy	15	Droperidol	200	372	105	330
6	1990	physiopathology	37	Apomorphine	12	57	17	74
5	1985	drug therapy	14	Mianserin	253	360	252	350
5	1987	blood	21	Platelet Activating Factor	374	877	268	224

5	1991	drug therapy	5	Buspirone	613	549	706	849
5	1992	drug therapy	5	Piroxicam	955	834	1402	875
5	1996	prevention & control	2	iprazochrome	1888	2193	823	2161
4	1986	drug therapy	20	Tamoxifen	187	134	402	230
4	1987	drug therapy	21	Phenelzine	669	1865	119	226
4	1992	drug therapy	4	Ketoprofen	1115	1532	692	1140
4	1992	drug therapy	1	oxetorone	2535	2923	1878	2923
4	1993	drug therapy	24	Diphenhydramine	108	222	111	180
4	1995	physiopathology	60	Cyclic GMP	53	275	25	14
4	1996	drug therapy	7	Acenocoumarol	535	184	1178	645
4	1999	chemically induced	2	Sertraline	1662	1872	918	1953
4	2004	blood	20	Octopamine	188	342	188	238
4	2004	blood	3	Synephrine	1642	1580	2352	1432
4	2004	drug therapy	24	Fentanyl	180	471	74	183
4	2005	drug therapy	2	Tramadol	1071	837	871	1616

7B. Migraine – Highest gold standard chemicals 1989-1990 order by descending Article Count (ArtCt).

					Rankings (out of 7,122 chems in HS)			
Art Ct	First Yr	DisQual	Prot Ct	ChemName	Avg Rank	Wt COS	Wt Prop	Prot Ct
41	1991	physiopathology	25	Nitric Oxide	311	647	462	264
19	1992	drug therapy	24	Fluoxetine	827	2563	28	284
13	1992	drug therapy	31	Acetazolamide	183	469	184	195
12	1995	drug therapy	47	Capsaicin	37	203	24	83
11	1991	drug therapy	11	Butorphanol	821	1931	208	725
10	2001	drug therapy	14	Dipyrone	344	559	394	519
9	1991	drug therapy	24	Magnesium Sulfate	190	363	234	279
8	1997	drug therapy	22	Droperidol	270	858	62	315
6	1990	physiopathology	49	Apomorphine	9	49	19	69
5	1991	drug therapy	13	Buspirone	439	934	267	583
5	1992	drug therapy	10	Piroxicam	1098	1524	1597	776
5	1993	drug therapy	1	Ketorolac	4674	3823	4373	3826
5	1993	drug therapy	6	Moclobemide	1661	2828	825	1340
5	1996	prevention & control	2	iprazochrome	2907	3318	1493	3394
5	1997	drug therapy	1	KB 2796	4655	3878	4355	3889
4	1992	drug therapy	7	Ketoprofen	665	492	976	1020
4	1992	drug therapy	1	oxetorone	3554	4204	2569	4160
4	1993	chemically induced	3	Ondansetron	2940	3518	2313	2518
4	1993	drug therapy	31	Diphenhydramine	60	129	80	192
4	1995	physiopathology	83	Cyclic GMP	4	47	10	7
4	1996	drug therapy	9	Acenocoumarol	701	320	1713	813
4	1999	chemically induced	7	Sertraline	1149	1873	745	1110
4	2004	blood	23	Octopamine	168	250	265	287

4	2004	blood	6	Synephrine	1573	2184	1587	1307
4	2004	drug therapy	37	Fentanyl	137	477	64	139
4	2004	drug therapy	2	zonisamide	2912	2396	4197	2910
4	2005	drug therapy	7	Tramadol	1220	2501	272	1135

7C. Migraine – Highest gold standard chemicals 1994-1995 order by descending Article Count (ArtCt).

					Rankings (out of 10,467 chems in HS)			
Art Ct	First Yr	DisQual	Prot Ct	ChemName	Avg Rank	Wt COS	Wt Prop	Prot Ct
12	1995	drug therapy	78	Capsaicin	29	239	21	42
12	1997	drug therapy	8	lamotrigine	2201	3628	1751	1533
10	2001	drug therapy	19	Dipyrone	304	404	400	593
8	1997	drug therapy	26	Droperidol	303	895	105	395
5	1995	drug therapy	1	dotarizine	6364	9261	5235	8883
5	1996	prevention & control	2	iprazochrome	3983	4238	2519	4704
5	1997	drug therapy	6	KB 2796	913	570	1040	1798
5	1998	prevention & control	1	venlafaxine	5272	6398	4183	5902
4	1995	physiopathology	108	Cyclic GMP	2	26	15	9
4	1996	drug therapy	13	Acenocoumarol	737	759	1223	909
4	1999	chemically induced	9	Sertraline	1568	3207	629	1372
4	2004	blood	30	Octopamine	179	351	244	314
4	2004	blood	7	Synephrine	2125	3287	1724	1732
4	2004	drug therapy	53	Fentanyl	81	334	37	123
4	2004	drug therapy	11	zonisamide	627	730	721	1070
4	2005	drug therapy	9	Tramadol	1203	2307	556	1349

Appendix 8. 5-HT6 binders and nonbinders used in the modeling sets

Binders	NonBinders
olanzapine	Ephedrine
Fluphenazine	Diclofenac
Haloperidol	Cocaine
Ketanserin	celecoxib
duloxetine	Aspirin
Loxapine	etoricoxib
Lysergic Acid Diethylamide	Ibuprofen
Amitriptyline	Ketorolac
ziprasidone	Methylphenidate
Mianserin	Naproxen
Molindone	nimesulide
Cyproheptadine	N-Methyl-3,4-methylenedioxyamphetamine
Ergotamine	Phenylpropanolamine
norclozapine	Piroxicam
Methysergide	pramipexol
atomoxetine	rofecoxib
Chlorpromazine	Rutin
Pimozide	Trazodone
venlafaxine	valdecoxib
Amoxapine	meloxicam
Bromocriptine	Ephedrine
quetiapine	Diclofenac
Risperidone	Cocaine
Perphenazine	celecoxib
Clozapine	Aspirin
Thioridazine	etoricoxib
Thiothixene	Ibuprofen
aripiprazole	Ketorolac
Trifluoperazine	Methylphenidate
	Naproxen
	nimesulide
	N-Methyl-3,4-methylenedioxyamphetamine
	Phenylpropanolamine
	Piroxicam
	pramipexol
	rofecoxib
	Rutin
	Trazodone
	valdecoxib
	meloxicam

Appendix 9. Dopamine Antagonists used in modeling sets

Chemical Name
Methotrimeprazine
Tiapride
Thiothixene
Thioridazine
Thiethylperazine
Sulpiride
Risperidone
Prochlorperazine
Pimozide
Perphenazine
Metoclopramide
Trifluoperazine
Loxapine
Amoxapine
Haloperidol
Fluphenazine
Flupenthixol
Droperidol
Domperidone
Clopenthixol
Chlorprothixene
Chlorpromazine
Benperidol
Perazine

REFERENCES

Adams, C. P., & Brantner, V. V. (2006). Estimating the cost of new drug development: is it really 802 million dollars? *Health Affairs (Project Hope), 25*(2), 420-428.

Ahlers, C. B., Hristovski, D., Kilicoglu, H., & Rindflesch, T. C. (2007). Using the literature-based discovery paradigm to investigate drug mechanisms. *AMIA Annual Symposium Proceedings / AMIA Symposium*, 6-10.

Anbar, R. D., Lapey, A., Khaw, K. T., Spragg, J., Strieder, D. J., Shaw, L. F., et al. (1990). Does lithium carbonate affect the ion transport abnormality in cystic fibrosis? *Pediatric Pulmonology, 8*(2), 82-88.

Andersen, M. (1991). Exacerbation of psoriasis during treatment with H2 antagonists. [Forvaerring af psoriasis under behandling med H2-antagonister] *Ugeskrift for Laeger, 153*(2), 132.

Aronson, A. R., Bodenreider, O., Chang, H. F., Humphrey, S. M., Mork, J. G., Nelson, S. J., et al. (2000). The NLM Indexing Initiative. *Proceedings / AMIA Symposium*, , 17-21.

Banville, D. L. (2006). Mining chemical structural information from the drug literature. *Drug Discovery Today, 11*(1-2), 35-42.

Bekhuis, T. (2006). Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. *Biomedical Digital Libraries, 3*, 2.

Bernstein, J. E., Parish, L. C., Rapaport, M., Rosenbaum, M. M., & Roenigk, H. H.,Jr. (1986). Effects of topically applied capsaicin on moderate and severe psoriasis vulgaris. *Journal of the American Academy of Dermatology, 15*(3), 504-507.

Bigal, M. E., & Krymchantowski, A. V. (2006). Emerging drugs for migraine prophylaxis and treatment. *MedGenMed : Medscape General Medicine, 8*(2), 31.

Blagosklonny, M. V., & Pardee, A. B. (2002). Conceptual biology: unearthing the gems. *Nature, 416*(6879), 373.

Blake, J. B. (1980). *Centenary of Index Medicus. NIH Publication 80-2068.* Bethesda, Maryland: U.S. Department of Health and Human Services.

Blake, J. B. (1986). From Surgeon General's bookshelf to National Library of Medicine: a brief history. *Bulletin of the Medical Library Association, 74*(4), 318-324.

Blaschke, C., Andrade, M. A., Ouzounis, C., & Valencia, A. (1999). Automatic extraction of biological information from scientific text: protein-protein interactions. *Proceedings*

International Conference on Intelligent Systems for Molecular Biology ; ISMB.International Conference on Intelligent Systems for Molecular Biology, , 60-67.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research, 32*(Database issue), D267-70.

Bodenreider, O. (2008). Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearbook of Medical Informatics, ,* 67-79.

Bradley, D. (2005). Why big pharma needs to learn the three 'R's. *Nature Reviews.Drug Discovery, 4*(6), 446.

Bray, D. (2001). Reasoning for results. *Nature, 412*(6850), 863.

Brown, J., Mellis, C. M., & Wood, R. E. (1985). Edetate sodium aerosol in Pseudomonas lung infection in cystic fibrosis. *American Journal of Diseases of Children (1960), 139*(8), 836-839.

Campillos, M., Kuhn, M., Gavin, A. C., Jensen, L. J., & Bork, P. (2008). Drug target identification using side-effect similarity. *Science (New York, N.Y.), 321*(5886), 263-266.

Chaffey, D., & Wood, S. (2005). *Business information management: Improving performance using information systems*. Harlow: FT Prentice Hall.

Chang, J. T., Schutze, H., & Altman, R. B. (2004). GAPSCORE: finding gene and protein names one word at a time. *Bioinformatics (Oxford, England), 20*(2), 216-225.

Chen, E. S., Hripcsak, G., Xu, H., Markatou, M., & Friedman, C. (2008). Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. *Journal of the American Medical Informatics Association : JAMIA, 15*(1), 87-98.

Chowdhury, G. (1992a). Automatic interpretation of the texts of chemical patent abstracts. 1. Lexical analysis and categorization. *Journal of Chemical Information and Computer Sciences, 32*(5), 463-467.

Chowdhury, G. (1992b). Automatic interpretation of the texts of chemical patent abstracts. 2. Processing and Results. *Journal of Chemical Information and Computer Sciences, 32*(5), 468-473.

Cimino, J. J., & Barnett, G. O. (1993). Automatic knowledge acquisition from MEDLINE. *Methods of Information in Medicine, 32*(2), 120-130.

Cooke, H. (2004). A historical review of the chemistry periodical literature until 1950. *Learned Publishing, 17*(2), 125-134.

Cooke-Fox, D. I., Kirby, G. H., & Rayner, J. D. (1989a). Computer translation of IUPAC systematic organic chemical nomenclature. 1. Introduction and background to a grammar-based approach. *Journal of Chemical Information and Computer Sciences*, *29*(2), 101-105.

Cooke-Fox, D. I., Kirby, G. H., & Rayner, J. D. (1989b). Computer translation of IUPAC systematic organic chemical nomenclature. 2. Development of a formal grammar. *Journal of Chemical Information and Computer Sciences, 29*(2), 106-112.

Cooke-Fox, D. I., Kirby, G. H., & Rayner, J. D. (1989c). Computer translation of IUPAC systematic organic chemical nomenclature. 3. Syntax analysis and semantic processing. *Journal of Chemical Information and Computer Sciences, 29*(2), 112-118.

Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., et al. (2008). ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research, 36*(Database issue), D344-50.

Derleth, D. P. (2003). A possible antiinflammatory treatment for cystic fibrosis. *American Journal of Respiratory and Critical Care Medicine, 167*(2), 278-9; author reply 279.

Di Giovanni, G., Di Matteo, V., Pierucci, M., & Esposito, E. (2008). Serotonin-dopamine interaction: electrophysiological evidence. *Progress in Brain Research, 172*, 45-71.

DiMagno, E. P. (2001). Gastric acid suppression and treatment of severe exocrine pancreatic insufficiency. *Best Practice & Research.Clinical Gastroenterology, 15*(3), 477-486.

Dogan, B., Karabudak, O., & Harmanyeri, Y. (2008). Antistreptococcal treatment of guttate psoriasis: a controlled study. *International Journal of Dermatology, 47*(9), 950-952.

Ehrlich, A., Booher, S., Becerra, Y., Borris, D. L., Figg, W. D., Turner, M. L., et al. (2004). Micellar paclitaxel improves severe psoriasis in a prospective phase II pilot study. *Journal of the American Academy of Dermatology, 50*(4), 533-540.

Elias, A. N., Goodman, M. M., Liem, W. H., & Barr, R. J. (1993). Propylthiouracil in psoriasis: results of an open trial. *Journal of the American Academy of Dermatology, 29*(1), 78-81.

Ellis, C. N., Berberian, B., Sulica, V. I., Dodd, W. A., Jarratt, M. T., Katz, H. I., et al. (1993). A double-blind evaluation of topical capsaicin in pruritic psoriasis. *Journal of the American Academy of Dermatology, 29*(3), 438-442.

Escobar, H., Perdomo, M., Vasconez, F., Camarero, C., del Olmo, M. T., & Suarez, L. (1992). Intestinal permeability to 51Cr-EDTA and orocecal transit time in cystic fibrosis. *Journal of Pediatric Gastroenterology and Nutrition, 14*(2), 204-207.

Ferrari, V. D., & Jirillo, A. (1996). Psoriasis and tamoxifen therapy: a case report. *Tumori, 82*(3), 262-263.

Flaxbart, D. (2007). The Chemical Abstracts Centennial: Whither CAS? *Issues in Science and Technology Librarianship, 49*(Winter)

Fliri, A. F., Loging, W. T., Thadeio, P. F., & Volkmann, R. A. (2005). Analysis of drug-induced effect patterns to link structure and side effects of medicines. *Nature Chemical Biology, 1*(7), 389-397.

Food and Drug Administration. (1989). *COSTART: Coding Symbols for Thesaurus of Adverse Reaction Terms*. Rockville, Maryland: Food and Drug Administration, Center for Drugs and Biologics, Division of Drug and Biological Products Experience.

Frerichs, C., & Smyth, A. (2009). Treatment strategies for cystic fibrosis: what's in the pipeline? *Expert Opinion on Pharmacotherapy, 10*(7), 1191-1202.

Fugmann, R. (1985). Peculiarities of Chemical Information from a Theoretical Viewpoint. *Journal of Chemical Information and Computer Sciences, 25*(3), 174-180.

Funk, M. E., & Reid, C. A. (1983). Indexing consistency in MEDLINE. *Bulletin of the Medical Library Association, 71*(2), 176-183.

Garfield, E. (1964). "Science Citation Index"--A New Dimension in Indexing. *Science (New York, N.Y.), 144*(3619), 649-654.

Garfield, E. (1985). History of Citation Indexes for Chemistry: A Brief Review. *Journal of Chemical Information and Computer Sciences, 25*(3), 170-174.

Garfield, E. (2001). From laboratory to information explosions... the evolution of chemical information services at ISI. *Journal of Information Science, 27*, 119-125.

Gasteiger, J., & Engel, T. (2003). *Chemoinformatics : A textbook*. Weinheim: Wiley-VCH.

Geldenhuys, W. J., & Van der Schyf, C. J. (2009). The serotonin 5-HT6 receptor: a viable drug target for treating cognitive deficits in Alzheimer's disease. *Expert Review of Neurotherapeutics, 9*(7), 1073-1085.

Giacomini, K. M., Krauss, R. M., Roden, D. M., Eichelbaum, M., Hayden, M. R., & Nakamura, Y. (2007). When good drugs go bad. *Nature, 446*(7139), 975-977.

Gibaldi, M. (1993). What is nitric oxide and why are so many people studying it? *Journal of Clinical Pharmacology, 33*(6), 488-496.

Gkoutos, G. V., Murray-Rust, P., Rzepa, H. S., & Wright, M. (2001). Chemical markup, XML and the World-Wide Web. 3. Toward a signed semantic chemical web of trust. *Journal of Chemical Information and Computer Sciences, 41*(5), 1124-1130.

Gkoutos, G. V., Rzepa, H., Clark, R. M., Adjei, O., & Johal, H. (2003). Chemical machine vision: automated extraction of chemical metadata from raster images. *Journal of Chemical Information and Computer Sciences, 43*(5), 1342-1355.

Green, M. R., Austin, S., & Weaver, L. T. (1993). Dual marker one day pancreolauryl test. *Archives of Disease in Childhood, 68*(5), 649-652.

Gunther, S., Kuhn, M., Dunkel, M., Campillos, M., Senger, C., Petsalaki, E., et al. (2008). SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Research, 36*(Database issue), D919-22.

Haas, S. W. (1997). Disciplinary variation in automatic sublanguage term identification. *Journal of the American Society for Information Science, 48*(1), 67-79.

Hajjo, R., Fourches, D., Roth, B. L., & Tropsha, A. (2009). In silico strategies to identify novel 5-HT6 receptor ligands as potential anti-alzheimer's and anti-obesity treatments. Paper presented at the *Abstracts of Papers, 238th ACS National Meeting,* Washington, DC, USA.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, *11*(1)

Halverstam, C. P., & Lebwohl, M. (2008). Nonstandard and off-label therapies for psoriasis. *Clinics in Dermatology, 26*(5), 546-553.

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research, 33*(Database issue), D514-7.

Harris, Z. S. (2002). The structure of science information. *Journal of Biomedical Informatics, 35*(4), 215-221.

Hattori, K., Wakabayashi, H., & Tamaki, K. (2008). Predicting key example compounds in competitors' patent applications using structural information alone. *Journal of Chemical Information and Modeling, 48*(1), 135-142.

Hatzivassiloglou, V., Duboue, P. A., & Rzhetsky, A. (2001). Disambiguating proteins, genes, and RNA in text: a machine learning approach. *Bioinformatics (Oxford, England), 17 Suppl 1,* S97-106.

Hodge, G. (1989). Automatic recognition of chemical names in natural-language texts. *Abstracts of Papers of the American Chemical Society, 197,* 17-CINF.

Hristovski, D., Friedman, C., Rindflesch, T. C., & Peterlin, B. (2006). Exploiting semantic relations for literature-based discovery. *AMIA Annual Symposium Proceedings / AMIA Symposium,* , 349-353.

Hristovski, D., Stare, J., Peterlin, B., & Dzeroski, S. (2001). Supporting discovery in medicine by association rule mining in Medline and UMLS. *Medinfo.MEDINFO, 10*(Pt 2), 1344-1348.

Ibison, P., Jacquot, M., Kam, F., Neville, A. G., Simpson, R. W., Tonnelier, C., et al. (1993). Chemical literature data extraction: The CLiDE Project. *Journal of Chemical Information and Computer Sciences, 33*(3), 338-344.

IUPAC. (2009). *International Union of Pure and Applied Chemistry.*. 2009, from www.iupac.org

Jensen, L. J., Saric, J., & Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews.Genetics, 7*(2), 119-129.

Karthikeyan, M., Krishnan, S., Pandey, A. K., & Bender, A. (2006). Harvesting chemical information from the Internet using a distributed approach: ChemXtreme. *Journal of Chemical Information and Modeling, 46*(2), 452-461.

Keiser, M. J., Setola, V., Irwin, J. J., Laggner, C., Abbas, A. I., Hufeisen, S. J., et al. (2009). Predicting new molecular targets for known drugs. *Nature, 462*(7270), 175-181.

Kerem, E. (2006). Mutation specific therapy in CF. *Paediatric Respiratory Reviews, 7 Suppl 1,* S166-9.

Koshland, D. E.,Jr. (1992). The molecule of the year. *Science (New York, N.Y.), 258*(5090), 1861.

Kostoff, R. N., Block, J. A., Stump, J. A., & Pfeil, K. M. (2004). Information content in Medline record fields. *International Journal of Medical Informatics, 73*(6), 515-527.

Kostoff, R. N., Briggs, M. B., Solka, J. L., & Rushenberg, R. L. (2008). Literature-related discovery (LRD): Methodology. *Technological Forecasting and Social Change, 75*(2), 186-202.

Krejsa, C. M., Horvath, D., Rogalski, S. L., Penzotti, J. E., Mao, B., Barbosa, F., et al. (2003). Predicting ADME properties and side effects: the BioPrint approach. *Current Opinion in Drug Discovery & Development, 6*(4), 470-480.

Kristensen, J. K., Petersen, L. J., Hansen, U., Nielsen, H., Skov, P. S., & Nielsen, H. J. (1995). Systemic high-dose ranitidine in the treatment of psoriasis: an open prospective clinical trial. *The British Journal of Dermatology, 133*(6), 905-908.

Kuhn, M., Campillos, M., Letunic, I., Jensen, L. J., & Bork, P. (2010). A side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology, 6*, 343.

Levine, D., & Gottlieb, A. (2009). Evaluation and management of psoriasis: an internist's guide. *The Medical Clinics of North America, 93*(6), 1291-1303.

Lindsay, R. K., & Gordon, M. D. (1999). Literature-based discovery by lexical statistics. *Journal of the American Society of Information Science, 50*(7), 574-587.

Magela Magalhaes, G., Coelho da Silva Carneiro, S., Peisino do Amaral, K., de Freire Cassia, F., Machado-Pinto, J., & Cuzzi, T. (2006). Psoriasis and pentoxifylline: a clinical, histopathologic, and immunohistochemical evaluation. *Skinmed, 5*(6), 278-284.

Manning, C. D., & Schuetze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, Mass.: MIT Press.

McDaniel, J. R., & Balmuth, J. R. (1992). Kekule: OCR-optical chemical (structure) recognition. *Journal of Chemical Information and Computer Sciences, 32*(4), 373-378.

Mendonca, E. A., & Cimino, J. J. (2000). Automated knowledge extraction from MEDLINE citations. *Proceedings: AMIA Annual Symposium*, 575-579.

Miano, S., Parisi, P., Pelliccia, A., Luchetti, A., Paolino, M. C., & Villa, M. P. (2008). Melatonin to prevent migraine or tension-type headache in children. *Neurological Sciences : Official Journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology, 29*(4), 285-287.

Mitchell, E. S., & Neumaier, J. F. (2005). 5-HT6 receptors: a novel target for cognitive enhancement. *Pharmacology & Therapeutics, 108*(3), 320-333.

Murray-Rust, P., & Rzepa, H. S. (2001). Chemical markup, XML and the World-Wide Web. 2. Information objects and the CMLDOM. *Journal of Chemical Information and Computer Sciences, 41*(5), 1113-1123.

Murray-Rust, P., & Rzepa, H. S. (2003). Chemical markup, XML, and the World Wide Web. 4. CML schema. *Journal of Chemical Information and Computer Sciences, 43*(3), 757-772.

Murray-Rust, P., Rzepa, H. S., Stewart, J. J., & Zhang, Y. (2005). A global resource for computational chemistry. *Journal of Molecular Modeling, 11*(6), 532-541.

Murray-Rust, P., Rzepa, H. S., Tyrrell, S. M., & Zhang, Y. (2004). Representation and use of chemistry in the global electronic age. *Organic & Biomolecular Chemistry, 2*(22), 3192-3203.

Murray-Rust, P., Rzepa, H. S., Williamson, M. J., & Willighagen, E. L. (2004). Chemical markup, XML, and the World Wide Web. 5. Applications of chemical metadata in RSS aggregators. *Journal of Chemical Information and Computer Sciences, 44*(2), 462-469.

Naldi, L., & Gambini, D. (2007). The clinical spectrum of psoriasis. *Clinics in Dermatology, 25*(6), 510-518.

Narayanasamy, V., Mukhopadhyay, S., Palakal, M., & Potter, D. A. (2004). TransMiner: mining transitive associations among biological objects from text. *Journal of Biomedical Science, 11*(6), 864-873.

National Library of Medicine. (2006). *Unified Medical Language System Fact Sheet.*http://www.nlm.nih.gov.libproxy.lib.unc.edu/pubs/factsheets/umls.html

National Library of Medicine. (2008). *MEDLINE.*http://www.nlm.nih.gov.libproxy.lib.unc.edu/pubs/factsheets/medline.html

National Library of Medicine. (2010). *Medical Subject Headings (MeSH) Fact Sheet.*http://www.nlm.nih.gov.libproxy.lib.unc.edu/pubs/factsheets/mesh.html

National Psoriasis Foundation. (2009). , December, 2009, from http://www.psoriasis.org/netcommunity/sublearn07_tools_iwfm

Neveol, A., Zeng, K., & Bodenreider, O. (2006). Besides precision & recall: exploring alternative approaches to evaluating an automatic indexing tool for MEDLINE. *AMIA Annual Symposium Proceedings / AMIA Symposium*, 589-593.

Nielsen, H. J., Nielsen, H., & Georgsen, J. (1991). Ranitidine for improvement of treatment-resistant psoriasis. *Archives of Dermatology, 127*(2), 270.

NIH. (2007). *Molecular Libraries Program.*, 2007, from http://mli.nih.gov.libproxy.lib.unc.edu/mli/

Omulecki, A., Broniarczyk-Dyla, G., Zak-Prelich, M., & Choczaj-Kukula, A. (1996). Is pentoxifylline effective in the treatment of psoriasis? *Journal of the American Academy of Dermatology, 34*(4), 714-715.

O'Sullivan, B. P., & Freedman, S. D. (2009). Cystic fibrosis. *Lancet, 373*(9678), 1891-1904.

Peikert, A., Wilimzig, C., & Kohne-Volland, R. (1996). Prophylaxis of migraine with oral magnesium: results from a prospective, multi-center, placebo-controlled and double-blind randomized study. *Cephalalgia : An International Journal of Headache, 16*(4), 257-263.

Petrič, I., Urbančič, T., Cestnik, B., & Macedoni-Lukšič, M. (2008). Literature mining method RaJoLink for uncovering relations between biomedical concepts. *Journal of Biomedical Informatics,* Polat, M., Lenk, N., Yalcin, B., Gur, G., Tamer, E., Artuz, F., et al. (2007). Efficacy of erythromycin for psoriasis vulgaris. *Clinical and Experimental Dermatology, 32*(3), 295-297.

Prandota, J. (2001). Clinical pharmacology of furosemide in children: a supplement. *American Journal of Therapeutics, 8*(4), 275-289.

Pruitt, K. D., Katz, K. S., Sicotte, H., & Maglott, D. R. (2000). Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends in Genetics : TIG, 16*(1), 44-47.

Rindflesch, T. C., Tanabe, L., Weinstein, J. N., & Hunter, L. (2000). EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pacific Symposium on Biocomputing.Pacific Symposium on Biocomputing,* , 517-528.

Robinson, M., Daviskas, E., Eberl, S., Baker, J., Chan, H. K., Anderson, S. D., et al. (1999). The effect of inhaled mannitol on bronchial mucus clearance in cystic fibrosis patients: a pilot study. *The European Respiratory Journal : Official Journal of the European Society for Clinical Respiratory Physiology, 14*(3), 678-685.

Rosenberg, E. W., Noah, P. W., Zanolli, M. D., Skinner, R. B.,Jr, Bond, M. J., & Crutcher, N. (1986). Use of rifampin with penicillin and erythromycin in the treatment of psoriasis. Preliminary report. *Journal of the American Academy of Dermatology, 14*(5 Pt 1), 761-764.

Roth, B. L., Lopez, E., Patel, S., & Kroeze, W. K. (2000). The Multiplicity of Serotonin Receptors: Uselessly Diverse Molecules or an Embarrassment of Riches? *Neuroscientist, 6*(4), 252.

Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science, 33*(2), 163-180.

Rubin, D. L., Thorn, C. F., Klein, T. E., & Altman, R. B. (2005). A statistical approach to scanning the biomedical literature for pharmacogenetics knowledge. *Journal of the American Medical Informatics Association : JAMIA, 12*(2), 121-129.

Sabat, R., Sterry, W., Philipp, S., & Wolk, K. (2007). Three decades of psoriasis research: where has it led us? *Clinics in Dermatology, 25*(6), 504-509.

Seki, K., & Mostafa, J. (2007). Discovering implicit associations between genes and hereditary diseases.*Pacific Symposium on Biocomputing* , 316-327.

Shatkay, H., & Feldman, R. (2003). Mining the Biomedical Literature in the Genomic Era: An Overview. *Journal of Computational Biology, 10*(6), 821-855.

Sinclair, S. (1999). Migraine headaches: nutritional, botanical and other alternative approaches. *Alternative Medicine Review : A Journal of Clinical Therapeutic, 4*(2), 86-95.

Skolnik, H. (1982). *The literature matrix of chemistry*. New York: Wiley.

Smalheiser, N. R., & Swanson, D. R. (1996a). Indomethacin and Alzheimer's disease. *Neurology, 46*(2), 583.

Smalheiser, N. R., & Swanson, D. R. (1996b). Linking estrogen to Alzheimer's disease: an informatics approach. *Neurology, 47*(3), 809-810.

Smalheiser, N. R., & Swanson, D. R. (1998). Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Computer Methods and Programs in Biomedicine, 57*(3), 149-153.

Sorensen, K. V. (1988). Valproate: a new drug in migraine prophylaxis. *Acta Neurologica Scandinavica, 78*(4), 346-348.

Srinivasan, P. (2004). Text mining: generating hypotheses from MEDLINE. *Journal of the American Society for Information Science and Technology, 55*(5), 396-413.

Stapley, B. J., & Benoit, G. (2000). Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *Pacific Symposium on Biocomputing*, 529-540.

Storey, S., & Wald, G. (2008). Novel agents in cystic fibrosis. *Nature Reviews.Drug Discovery, 7*(7), 555-556.

Swanson, D. R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine, 30*(1), 7-18.

Swanson, D. R. (1988). Migraine and magnesium: eleven neglected connections. *Perspectives in Biology and Medicine, 31*(4), 526-557.

Swanson, D. R. (1990). Medical literature as a potential source of new knowledge. *Bulletin of the Medical Library Association, 78*(1), 29-37.

Szeifert, G. T., Varga, E., Damjanovich, L., & Gomba, S. (1987). The chronically furosemide-treated mouse as a possible ultrastructural model for cystic fibrosis. *Acta Morphologica Hungarica, 35*(3-4), 207-210.

Torvik, V. I., Renear, A., Smalheiser, N. R., & Marshall, C. (2009). Beyond (simple) reading: Strategies, discoveries, and collaborations. Paper presented at the Vancouver, British Columbia, Canada.

Townsend, J. A., Adams, S. E., Waudby, C. A., de Souza, V. K., Goodman, J. M., & Murray-Rust, P. (2004). Chemical documents: machine understanding and automated information extraction. *Organic & Biomolecular Chemistry, 2*(22), 3294-3300.

Tsankov, N., & Grozdev, I. (2009). Rifampicin in the treatment of psoriasis. *Journal of the European Academy of Dermatology and Venereology : JEADV, 23*(1), 93-95.

van Rijsbergen, C. J., Robertson, S. E., & Porter, M. F. (1980). *New models in probabilistic information retrieval* No. 5587)British Library.

Vander Stouw, G. G., Naznitsky, I., & Rush, J. E. (1967). Procedures for Converting Systematic Names of Organic Compounds into Atom-Bond Connection Tables. *Journal of Chemical Documentation, 7*(3), 165-169.

Weeber, M., Klein, H., de Jong-van den Berg, L. T. W., & Vos, R. (2001). Using Concepts in Literature-Based Discovery: Simulating Swanson's Raynaud-Fish Oil and Migraine-Magnesium Discoveries. *Journal of the American Society for Information Science and Technology, 52*(7), 548-557.

Weeber, M., Vos, R., Klein, H., De Jong-Van Den Berg, L. T., Aronson, A. R., & Molema, G. (2003). Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. *Journal of the American Medical Informatics Association : JAMIA, 10*(3), 252-259.

Weisgerber, D. W. (1997). Chemical Abstracts Service Chemical Registry System: History, scope, and impacts. *Journal of the American Society for Information Science, 48*(4), 349-360.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., et al. (2008). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research, 36*(Database issue), D13-21.

Wilbur, W. J., Hazard, G. F.,Jr, Divita, G., Mork, J. G., Aronson, A. R., & Browne, A. C. (1999). Analysis of biomedical text for chemical names: a comparison of three methods. *Proceedings / AMIA Annual Symposium.* 176-180.

Willett, P., Barnard, J. M., & Downs, G. M. (1998). Chemical Similarity Searching. *Journal of Chemical Information and Computer Sciences, 38*(6), 983-996.

Williams, A. J. (2008). Public chemical compound databases. *Current Opinion in Drug Discovery & Development, 11*(3), 393-404.

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., et al. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research, 36*(Database issue), D901-6.

Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., et al. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research, 34*(Database issue), D668-72.

Wren, J. D., Bekeredjian, R., Stewart, J. A., Shohet, R. V., & Garner, H. R. (2004). Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics (Oxford, England), 20*(3), 389-398.

Yetisgen-Yildiz, M., & Pratt, W. (2006). Using statistical and knowledge-based approaches for literature-based discovery. *Journal of Biomedical Informatics, 39*(6), 600-611.

Yetisgen-Yildiz, M., & Pratt, W. (2009). A new evaluation methodology for literature-based discovery systems. *Journal of Biomedical Informatics, 42*(4), 633-643.

Yu, H., Kim, W., Hatzivassiloglou, V., & Wilbur, W. J. (2007). Using MEDLINE as a knowledge source for disambiguating abbreviations and acronyms in full-text biomedical journal articles. *Journal of Biomedical Informatics, 40*(2), 150-159.

Zamora, E. M., & Blower, P. E. (1984a). Extraction of chemical reaction information from primary journal text using computational linguistics techniques. 1. Lexical and syntactic phases. *Journal of Chemical Information and Computer Sciences, 24*(3), 176-181.

Zamora, E. M., & Blower, P. E. (1984b). Extraction of chemical reaction information from primary journal text using computational linguistics techniques. 2. Semantic phase. *Journal of Chemical Information and Computer Sciences, 24*(3), 181-188.

Zimmermann, M., Fluck, J., Thi le, T. B., Kolarik, C., Kumpf, K., & Hofmann, M. (2005). Information extraction in the life sciences: perspectives for medicinal chemistry, pharmacology and toxicology. *Current Topics in Medicinal Chemistry, 5*(8), 785-796.

Zonneveld, I. M., Meinardi, M. M., Karlsmark, T., Johansen, U. B., Kuiters, G. R., Hamminga, L., et al. (1997). Ranitidine does not affect psoriasis: a multicenter, double-blind, placebo-controlled study. *Journal of the American Academy of Dermatology, 36*(6 Pt 1), 932-934.