COMPUTATIONAL STUDY OF THERAPEUTIC TARGETS AND ADME-ASSOCIATED PROTEINS AND APPLICATION IN DRUG DESIGN

ZHENG CHANJUAN
(M.Sc. ChongQing Univ.)

A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
DEPARTMENT OF PHARMACY
NATIONAL UNIVERSITY OF SINGAPORE

2006

ACKNOWLEDGEMENTS

This thesis would not have been possible to be completed without the kind support, help, and guidance by lots of people. First of all, I would like to express my deep gratitude to my thesis advisor Dr. Chen Yuzong. He provides me with the guidance, support, and encouragement during my years at National University of Singapore. His advice and insights guided me throughout my doctoral studies. Likewise, his professional knowledge and kind patience kept me motivated to complete my Ph.D. thesis. His commentary and counsel I retain in my mind will continue to guide me through my professional career in future.

Also, I would like to thank my current colleagues and friends for their support and collaboration in my academic research and daily life: Mr. Yap Chun Wei, Mr. Han Lianyi, Mr. Lin Honghuang, Mr. Zhou Hao, Mr. Xie Bin, Ms. Cui Juan, Ms. Zhang Hailei, Ms. Tang Zhiqun, Ms. Jiang Li, Mr. Li Hu, Mr. Ung Choong Yong. We shared lots of precious experience and happy life in Singapore, which are the treasures in my life. Although my doctoral study has come to an end, the friendship between us will remain. In addition, I would also like to thank my former colleagues for their helpful discussion, advice, guidance and encouragement on my studies and research: Dr. Cao Zhiwei, Dr. Ji Zhiliang, Dr. Chen Xin, Mr. Wang Jifeng, Ms. Sun Lizhi, Ms. Yao Lixia, and Dr. Xue Ying.

I would also like to give special thanks to my husband and my parents for their endless love, support, and encouragement. I dedicate this thesis to them with all my love.

TABLE OF CONTENTS

SUMMARY

With the exponential growth of genomic data, the pharmaceutical industry enter the post-genomic era and adopts a multi-disciplinary strategy is increasingly used to advance drug discovery. A large variety of specialties and general-purpose bioinformatics databases have been developed to store, organize and manage vast amounts of biomedical and genomic data. The first aim of this thesis is to develop or update three pharmainformatics databases: Therapeutic Target Database (TTD), Therapeutically Relevant Multiple Pathways (TRMP) database, and ADME-Associated Proteins (ADME-AP) database. These databases may serve as the basis for further knowledge discovery in drug target search analysis; drug pharmacokinetics and pharmacogenetics studies; and drug design and testing.

TTD (http://bidd.nus.edu.sg/group/cjttd/ttd.asp) may be the world's first public resource for providing comprehensive information about the reported targets of marketed and investigational drugs. There is a significant increase from that of ~500 targets reported in a 1996 survey [1] to 1,535 targets in latest TTD version, indicating that more therapeutic targets and related information recorded in recent publications. This part of work is important for laying the foundations to more advanced studies about therapeutic targets. By using similar developing strategies, a database of known therapeutically relevant multiple pathways (TRMP, http://bidd.nus.edu.sg/group/trmp/trmp.asp), was developed to facilitate a comprehensive understanding of the relationship between different targets of the same disease and also to facilitate mechanistic study of drug actions. It contains multiple and individual pathways information, and also include those relevant targets, disease, drugs information. Moreover, a new version of another pharmainformatics database, ADME-AP database

(http://bidd.nus.edu.sg/group/admeap/admeap.asp) has been updated in this work. A great number of polymorphisms and drug response information have been integrated into the old version. By analysis of this kind of information, we assess the usefulness of the relevant information for facilitating pharmacogenetic prediction of drug responses, and discuss computational methods used for predicting individual variations of drug responses from the polymorphisms of ADME-APs.

With the completion of human genome sequencing and the rapid development of numerous computational approaches; continuous effort and increasing interest have been directed at the search of new targets, which has led to the identification of a growing number of new targets as well as the exploration of known targets. As a result, the second aim of this thesis is to carry out a computational study of therapeutic targets.

Firstly, the progress of target exploration is studied and some characteristics of currently explored targets, including their sequence, family representation, pathway association, tissue distribution, genome location are analyzed. Moreover, from these target features, some simple rules can be derived for facilitating the search of druggable proteins and for estimating the level of difficulty of their exploration, including (1) Protein is from one of the limited number of target families; (2) Sequence variation between protein's drug-binding domain and those of the human proteins in the same family allows differential binding of a "rule-of-five" molecule; (3) Protein preferably has less than 15 human similarity proteins outside its family (HSP); (4) Protein is preferably involved in no more than 3 human pathways (HP); (5) For organ or tissue specific diseases, protein is preferably distributed in no more than 5 human tissues (HT); (6) A higher number of HSP, HP and HT does not preclude the

protein as a potential target, it statistically increases the chance of undesirable interferences and the level of difficulty for finding viable drugs. The results indicate that some simple rules can be derived for facilitating the search of druggable proteins and for estimating the level of difficulty of their exploration.

Secondly, to test the feasibilities of target identification by using Artificial Intelligent (AI) methods from protein sequence, an AI system is trained by using sequence derived physicochemical properties of the known targets. Furthermore, this prediction system is evaluated by using 5-fold cross validation and scanning human, yeast, and HIV genomes. The prediction results are consistent with previous studies of these genomes, which suggest that AI methods such as Support Vector Machines (SVMs) may be potentially useful for facilitating genome search of druggable proteins. With more biomedical data added in, the preliminary prediction system of druggable proteins will be extended and consolidated for speeding up the process of drug discovery.

LIST OF TABLES

LIST OF FIGURES

ACRONYMS

ABC	ATP-Binding Cassette
ADME	Absorption, Distribution, Metabolism and Excretion
ADME-AP	ADME-Associated Proteins
ADR	Drug Adverse Reaction
AI	Artificial Intelligent
ANN	artificial neural networks
CBI	Center for Information Biology
CYP	Cytochrome P450
DA	Discriminant Analysis
DBMS	Database Management System
EBI	European Bioinformatics Institute
EMBL	European Molecular Biology Laboratory
FDA	Food and Drug Administration
GPCR	G-protein coupled receptor
HGP	Human Genome Project
HP	Human Pathways
HSP	Human Similarity Proteins
HT	Human Ttissues
HUGO	Human Genome Organization
KEGG	Kyoto Encyclopedia of Genes and Genomes database
MBD	Molecular Biology Database
MMPs	Matrix Metalloproteinases
NAR	Nucleic Acids Research
NCBI	National Center for Biotechnology Information
NIH	National Institutes of Health
OODB	Object-Oriented Database
OOPL	Object-Oriented Programming Language
OSH	Optimal Separation Hyperplane
PDB	Protein Data Bank
SIB	Swiss Institute of Bioinformatics
SNP	Single-Nucleotide Polymorphisms
SQL	Structured Query Language
SRM	Structural Risk Minimization
SVMs	Support Vector Machines
TCDB	Transporter Classification Database
TET	Target Exploration Time
TRMP	Therapeutically Relevant Multiple Pathways
TTD	Therapeutic Target Database
VC	Vapnik-Chervonenkis
WHO	World Health Organization

1 Introduction

1.1 Overview of target discovery in pharmaceutical research

Due to the modern life style, an increasing number of people are suffering from various health problems. How to deal with those problems has become the research focus of many biomedical scientists in both academic and pharmaceutical industry [2]. Thus, most scientists pay close attention to drug discovery. It is generally agreed that finding effective drugs for specific disease is an essential way to solve the health problems [2]. In addition, with the advent of molecular biology, the completion of human genome project and the rapid development of numerous computational approaches, more innovative biological concepts and technologies have been introduced into drug discovery [3-5]. These innovations are essential for constructing modern drug discovery programs in which target discovery plays an important role [3].

1.1.1 Process of drug discovery

Drug development is generally a long, costly and uncertain process. Figure 1-1 illustrates the process of drug discovery, which can be roughly divided into two phases [6]. One is the early pharmaceutical research phase and the other is the late phase. The former mainly comprises preliminary investigations, target discovery and lead discovery. The latter consists of preclinical and clinical evaluation. According to the Tufts Center for the study of drug development (November, 2001), by using traditional drug discovery methods, developing a new marketed drug takes 10-15 years, and spends about $800 million USD.

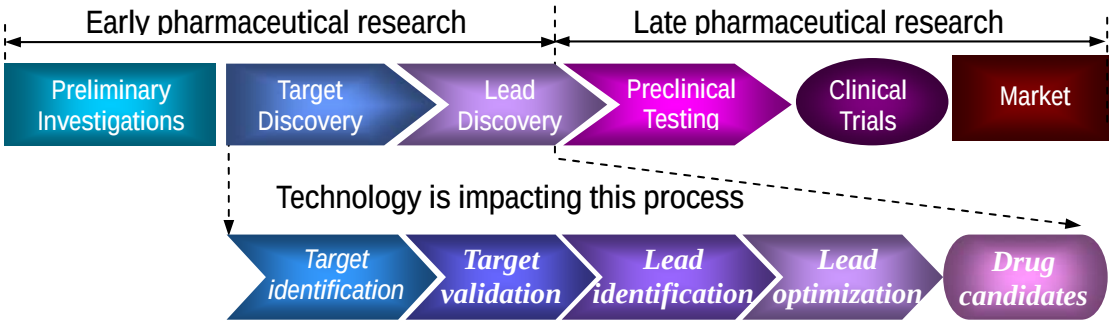


Figure 1-1: Overview of drug discovery process [6]

How to efficiently reduce the cost and the time of drug discovery is a major task of current research. As revealed by Figure 1-1, at certain drug design stages, the use of computational technologies would be a feasible way to solve this problem. Moreover, most drug discovery activities begin with target discovery, which involve the identification and early validation of disease modifying targets. Therefore, computational study of the target characteristics and developing computer target prediction methods are significant for understanding the mechanism of drug action and thus speeding up new target discovery [3, 7].

1.1.2 Brief introduction to target discovery

Generally, target discovery includes two parts: target identification and target validation [6]. Target identification attempts to find new targets, normally proteins, which can be modulated by modulators, such as small molecules and peptides, and thus inhibit or reverse disease progression. For target validation, it plays a crucial role in demonstrating the function of potential targets in the disease phenotype. The various techniques applied to target discovery can be grouped into two broad strategies: system and molecular approaches [8]. In terms of system approach, the

focus is on the study of disease in whole organisms. The information used in this approach is derived from the clinical science and *in vivo* animal studies. Thus the system approach has traditionally been the primary target discovery strategy in drug discovery. By contrast, molecular approach attempts to identify the novel targets through an understanding of the cellular mechanisms. This approach has been driven by the development of molecular biology, genomics and proteomics in recent decades. As a result, it has become an important strategy in modern target discovery.

1.1.2.1 Traditional target discovery

Historically, traditional target discovery, in which classical system approaches are usually used, predominated in the 1950s and 1960s [9]. To date, it is still relevant for many disease cases in which the related disease phenotypes can only be detected in the organism, such as some complex diseases responsible for phenotypic differences in genetically identical organisms [10]. In traditional routes, therapeutic target identification is just performed in two ways, either from randomly screening possible targets known or from clues given by traditional remedies [9]. Obviously, finding a good therapeutic target only by chance or experience makes target identification uncertain and inefficient. In addition, traditional target validation relies predominantly on experimental work in the laboratory by studying animal models *in vivo*. This is also a long-term work and needs continuous investment. Since the whole traditional process is expensive and time-consuming, construction of new modern target discovery system has become an urgent focus in drug research and development.

1.1.2.2 Modern target discovery

Since the late 1990s, as new molecular biology, especially genomic science, novel

genetic techniques, bioinformatics tools and *in silico* analysis have been integrated into drug research and development. Target discovery has gradually become a cross-disciplinary science, driven not only by biomedical science, pharmacology and chemistry but also by computational technology [4]. In modern target discovery, scientists mainly focus on specific molecular targets encoded by disease related essential genes of known sequence with novel, proven physiological function [5]. Instead of following traditional routes, in which an animal model of disease to yield a target is applied, current target discovery takes advantage of genomics data and bioinformatics techniques. For instance, the genomics information of therapeutic targets is analyzed by computational approaches from which useful information is generated, which is applied to improve the process of target discovery and ultimately to reduce the cost and time needed for drug discovery.

1.2 Overview of bioinformatics and its role in facilitating drug discovery

In 1988, the Human Genome organization (HUGO), an international organization of scientists involved in Human Genome Project, was founded. Just two years later, the Human Genome Project (HGP) was started. By referring to the international 13-year effort, this project was completed in 2003 successfully. All of the estimated 20,000-25,000 human genes were discovered and made accessible for further biological study. In addition, another goal of HGP, determination of the complete sequence of the 3 billion DNA subunits (bases in the human genome), is currently under way.

Undoubtedly, the completed human genome sequence, a grand achievement of HGP, provides tremendous opportunities for pharmaceutical research. Despite the

opportunities, there are many challenges, such as identifying the genes (protein-coding regions, structural RNAs, enzymatic RNAs and regulatory sequences) and other functional fragments (DNA-binding sites, promoters, termination sites, etc.) from the vast raw genome sequence, understanding physiological function of the proteins or peptides coded by those genes, correlating disease states to certain genes and figuring out the potential protein-protein interactions and their pathways in various situations including pathological conditions. So many promising challenges excite everyone in post-genomic era. However, the problem is that a vast amount of biological data has been generated by mapping human genome. Now, more than ever, scientists need sophisticated computational techniques to store, organize, manage, and analyze these genomic data, which belongs to a new discipline named bioinformatics.

1.2.1 Brief introduction to bioinformatics

Bioinformatics is an interdisciplinary research area that crosses between biology, computer science, physics, mathematics and statistics. As described by National Institutes of Health (NIH), bioinformatics is the "*research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data*" [11]. In brief, bioinformatics are used to "*address problems related to the storage, retrieval and analysis of information about biological structure, sequence and function*" [12]. Even if bioinformatics is a new term, some of the major events in bioinformatics occurred long before it was coined. Generally, the development of bioinformatics passed through several phases (Table 1-1).

Table 1-1: A brief history of bioinformatics

Phases	Important events	Year
Before 1950s	Gregory Mendel: “Genetic inheritance” theory	1865
1950s	Alfred Day Hershey & Martha Chase: Proving that DNA alone carries genetic information	1952
	Watson&Crick: Proposing the double helix model for DNA based x-ray data obtained by Franklin & Wilkins	1953
	Perutz's group: Developing heavy atom methods to solve the phase problem in protein crystallography	1954
	Frederick Sanger: analyzing the sequence of the first protein “bovine insulin”	1955
1960s	Sidney Brenner, Franšois Jacob, Matthew Meselson: identifying messenger RNA	1961
	Pauling: theory of molecular evolution	1962
	Margaret Dayhoff: Atlas of Protein Sequences	1965
	The ARPANET: created by linking computers at Standford and UCLA	1969
1970s	Needleman-Wunsch algorithm developed: sequence comparison	1970
	Paul Berg’s group: creating the first recombinant DNA molecule	1972
	The Brookhaven Protein DataBank is announced	1973
	Vint Cerf & Robert Khan: developing the concept of connecting networks of computers into an "internet" and developing the Transmission Control Protocol (TCP)	1974
	Bill Gates and Paul Allen: Microsoft Corporation (Popularization of personal computers from 1980s)	1975
	P.H.O'Farrel: Two-dimensional electrophoresis, where separation of proteins on SDS polyacrylamide gel is combined with separation according to isoelectric points.	1975
	Staden: DNA sequencing and software to analyze it	1977
1980s	Smith-Waterman algorithm developed	1981
	Doolittle: The concept of a sequence motif	1981
	GenBank	1982
	Phage lambda genome sequenced	1982
	Wilbur-Lipman algorithm developed: Sequence database searching algorithm	1983
	FASTP/FASTN: fast sequence similarity searching	1985
	The Human Genome Organization (HUGO) founded	1988
	National Center for Biotechnology Information (NCBI) created at NIH/NLM	1988
	EMBnet network for database distribution	1988
	Pearson and Lupman: The FASTA algorithm for sequence comparison	1988
	The genetics Computer Group (GCG) becomes a private company.	1989
1990s	The Human Genome Project: Mapping and sequencing the Human Genome	1990
	Altschul,et.al.: The BLAST program for fast sequence similarity searching	1990
	ESTs: expressed sequence tag sequencing	1991
	The research institute in Geneva (CERN): announcing the creation of the protocols which make -up the World Wide Web.	1991
	Sanger Centre, Hinxton, UK	1993
	EMBL European Bioinformatics Institute, Hinxton, UK	1994
	Netscape Communications Corporation founded and releases Naviagator, the commercial version of NCSA's Mozilla.	1994
	Attwood and Beck: The PRINTS database of protein motifs	1994
	First bacterial genomes completely sequenced: Haemophilus influenza genome (1.8 Mb) and Mycoplasma genitalium genome	1995
	Yeast genome completely sequenced: Saccharomyces cerevisiae (baker's yeast, 12.1 Mb)	1996
	Bairoch, et.al.: The prosite database	1996
	Affymetrix produces the first commercial DNA chips	1996
	PSI-BLAST	1997
	The genome for E.coli (4.7 Mbp) is published	1997
	deCode genetics publishes a paper that described the location of the FET1 gene, which is responsible for familial essential tremor, on chromosome 13 (Nature Genetics).	1997
	Worm (multicellular) genome completely sequenced	1998
	The genomes for Caenorhabitis elegans and baker's yeast are published	1998
	The Swiss Institute of Bioinformatics	1998

	First Human Chromosome 22 to be sequenced: Human Chromosome 22 completed	1999
	Fly genome completely sequenced	1999
	deCode genetics maps the gene linked to pre-eclampsia as a locus on chromosome 2p13.	1999
2000s	Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL: The large-scale organization of metabolic networks	2000
	Drosophila genome completed: D.melanogaster genome (180 Mb)	2000
	The genome for Pseudomonas aeruginosa (6.3 Mbp) is published	2000
	Draft Sequences of Human Chromosomes 5, 16, 19 Completed	2000
	Human Chromosome 21 Completed	2000
	The completion of a "working draft" DNA sequence of the human genome	2000
	The initial analysis of the working draft of the human genome sequence	2001
	Human Chromosome 20 Completed	2001
	Draft sequence of *Fugu rubripes*	2002
	Draft sequence of mouse genome	2002
	Human genome project completion (1990-2003)	2003
	Human Chromosome 14, Y, 7, 6 Completed	2003
	Human Chromosome 13, 19, 10, 9, 5 Completed	2004
	Human Gene count estimates changed from 20,000 to 25,000	2004
	…	…

The entries in Table 1-1 shows that the most significant progress in bioinformatics has been made remarkably in the last thirty years. With the invention of various sequence retrieval methods in 1970-80s, increasingly sophisticated sequence alignment algorithms were developed. In 1980s, scientists used computational tools to predict RNA secondary structure, and then began to predict protein secondary structure or 3D structure. In addition, the FASTA for sequence comparison and BLAST algorithm for fast sequence similarity searching were published in 1980-90s and they dramatically impelled the bioinformatics forward. Since 1990, many of new biotechnologies, including automatic sequencing, DNA chips, protein identification, mass spectrometers, etc., have been applied more and more widely. Numerous biological data have been produced continuously. Furthermore, large quantities of sequence data have also been generated by mapping and sequencing genomes of the human and other species. Table 1-2 gives some examples about the statistic data of the biological information space as of Feb 2005.

Table 1-2: The biological information space as of Feb 11th, 2005

Type of information	Number of entries/records
Nucleotides	44,575,745,176
Nucleotide records	49,127,925
Protein sequences	5,785,962
3D structures in PDB	28,905
BIND Interactions	134,886
Human Unigene Cluster	52,888
Completed Genome project	238
Different taxonomy Nodes	249,219
dbSNP records	18,883,945
RefSeq Genomic records	180,770
RefSeq RNA Records	352,275
RefSeq Protein Records	1,310,899
GenSAT images	98,680
GEO profiles	11,288,275
Homologene gene	38,137
PubChem compounds	897,246
PubMed records	15,382,675
PubMed Central records	341,602
OMIM records	16,521

Obviously, it is impossible to deal with these data manually. These huge data sets contain vital information for quantitative study of biology which is expected to revolutionize biology and medical research. On the one hand, the biology and medicine should not only be treated as specific biochemical technologies, but also as an information science. On the other hand, as more biological information becomes available and laboratory equipment becomes more automated, it is necessary to explore the use of computers and computational methods for facilitating experimental design, data analysis, simulation and prediction of biological phenomena and processes. Meanwhile, the use of computational methods can also improve the speed and efficacy, and reduce the cost of experimental studies.

At present, there are three primary public domain bioinformatics servers (Figure 1-2): National Center for Biotechnology Information (NCBI: http://www.ncbi.nlm.nih.gov/), European Bioinformatics Institute (EBI: http://www.ebi.ac.uk/), and Center for Information Biology (CBI: http://www.ddbj.nig.ac.jp/). Basically, each server

performs two parts of task. One is to develop and provide databases to efficiently store and manage data. The other is to invent useful bioinformatics algorithms and tools to analyze the data and generate new knowledge for biological and medical use. With the exponential growth of sequences, structures, and literature, bioinformatics databases are playing an increasingly crucial role in biological data management and knowledge discovery [13-16].

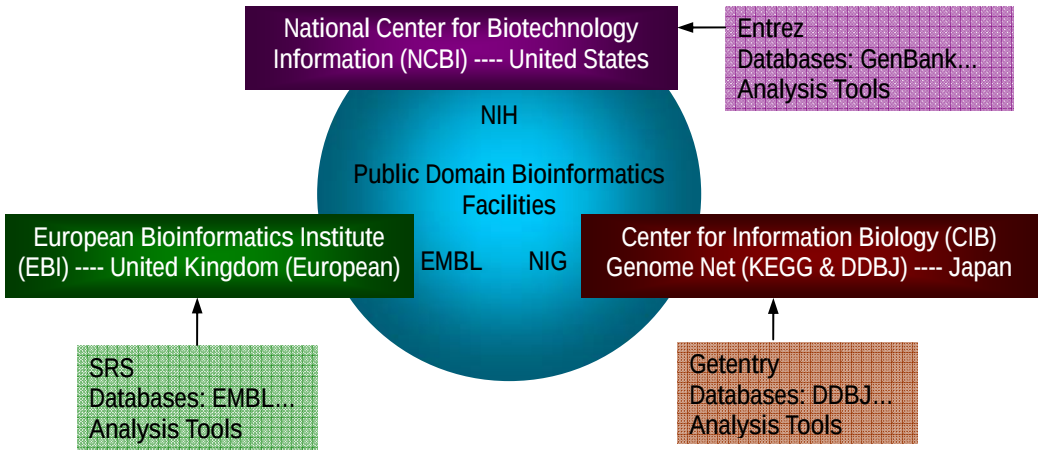


Figure 1-2: Primary public domain bioinformatics servers

1.2.2 Brief introduction to bioinformatics databases

Bioinformatics is the science of using information to understand biology [17]. The core of bioinformatics is the organization of information into databases. Bioinformatics database is an organized, integrated and shared collection of logically related bioinformatics data, which represent any meaningful objects and events in life science. These data can be transformed into information through data modeling, and thus provide useful knowledge to viewers.

Historically, the first bioinformatics database was established a few years after the first protein sequences became available. The first protein sequence (bovine insulin) was reported by Frederick Sanger at the end of 1950s [18]. It just consists of 51 residues. In 1963, the first tRNA molecule to be sequenced was the yeast alanine tRNA with 77 bases by Robert Holley and co-workers [19]. After that, Margaret Dayhoff gathered all the available sequence data to create the first bioinformatics database–Atlas of Protein Sequence and Structure [20-22], which is the origin of PIR-International Protein Sequence Database [23]. The Brookhaven National Laboratory's Protein Data Bank (PDB) followed in 1972 with a collection of the X-ray crystallographic protein structures [24] and it was considered as the first bioinformatics database, which stored and managed 3D protein structure data by using computational and mathematical techniques. In 1980s, due to the invention of automated DNA sequencing technology, the exponential growth of large quantities of DNA sequence data and associated knowledge came into being, and finally became the significant driving force for the development of bioinformatics database. The biological data and knowledge needs to be stored in a computationally amenable form, which can be shared by the bioinformatics community for both humans and computers. The Swiss-Prot, an important annotated protein sequence database, was established in 1986 and maintained collaboratively, since 1987, by the group of Amos Bairoch first at the Department of Medical Biochemistry of the University of Geneva and now at the Swiss Institute of Bioinformatics (SIB) and the European Molecular Biology Laboratory (EMBL) Data Library [25].

Subsequently, a huge variety of diverse bioinformatics databases have been growing either in the public domain or commercial third parties. Figure 1-3 summarizes the development trend of Molecular Biology Database (MBD) collected by Nucleic Acids

Research from 1999 to 2005. In comparison with 202 MBDs in 1999, the total number of MBD in 2005 was 719. It was about 3.5 times than that of in 1999 and the increase rate reached 256%. The data indicates that the development of MBD is likely to have a continuous upward tendency in the following years. According to the latest database issue of Nucleic Acids Research (NAR) [26], to date, more than 700 different databases covering diverse areas of biological research, including sequence, structure, genetics, genomes, proteomics, intermolecular interactions, pathways, diseases, microarray data and other gene expression information.

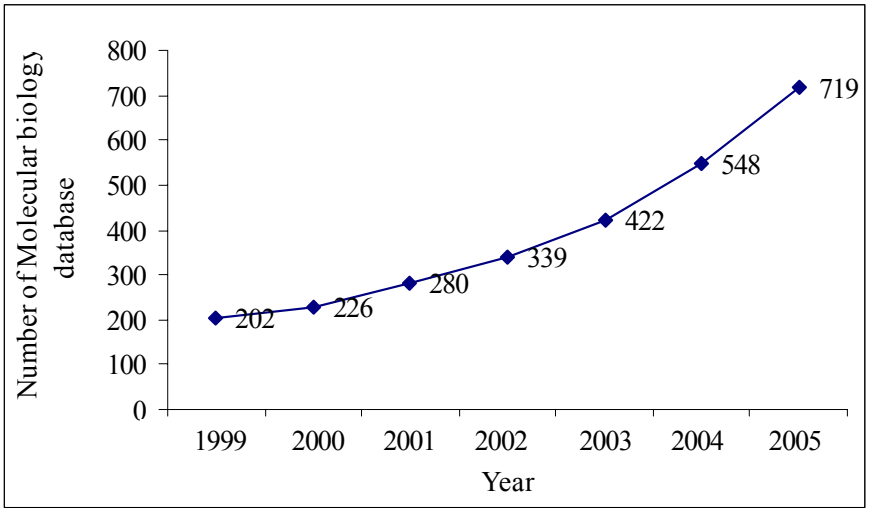


Figure 1-3: Molecular biology database collection in NAR (1999~2005) [26]

On the basis of the scope of databases, a biological database can be grouped into three categories [27]: general biological databases, which store the raw data of DNA/protein sequence, structure, biological and medical literature; derived databases, whose data are derived from the general biological databases, however, contain novel information; and subject-specialized databases, which collect individual, specialized information for the communities of particular interests. Besides the diverse area

covered by different kinds of bioinformatics databases, the application of biological databases is broad, both in the academia and industries. In our research, three pharmainformatics* databases: Therapeutic Target Database (TTD), Therapeutically Relevant Multiple Pathways (TRMP) database, and ADME-associated Proteins (ADME-AP) database, which are specific bioinformatics databases applied in biomedical science, are developed or updated and their applications in drug discovery are also discussed.

1.3 The need for computational study of therapeutic targets and ADME-associated proteins

Usually, general bioinformatics databases are useful for studying general genetics, proteomics, and structural problems, but they are not designed for providing information of proteins relevant to drug discovery. However, for many pharmaceutical researchers, sometimes they are more interested in specific knowledge in their research area. For instance, which kinds of proteins could be considered as potential therapeutic targets? Is there any specific databases providing information about drug absorption, distribution, metabolism and excretion associated proteins (ADME-APs) or disease relevant therapeutic pathways? Obviously, there is a need to develop special pharmainformatics databases dedicated to drug studies.

1.3.1 The need for development of pharmainformatics databases

1.3.1.1 Therapeutic target database

Researches have shown that the paradigm of modern drug discovery is built on the

*Pharmainformatics is the integration of Bioinformatics & Cheminformatics.

search of drug leads against a pre-selected therapeutic target, which is followed by testing of the derived drug candidates [9, 28, 29]. So far, continuous efforts in target discovery have been made in the exploration of the targets of highly successful drugs, and identification of new targets [1, 6, 9, 28, 29]. Furthermore, the search for new targets and the study of existing targets are facilitated by rapid advances in protein structures [30], proteomics [31], genomics [32, 33], and molecular mechanism of diseases [34, 35]. Currently, scientists mainly use these technologies for finding clues to new target identification and for probing the molecular mechanisms of drug action, adverse drug reactions, and pharmacogenetic implication of variations. Undoubtedly, the advances and development of target identification and validation technologies will lead to the discovery of a growing number of new and novel targets. Drews and Ryser [36] reported that there were ~500 targets underlying current drug therapy undertaken in 1996, 120 of which have been reported to be the identifiable targets of currently marketed drugs [37]. In the subsequent few years, Drews [9] and other researchers [37] made some analysis based on the ~500 targets, including distribution of target biochemical class and estimation of possible target number of human species.

Due to increasing exploration of disease-specific protein subtypes of existing targets and new information about previously unknown or un-reported targets of existing drugs and investigational agents, the number of successful and research targets should significantly increase. However, there is no updated list available on therapeutic target. Up to date, almost all review articles about therapeutic targets are based on the targets list reported by Drews and Ryser in 1997 [36]. Thus, it is necessary to develop a specific pharmainformatics database for providing timely information of the known and newly proposed therapeutic protein and nucleic acid targets described in the established publications.

1.3.1.2 Therapeutically relevant multiple pathways database

Proteins and nucleic acids that play key roles in disease processes have been explored as therapeutic targets for drug development [9, 29]. Knowledge of these therapeutically relevant proteins and nucleic acids has facilitated modern drug discovery by providing platforms for drug screening against a pre-selected target [9]. It has also contributed to the study of the molecular mechanism of drug actions, discovery of new therapeutic targets, and development of drug design tools [37, 38]. Information about non-target proteins and natural small molecules involved in these pathways is also useful in the search of new therapeutic targets and in understanding how therapeutic targets interact with other molecules to perform specific tasks.

A number of web-based resources of therapeutically-targeted proteins and nucleic acids are available [39, 40], which provide useful information about the targets of drugs and investigational agents. While information about multiple pathways can be obtained from the existing individual pathway databases, interfaces that integrate multiple pathway maps may provide more convenient platforms for facilitating the analysis of the collective effects of different proteins in separate pathways. Moreover, the existing databases either include significantly more number of pathways than therapeutic ones or they are intended for specific types of pathways that do not cover all of the therapeutic ones, which can sometimes make the search of therapeutically relevant constituents less convenient. It is thus desirable to have a database specifically designed as a convenient source of information about therapeutically relevant multiple pathways to complement existing databases.

In addition, crosstalk between proteins of different pathways is common phenomena

and these often have therapeutic implications [41-48]. Cocktail drug combination therapies directed at multiple targets have been explored for a number of diseases including AIDS [49], cancer [50, 51], Alzheimer disease [52], amyotrophic lateral sclerosis [53], and dyslipidemia [54]. These prompted interest for more extensive exploration of synergistic targeting of multiple targets in drug discovery [55]. Potentially harmful interactions arising from multiple targeting are also closely watched and studied [56]. Effective drugs with robust phenotypic effects are known to simultaneously affect many proteins in different pathways [55]. For instance, in addition to interacting with its main target protein cyclooxygenase, anti-inflammatory drug aspirin is known to affect NF-kappa B pathway and other connected cellular targets that normally contribute to perpetuate the inflammatory state [57, 58]. Therefore, it is necessary for us to develop a therapeutically relevant multiple pathway database to facilitate the analysis of the potential implications of multiple target-based therapies and for mechanistic study of drug effects.

1.3.1.3 ADME-associated protein database

Inter-individual variations in drug response are well recognized and these variations are frequently associated with polymorphisms in the proteins involved in ADME-APs [59-61] as well as those in therapeutic targets and drug adverse reaction (ADR) related proteins [62, 63]. Pharmacogenetic study with respect to these proteins and their regulatory sites is important for the understanding of molecular mechanism of drug responses and for the development of personalized medicines and optimal dosages for individuals [59, 64-67]. Nearly 100,000 putative single-nucleotide polymorphisms (SNP) have been identified in the coding regions of human genome [68, 69], some of which have been linked to substantial changes in drug response and

used for the analysis of individual variations to drug therapies [59-61, 70, 71]. Sequence polymorphism is only one of the factors for variations of drug responses. Other factors include altered methylation of genes, differential splicing of mRNAs, and differences in post-transcriptional processing of proteins such as protein folding, glycosylation, turnover and trafficking [63]. Thus, in addition to polymorphisms, there is a need to investigate the effects of transcriptional and post-transcriptional profiles of ADME-APs as well as therapeutic targets and ADR-related proteins.

Knowledge of ADME-APs is not only useful for the identification of pharmacogenetic polymorphisms, but also enables a focused study of polymorphisms, transcriptional and post-transcriptional profiles that alter the function or drug affinity of the target [66]. However, for most drugs, not all of the ADME-APs responsible for their metabolism and disposition are known. As a result, in many cases, molecular study of the pharmacokinetic aspect of pharmacogenetics may need to be based on the study of ADME-APs to find out which proteins are responsible for the metabolism and disposition of a particular drug, and how the polymorphisms, transcriptional and post-transcriptional profiles of these proteins determine the individual variations to that drug.

Up to date, a number of freely-accessible internet databases have appeared which provide useful information about drug ADME-APs as well as therapeutic and drug toxicity targets [40, 72, 73]. Although they provide comprehensive knowledge about ADME-APs, most of these databases are just for specific groups of ADME-APs. Moreover, information about reported polymorphisms and pharmacogenetic effects of ADME-APs is seldom mentioned. Thus, it is desirable to complete the ADME-AP database, which can provide basic biological information about ADME-APs and also

reported pharmacogenetic relevant information. Such information contained in ADME-AP database can reach a meaningful level for facilitating biomedical research. As a result, ADME-AP database may serve as a useful resource for comprehensively understanding pharmacogenetics.

1.3.2 In silico mining of therapeutic targets

As described in previous section, it is important for the drug discovery communities to explore the current targets in the literature, which is a good way to find new therapeutics and more effective treatment options. According to computational analysis of therapeutic target, at present, the major concern of many researchers is about the estimation of the total number of human targets [37, 74, 75]. Hophins and Groom [37] statistically analyzed the disease genes and related proteins and suggested that the total number of the estimated potential targets in the human genome ranges from 600 to 1,500. Moreover, by investigating the yeast genome, they found that antifungal targets constitute 2-5% of the whole genome in yeast. Assuming a similar percentage of targets in disease-related microbial genomes, the number of potential targets in disease-related microbial genomes can be roughly estimated as >1,000. Miller and Hazuda [74] pointed out that a typical viral genome contains 1-4 targets, which gives a crude estimate of >100 potential targets in disease-related viral genomes. According to this, the total number of distinct targets is likely to be within range of 1,700~3,000. In another research done by Wen and Lin [75] in 2003, a similar estimation was obtained.

One way to assess the opportunities available for pharmaceutical industry is to begin by studying human genome and searching those genes relevant to drugs and diseases.

However, in the human genome, there are up to 22,300 or so genes currently [76]. Mining useful information from such large data set may be an extremely tough work for pharmaceutical scientists. As a result, knowledge discovery from current known targets is very important. Some meaningful work, such as generating some common rules describing targets and druggable proteins prediction by computational approach, would be done for facilitating to cut down the range of genes needed to be studied and speeding up the target discovery.

1.4 Objective and scope of the thesis

Generally, the research was planned to complete two main aspects of work. The first aspect was concerned development of pharmainformatics databases; the second aspect of this research involved *in silico* mining the therapeutic targets and ADME-AP data by using bioinformatics tools. Therefore,

- The first objective was to launch the new version of TTD, which was first published in 2002 [39]. Accordingly, we optimized the database structure, completed data validation and updating, and provided some more important information on the current therapeutic targets. In addition, the web interface was improved to be more user-friendly and the query methods were enhanced to support complex searching.
- The second objective was to develop a TRMP database, which was to give information about inter-related multiple pathways of a number of diseases and physiological processes.
- The third objective was to update the database of ADME-APs, which was first launched in 2002 [73]. Especially, information about reported polymorphisms and pharmacogenetic effects were integrated into the ADME-AP database.

Furthermore, we also statistically analyzed reported polymorphisms and drugs with altered responses linked to protein.

As we know, target discovery is highly dependent upon a correct understanding of the information generated from lots of therapeutic targets and drug ADME-APs. Therefore, another significant objective of this research was to carry out computational analysis of therapeutic targets and drug ADME-APs data. Regarding the pharmacogenetic information of ADME-APs, the purpose of this part of study was to discuss how to use the relevant information of ADME-APs for facilitating pharmacogenetics research. Particularly, we studied the feasibility of predicting pharmacogenetic response to drugs. The other important part of the study aimed to provide an overview of the progress in the exploration of therapeutic targets and to investigate the characteristics of these targets for finding some useful clues which could facilitate the search of new targets. Basically, this objective was planned to be achieved in two steps.

- Firstly, based on the primary information provided by TTD, secondary information could be retrieved from other general biological databases, including the sequence, structure, family representation, pathway association, tissue distribution, genome location features, etc. Subsequently, the main characteristics of all successful and research targets could be generated by taking advantage of the secondary information.
- Secondly, we studied the possible rules for guiding the search of druggable proteins and discussed the feasibility of using a statistical learning method, Support Vector Machines (SVMs), for predicting druggable proteins directly from their sequences.

At present, TTD may be the world's first public comprehensive database for

therapeutic targets. It may serve as an essential data resource for target research and development in drug discovery area. Results of this study may suggest several common rules for therapeutic targets. The clues based on the knowledge of existing targets are useful for new target identification. It is also important for the molecular dissection of the mechanism of action of drugs, the prediction of features that guide new drug design, and the development of tools for these tasks. Moreover, this research may provide an alternative solution rather than BLAST to predict druggable proteins. Principally, analysis of these targets may provide useful information about general trends, current focuses of research, areas of successes and difficulties in the exploration of therapeutic targets for the discovery of drugs against specific diseases.

About the scope of the thesis, therapeutic target data used here depend mainly on the collections in the TTD, and unavoidably we may miss some therapeutic targets, which have not been collected by TTD yet. Furthermore, computational analysis of therapeutic targets focuses mainly on the ones whose annotations are adequate. In addition, this thesis considers the problem of data classification in high dimensional space. Generally, there are two different strategies for protein data classification. One is structure based approach, including molecular dynamics, molecular mechanics, and geometry methods. The other is sequence based approach, including decision tree, artificial neural networks, and SVMs. In this thesis, we made use of only SVMs to predict druggable proteins.

1.5 Layout of the thesis

As introduced above, the problems addressed in this thesis have been focused on pharmainformatics database development, computational study of therapeutic targets and ADME-APs. In the coming chapters, a brief introduction to the methods used in

this study was discussed, and this included the strategy of database development and basic theory for SVMs, a computational methods used for data analysis. In chapter 3, by using the similar database developing strategy, two pharmainformatics databases were constructed and presented. Due to similar developing strategy, the detail about how the ADME-AP database was constructed was omitted and integrated its brief introduction into the computational analysis section.

Moreover, applications based on the TTD were also carried out to facilitate target discovery. In chapter 4, on the basis of therapeutic target data, the progress of target exploration was summarized and the characteristics of the currently explored targets were analyzed. Subsequently, chapter 5 described how to use SVMs to *in silico* predict druggable proteins. Chapter 4 and 5 would be considered as the most important chapters in this study. In chapter 6, ADME-AP database was updated and a discussion on how to use the ADME-APs data to facilitate pharmacogenetics research was presented. Finally, conclusion was made in the final chapter.

2 Methodology

2.1 Strategy of pharmainformatics database development

Even though pharmainformatics databases have different sorts of applications in scientific research, the strategy of database development follows similar basic ideas. Thus, this chapter describes general strategy of knowledge-based pharmainformatics database development. The similar strategies have been extended to the construction of TTD, TRMP database, and ADME-AP database, which are discussed in later. Generally, the development of a database is a complicated and time-consuming process, including preliminary planning, information collection, database construction, and database access and representation. Here a stage by stage development of the database is discussed.

2.1.1 Preliminary plan of the pharmainformatics database

Making a preliminary plan before the start of the database development may help to focus on relevant points and not gather unnecessary information. In this stage, the objective and content of the database should be seriously considered and determined.

As described in previous chapter, target discovery plays a very important role in drug research and development. It is essential for biomedical researcher to know more about therapeutic targets, therapeutic relevant pathways, and ADME-APs. However, up to date, there is no similar pharmainformatics database that provides this specific information. Thus, the development of such a kind of knowledge-based pharmainformatics databases will be meaningful. To conclude, the database will meet the expectations of those corresponding researchers, afford them what they want, and

help them find further information they need. After preliminary consideration of the whole database, a detail description of the database development will be presented.

2.1.2 Collection of pharmainformatics database information

Normally, a knowledge-based pharmainformatics database is supposed to provide enough domain knowledge around a specific subject in pharmacology. For instance, therapeutic target database will let users know about some biological information for specific therapeutic target, relevant disease conditions, and drugs/ligands corresponding to this target, and so on. Thus, for every pharmainformatics database entry, there are several different knowledge domains. Some of them provide basic introduction to entries themselves, and some others give information derived from entries or relevant to entries.

The information mentioned above can be selected from a comprehensive search of available literatures including pharmacology textbooks, review articles and a large number of other publications. With respect to different type of information, we use different collecting methods. The subject of database, such as therapeutic target, therapeutic pathways, and ADME-APs, is the primary focus. Thus, in the first step, we collect reliable subject information. At present, no ready index or library is available and almost all the relevant information is scattered in various biological and medical literatures. Therefore, literature information extraction is the only feasible way to collect the essential biological and medical information. It is generally agreed that literatures are typically unstructured data source. In addition, the names of the subject, which may be in some synonymous terms, various abbreviations, or totally different expression, are difficult to be recognized by automatic language processing.

A fully automated literature information extraction system, thus, cannot be invented to gather useful information from literature efficiently.

In this study, automatic text mining methods with manual reading process was combined. Simple automated text retrieval programs developed in PERL were used to screen the literature that contained the key word related to searching the subject in local Medline abstract packages [77]. Then, the useful subject information was picked up manually from these matched Medline abstract. If necessary, the full literature was referred to facilitate information searching. Meanwhile, in many cases, the relevant information about the same subject could also be found in the same literature. Thus, in the first step, not only subject but also relevant information could be obtained and recorded. In the second step, detail biological information of subject was automatically selected from some relevant general or specific biological databases, such as SwissProt, GeneCard, etc., by text mining programs. Likewise, some other information derived from the subject was also extracted from the corresponding databases in the same way. After information collection, a consideration how to store, organize and manage the data by using database techniques was discussed. In the next section, the database construction is described.

2.1.3 Organization and structure of pharmainformatics database

A good database system enables the user create, store, organize, and manipulate data efficiently. By integrating databases and web sites, users and clients can open up possibilities for data access and dynamic web content. An integrated information system of our pharmainformatics database is constructed according to some

standardization strategies as follows:

- Establishment of standardized data format and appropriate data model
- Database structure construction
- Development of Database Management System (DBMS)

Since the original data information collected in previous section is independent, the first major activity of a database construction process includes creation of digital files from these information fragments and construction of an appropriate data model.

2.1.3.1 The data model

The data model is an integrated collection of concepts for describing data, relationships between data, and constraints on the data [78]. An organized collection of data and relationships among data items is the database. Over the years there have been several different basic ways of constructing databases, among which have been listed as follow:

- The flat file model
- The hierarchical model
- The network model
- The relational model
- The object-oriented model

The flat-file model is the simplest data model, which is essentially a plain table of data. Each item in the flat file, called a record, corresponds to a single, complete data entry. A record is made up by data elements, which is the basic building block of all data models, not just flat files. The flat-file data model is relatively simple to use; however, it is inefficient for large databases.

The hierarchical data model organizes data in a tree structure (Figure 2-1). It has been used in many well-known database management systems. The basic idea of hierarchical systems is to organize data into different groups, which can be divided into different subgroups. In a subgroup, there may be some sub-subgroups, among which the sub-subgroups may have sub-sub-subgroups, and so on. That is to say, there is a hierarchy of parent and child data segments. In a hierarchical database the parent-child relationship is one to many. The hierarchical data model will be convenient to use and run very efficiently only if the nature of the application remains strictly hierarchical. Actually, in real world application, few database management problems remain strictly hierarchical. It is the major failing of this kind of data model.

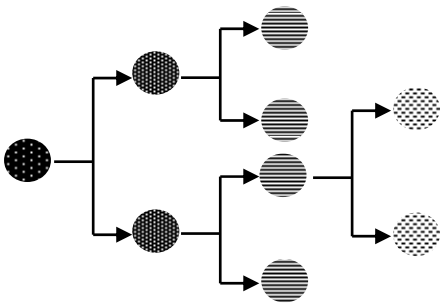


Figure 2-1: The Hierarchical Data Model

In most cases, the relationships of data would be arbitrarily complex (Figure 2-2). The circles in triangle (left) represent “children” and the circles in square (right) represent “parents”. The broken line links the children to their parents. In this model, some data were more naturally modeled with multiple parents per child. So, the network model permitted the modeling of many-to-many relationships in data. This model, thus, can handle varied and complex information while remaining reasonably efficient. Even so, the biggest problem with the network data model is that databases can get excessively complicated.

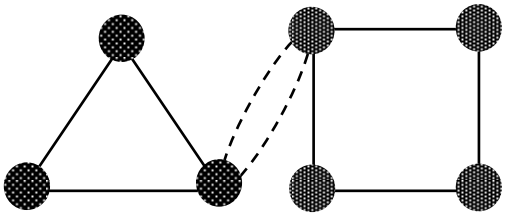


Figure 2-2: The Network Data Model

The relational model was formally introduced by E. F. Codd in 1970 and has been extensively used in biological database development (Figure 2-3). The model is a much more versatile form of database. On the basis of this kind of data model, a novel system named relational database management system is established. A relational database allows the definition of data structures, storage and retrieval operations and integrity constraints. In such a database the data and relations between them are organized in tables.

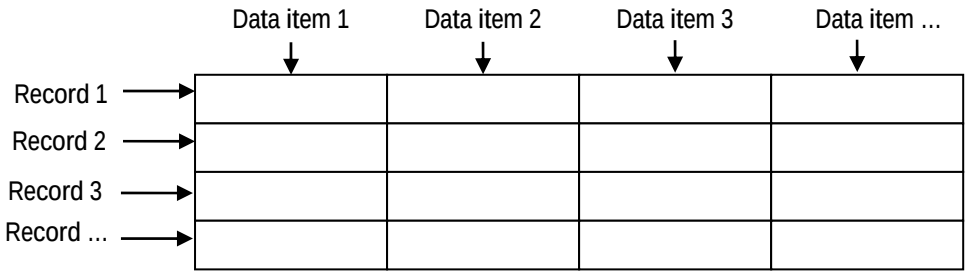


Figure 2-3: The Relational Data Model

Every relational database consists of multiple tables of data, related to one another by columns that are common among them. Every table is a collection of records and each record in a table contains the same fields. Therefore, if the database is relational, we can have different tables for different information. And the common columns, such as entry ID, can be used to relate the different tables. Relational database is the

predominant form of database in use today, especially in biological research field. It is the type which has been used in this research work.

The object-oriented database (OODB) paradigm is "*the combination of object-oriented programming language (OOPL) systems and persistent systems*" [79]. "The power of the OODB comes from the seamless treatment of both persistent data, as found in databases, and transient data, as found in executing programs" [79]. The database functionality is added to object programming languages in object database management systems, which extend the semantics of the C++, Smalltalk and Java object programming languages to provide full-featured database programming capability. The combination of the application and database development with a data model and language environment is a major advantage of the object-oriented model. As a result, applications require less code, use more natural data modeling, and code bases are easier to maintain.

2.1.3.2 Relational pharmainformatics database structure construction

The relational model has been used in our pharmainformatics databases. It represents relevant data in the form of two-dimension tables. Each table represents relevant information collected. The two-dimensional tables for the relational database include entry ID list table (Table 2-1), main information table (Table 2-2), which contains a record for the basic information of each entry, data type table (Table 2-3), which demonstrates the meaning represented by different number, and reference information table (Table 2-4), which gives the general reference information following by different PubMed ID in Medline [77].

Table 2-1: Entry ID list table

Entry ID	Entry name
…	…

Table 2-2: Main information table

Entry ID	Data type ID	Data content	Reference ID
…	…	…	…

Table 2-3: Data type table

Data type ID	Data type
…	…

Table 2-4: Reference information table

Reference ID	Reference
…	…

Figure 2-4 is the general logical view of database we developed. It shows the organization of relevant data into relational tables. In these tables, certain fields may be designated as keys, by which the separated tables can be linked together for facilitating to search specific values of that field. Commonly, in relational table, the key can be divided into two types. One is primary key, which uniquely identifies each record in the table. Here it is a normal attribute that is guaranteed to be unique, such as entry ID in Table 2-1 with no more than one record per entry. The other is foreign key, which is a field in a relational table that matches the primary key column of another table. The foreign key can be used to cross-reference tables. For example, in tables of our databases, there are two foreign keys: Data type ID and Reference ID. According to Figure 2-4, a connection between a pair of tables is established by using a foreign key. The two foreign keys make three tables relevant. Generally, there are three basic types of relationships of related table: one-to-one, one-to-many, and many-to-many. In our case, these databases belong to one to many relationships.

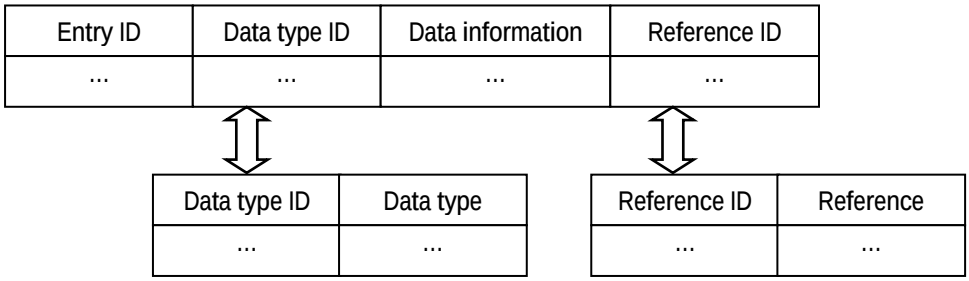


Figure 2-4: Logical view of the database

2.1.3.3 Development of Database Management System

By using relational database software (e.g. Oracle, Microsoft SQL Server) or even personal database systems (e.g. Access, Fox), the relational database can be organized and managed effectively. This kind of data storage and retrieval system is called Database Management System (DBMS). An Oracle 9i DBMS is used to define, create, maintain and provide controlled access to our pharmainformatics databases and the repository. All entry data from the related tables described in previous section are brought together for user display and output using SQL queries.

2.2 Computational methods for the prediction of druggable proteins

Besides pharmainformatics database development, another significant work of this study was focused on computational analysis of therapeutic targets and ADME-APs. A well known machine learning method, SVMs, has been used. Thus, in this section, a general introduction to SVMs is discussed.

2.2.1 Introduction to machine learning

Learning is the most typical way in which humans "*acquire knowledge,*

comprehension or mastery of (a subject) or skill through experience or study" (Oxford English Dictionary, 1989). Human beings, according to evolutionary theory, have developed big brains that enable them to observe, interpret, and understand the complex world. As a result, in the past few centuries, human learning has been used in traditional routes. During the process of human learning, human beings take advantage of their intuition to characterize and represent the data. Certainly, there are many difficulties, misunderstandings or low efficiencies. Moreover, humans do not enough have comprehensive knowledge to enable them to analyze each phenomenon in a reasonable way.

With the invention of computers, it is possible to combine human learning with computational technology. Computers have been designed to simulated human's brains to learn about multifarious data coming from various research fields. Furthermore, computers are capable of doing "automatic programming". That is to say, a computer program can learn from experience with respect to some class of data, knowledge, or experimentation. Such kinds of things are called machine learning, whose common tasks include concept learning for prediction, data clustering, and association rule mining. In the information age, particularly in biological research area, a huge volume of genomic information has been generated increasingly resulting from large scale genome sequencing projects. It is obviously beyond the capability of human beings to effectively explore the information without the aid of intelligent computer technology. One way to match the need for analysis and interpreting huge information of biology systems is to utilize the artificial intelligence which aims to mimic how the brain works. Statistical machine learning was designed for computers to learn from observations, and subsequently the learned knowledge could be used in decision making process for the new discovery. It has a long history and has been

successfully applied to solve many biological problems in real life.

Particularly in solving biology information-intensive problems, many statistical learning methods such as discretized naïve Bayes [80], C4.5 decision trees [81], and instance-based leaning [82], neural networks [83] and SVMs [82, 84-89], have shown the potential to predict the unknown characteristic from the observed knowledge. In this work, we are going to focus on one of the machine learning methods, support vector machines, which is one of the most important machine learning methods and are regarded as a main example of "kernel methods".

2.2.2 Introduction to support vector machines

SVMs introduced by Vapnik [90] in 1979, are a set of related supervised learning methods used as robust tools for classification and regression in noisy and complex domains. Since it was further explained by Vapnik [91] in 1995 and more theoretically elaborated by Burges [92] in 1998, increasing effort have been directed in both the theory study and application in real life problems, such as text categorization [93-95], tone recognition [96], image detection [97-100]; flood stage forecasting [101]; cancer diagnosis [102-104], microarray gene expression data analysis [105], protein secondary structure prediction [106, 107], identification of protein-protein interaction [108] and many other classification problems. Basically the tasks of SVMs can be described into two ways: i) extraction of valuable information from datasets and ii) construction of fast classification algorithms for massive data, which it is based on the structural risk minimization (SRM) principle from statistical learning theory [91].

During the process of classification, SVMs construct a hyperplane which could

separate two groups of examples with a maximum margin (Figure 2-5). New data is subsequently tested by labeling their comparative position to the separation hyperplane, where the two sides of separation hyperplane represent different classes. In SVMs theory, this separation hyperplane has been proved to be unique if the feature space is fixed. Real life problems are not always straightforward in a linear form; the SVMs extrapolated the same idea to the non-linear problem domains by introducing kernel mappings which are able to project the input data from input space into a high-dimensional feature space in which the training examples can be linearly separated. In following section, we will have a closer look on the theory of SVMs.

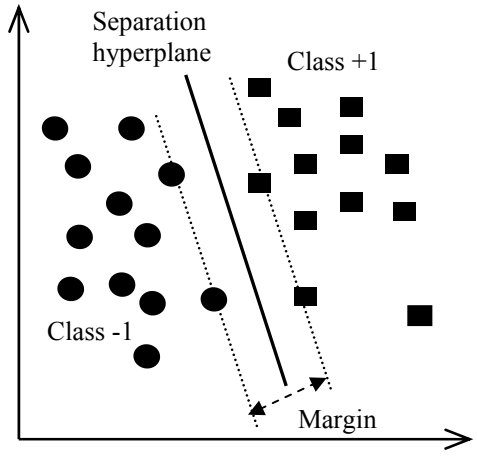


Figure 2-5: Separating hyperplanes in SVMs (the circular dots and square dots represent samples of class -1 and class +1, respectively.)

2.2.3 The theory and algorithms of support vector machines

The mathematical foundation of SVMs is based on the structural risk minimization principle from statistical learning theory [92]. The structural risk expresses an upper bound on the generalization error. There are two types of SVMs algorithms, linear and nonlinear SVMs. The basic idea of SVMs, both linear and nonlinear SVMs, for

pattern reorganization is to construct an optimal separation hyperplane (OSH) separating two different classes of feature vectors with a maximum margin [91, 109].

2.2.3.1 Linear case

The training data of two separable classes with n samples can be represented by:

$$(x_1, y_2), (x_2, y_2), ... (x_l, y_l), \quad i = 1,2,...,l \tag{2-1}$$

The aim of SVMs is to establish a function map from the training examples (x_i, y_i) for discriminant patterns:

$$f : R^N \rightarrow \{\pm 1\} \tag{2-2}$$

where x_i is the N-dimensional feature vectors and $x_i \in R^N$ is an N dimensional space, y_i is the corresponding class label and $y_i \in \{-1,+1\}$ is the class index. And (x_i, y_i) is under the same probability distribution $p(x, y)$, $(x_1, y_2), (x_2, y_2), ... (x_l, y_l) \in R^N X \{\pm 1\}$. The function f is considered to be well generalized so that the training dataset (x_i, y_i), $i = 1,2,...,l$, satisfy $f(x_i) = y_i$.

As indicated in Figure 2-6, the hyperplane in SVMs is constructed by finding a weight vector w and bias b that minimizes $\|w\|^2$ which satisfies the following conditions:

$$w \cdot x_i + b \geq +1, \text{ for } y_i = +1 \text{ (positive class)} \tag{2-3}$$

or

$$w \cdot x_i + b \leq -1 \text{ , for } y_i = -1 \text{ (negative class)} \tag{2-4}$$

Here w is a vector normal to the hyperplane, $|b|/\|w\|$ is the perpendicular distance

from the hyperplane to the origin and $\|w\|^2$ is the Euclidean norm of w. After the determination of w and b, a given vector x can be classified by using the decision function $sign[(w \cdot x) + b]$, a positive or negative value indicates that the vector x belongs to the positive or negative class respectively.

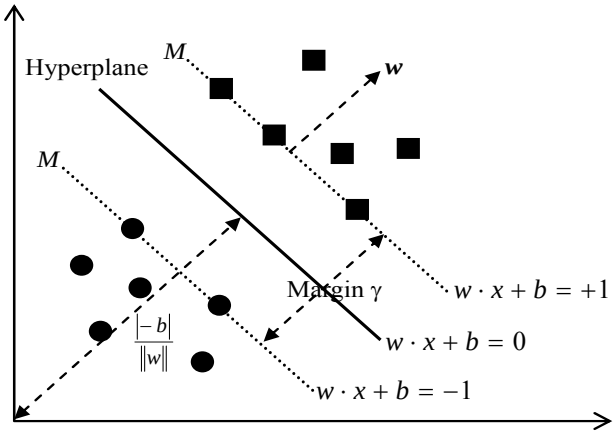


Figure 2-6: Construction of hyperplane in linear SVMs (the circular dots and square dots represent samples of class -1 and class +1, respectively.)

The hyperplane constructed based on the feature space usually works for the problems which can be linearly solved, where the above implementation of SVMs is called linear SVMs (Figure 2-6). However, many real-world problems are much more complicated and usually cannot be solved in a linear form, for instance the protein function classification [85, 110], hand-writing identification [93-95] and therapeutic regimen diagnosis [102-104].

2.2.3.2 Nonlinear case

The capability of SVMs to solve non-linear separable problems is extended by projecting the input data from feature space to higher dimension space through kernel function $K(x_i, x_j)$. The use of kernel functions for the feature transformation is to

convert the non-liner problem in lower feature dimension to the higher dimension where the problem becomes linearly solvable. An example of a kernel function is the Gaussian kernel, which has been extensively used in a number of protein classification studies [84, 86, 92, 106-108, 111]:

$$K(x_i, x_j) = e^{-\|x_j - x_i\|^2 / 2\sigma^2} \tag{2-5}$$

The same SVMs procedure is then applied to the feature vectors in this feature space and the decision function for their classification is given by:

$$f(x) = sign(\sum_{i=1}^{l} \alpha_i^0 y_i K(x, x_i) + b) \tag{2-6}$$

Where the coefficients α_i^0 and b are determined by maximizing the following Langrangian expression:

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{2-7}$$

Under conditions: $\alpha_i \geq 0$ and $\sum_{i=1}^{l} \alpha_i y_i = 0$. A positive or negative value from Eq. (2-6) indicates that the vector x belongs to the positive or negative group respectively.

2.2.4 Model evaluation of support vector machines

As in the case of all discriminative methods [112-114], the performance of SVMs classification can be measured by the quantity of true positive *TP* (correctly predicted members), false negative *FN* (members incorrectly predicted as non-members), true negative *TN* (correctly predicted non-members), and false positive *FP* (non-members incorrectly predicted as members). Because the number

of members and non-members is imbalanced, two unique quantities [115], sensitivity and specificity, are used to measure the accuracy for the members and non-members of a specific class.

Sensitivity: $Q_p = TP/(TP + FN)$ (2-8)

Specificity: $Q_n = TN/(TN + FP)$ (2-9)

The overall accuracy is:

$$Q = (TP + TN)/(TP + FN + TN + FP) \quad (2\text{-}10)$$

Here the positive prediction accuracy Q_p is for proteins that have a specific property; the negative prediction accuracy Q_n is for proteins without that property. In some cases, Q, Q_p, and Q_n are insufficient to provide a complete assessment of the performance of a discriminative method [113, 116]. Thus the Matthews correlation coefficient:

$$MCC = (TP \times TN - FP \times FN)/\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)} \quad (2\text{-}11)$$

is used to evaluate the randomness of the prediction, where $MCC \in [-1,1]$.

3 Therapeutic target database and therapeutically relevant multiple-pathways database development

As mentioned in chapter 1, this thesis consists of two parts of work. One is about databases development, which is introduced in this chapter; the other is about computational analysis of therapeutic targets and ADME-APs, which will be discussed in following chapters. With respect to databases development, the TTD was reconstructed and updated. A new TRMP database was developed. In addition, another pharmainformatics database, ADME-AP database, was updated. Because the database structure has no major modification and its development followed similar strategy, the detail of ADME-AP database development was omitted and integrated its brief introduction into the chapter on computational analysis of ADME-APs.

3.1 Therapeutic target database development

3.1.1 Preliminary plan of therapeutic target database

It is widely known, target discovery is one of the key processes in drug discovery. Furthermore, knowledge about known or investigated therapeutic targets is essential for target discovery and it may facilitate biomedical researcher to find more potential targets. However, up to date, there is no similar publicly accessible web-based database providing specific information about therapeutic target. Thus, it is meaningful to develop such kind of target information database, which can provide timely information of the known and newly proposed therapeutic protein and nucleic acid targets described in the established publications. As a repository of specific pharmainformatics database, it is helpful in catering for the need and interest of the

biological and medical research communities. Therefore, the relevant information of targets, such as drug/ligands information, corresponding disease conditions, is essential. For facilitating to understand these targets in further, extra information, such as cross links to other databases, is also included to make TTD comprehensive and applicable.

3.1.2 Collection of therapeutic target information

As indicated in previous section, three important information communities should be included in TTD: therapeutic target information, information about targets binding ligands and therapeutic effects. Detail information items are given here:

- Therapeutic target information: Target name, Synonyms of target, Type of target (if successful target, the example of relevant market drug is given.), Biological function of target
- Disease information: Disease name, Relevant PubMed references
- Drugs/ligands information: Drugs/ligands name, Drugs/ligands function (agonist, antagonist, inhibitor, blocker, etc.), Drugs/ligands detail information (name, synonyms, CAS number, etc.)
- Others: Relevant US patent information (register number, patent title, author, issued year, corresponding diseases, etc.), some useful cross links (3D structure, on-line medical dictionary, etc.)

The information mentioned above is obtained by two steps. Firstly, therapeutic target was selected from a large number of relevant literatures by combining automatic text mining techniques and manual reading process. Some automated text retrieval PERL programs were developed to retrieval the literature containing the key work “target”

in local Medline abstract packages [77]. Next, useful therapeutic target and relevant disease and drugs/ligands information were collected manually from the matched Medline abstracts. In most cases, the full literature was referred to find more detail and exact information. After the first step, not only target information but also relevant information about disease conditions and possible corresponding drug/ligands were recorded. In the second step, detail information of targets was automatically selected from some relevant general or specific biological databases, by using text mining PERL programs. Likewise, related US patent information was extracted from US Patent and Trademark Office USPTO Web Patent Databases by accessing the following website http://www.uspto.gov/patft/. Moreover, according to mechanism of drug action published by US Food and Drug Administration (FDA, published at http://www.centerwatch.com/patient/drugs/druglsal.html), targets were roughly divided into two broad groups: successful target, which target by at least one marketed drug, and research target, which is targeted by investigational agents. Regarding the successful target, some examples of corresponding marketed drugs were given. When information collection was completed, TTD construction followed.

3.1.3 Construction of therapeutic target database

TTD adopts the relational data model, which represents therapeutic target data in the form of two-dimension tables. The two-dimensional tables here include therapeutic target ID table (Table 3-1), main information table (Table 3-2), data type table (Table 3-3), and reference information table (Table 3-4). In these tables, TTD ID serves as the primary key; Data type ID and Reference ID are considered as foreign keys.

Table 3-1: Therapeutic target ID list table

TTD ID	Target name
TTT0000001	Placenta growth factor
TTT0000002	P2Y purinoceptor 1
…	…

Table 3-2: Target main information table

TTD ID	Data type ID	Data content	Reference ID
TTT0000001	101	Placenta growth factor	
TTT0000001	102	Research target	
TTT0000001	103	PlGF-131	
TTT0000001	104	Cancers	12678905
…	…	…	…

Table 3-3: Data type table

Data type ID	Data type
101	Target name
102	Target category
103	Synonyms
…	…

Table 3-4: Reference information table

Reference ID	Reference
12678905	Vascular endothelial cell growth factor (VEGF), an emerging target for cancer chemotherapy. Curr Med Chem Anti-Canc Agents. 2003 Mar;3(2):95-117. Review.
…	…

3.1.4 Therapeutic target database structure and access

Basically, TTD web interface comprises four layers. The top layer is the main graphical user interface with a querying tool for finding specific entries of therapeutic target (Figure 3-1). The searching results followed by some specific matching rules will be listed in the second layer (Figure 3-2). By clicking into each entry, the browser can access the detail information for specific target, which is displayed in the third layer (Figure 3-3). More information is given in the fourth layer (Figure 3-4). The detail information about each layer will be discussed in the following parts.

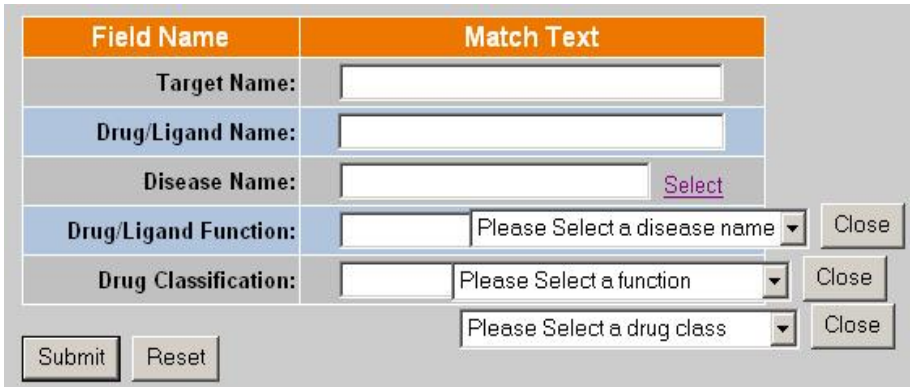


Figure 3-1: The web interface of TTD. Five types of search mode are supported

TTD can be accessed at http://bidd.nus.edu.sg/group/cjttd/ttd.asp. Its web interface is shown in Figure 3-1. The new version of TTD is searchable by five types of search mode: target name, drugs/ligands name, disease name, drugs/ligands function, and drug classification. Queries can be submitted by entering or selecting the required information in any one or combination of the five fields in the form. Users can specify full name or any part of the name in a text field, or choose one item from a selection field. Wild character of '*' and '?' is supported in text field. The relevant disease conditions are classified according to international statistical classification of diseases of World Health Organization (WHO) [117] listed in Table 3-5. In addition, the lists of drug classification are given in Table 3-6.

Table 3-5: Disease class and associated diseases

Disease class	Associated diseases	Disease class	Associated diseases
Blood and blood-forming organs diseases	Coagulation disorders	Musculoskeletal system and connective tissue diseases	Arthritis
	Platelet disorders		Connective tissue disorders
	Red blood cell disorders		Movement disorders
Circulatory system diseases	Blood vessel disorders		Muscular disorders
	Cardiac dysrhythmias		Skeletal disorders
	Cardiovascular disorders	Neoplasms	Bladder cancers
	Circulation disorders		Brain cancers
	Heart disorders		Breast cancers
	Water-retaining diseases		Cancer metastasis
Congenital anomalies	Adrenal glands disorders		Endocrine cancers
	Cerebral disorders		Gastrointestinal cancers
Digestive system diseases	Gallbladder disorders		Hepatic cancers
	Gastric disorders		Leukaemia
	Gastrointestinal disorders		Lung cancers
	Gastrointestinal motility disorders		Lymphoma
	Hepatic disorders		Mesothelioma
	Intestinal disorders		Muscular cancers
	Pancreatic disorders		Myeloma
Endocrine disorders	Aldosterone disorders		Neuronal cancers
	Antidiuretic hormone disorders		Pancreatic cancer
	Glucocorticoid hormone disorders		Renal cancers
	Growth hormone disorders		Reproductive organ cancers
	Insulin disorders		Skeletal cancers
	Neurotransmitter disorders		Skin cancers
	Parathyroid hormone disorders		Thyroid cancers
	Sex hormone disorders		Vascular cancers
	Thyroid hormone disorders	Nervous system and sense organs diseases	Alzheimer's disease
Genitourinary system diseases	Female reproductive organ disorders		Eye disorders
	Lower urinary tract disorders		Headache
	Male reproductive organ disorders		Huntington's disease
	Renal disorders		Neuronal disorders
	Reproductive organ disorders		Parkinson's disease
Immunity disorders	Allergic disorders		Seizure disorders
	Autoimmune disorders		Sensory disorders
	Immunologic disorders	Nutritional and metabolic diseases	Deficiency disorders/ Nutritional disorders
	Transplant disorders		
Infectious and parasitic diseases	Bacterial infections		Electrolyte disorders
	Fungal infections		Lipid disorders
	Helminth infections		Metabolic disorders
	Infections	Respiratory system diseases	Bronchus disorders
	Parasitic infections		Lung disorders
	Prion infections		Nasal disorders
	Viral infections		Respiratory disorders
Inflammation	Inflammation/pain	Skin and subcutaneous tissue diseases	Hair disorders
Injury and poisoning	Injuries		Lupus erythematosus
	Poisoning		Pruritus
Mental disorders	Cognitive deficits		Skin disorders
	Drug dependence	Symptoms, signs, and ill-defined conditions	Cellular disorders
	Eating disorders		Ill-defined disorders
	Mental disorders		Multisystem disorders
	Mood disorders		Nausea/vomiting
			Pain

Table 3-6: Drug classification listed in TTD

Drug Class Name			
Alzheimer's	Antidiarrheal	Antitussive	Immunostimulant
Analgesic	Antidote	Antiviral	Immunosuppressant
Anthelminthic	Anti-emetic	Anxiolytic	Lipid-lowering
Anti-acne	Antifungal	Bronchodilator	Muscular agents
Anti-allergy	Anti-gastric secretion	Cardiotonic	Nasal decongestion
Anti-androgen	Antihypertensive	Cardiovascular agents	Neurologic agents
Anti-angiogenic	Anti-infectives	Central nervous system agents	Ocular agent
Anti-arrhythmia agents	Anti-inflammatory	Dermatologic agents	Parkinson's
Anti-asthmatic	Antimalarial	Diuretics	Procoagulant
Antibacterial	Anti-migraine	Drug dependence (narcotics)	Urinary agents
Anticancer	Anti-obesity	Electrolyte	Vasoconstrictor
Anti-cholesterol	Antiparasitic	Endocrinologic agents	Vasodilator
Anti-coagulant	Antiplatelet	Gastrointestinal agent	Vitamin
Anticonvulsant /Antiepileptic	Antipruritic	Glaucoma treatment	Misc
Antidepressant	Antipsychotic	Gout medicines	
Antidiabetic	Antipyretic	Hormone	

The result of a typical search (e.g. Leukemia) is illustrated in Figure 3-2. All of therapeutic targets that satisfy the search criteria are listed along with the disease conditions to be treated, drugs or ligands directed at the target, and the drug class.

Figure 3-2: Interface of a search result on TTD

More detailed information of a target can be obtained by clicking the corresponding target name (e.g. 5-HT 2B receptor). The result is displayed in an interface shown in Figure 3-3. From this interface, information related to type of target (successful target or research target), target synonyms (for facilitating search), target function, relevant diseases, drugs/ligands and their functions (such as agonist, activator, antagonist, inhibitor, blocker, etc.), related US patent and some of the cross-database shortcuts are provided. If the type of target is marked as successful target, the corresponding drug(s) is listed in this sheet. For an enzymatic target, its EC number is also given here.

Target Information	
Name	5-hydroxytryptamine 2B receptor
Type of target	Research target
Synonyms	5-HT 2B
	5-HT-2B
	Serotonin receptor
	Serotonin receptor 2B
Function	This is one of the several different receptors for 5- hydroxytryptamine (serotonin), a biogenic hormone that functions as a neurotransmitter, a hormone, and a mitogen. This receptor mediates its action by association with G proteins that activate a phosphatidylinositol-calcium second messenger system.
Disease	Anxiety disorder, unspecified [1], Search Karolinska
	Migraine [2], Search Karolinska
	P-chloroamphetamine-induced hyperglycemia [3], Search Karolinska
Antagonist	LY53857 [3]
	MT-500
	Ritanserin [3]
	SB 206553 [1]
Agonist	Sumatriptan [2]
Related US Patent	6,444,477
	6,638,953
Cross References	3D Structure Related Literature On-Line Medical Dictionary

References:

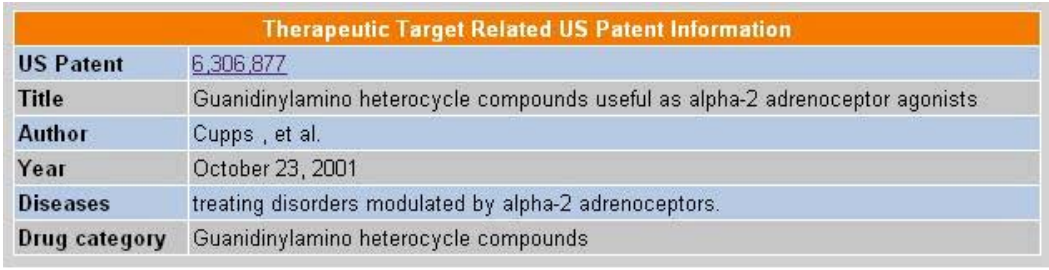
1: Strain-dependent effects of diazepam and the 5-HT2B/2C receptor antagonist SB 206553 in spontaneously hypertensive and Lewis rats tested in the elevated plus-maze. Braz J Med Biol Res. 2001 May;34(5):675-82. PubMed

2: Serotonin in migraine: theories, animal models and emerging therapies. Prog Drug Res. 1998;51:219-44. Review. PubMed

3: p-Chloroamphetamine, a serotonin-releasing drug, elicited in rats a hyperglycemia mediated by the 5-HT1A and 5-HT2B/2C receptors. Eur J Pharmacol. 1998 Oct 23;359(2-3):185-90. PubMed

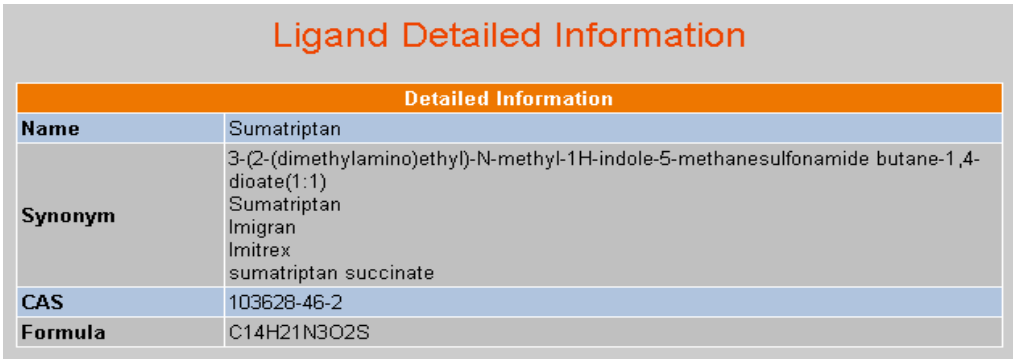
Figure 3-3: Interface of the detailed information of target in TTD

According to those targets with US patent number, information relevant to US patent can be found by clicking the corresponding register number of corresponding US patent (Figure 3-4). Meanwhile, the details about those ligands used as drugs are also given in further information layers (Figure 3-5).



Therapeutic Target Related US Patent Information	
US Patent	6,306,877
Title	Guanidinylamino heterocycle compounds useful as alpha-2 adrenoceptor agonists
Author	Cupps , et al.
Year	October 23, 2001
Diseases	treating disorders modulated by alpha-2 adrenoceptors.
Drug category	Guanidinylamino heterocycle compounds

Figure 3-4: Interface of the detailed information of target related US patent in TTD



Ligand Detailed Information

Detailed Information	
Name	Sumatriptan
Synonym	3-(2-(dimethylamino)ethyl)-N-methyl-1H-indole-5-methanesulfonamide butane-1,4-dioate(1:1) Sumatriptan Imigran Imitrex sumatriptan succinate
CAS	103628-46-2
Formula	C14H21N3O2S

Figure 3-5: Interface of the ligand detailed information in TTD

3.1.5 Statistics of therapeutic targets database data

TTD is now a publicly accessible web-based database that provides comprehensive information about the therapeutic targets, which includes both therapeutic protein and nucleic acid targets together with the targeted disease conditions, the corresponding drugs/ligands, and related US patent information. Cross-links to other databases are introduced to facilitate the access of information regarding the function, 3D structure, and relevant literatures of each target.

The first version of TTD was launched in 2002 [39], which contained only 433 entries of protein and nucleic acid targets, 809 different drugs/ligands, and around 800 disease and literature entries. On the basis of the old version, we collected more target data by key words searching comprehensively relevant abstracts on *Medline* [77]. At present, the number of targets described in the new version of TTD has reached to a total of 1,535 distinct proteins (including subtypes). 268 successful targets, which are confirmed to be targeted by a marketed drug, and 1,267 research targets, which are specifically described as a therapeutic target in a referred journal publication, have been categorized in this database. Both the human and non-human targets are collected. Protein subtypes targeted by subtype-specific agents are counted as separate targets. So far, the TTD is considered as the first comprehensive database for therapeutic targets. It may serve as an essential data resource for target research and development in drug discovery area.

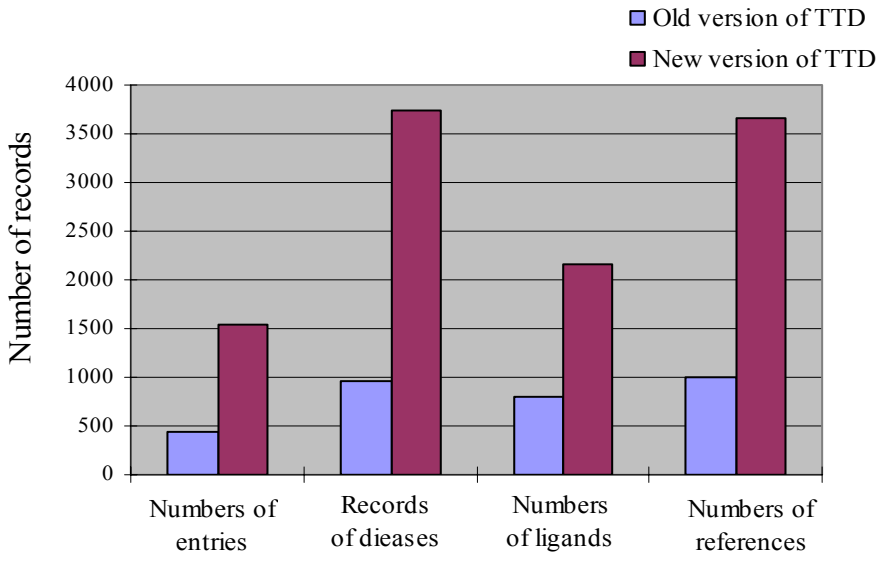


Figure 3-6: Comparison between old and new version of TTD data

Figure 3-6 is a comparison between the old and the new version of TTD data records. As revealed by this figure, there is a sharp increase in target data. Currently, the total

number of therapeutic target entries amounts to 1,535. In addition, the number of disease and reference records is about 3.5 times than that of old ones respectively. Obviously, the abundant data make further knowledge discovery from therapeutic target possible. The comprehensive analysis of therapeutic target is one of the most important parts in this thesis and it will be discussed in the next section.

3.2 Therapeutically relevant multiple-pathways database development

In this section, the Therapeutically Relevant Multiple Pathway (TRMP) database, which gives information about inter-related multiple pathways of a number of diseases and physiological processes, is introduced. As mentioned in previous chapter, the database development strategy used in TTD has been extended to construct the TRMP database. Therefore, the procedure of TRMP database development will be described in the same way.

3.2.1 Preliminary plan of therapeutically relevant multiple-pathways database

Most of the existing pathway databases are focused on describing the whole biological relationships or some protein in a specific pathway. According to therapeutically relevant multiple pathways, the primary concern is centered around important proteins, which are mostly considered as drug targets in relevant diseases and physiological processes. TRMP database, thus, will collect not only therapeutic pathways but also information for those key proteins, which play important roles in relevant disease and physiological conditions. The comprehensive information provided in TRMP database will serve as useful resources for facilitating the analysis

of the potential implications of multiple target-based therapies and the study of mechanism of drug actions.

3.2.2 Collection of therapeutically relevant pathway information

Two groups of information need to be gathered in TRMP database. One is pathways information, which gives the user useful inter-related multiple pathways information; the other is key protein information, which explains some key protein of pathways. Information about pathways is listed as follows:

- Information of multiple pathways
- Information of individual pathways
- Related therapeutic targets (colored in chart)
- Relevant disease conditions or physiological processes

In addition to specific protein, its corresponding information is displayed in similar pattern to TTD. The information including:

- Protein name, synonyms
- SwissProt access number
- Species
- Gene information: gene name, gene location
- Sequence information: protein sequence (AASEQ), gene sequence (NTSEQ)
- Potential therapeutic implications while applicable
- Cross-links to other databases (GeneCard, GDB, Locuslink, NCBI, KEGG, OMIM, SwissProt)

Pathway relevant information can be obtained or extracted from various internet

pathway databases and protein databases. These include Expasy Biochemical pathways (http://www.expasy.ch/cgi-bin/search-biochem-index) [118], Kyoto Encyclopedia of Genes and Genomes (http://www.genome.ad.jp/kegg) [119], Metabolic Pathways of Biochemistry (http://www.gwu.edu/~mpb/) [120], Signaling Pathway database (http://www.grt.kyushu-u.ac.jp/eny-doc/spad.html) [121], Cell Signaling Networks database (http://geo.nihs.go.jp/csndb) [122], Enzymes and Metabolic Pathways (http://emp.mcs.anl.gov) [123], PathDB system for pathways (http://www.ncgr.org/pathdb/) [124], Encyclopedia of E. Coli Genes and metabolism (http://www.biocyc.org) [125], Biocarta (http://www.biocarta.com) [126], the University of Minnesota Biocatalysis/Biodegradation database (http://umbbd.ahc.umn.edu) [127], Soybean metabolic pathways (http://cgsc.biology.yale.edu/metab.html) [128], Nicholson minimaps (http://www.tcd.ie/Biochemistry/IUBMB-Nicholson) [129], Database of Interacting Proteins (http://dip.doe-mbi.ucla.edu/) [130], Biomolecular interaction network database (http://www.blueprint.org/bind/bind.php) [131], TRANSPATH (http://www.biobase.de/pages/products/databases.html) [132], and Signal transduction knowledge environment (http://stke.sciencemag.org/index.dtl) [133]. Moreover, detail information for pathway entries are obtained by comprehensive searching of related publications in Medline [77]. Combination of three keywords (disease name, target name, and “pathway”) is used in searching the relevant publications. The relevant information is derived primarily from review articles and pharmacology textbooks. Primary articles are also used for clarification purpose. The extracted information is double checked against the referenced articles independently by different persons. All of the references used for generating the pathways are provided in the database. Human data are used for human pathways and proteins. Likewise, the corresponding species data are used for bacterial or viral

pathways and proteins.

The data collecting methods were similar to those mentioned in TTD development. Generally, by using PERL programs, the literature within "pathway" information was retrieved automatically. Next, more details were obtained by manually reading the downloaded literature. The preliminary data collection included the gathering of pathways information, relevant therapeutic targets and disease information, and corresponding drugs/ligands directed at each of these targets. Furthermore, information of specific protein was automatically generated by text mining programs, which picked up relevant information from other general or specific biological databases.

3.2.3 Construction of therapeutically relevant multiple-pathways database

Differing from TTD, TRMP database has not only literal data, but also graphic data. Thus, two data models are applied in TRMP database. The pathway graphic data were recorded by simple flat-file model. Each multiple pathway map was considered as one flat file. All of the flat files were displayed by means of HTML web pages. The interactive maps of each pathway entry of TRMP database were constructed by using Macromedia FLASH. The corresponding database architect associated with the pathway interactive maps is developed by using Active Server Page with Oracle 9i support, which was the same as the one used in TTD Database management system. In additional, the protein information data of TRMP database used a relational data model. The relational data tables are designed as same as those of TTD (Table 3-7, Table 3-8, Table 3-9). Here, TRMP database ID is designed as primary key and data

type ID is foreign key.

Table 3-7: Pathway related protein ID table

TRMP ID	Protein name
TRMP01001	AccA
TRMP01002	AccD
…	…

Table 3-8: Pathway related protein main information table

TRMP ID	Data type ID	Data content
TRMP01001	100	AccA
TRMP01001	102	Acetyl-coenzyme A carboxylase carboxyl transferase subunit alpha
…	…	…

Table 3-9: Data type table

Data type ID	Data type
100	Entry
101	Potential therapeutic implications
102	Protein name
103	Synonyms
…	…

3.2.4 Therapeutically relevant multiple-pathways database structure and access

During the development of TRMP database, three layers were used. The top layer was the main graphical user interface with a querying tool for finding specific entries of the multiple pathways (Figure 3-7). The second layer was the graphical interface for the interactive maps of multiple pathways with a browser tool for retrieving additional information (Figure 3-8). The browser tool was used both for accessing information about the constituent individual pathways from other databases and for retrieving information about individual targets or non-target proteins directs a retrieving request from TRMP database. The third layer is the graphic interface for entries of individual

targets or non-target proteins with a browser tool for accessing additional information from other databases (Figure 3-9).

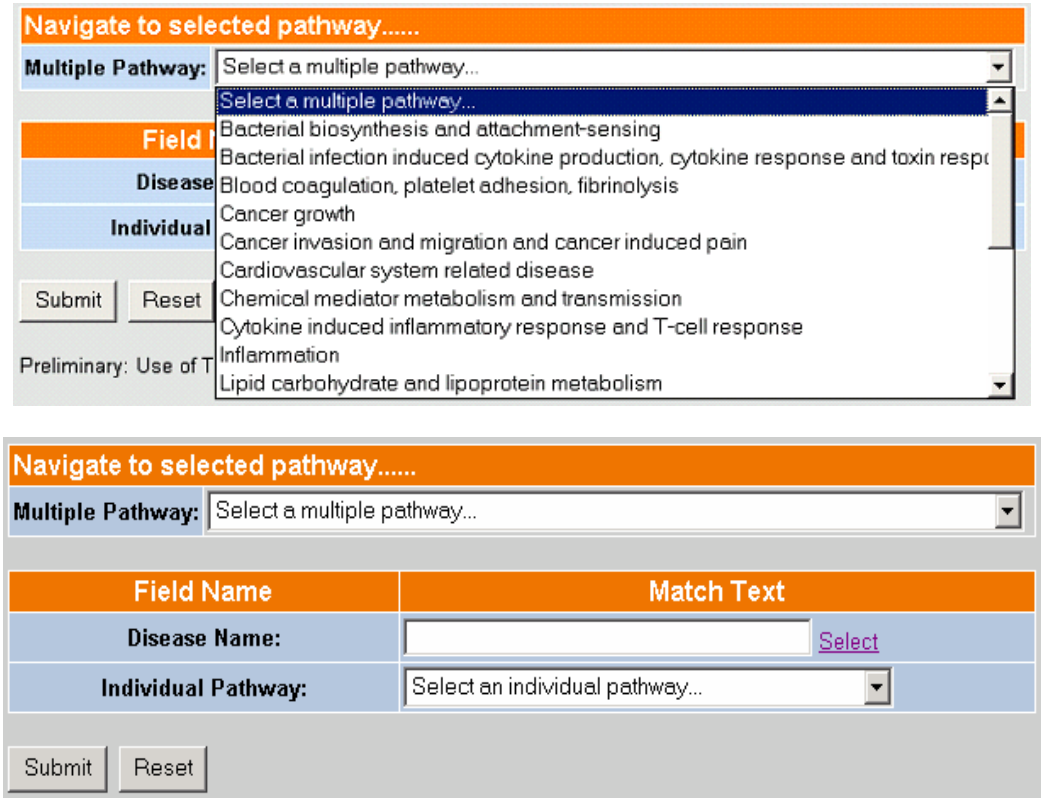


Figure 3-7: Web interface of TRMP database

TRMP database can be accessed at http://bidd.nus.edu.sg/group/trmp/trmp.asp. Its web interface is shown in Figure 3-7. Three types of search mode are supported. Firstly, this database is searchable by selecting the name of a particular entry of multiple-pathways. Also, it can be accessed by selection of a disease or an individual pathway name from the list provided in the corresponding selection field.

Moreover, searches involving any combination of these three selection fields are also supported. The pathways are indexed according to multiple pathway or individual pathway, which are listed in Table 3-10. Meanwhile, the list of pathway related diseases or conditions is given in Table 3-11.

Table 3-10: Multiple pathways and corresponding individual pathways

Multiple Pathway		Individual Pathway
Bacterial biosynthesis and attachment-sensing		Cpx system pathway; Lipid synthesis pathway; Peptidoglycan synthetic pathway
Bacterial infection induced cytokine production, cytokine response and toxin response		Chemokine signaling pathway; MAPK pathway; NF-kB activation pathway; Pathway of growth factor induced microfilament bundling; Pathway of pseudomonas exotoxin induced cell death; Pathway of toxins induced block of actin polymerization; TLR signaling pathway
Blood coagulation, platelet adhesion, fibrinolysis		Fibrin formation pathway; Platelet activation pathway
Cancer growth		Adrenaline pathway; Apoptosis pathway; COX pathway; EGF pathway; Estrogen pathway; GnRH pathway; Hypersensitive pathway; IGF pathway; MAPK pathway; Myc pathway; NF-kB pathway; p53 Pathway; PI3K-AKT pathway; RAS-signaling pathway; Rb Pathway; RHO regulated cell-cycle pathway; TGF pathway; TRADD pathway; TRAIL induced apoptosis pathway; Wnt signaling pathway
Cancer invasion and migration and cancer induced pain		Integrin-dependent intracellular signaling pathway; MET down regulation pathway; MET-dependent invasive growth signaling pathway; Nociceptor signaling pathway; Plexin-B-mediated pathway; RAC pathway; RAS pathway
Cardiovascular system related disease		Acetylcholine pathway; ATII pathway; Bradykinin pathway; Bradykinin synthesis pathway; CNP/NPR-B/cGMP pathway; Endothelin pathway; Noradrenalin pathway; PLC-IP3 pathway; Rho-Rho-kinase pathway; Serotonin pathway; Serotonin synthesis pathway
Chemical mediator metabolism and transmission		Acetylcholine pathway; Acetylcholine synthesis pathway; Adrenaline pathway; MAO pathway; Noradrenalin synthesis pathway; Serotonin pathway; Serotonin synthesis pathway
Cytokine induced inflammatory response and T-cell response		CD14 pathway; Interleukin-1 pathway; Interleukin-18 pathway; TLR signaling pathway; TNF signaling pathway
Inflammation		Bradykinin pathway; Bradykinin synthesis pathway; COX pathway; Glucocorticoid pathway; Glucocorticoid synthesis pathway; Histamine pathway; Leukotriene synthesis pathway; Noradrenalin Pathway; Noradrenalin synthesis pathway; NOS pathway; Pathway of arachidonate release; TNF signaling pathway
Lipid carbohydrate and lipoprotein metabolism	Lipid, carbohydrate metabolism in adipose tissue cells	Beta-oxidation pathway; cAMP/PKA/CREB pathway; cAMP-PKA pathway; CRH pathway; Glucose transport pathway; Glycolysis pathway; Hexose monophosphate pathway; Insulin pathway; Interleukin-6 pathway; Leptin pathway; Lipid fatty acid synthesis pathway; Lipid synthesis pathway; LPL pathway; Melanocortin pathway; NPY pathway; TCA pathway; TNF signaling pathway
	Lipid, carbohydrate metabolism in liver cells	Bile acid recovery pathway; Cholesterol synthesis pathway; Fatty acid synthesis pathway; Glycogen synthesis; Glycolysis pathway; Hexose monophosphate pathway; Lipid synthesis pathway
	Lipid, carbohydrate metabolism in muscle tissue cells	Beta-oxidation pathway; cAMP-PKA pathway; Glucose transport pathway; Glycolysis pathway; Insulin pathway ; Interleukin-6 pathway; Leptin pathway; LPL pathway; TNF signaling pathway
	Lipoprotein metabolism	Lipoprotein metabolism pathway
	Protein, carbohydrate, lipid digestion and absorption	Carbohydrate absorption pathway; Carbohydrate digestion pathway; Lipid absorption pathway; Lipid digestion pathway; Protein absorption pathway; Protein digestion pathway
Viral infection induced effect on cytokine, RNA, viral protein and genome synthesis	Viral infection induced cytokine production, RNA translation inhibition, viral protein and genome synthesis	DNA synthesis pathway; IRF-3 activation pathway; MAPK pathway; NF-kB activation pathway; Protein synthesis pathway; RNA replication pathway; RNA synthesis pathway
	Viral infection induced cytokine production 2	Death receptor pathway; MAPK pathway; MyD88-dependent pathway; MyD88-independent pathway; NF-kB activation pathway; TLR signaling pathway; TNF signaling pathway
	Cytokine induced RNA degradation and translation inhibition	IFN pathway; STAT pathway
	Cytokine response and viral counteraction	IFN pathway

Table 3-11: Therapeutically relevant multiple pathways related disease or conditions

Therapeutically Relevant Multiple Pathways related Disease Name		
Acute Heart attack/Myocardial infarction	Fever(Syndromes)	Nasal decongestion (conditions)
Alzheimer's disease	Gastrointestinal disorder	Neuromuscular disorders
Angina	Glaucoma	Obesity
Anxiety tremor	Heart failure	Pancreatic cancer
Asthma	Hemorrhage	Parkinson
Bacterial infection	AIDS	Phaeochromocytoma
Begin prostatic hypertrophy	Hypercholesterolemia	Postmenopausal breast cancer
Bradycardia	Hyperlipidemia	Progressive renal insufficiency
Breast cancer	Hypertension	Prostate cancer
Cancer	Hypertension in pregnancy	Prostatic hyperplasia
Cancer pain	Hypotension	Psychosis
Cardiac arrhythmias	Hypoxia (Syndromes)	Purpura (conditions)
Cardiac dysrhythmias	Inflammation	Refractory angina
Cardiac failure	Insulin resistance	Rheumatoid arthritis
Cardiogenic shock	Ketoacidosis	Schizophrenia
Colon cancer	Lactic acidosis	Sepsis
Coronary artery spasm	Lung cancer	Septic shock
Cushing (Syndromes)	Melanoma	Thrombosis and embolism
Depression	Metastasis	Type 2 diabetes
Diabetes mellitus	Migraine	Viral infection
Diabetic nephropathy	Migraine prophylaxis (conditions)	Vomiting/Nausea (Syndromes)
Disseminated intravascular coagulation	Mood disorders	Von willebrand
Endotoxin shock	Multiple myeloma	
Erectile dysfunction	Myasthenia gravis	

Figure 3-8 illustrates the interface for an entry of therapeutically relevant multiple pathways. A therapeutically targeted protein is represented by a red rectangle box and a non-target protein by a yellow or blue rectangle box respectively, with the name of the target or protein included in each respective box. Genes and RNAs are represented by orange boxes so that they can be easily distinguished from proteins. More detailed information about each target can be obtained by clicking the respective red rectangle box which is linked to the corresponding target information page provided in our database. Proteins in the yellow boxes are those with detailed information available. The relevant information can be accessed by clicking a yellow box which is linked to the corresponding protein information page provided in our database. Proteins in the

blue boxes are those with only a general name specified in the literature which is not specific enough to determine their identity. As a result, no detailed information about these proteins is available in the current version of our database.

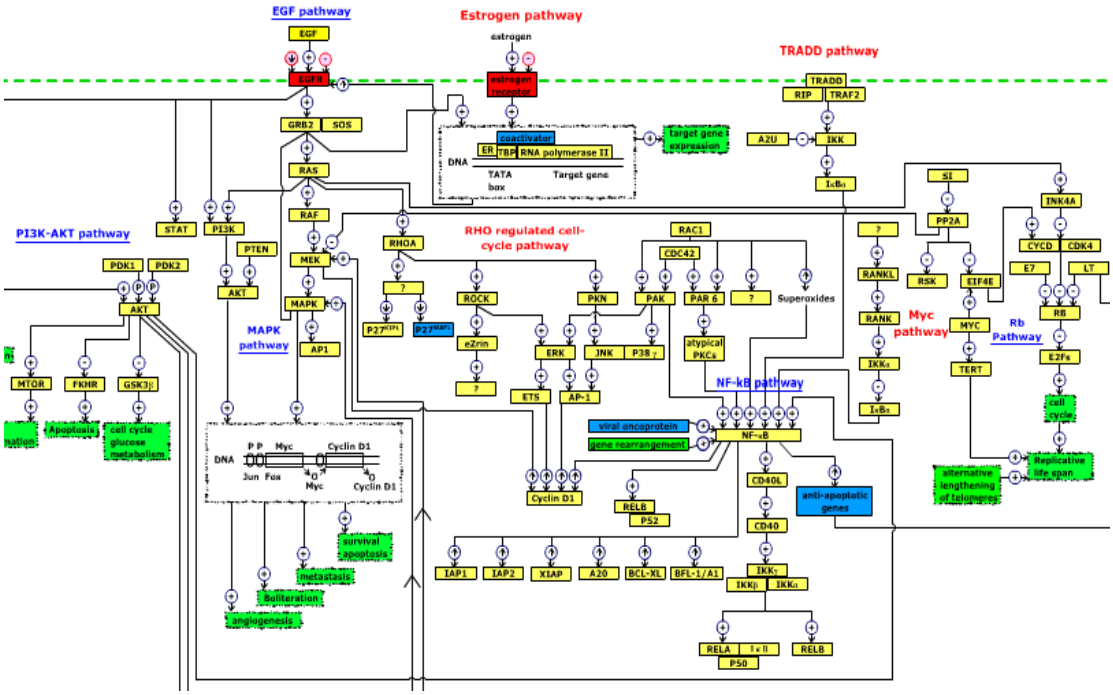
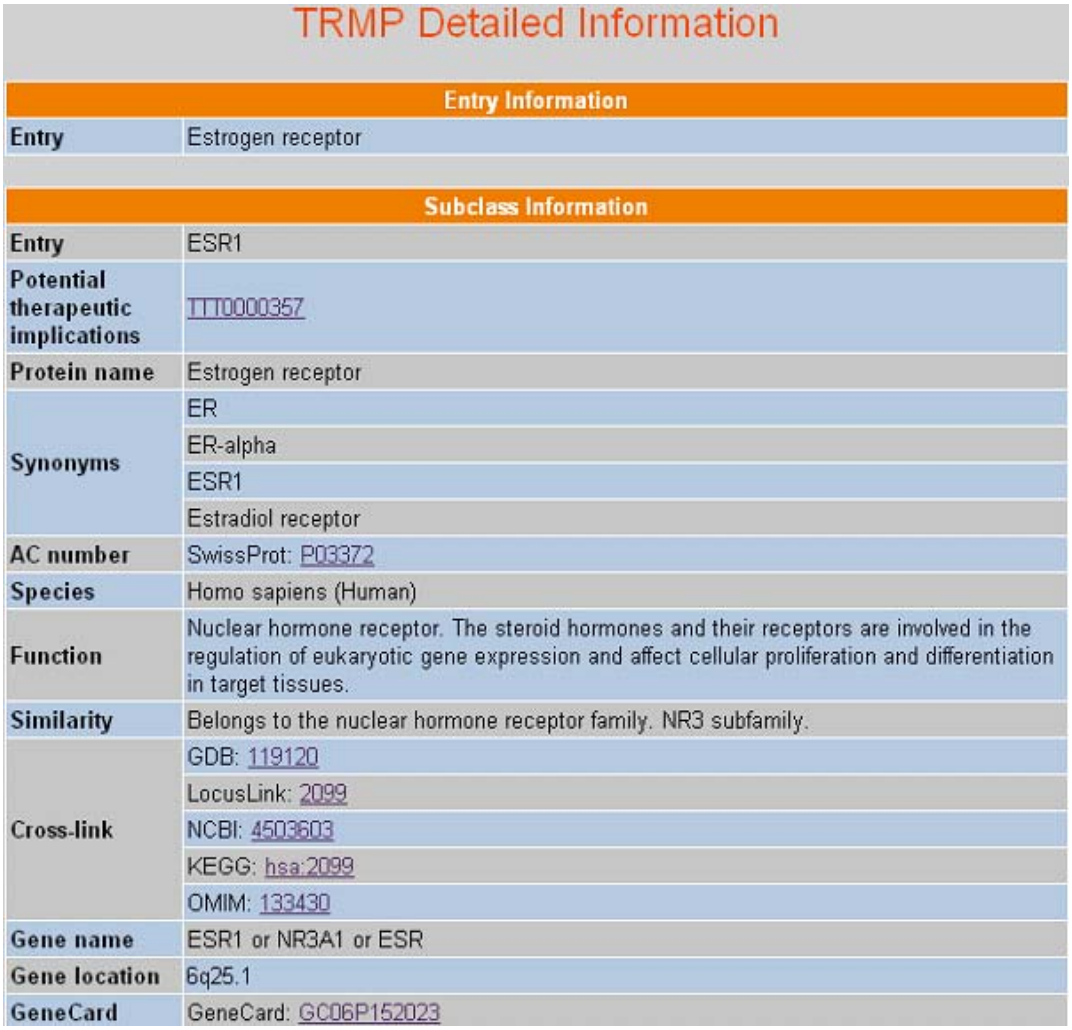


Figure 3-8: Interface of a multiple pathways entry of TRMP database

A small molecule ligand is represented by its name and its action of a protein is indicated by a white circular box with one of the following symbols inside. These symbols are +, -, ↑, ↓, P, R, B, D and A which represents activation of the protein, inhibition of the protein, increase of the protein level, decrease of the protein level, protein phosphorylation, release of the protein to extra-cellular environment, binding to the protein with unknown effect, binding to the protein leading to its dimerization, and binding to the protein as an antibody respectively. A pink circular box indicates the site and action of a drug or investigative agent and the type of drug action is represented by the same set of symbols as that for small molecule ligands. More detailed information about the corresponding drugs is represented through a

mouse-over-effect upon clicking a relevant pink drug action circular box. The names of the constituent individual pathways contained in each entry are given. Cross links to other pathway databases are provided for those individual pathways that are described in other pathway databases. The effects of the pathways are given by the green boxes with unregulated dot line which include the description about each effect. There are cases that the actual protein involved in a particular process in a pathway is unidentified. Thus, instead of the actual protein, the related process is described in the same way. In addition, the complex of several proteins is demonstrated by the light-blue boxes with dot line.



TRMP Detailed Information

Entry Information	
Entry	Estrogen receptor

Subclass Information	
Entry	ESR1
Potential therapeutic implications	TTT0000357
Protein name	Estrogen receptor
Synonyms	ER
	ER-alpha
	ESR1
	Estradiol receptor
AC number	SwissProt: P03372
Species	Homo sapiens (Human)
Function	Nuclear hormone receptor. The steroid hormones and their receptors are involved in the regulation of eukaryotic gene expression and affect cellular proliferation and differentiation in target tissues.
Similarity	Belongs to the nuclear hormone receptor family. NR3 subfamily.
Cross-link	GDB: 119120
	LocusLink: 2099
	NCBI: 4503603
	KEGG: hsa:2099
	OMIM: 133430
Gene name	ESR1 or NR3A1 or ESR
Gene location	6q25.1
GeneCard	GeneCard: GC06P152023

Figure 3-9: Interface of a target entry of TRMP database

Figure 3-9 gives the interface of a target entry of TRMP database, which is similar to that of a non-target protein entry with the exception that the former contain a section for potential therapeutic implications. Information provided include protein name, synonyms, SwissProt AC number, species, gene name and location, protein sequence (AASEQ) and gene sequence (NTSEQ) as well as potential therapeutic implications while applicable. Cross-links to other databases are provided which include GeneCard, GDB, Locuslink, NCBI, KEGG, OMIM and SwissProt to facilitate the access of more detailed information about various aspects of particular target or non-target protein.

3.2.5 Statistics of therapeutically relevant multiple-pathways database data

TRMP database is also a publicly accessible web-based database, which is designed to provide information about known therapeutic targets within each network of multiple pathways, the corresponding drugs/ligands directed at each of these targets, the constituent individual pathways, and information about the proteins involved in these pathways. Cross-links to other databases are introduced to facilitate the access of information about the constituent individual pathways: the function, sequence, nomenclature, and related literatures of each protein in the pathways.

The TRMP database currently contains 11 entries of multiple pathways that include 97 distinct individual pathways and 120 therapeutic targets covering 72 different disease conditions. A total of 32 of the 97 distinct individual pathways are included in other pathway databases. Apart from multiple pathways and distinct individual pathways, the related diseases, the number and examples of associated known

therapeutic targets, and examples of corresponding drugs directed at these targets are also included. With rapid advances in proteomics [31], pathways [55] and systems [134], new information about therapeutically relevant multiple pathways can be incorporated or the corresponding databases can be cross-linked to TRMP database to provide more comprehensive information about the therapeutically relevant pathways, related targets and their relationship to other biomolecules and cellular processes.

Furthermore, the approach of linking human proteins to the human pathway constituents (with the exception of the viral and bacterial specific ones) was critically based on the assumption that all of the shown pathways were found in human although this might not have been experimentally verified. The pathway models in the review articles and textbooks were often based on the results of a patchwork of experimental systems involving genes and proteins of different species origin. Thus caution was needed to interpret the molecular interactions and pathway constituents in TRMP database. Effort would be made to promptly update newly reported results in the database. So far, the part of pharmainformatics databases development has been described. In next chapter, the focus on computational study of therapeutic targets, which was one of the most important parts in this thesis.

4 Computational analysis of therapeutic targets

Therapeutic targets can be divided into successful targets, which are targeted by at least one marketed drug [9, 135], and research targets, which are targeted only by investigational agents [136-140]. The search for new targets has been facilitated by advances in genomics [32, 33] and proteomics [31], a deeper understanding of molecular mechanism of diseases [34, 35], and the development and improvement of technologies for target identification and validation [4, 5, 8, 9, 141, 142]. Since 1996, a growing number of new and novel research targets have emerged [136-138, 140]. Drug design effort has increasingly been focused on disease-specific protein subtypes [143, 144]. Progress has been made in probing some of the previously unknown targets of marketed and investigational drugs [9, 28, 29, 145, 146]. While a relatively small number of research targets are known to have become successful targets since 1996, the number of successful targets collected in the TTD appears to have substantially increased since previous reports [9, 37, 135]. This could be due in part to a variety of factors such as the inclusion of nonhuman targets and protein subtype targets in the new report, the approval of a growing number of subtype-specific drugs since the publication of previous reports and the gain of new knowledge about previously unknown targets of marketed drugs.

This chapter provides a comprehensive analysis of these targets so as to provide useful hints about the current trends of exploration of therapeutic targets and the focus of interest for drug discovery for various diseases.

4.1 Distribution of therapeutic targets with respective disease classes

4.1.1 Distribution pattern of successful target

Distribution of successful targets with respect to different disease classes is given in Table 4-1. The total number of distinct successful targets is 268, 120 of which are for more than one disease classes. Because of this redundancy of targets, the sum of the number of targets in these classes is greater than 268. The number of targets shared between different disease classes is also given in the Table 4-1. Disease classes are based on the international statistical classification of diseases of WHO [117].

Targets for neoplasms, infectious and parasitic diseases, nervous system and sense organs disorders, circulatory system diseases, and mental disorders, which contain 78, 78, 56, and 46 targets respectively, constitute the groups with the largest number of targets. Other groups consisting of substantial numbers of targets are those of respiratory system diseases, genitourinary system diseases, musculoskeletal system and connective tissue diseases, and endocrine disorders. The number of targets for each of these classes is 35, 24, 23, and 21, respectively.

Examples of successful targets in the class of neoplasms are estrogen receptor and aromatase (breast cancer), thymidylate synthase and DNA topoisomerase I (colorectal cancer), leutinizing-hormone-releasing hormone (prostate cancer) and *BCR-ABL* tyrosine kinase (chronic myeloid leukemia). Examples in the class of infectious and parasitic diseases are HIV-1 protease (AIDS), influenza A virus M2 protein (influenza A), HBV polymerase (Hepatitis B), penicillin-binding proteins and DD-carboxypeptidase (bacterial infections), histamine N-methyltransferase and

dihydropteroate synthetase (malaria), 1,3-beta-glucan synthase and lanosterol 14-alpha-demethylase (fungal diseases). Those in the class of nervous system and sense organs disorders are acetylcholinesterase and NMDA receptor (Alzheimer's disease), catechol-O-methyl-transferase and D2 dopamine receptor (Parkinson's disease), alpha-2 and beta-1 adrenoceptor (glaucoma and ocular hypertension), 5-HT1D receptor (migraine), and mu/kappa opioid receptor (drug dependence).

Additional examples of successful targets are platelet glycoprotein IIb/IIIa receptor (acute coronary syndrome), angiotensin-converting enzyme, angiotensin receptor AT1, beta-1 and alpha adrenoceptor (hypertension, cardiac failure, arrhythmias), endothelin receptor (primary pulmonary hypertension) for circulatory system diseases; monoamine oxidase A and serotonin transporter (depression), D2 dopamine receptor (schizophrenia), GABA receptor and beta adrenergic receptor (insomnia, anxiety) for mental disorders; beta-2 adrenergic receptor, 5-lipoxygenase and leukotriene receptor (asthma) and sigma-type opioid receptor (cough) for respiratory system diseases; phosphodiesterase type 5 (erectile dysfunction) and muscarinic receptor M3 (overactive bladder) for genitourinary system diseases; cyclooxygenase 2, tumor necrosis factor-alpha, interleukin 1 receptor (rheumatoid arthritis, osteoarthritis) and farnesyl pyrophosphate synthetase (osteoporosis) for musculoskeletal system and connective tissue diseases; gastrointestinal lipases, fatty acid synthase (obesity) and farnesyl pyrophosphate synthetase (hypercalcemia) for nutritional and metabolic diseases; and insulin receptor and peroxisome proliferator activated receptor-gamma (diabetes) for endocrine disorders.

Table 4-1: Number of successful targets in different disease classes

Indications	Disease Classes	Number of therapeutic targets		Shared therapeutic targets																	
		All related targets	Non-redundant targets	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(l)	(m)	(n)	(o)	(p)	(q)	(r)
(a)	Blood and Blood-Forming Organs Diseases	13	2	-	8	1	1	1	2	2	0	0	1	0	2	0	4	2	3	1	1
(b)	Circulatory System Diseases	54	9	8	-	11	10	10	24	15	6	7	6	2	6	12	19	6	8	8	2
(c)	Digestive System Diseases	19	4	1	11	-	5	3	8	9	4	3	5	1	2	5	5	3	6	1	1
(d)	Genitourinary System Diseases	24	0	1	10	5	-	6	11	7	3	6	1	1	2	6	12	1	2	2	3
(e)	Musculoskeletal System and Connective Tissue Diseases	23	4	1	10	3	6	-	10	6	2	2	5	4	6	6	12	1	5	2	3
(f)	Nervous System and Sense Organs Diseases	56	7	2	24	8	11	10	-	17	4	6	3	2	7	27	13	3	14	7	2
(g)	Respiratory System Diseases	35	5	2	15	9	7	6	17	-	5	3	8	2	5	12	10	2	8	4	1
(h)	Skin and Subcutaneous Tissue Diseases	13	2	0	6	4	3	2	4	5	-	3	3	1	1	2	7	2	2	2	1
(i)	Endocrine Disorders	21	6	0	7	3	6	2	6	3	3	-	3	0	3	3	8	4	1	1	1
(j)	Immunity Disorders	18	2	1	6	5	1	5	3	8	3	3	-	3	6	2	9	2	3	2	1
(k)	Infectious and Parasitic Diseases	78	57	0	2	1	1	4	2	2	1	0	3	-	4	1	17	4	1	1	2
(l)	Inflammation	15	1	2	6	2	2	6	7	5	1	3	6	4	-	2	8	1	4	1	1
(m)	Mental Disorders	46	10	0	12	5	6	6	27	12	2	3	2	1	2	-	5	3	10	2	0
(n)	Neoplasms	78	29	4	19	5	12	12	13	10	7	8	9	17	8	5	-	5	5	6	4
(o)	Nutritional and Metabolic Diseases	21	5	2	6	3	1	1	3	2	2	4	2	4	1	3	5	-	1	0	0
(p)	Symptoms, Signs, and Ill-Defined Conditions	22	2	3	8	6	2	5	14	8	2	1	3	1	4	10	5	1	-	1	2
(q)	Injury and Poisoning	15	3	1	8	1	2	2	7	4	2	1	2	1	1	2	6	0	1	-	0
(r)	Congenital Anomalies	4	0	1	2	1	3	3	2	1	1	1	1	2	1	0	4	0	2	0	-
Total successful therapeutic targets based on disease classes		555(duplicate) 268(distinct)	148	Redundancy of therapeutic targets =120 ; Non-redundancy of therapeutic targets =148																	

There are a number of innovative targets emerged since 1996 that are based on new mechanisms or new targets for treating diseases, which usually finds large market and become highly successful [38]. These targets, together with the year of first FDA approval and the name of the approved drug, are vascular endothelial growth factor (2004, Bevacizumab) for the treatment of colorectal cancer, NMDA receptor (2003, Memantine) for Alzheimer's disease, HIV gp41 envelope glycoprotein (2003, Enfuvirtide) for HIV infection, HBV DNA polymerase (2002, Adefovir dipivoxil) for hepatitis B, mineralocorticoid receptor (2002, Eplerenone) for hypertension, endothelin receptor (2001, Bosentan) for primary pulmonary hypertension, *BCR-ABL* tyrosine kinase (2001, Imatinib) for chronic myeloid leukemia, retinoid receptors (1999, Bexarotene) for cutaneous T cell lymphoma, gastrointestinal lipase (1999, Orlistat) for obesity, FK-binding protein 12 (1999, Sirolimus) for the prevention of organ rejection following renal transplantation, Receptor protein-tyrosine kinase erbB-2 (HER2/neu) (1998, Trastuzumab) for HER2 positive metastatic breast cancer, phosphodiesterase 5 (1998, Sildenafil) for erectile dysfunction, platelet glycoprotein IIb/IIIa receptor (1998, Tirofiban, Eptifibatide) for severe chest pain and small heart attacks, cyclooxygenase 2 (1998, Celecoxib) for arthritis, peroxisome proliferator activated receptor (1997, Troglitazone) for type 2 diabetes mellitus, and platelet P2Y12 receptor (1997, Clopidogrel) for stroke and heart attack.

4.1.2 Targets for the treatment of diseases in multiple classes

Some targets have been explored for the treatment of diseases from more than one class. Disease classes with higher concentration of shared targets are those for circulatory system diseases, neoplasms, and nervous system and sense organs disorders. For instance, there are 24, 19, and 15 targets of circulatory system diseases

that are shared with those of nervous system and sense organ disorders, neoplasms, and respiratory diseases respectively. High concentration of shared targets in this class may be partly attributed to the involvement of circulatory system in various disease conditions. For example, there are strong interactions between nervous systems and cardiovascular systems, and it is not surprising that targets involved in the crosstalk between these systems are used for both diseases [147]. Moreover, tumor growth relies on the formation of new blood vessels, and proteins involved in angiogenesis have been targeted for anticancer drug development as well as circulatory system diseases [138]. In addition, sensory receptors in the respiratory system are known to respond to irritants and subsequently induce cardiovascular responses, and targets involved in these responses are used for symptom relief of respiratory diseases as well as for the treatment of cardiovascular diseases [148].

An example of a shared target is beta-adrenoceptor for circulatory system diseases, nervous system disorders, and respiratory system diseases. Heart failure is known to harmfully activate sympathetic nervous system as well as the rennin-angiotensin system, and these circulatory system disease-associated disorders can be treated by beta-adrenoceptor antagonists [149]. Meanwhile, beta-adrenoceptor blocking drugs have been used in the central nervous system related disorders, such as psychiatry and neurology [150, 151]. In addition, beta-adrenoceptor agonists have also been used for the treatment of asthma, a typical respiratory system disease, by dilating the bronchial smooth muscle [152].

Another example of a shared target is dual-specificity protein phosphatases (DSPases), which represent a subclass of the protein tyrosine phosphatases with highly conserved phosphatase active site motifs. DSPases dephosphorylate serine, threonine, and

tyrosine residues in the same protein substrate, and they play important roles in multiple signaling pathways and appear to be deregulated in cancer and Alzheimer's disease [153]. Because of their roles and properties, there has been increasing effort for identifying DSPase inhibitors that are more potent and selective than the general tyrosine phosphatase inhibitor sodium orthovanadate, for the treatment of both diseases, which has led to the discovery of several promising leads [154].

4.1.3 Distribution pattern of research targets

Table 4-2 lists the distinct research target distribution in different disease classes. The majority of the research targets are distributed in the class of neoplasms and infectious and parasitic diseases, which accounts for 37% and 23% respectively. Moreover, four other disease class, namely nervous system and sense organs disorders, circulatory system diseases, nutritional and metabolic disorders and inflammation, are also important in research target discovery.

According to Table 4-2, 13%, 13%, 9%, and 9% of the research targets are distributed in these classes respectively. Overall, the number of non-redundant research targets in these six disease classes is 708 which accounts for 56% of the total number of research targets. This reflects the intensive efforts directed at the search for effective therapeutics for cancer treatment and prevention [155-157], cardiovascular diseases [158, 159], inflammatory diseases [160], obesity [161-164] and high cholesterol [165, 166].

Table 4-2: Distinct research target distribution in different disease classes

Disease Classes	Number of therapeutic targets	Distinct research target in different disease classes %
Blood and Blood-Forming Organs Diseases	41	3%
Circulatory System Diseases	168	13%
Digestive System Diseases	45	4%
Genitourinary System Diseases	50	4%
Musculoskeletal System and Connective Tissue Diseases	92	7%
Nervous System and Sense Organs Diseases	171	13%
Respiratory System Diseases	63	5%
Skin And Subcutaneous Tissue Diseases	32	3%
Endocrine Disorders	91	7%
Immunity Disorders	70	6%
Infectious and Parasitic Diseases	287	23%
Inflammation	111	9%
Mental Disorders	61	5%
Neoplasms	468	37%
Nutritional and Metabolic Diseases	120	9%
Symptoms, Signs, and Ill-Defined Conditions	62	5%
Injury and Poisoning	51	4%
Congenital Anomalies	2	0%
Total research therapeutic targets based on disease classes	total distinct research targets=1267	100%
	total duplicate research targets=1989	

4.1.4 General distribution pattern of therapeutic targets

The number of research targets of each disease class is given in Figure 4-1 along with that of successful targets. With the exception of the class of congenital anomalies, there appears to be a significant increase in the level of exploration of targets for every disease class, as evidenced by the significantly larger number of research targets than that of successful targets, which reflects intensive efforts for finding effective treatment options against all diseases. Little success seems to have been made in the identification of useful targets for congenital anomalies. The low target distribution of this disease class may be due partly to the use of surgical therapies as primary treatment options [167, 168] and partly to the lack of knowledge of the mechanism of

the relevant diseases [169].

The disease classes with the largest increase of targets are those of neoplasms with 468 research targets versus 78 successful targets, infectious and parasitic diseases with 287 research targets versus 78 successful targets, nervous system and sense organs disorders with 171 research targets versus 56 successful targets, circulatory system diseases with 168 research targets versus 54 successful targets, nutritional and metabolic disorders with 120 research targets versus 21 successful targets, inflammation with 111 research targets versus 15 successful targets, musculoskeletal system and connective tissue diseases with 92 research targets versus 23 successful targets, and endocrine disorders with 91 research targets versus 21 successful targets.

Examples of specific diseases in these key classes that have a substantial number of research targets are various cancers with 468 targets [155-157], cardiovascular diseases with 120 targets [158, 159], diabetes with 65 targets[170], arthritis with 64 targets[171], obesity with 57 targets [35, 161-164], Alzheimer's disease with 44 targets [172, 173], and high cholesterol with 12 targets [165, 166]. These diseases affect a significant number of patients and thus have received substantial interest in the development of new therapeutics for their treatment. Another class with high ratio of research versus successful targets is that of infectious and parasitic diseases, which has a ratio of 78/287. The significant increase in the number of research targets for this disease class primarily stems from the pursuit for new generation of antibiotics [174], antifungal agents [175], and anti-HIV drugs [176] as well as for the development of effective drugs for malaria [177] and a variety of viral infections such as hepatitis, herpes simplex virus, and respiratory syncytial virus [176].

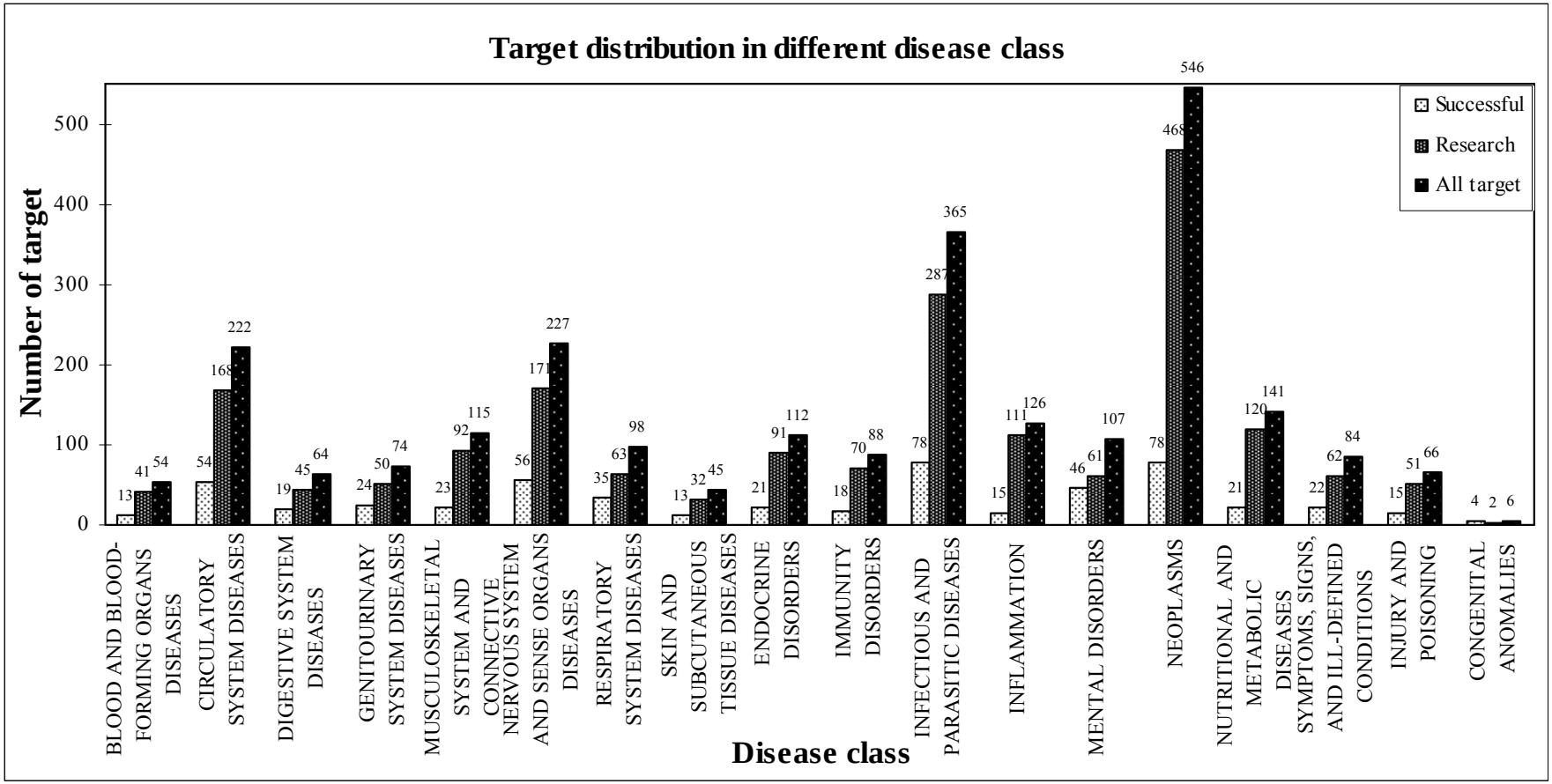


Figure 4-1: Distribution of therapeutic targets against disease classes

4.2 Current trends of exploration of therapeutic targets

4.2.1 Targets of investigational agents in the US patents approved in 2000-2004

Clues about the current trend of target exploration can be obtained from the targets described in the recently approved patents of investigational agents. Most of these patents describe molecular mechanism and many of them provide the identifiable target for each group of patented agents. Table 4-3 and Table 4-4 give some of the successful targets and research targets described in the US patents approved between January 2000 and September 2004. A total of 2,080 US patents of investigational agents have been approved during this period, 1,606 or 77.2% of which have an identifiable target.

There are 395 identifiable targets described in these 1,606 patents. Of these targets, 264 have been found in more than one patent and 50 appear in more than 10 patents. The number of patents associated with a target can be considered to partly correlate with the level of effort and intensity of interest currently being directed at it. Approximately 1/3 of the patents with an identifiable target were approved in the past year. This suggests that the effort for the exploration of these targets is on going and there has been steady progress in the discovery of new investigational agents directed at these targets.

Table 4-3: Some of the successful targets explored for the new investigational agents described in the US patents approved in 2000-2004.

Therapeutic Target		Number of US patents	Targeted Diseases
Protein	Subgroup		
Adrenergic receptors (63)	Alpha adrenoceptor	1	Nasal Congestion, Glaucoma, Asthma, Migraine, Diarrhea
	Alpha-1 adrenoceptor	4	Congestive Heart Failure, Hypertension, Benign Prostatic Hyperplasia, Eye Disorders
	Alpha-1D adrenoceptor	5	Benign Prostatic Hyperplasia, Peripheral Vascular Disease, Congestive Heart Failure, Hypertension
	Alpha-1B adrenoceptor	3	CNS Disorders, Anxiety, Sleep Disorders, Schizophrenia, Hypertension, Sexual Dysfunction
	Alpha-2 adrenoceptor	8	Nasal Congestion, Glaucoma, Asthma, Migraine, Diarrhea
	Alpha-2C adrenoceptor	1	Mental Illnesses
	Beta adrenoceptor	6	Airway Inflammatory Disorders, Asthma, Obstructive Lung Disease, Ocular Hypertension, Glaucoma
	Beta-2 adrenoceptor	5	Pulmonary Disorders, Asthma, Emphysema, Neurological Disorders, Cardiac Disorders
	Beta-3 adrenoceptor	30	Metabolic Disorders, Atherosclerosis, Gastrointestinal Disorders, Type II Diabetes
HIV protease (58)			Retroviral Infection, Viral Infections (HIV), Viral Infections (EHV)
Serotonin receptors (43)	5-HT receptor	1	Headaches
	5-HT 1 receptor	7	Depression, Anxiety, Eating Disorders, Obesity, Drug Abuse, Cluster Headache, Migraine, Pain
	5-HT 1A receptor	4	Mood Disorders, Pain, Neuronal Disorders
	5-HT 1B receptor	4	Migraine, Depression, Psychological Disorders
	5-HT 1D receptor	6	Depression, Psychological Disorders
	5-HT 1F receptor	2	Headaches
	5-HT 2 receptor	3	Cardiovascular Disorders, Central Nervous System Disorders, Gastrointestinal Disorders, Glaucoma
	5-HT 2A receptor	5	Psychotic Disorders, Schizophrenia, Sleep-Disordered Breathing, Sleep Apnea Syndrome
	5-HT 2B receptor	2	Irritable Bowel Syndrome
	5-HT 2C receptor	6	Obesity, Obsessive Compulsive Disorder, Depression
	5-HT 3 receptor	8	Gastrointestinal Motility Disorders, Headache, Anxiety, Depression, Psychosis, Rheumatoid Disease
	5-HT 6 receptor	2	Hyperactivity Disorders, Attention Deficit Hyperactivity Disorder
	5-HT 7 receptor	4	CNS Disorders, Aforementioned Disorders, Disorders Of The Bladder, Urinary Retention
Factor Xa (47)			Thrombotic Disorders, Coronary Artery, Cerebro-Vascular Disease, Inflammatory Diseases, Cancers
Substance-P receptor (39)			Asthma, Cough, Bronchospasm, Depression, Emesis, Imflammatory Diseases, Gastrointestinal Disorders
Tyrosine kinases (39)	Tyrosine-protein kinase	28	Cancers, Atherosclerosis, Restenosis, Endometriosis, Psoriasis
	Tyrosine-protein kinase SRC	5	Immune Diseases, Cancers, Atheroscelerosis, Graft Rejection, Rheumatoid Arthritis
	Tyrosine-protein kinase JAK3	3	Allergic Disorders
	Tyrosine-protein kinase SYK	1	Inflammatory Diseases, Obstructive Airways Disease
	Tyrosine-protein	2	Cancers, Immune Diseases

	kinase BTK		
Cyclooxygenase 2 (38)			Alzheimer's Disease, Osteoporosis, Glaucoma, Inflammation, Asthma, Cancers, Heart Diseases
Thrombin (36)			Blood Coagulation, Cardiovascular Disorders, Thrombosis, Ischemia, Stroke, Restenosis, Inflammation
NMDA receptor (27)	NMDA receptor	14	Central Nervous System Disorders, Inflammatory Diseases, Allergic Diseases, Depression, Drug Abuse
	NMDA receptor NR2B	13	Pain, Migraine, Depression, Anxiety, Schizophrenia, Parkinson's Disease, Stroke
Opioid receptors (25)	opioid receptor	4	Eating Disorders, Narcotic Dependence, Alcoholism, Pain, Drug Dependence
	Mu-type opioid receptor	2	Constipation, Vomiting And/Or Nausea, Pain, Anxiety
	Delta-type opioid receptor	3	Central Nervous System Disorders, Peripheral Nervous System Diseases, Pain
	Kappa-type opioid receptor	16	Depression, Headaches, Inflammation, Arthritis, Stroke, Abdominal Pain, Pruritus
Inducible NOS (24)			CNS Disorders, Inflammation, Shock, Immune Disorders, Disorders Of Gastrointestinal Motility
Muscarinic receptors (22)	Muscarinic receptor	10	Cognitive Disorders, Alheimer's Disease, Neurologic, Psychiatric Disorders, Pain
	M1 receptor	3	Cognitive Disorders, Alheimer's Disease, Glaucoma
	M2 receptor	8	Cognitive Disorders, Alheimer's Disease, Smooth Muscle Disorders
	M3 receptor	5	Smooth Muscle Disorders
	M4 receptor	4	Mental Disorders, Parkinson's Disease, Glaucoma
Adenosine receptors (22)	Adenosine receptor	1	Cardiac And Circulatory Disorders, CNS Disorders, Respiratory Disorders
	A1 receptor	6	Allergic Disorders, CNS Disorders, Asthma
	A2a receptor	10	Central Nervous System Disorders, Parkinson's Disease
	A2b receptor	3	Airway Diseases, Asthma, Inflammation, Diabetes Mellitus
	A3 receptor	6	Bronchus Disorders, Inflammation, Allergosis
HIV RT (20)			Viral Infections (HIV)
PDE5 (19)			Cardiovascular And Cerebrovascular Disorders, Urogenital System Disorders, Erectile Dysfunction
Histamine receptors (16)	H1 receptor	8	Allergy, Rhinitis, Congestion, Inflammation, CNS Diseases, Respiratory Disorders, Viral Infections
	H2 receptor	6	Dry Eye, Duodenal Ulcer, Gastro-Esophogeal Reflux Disease, Gastrointestinal Disorders
	H3 receptor	5	Allergy, Congestion, Inflammation, CNS-Related Diseases
TNF (16)			Inflammatory Diseases, Allergic Diseases, Cytokine-Induced Toxicity, Muscular Disorders
Serotonin re-uptake (16)			CNS-Related Diseases, Anxiety
GnRH receptor (16)			Sex-Hormone Related Disorders, Steroid-Dependent Tumors, Prostate Cancer
Endothelin receptors (15)	Endothelin receptor	11	Angina, Pulmonary Hypertension, Raynaud's Disease, Migraine, Blood Vessel Disorders, Renal Diseases
	Endothelin A receptor	4	Hypertension, Acute Myocardial Infarct, Raynaud's Syndrome, Atherosclerosis, Asthma , Prostate Cancer
	Endothelin B receptor	2	Hypertension, Acute Myocardial Infarct, Stroke, Benign Prostate Hypertrophy, Atherosclerosis, Asthma
HMG-CoA reductase (13)			Atherosclerosis, Lipid Disorders, Hypercholesterolemia, Hypertriglyceridemia, Combined Hyperlipidemia
Gastric H+/K+ ATPase (12)			Bacterial Infections, Gastric Acid Related Diseases, Nasal Disorders, Bronchus Disorders, Osteoporosi

U-plasminogen activator (12)			Angiogenic Disorders, Arthritis, Inflammation, Osteoporotic, Cancers, Lymphomas, Chronic Dermal Ulcers
LHRH receptor (12)			Hormone-Dependent Tumours and Disorders, Benign Prostate Hyperplasia, Endometriosis
LH-RH (12)			Hormone-Dependent Tumours and Disorders, Benign Prostate Hyperplasia, Endometriosis
RARs (11)	Retinoic acid receptor	5	Acne, Psoriasis, Rheumatoid Arthritis, Viral Infections
	RAR-alpha	1	Systemic Erythematosus, Glomerulonephritis, Lupus Nephritis, Autoimmune Anemia
	RAR- gamma	5	Emphysema And Associated Pulmonary Diseases, Dermatological Disorders, Epithelial Lesions, Tumors
PPARs (10)	PPAR-alpha	4	Abnormality Of Lipidmetabolism, Type II Diabetes
	PPAR-gamma	5	Diabetes, Obesity, Metabolic Syndrome, Cardiovascular Diseases, Dysplidemia, Cancers
	PPAR-delta	2	Dyslipidemia, Syndrome X, Cardiovascular Diseases, Diabetes, Obesity, Anorexia Bulimia
Glycogen synthase kinase-3 (9)			Cancers, Diabetes, Alzheimer's Disease
Prostanoid FP receptor (9)			Bone Disorders, Glaucoma, Ocular Hypertension
calcium channel (9)			Cardiovascular Disorders, Angina, Hypertension, Ischemia
5-lipoxygenase (8)			Asthma, Atherosclerosis, Cancers
GP IIb/IIIa receptor (8)			Cancers, Osteoporosis, Arteriosclerosis, Restenosis, Opthalmia
DNA topoisomerases (7)	DNA Topo I	3	Cancers
	DNA Topo II	6	Bacterial Infections, Cancers
ACE (7)			Diabetic Complications, Diabetic Retinopathy, Diabetic Neuropathy, Diabetic Nephropathy
Glucocorticoid receptor (7)			Cocaine Addiction , Depression, Alzheimer's Disease, Aforementioned Diseases, Dibetes
serine protease (7)			Cardiovascular Disorders, Thrombosis, Asthma
Angiotensin II receptor (7)	AT1	6	Acute Myocardial Infarction, Cancers, Hypertension, Qt Dispersion
	AT2	6	Acute Myocardial Infarction, Cancers, Hypertension, Qt Dispersion, Wounds Healing
Estrogen receptors (6)	Estrogen receptor	2	Breast Cancer, Inflammatory Diseases, Sepsis, Viral Infections, Cardiovascular Diseases
	Estrogen receptor beta	4	Uterine Cancer, Adjuvant Breast Cancer, Prostate Cancer, Benign Prostatic Hyperplasia, Ovarian Cancers
Tryptase (6)			Cardiovascular Disorders, Inflammatory Diseases, Cancers
Dopamine receptors (5)	Dopamine receptor	2	Cancers, Parkinson's Disease
	D(2) receptor	2	Fibromyalgia, Musculoskeletal Pain Symptoms Associated With Fibromyalgia
	D(3) receptor	2	Fibromyalgia, Musculoskeletal Pain Symptoms Associated With Fibromyalgia
	D(4) receptor	1	Central Nervous System Disorders, Psychotic Disorders, Schizophrenia
Interleukin 1 receptor (5)	IL-1R	4	Hypotension, Tachycardia, Lung Edema, Renal Failure
	IL-1R-beta	1	Allergic Rhinitis, Allergic Asthma, Allergic Inflammatory Diseases
Neuraminidase (5)			Influenza A, Influenza B, Viral Infections, Bacterial Infections
Histone deacetylase (5)			Cancers, Hematological Disorders, Metabolic Disorders, Cystic Fibrosis, Adrenoleukodystrophy

Table 4-4: Research targets explored for the new investigational agents described in the US patents approved in 2000-2004.

Therapeutic Target		Number of US patents	Targeted Diseases
Protein	Subgroup		
MMPs (79)	Matrix metalloproteinase	62	Arthritis, Cancers, Tissue Ulceration, Periodontal Disease, Bone Disease, Diabetes
	MMP-1	1	Pulmonary Emphysema
	MMP-2	12	Cancers
	MMP-3	9	Multiple Sclerosis, Heart Failure, Cancers, Inflammation, Arthritis, Autoimmune Disorders
	MMP-4	1	Arthritis, Cancers
	MMP-7	1	Inflammatory Diseases, Rheumatoid Arthritis, Tumours
	MMP-8	1	Inflammatory Diseases, Cancers
	MMP-9	5	Cancers, Arthritis
	MMP-11	1	Cancers
	MMP-12	1	Ulcerative Colitis, Crohn's Disease, Atherosclerosis, Gastro-Intestinal Ulcers, Emphysema
	MMP-13	9	Osteoarthritis, Rheumatoid Arthritis, Cancers, Inflammation, Heart Failure
PDEs (78)	Phosphodiesterase	3	Erectile Dysfunction, Sexual Dysfunction
	PDE1A	2	Cardiovascular And Cerebrovascular Disorders, Erectile Dysfunction
	PDE2A	2	Cardiovascular And Cerebrovascular Disorders, Disorders of Urogenital System
	PDE3	5	Airway Obstructions, Inflammatory Diseases, Premature Ejaculation, Sexual Dysfunction
	PDE4	49	Airway Obstructions, Inflammatory Diseases, Allergic Disorders
	PDE4A	1	Respiratory Disorders, Asthma
	PDE7	4	Asthma, Rheumatoid Arthritis, Psoriasis, Atopic Dermatitis, Chronic Bronchitis
Alpha v integrin receptors (40)	Alpha v beta 3 integrin receptor	39	Cancers, Arteriosclerosis, Restenosis, Osteolytic Disorders, Osteoporosis, Ophthalmic Diseases
	alpha v beta 5 integrin receptor	16	Cancers, Osteoporosis, Arteriosclerosis, Restenosis, Opthalmia
Farnesyl-protein transferase (26)			Cancers, Restenosis, Atherosclerosis
ADAM 17 (25)			Arthritis, Tumor Metastasis, Tissue Ulceration, Bone Disease, Diabetes, HIV Infection
Cathepsin K (23)			Autoimmune Diseases, Cartilage Degradation, Osteoporosis, Pulmonary Disorders
Substance-K receptor (22)			Asthma, Cough, Bronchospasm, Depression, Imflammation, Gastrointestinal Disorders
Tachykinin NK(3) receptor (19)			CNS Disorders, Inflammation, Pain, Migraine, Asthma, Emesis, Gastrointestinal Disorders
Neuropeptide Y receptor (18)	Neuropeptide Y receptor	10	Eating Disorders, Feeding Disorders, Cardiovascular Disorders, Physiological Disorders
	NPY-Y5 receptor	8	Eating Disorders, Diabetes, Nutritional Disorders, Obesity
CDKs (17)	Cyclin-dependent kinase	12	Cancers, Inflammation, Arthritis, Alzheimer's Disease, Cardiovascular Disorders
	Cell division protein kinase 2	4	Alopecia, Cancers
	Cell division protein kinase 4	1	Cancers
Stress kinase p38			Chronic Inflammatory, Autoimmune Diseases,

(17)			Hypercholesterolemia
Hexokinase D (17)			Type II Diabetes
Phospholipase A2 (16)	Phospholipase A2	12	Inflammatory Diseases, Allergic Diseases, Pancreatitis, Septic Shock
	Cytosolic phospholipase A2	4	Inflammation, Asthma, Arthritis, Inflammatory Bowel Disease, Neurodegenerative Diseases
Cytochrome P450RAI (15)			Skin Diseases, Cancers, Cardiovascular Diseases, Inflammation, Neurodegenerative Diseases
Cathepsin S (14)			Osteoporosis, Autoimmune Disorders
Vasopressin receptor (13)	Vasopressin receptor	4	Cerebrovascular Disease, Cerebral Edema, Cerebral Infarction, Depressant, Anxiety
	Vasopressin V1a receptor	4	Obsessive-Compulsive Disorder, Aggressive Disorders, Depression, Anxiety
	Vasopressin V2 receptor	7	Diabetes Insipidus, Nocturnal Enuresis, Nocturia, Urinary Incontinence, Coagulation Disorders
Trypsin-like serine protease (13)			Thrombosis, Ischemia, Stroke, Restenosis, Inflammation
Interleukin-8 receptor (13)			Inflammation
Corticoliberin (12)			Circadian Rhythm Disorders, Congestive Heart Failure, Hypertension, Metabolic Disorders, Stroke
cathepsin B (12)			Autoimmune Diseases, Pancreatitis, Inflammatory Airway Disease, Bone And Joint Disorders
cathepsin L (12)			Autoimmune Diseases, Myocardial Infarct, Inflammation, Muscular Dystrophies, Alzheimer's Disease
Caspases (12)	Caspase	4	Cancers
	Caspase-8	8	Inflammation, Cancers, Autoimmune Disorders, Neuronal Disorders
	caspase-9	1	Inflammation, Cancers, Autoimmune Diseases, Ischemic Diseases, Neurodegenerative Disorders
CCRs (12)	CCR1	2	Inflammation, Immune Diseases
	CCR2	2	Atherosclerosis, Inflammatory Diseases, Immune Disorders, Transplant Rejection, Aids
	CCR3	3	Respiratory Disorders, Bronchus Disorders, Inflammatory Diseases, Allergy
	CCR5	5	Inflammatory Diseases, Viral Infections (HIV)
Prenyl-protein transferase (12)			Cancers
Prostaglandin E receptor (11)	Prostaglandin E receptor	3	Dry Eye, Keratoconjunctivitis, Sjogren's Syndrome, Ocular Surface Diseases, Glaucoma
	Prostanoid EP2 receptor	4	Ocular Hypotensive, Glaucoma, Mesangial Proliferative Glomerulonephritis
	Prostanoid EP4 receptor	4	Renal Failure, Dry Eye
PTP-1B (10)			Diabetes, Obesity, Autoimmune Diseases, Acute And Chronic Inflammation, Osteoporosis, Cancers
Serine/threonine protein kinase (10)	Serine/threonine protein kinase	2	Tumor Growth, Restenosis, Atherosclerosis, Cancers
	Serine/threonine protein kinase 12	8	Cancers, Diabetes, Alzheimer's Disease
Endothelin (9)	Endothelin	7	Angina, Pulmonary Hypertension, Raynaud's Disease, Migraine, Heart Failure, Respiratory Disorders
	Endothelin-1	2	Pulmonary Hypertension, Cerebral Infarction, Cerebral Ischemia, Congestive Heart Failure
Beta-lactamase (9)			Bacterial Antibiotic Resistance, Bacterial Infections
Metabotropic glutamate receptor (9)	mGLUR	7	Neurological Disorders, Psychosis, Schizophrenia, Alzheimer's Disease, Cognitive and Memory Disorders
	mGLUR1	1	Neurological Diseases, Neurodegenerative Diseases,

			Psychotic Diseases
	mGLUR5	1	Neurological Disorders, Psychiatric Disorders
Interleukin-1 beta convertase (8)			Inflammatory and Autoimmune Diseases, Bone Disorders, Proliferative Disorders, Infectious Diseases
GluRs (8)	Glutamate receptor, ionotropic kainate 1	3	Headaches, Neuronal Disorders
	Glutamate receptor AMPA	5	Epilepsy, Diseases Resulting In Muscle Spasm, Various Neurodegenerative Diseases, Stroke
Aldose reductase (8)			Diabetic Neuropathy, Diabetic Nephropathy, Diabetic Retinopathy, Diabetic Cardiomyopathy
Protease activated receptor 1 (8)			Aggregation Of Blood Platelets, Thrombosis, Thromboembolism, Myocardial Infarction

Many of the highly explored targets (those described in a large number of patents) are successful targets, which seem to indicate continuous effort and prolonged interest in the exploration of the targets of highly successful drugs for deriving new therapeutic agents. Successful targets described in a higher number of patents are adrenoceptor subtypes (63 distinct patents, 41 beta- and 22 alpha- subtypes, for cardiovascular diseases, depression, hypertension, asthma, diabetes, and obesity etc.), HIV protease (58 patents, for HIV infections), 5-HT receptor subtypes (43 distinct patents, 23 5-HT1, 16 5-HT2, 8 5-HT3 , 2 5-HT6 and 4 5-HT7 subtypes, for depression, anxiety, eating disorders, obesity, irritable bowel syndrome, attention deficit hyperactivity disorder, bladder disorder etc.), coagulation factor Xa (47 patents, for thromboembolic disorders), Substance-P receptor (39 targets, for asthma, bronchitis, migraine etc.), tyrosine kinases (39 patents, for angiogenic disorders, cancer, inflammatory diseases, allergic diseases etc.), cyclooxygenase 2 (38 patents, for inflammation, senile dementia, cancer, asthma, and congestive heart failure), thrombin (36 patents, for thrombosis, myocardial ischemia, myocardial infarction etc.), NMDA receptors (27 patents, for central nervous system disorders), opioid receptors (25 patents, for depression, pain, inflammation, arthritis, pruritus, alcohol and drug dependence etc.), inducible nitric oxide synthase (24 patents, for

inflammation, pain, arthritis, asthma, bronchitis etc.), muscarinic receptors (22 patents, for Alzheimer's disease, pain, glaucoma etc), and adenosine receptors (22 patents, for asthma, inflammation, diabetes, coronary artery disease, hepatic fibrosis, renal dysfunction etc.).

Research targets that are described in a higher number of patents are matrix metalloproteinase (79 patents, for cancers, tissue ulceration, abnormal wound healing, periodontal disease, bone disease, diabetes, arthritis, atherosclerosis, inflammation etc.), phosphodiesterase 4 (49 patents, for inflammation, asthma, prostate diseases, osteoporosis etc.), alpha v beta 3 integrin receptor (39 patents, for angiogenic disorders, inflammation, bone degradation, cancer, diabetic retinopathy, thrombosis etc.), farnesyl-protein transferase (26 patents, for arthropathies, arthritis, gout, cancers, restenosis etc.), ADAM 17 (25 patents, for arthritis, cancers, tissue ulceration, abnormal wound healing, periodontal disease, bone disease etc.), cathepsin K (23 patents, autoimmune diseases, cartilage degradation, osteoporosis, pulmonary disorders), and substance-K receptor (22 patents, for asthma, cough, bronchospasm, inflammatory diseases, arthritis, central nervous system disorders etc.).

4.2.2 Known targets of the FDA approved drugs in 2000-2004

Analysis of the known targets of recently approved drugs provides a useful hint about how therapeutic targets have been successfully explored. Drug discovery typically takes 10~15 years for a successful drug to move from the initial designing stage to the market [9, 28, 29]. Thus these targets also offer some picture about where some of efforts and resources have been directed by the pharmaceutical industry and research communities since the early 1990s.

Table 4-5: Known therapeutic targets of the FDA approved drugs in 2000-2004. There are a total of 66 targets targeted by 100 approved drugs

Therapeutic Target (Drug Action)			Number of FDA approved drugs	Drug Name	Targeted Diseases	Company
Protein	Subgroup	Action				
5-hydroxytryptamine receptor (11)	5-HT1A receptor	partial agonist	1	Abilify (aripiprazole)	Oral drug for the treatment of schizophrenia	Bristol-Myers Squibb and Otsuka America Pharmaceutical
	5-HT1B receptor	agonist	4	Axert (almotriptan malate) tablets	For the treatment of migraine attacks	Pharmacia
				Zomig-ZMT (zolmitriptan)	Orally disintegrating tablet for the treatment of acute migraine in adults	AstraZeneca
				Frova (frovatriptan succinate)	Tablets for the acute treatment of migraine attacks	Elan
				Relpax (eletriptan hydrobromide)	For the acute treatment of migraine headaches	Pfizer
	5-HT1D receptor	agonist	4	Axert (almotriptan malate) tablets	For the treatment of migraine attacks	Pharmacia
				Zomig-ZMT (zolmitriptan)	Orally disintegrating tablet for the treatment of acute migraine in adults	AstraZeneca
				Frova (frovatriptan succinate)	Tablets for the acute treatment of migraine attacks	Elan
				Relpax (eletriptan hydrobromide)	For the acute treatment of migraine headaches	Pfizer
	5-HT2 receptor	antagonist	2	Geodon (ziprasidone mesylate)	To control agitated behavior and psychotic symptoms in schizophrenia patients	Pfizer
				Ziprasidone (ziprasidone HCl)	Oral capsule for the treatment of schizophrenia	Pfizer
	5-HT3 receptor	antagonist	3	Aloxi (palonosetron)	For the prevention of nausea and vomiting associated with emetogenic cancer chemotherapy	MGI Pharma / Helsinn Healthcare
				Kytril (granisetron) Solution	For the prevention of nausea and vomiting associated with cancer therapy	Hoffmann-La Roche
				Lotronex (alosetron HCl) Tablets	Indicated for Irritable Bowel Syndrome (IBS) in females with diarrhea-predominant IBS	Glaxo Wellcome, Inc.
	5-HT4 receptor	agonist	1	Zelnorm (tegaserod maleate)	For the short-term treatment of irritable bowel syndrome in women whose primary bowel symptom is constipation	Novartis Pharmaceuticals
Adrenergic receptor (8)	Alpha-1	antagonist	1	UroXatral (alfuzosin HCl	For the treatment of the signs and symptoms of	Sanofi-Synthelabo

	receptor			extended-release tablets)	benign prostatic hyperplasia	
	Alpha-2 receptor	antagonist	1	Remeron SolTab (mirtazapine)	Orally disintegrating tablet for the treatment of depression	Organon
	Beta-1 receptor	blocker	2	Betapace AF Tablet (Sotalol)	For treatment of the irregular heartbeats in patients with atrial fibrillation	Berlex Laboratories, Inc.
				Betaxon (levobetaxolol)	For lowering IOP in patients with chronic open-angle glaucoma or ocular hypertension	Alcon Laboratories, Inc.
	Beta-2 receptor	agonist	4	DuoNeb (albuterol sulfate and ipratropium bromide)	For the treatment of bronchospasm associated with COPD	Dey Laboratories
				Foradil Aerolizer (formoterol fumarate inhalation powder)	Bronchodilator for COPD, asthma and bronchospasm	Novartis
				Ventolin HFA (albuterol sulfate inhalation aerosol)	For the treatment or prevention of bronchospasm	GlaxoSmithKline
				Xopenex (levalbuterol HCl)	For treatment of the reversible obstructive airway disease	Sepracor
Serotonin re-uptake		inhibitor	6	Cymbalta (duloxetine)	Depression	Eli Lilly
				Lexapro (escitalopram oxalate)	An orally administered selective serotonin reuptake inhibitor useful for the treatment for major depressive disorder	Forest Laboratories
				Paxil CR	Oral tablet for the treatment of depression and panic disorder	GlaxoSmithKline
				Prozac Weekly (fluoxetine HCl)	For the treatment of depression	Eli Lilly and Company
				Ultracet (acetaminophen and tramadol HCl)	For the short-term management of acute pain	Ortho-McNeil Pharmaceutical
				Zoloft (sertraline HCl)	Oral tablets for the treatment of premenstrual dysphoric mood disorder (PMDD)	Pfizer
Coagulation Factor (5)	Thrombin	inhibitor	3	Angiomax (bivalirudin)	As an anticoagulant in conjunction with aspirin in patients with unstable angina undergoing percutaneous transluminal coronary angioplasty	Medicines Company
				Argatroban Injection	Anticoagulant for prophylaxis or treatment of thrombosis in patients with heparin-induced thrombocytopenia.	Texas Biotechnology Corporation and SmithKline

						Beecham
				Innohep (tinzaparin sodium) injectable	For the treatment of acute symptomatic deep vein thrombosis	Dupont Pharmaceuticals CO.
	Factor Va	inhibitor	1	Xigris (drotrecogin alfa [activated])	For the treatment of severe sepsis	Eli Lilly
	Factor VIIIa	inhibitor	1	Xigris (drotrecogin alfa [activated])	For the treatment of severe sepsis	Eli Lilly
	Factor Xa	inhibitor	2	Arixtra	Injectable solution for the prevention of deep vein thrombosis	Organon and Sanofi-Synthelabo
				Innohep (tinzaparin sodium) injectable	For the treatment of acute symptomatic deep vein thrombosis	Dupont Pharmaceuticals CO.
Muscarinic acetylcholine receptor (5)	(non-selective)	antagonist	2	Detrol LA (tolterodine tartrate)	For the treatment of overactive bladder with symptoms of urge urinary incontinence, urgency and frequency	Pharmacia and UpJohn
				Sanctura (trospium chloride)	Overactive bladder	Indevus Pharmaceuticals
	M1 receptor	agonist	1	Evoxac (cevimeline HCl)	For the treatment of symptoms of dry mouth in patients with Sjogren's Syndrome	SnowBrand Pharmaceuticals
	M3 receptor	antagonist	1	Vesicare (solifenacin succinate)	For the treatment of overactive bladder with symptoms of urge urinary incontinence	Yamanouchi, GlaxoSmithKline
		agonist	1	Evoxac (cevimeline HCl)	For the treatment of symptoms of dry mouth in patients with Sjogren's Syndrome	SnowBrand Pharmaceuticals
		inhibitor	1	Spiriva HandiHaler (tiotropium bromide)	Chronic Obstructive Pulmonary Disease (COPD)	Boehringer Ingelheim
3-hydroxy-3-methylglutaryl-coenzyme A reductase		inhibitor	5	Advicor (extended-release niacin/lovastatin)	For the treatment of cholesterol disorders	Kos Pharmaceuticals
				Altocor (lovastatin) Extended-Release Tablets	Oral tablets for the adjuctive treatment of hypercholesterolemia	Andrx
				Caduet (amlodipine/atorvastatin)	Hypertension/Angina	Pfizer
				Crestor (rosuvastatin calcium)	For the treatment of primary hypercholesterolemia (heterozygous familial and nonfamilial) and mixed dyslipidemia	AstraZeneca

				Lescol XL (fluvastatin sodium) tablet, extended release	For the use as an adjunct to diet to reduce elevated total cholesterol	Novartis Pharmaceuticals Corp.
Cyclooxygenase (5)	COX	inhibitor	3	Bayer Extra Strength Aspirin	Mild to moderate migraine pain	Bayer Corporation
				Children's Motrin Cold	Common cold	McNeil Consumer Healthcare
				Mobic (meloxicam)	Osteoarthritis	Boehringer Ingelheim Pharmaceuticals Inc
	COX-2	inhibitor	2	Bextra	Oral tablet for the treatment of osteoarthritis, rheumatoid arthritis and menstrual pain	Pharmacia and Pfizer
				Vioxx (rofecoxib)	For the treatment of rheumatoid arthritis	Merck
Dopamine receptor (4)	D(1B) receptor	agonist	1	Apokyn (apomorphine HCl)	Parkinson's Disease	Mylan Bertek Pharmaceuticals
	D(2) receptor	partial agonist	1	Abilify (aripiprazole)	Oral drug for the treatment of schizophrenia	Bristol-Myers Squibb and Otsuka America Pharmaceutical
		agonist	1	Apokyn (apomorphine HCl)	Parkinson's Disease	Mylan Bertek Pharmaceuticals
		antagonist	2	Geodon (ziprasidone mesylate)	To control agitated behavior and psychotic symptoms in schizophrenia patients	Pfizer
				Ziprasidone (ziprasidone HCl)	Oral capsule for the treatment of schizophrenia	Pfizer
	D(3) receptor	agonist	1	Apokyn (apomorphine HCl)	Parkinson's Disease	Mylan Bertek Pharmaceuticals
	D(4) receptor	agonist	1	Apokyn (apomorphine HCl)	Parkinson's Disease	Mylan Bertek Pharmaceuticals
Noradrenergic re-uptake		inhibitor	4	Cymbalta (duloxetine)	Depression	Eli Lilly
				Ritalin LA (methylphenidate HCl)	Oral capsules for the treatment Attention-Deficit/Hyperactivity Disorder (ADHD)	Novartis Pharmaceuticals
				Strattera (atomoxetine HCl)	For the treatment of Attention-Deficit/Hyperactivity Disorder (ADHD) in children, adolescents and adults.	Eli Lilly

				Ultracet (acetaminophen and tramadol HCl)	For the short-term management of acute pain	Ortho-McNeil Pharmaceutical
DNA topoisomerase (3)	DNA TopII	inhibitor	3	Avelox I.V. (moxifloxacin HCl)	Injectable antibacterial agent for adults with susceptible strains of bacterial infections	Bayer
				Novantrone (mitoxantrone HCl)	For reducing neurologic disability and/or the frequency of clinical relapses in patients with multiple sclerosis	Immunex Corporation
				Quixin (levofloxacin)	For treatment of bacterial conjunctivitis	Santen
	DNA TopIV	inhibitor	2	Avelox I.V. (moxifloxacin HCl)	Injectable antibacterial agent for adults with susceptible strains of bacterial infections	Bayer
				Quixin (levofloxacin)	For treatment of bacterial conjunctivitis	Santen
Gonadotropin-releasing hormone (3)	GnRH	agonist	2	Eligard (leuprolide acetate)	For the palliative treatment of advanced prostate cancer	Atrix Laboratories
				Viadur (leuprolide acetate implant)	For pain relief in men with advanced prostate cancer	ALZA Corporation
		antagonist	1	Plenaxis (abarelix for injectable suspension)	For treatment of advanced prostate cancer	Praecis Pharmaceuticals
HIV-1 protease		inhibitor	3	Kaletra Capsules and Oral Solution	For the treatment of HIV-1 infection	Abbott Laboratories
				Lexiva (fosamprenavir calcium)	For the treatment of HIV infection in adults in combination with other antiretroviral agents.	GlaxoSmithKline
				Reyataz (atazanavir sulfate)	For the treatment of HIV-1 infection in combination with other antiretroviral agents	Bristol-Myers Squibb
HIV-1 reverse transcriptase		inhibitor	3	Sustiva	Once-daily oral tablet for the the treatment of HIV infection	Bristol-Myers Squibb
				Trizivir (abacavir sulfate; lamivudine; zidovudine AZT) Tablet	For the treatment of HIV-1 infection	Glaxo Wellcome
				Viread	Once-daily oral tablet for the treatment of human immunodeficiency virus (HIV) infection	Gilead Sciences
Proton pump		inhibitor	3	Aciphex (rabeprazole sodium)	For the treatment of symptomatic gastroesophageal reflux disease	Eisai
				Nexium (esomeprazole magnesium)	For the eradication of Helicobacter pylori, the healing of erosive esophagitis, and the treatment	AstraZeneca

					of symptomatic GERD	
				Protonix (pantoprazole sodium) Intravenous Formulation; Delayed Release Tablets	For the short-term treatment of gastroesophageal reflux disease; Oral tablets for the treatment of gastroesophageal and pathological hypersecretory conditions	Wyeth-Ayerst Laboratories; Wyeth Pharmaceuticals
Epidermal growth factor receptor (3)		inhibitor	2	Erbitux (cetuximab)	Colorectal Cancer	Imclone, Bristol-Myers Squibb
				Iressa (gefitinib)	For the second-line treatment of non-small-cell lung cancer	AstraZeneca
	HER1/EGFR	inhibitor	1	Tarceva (erlotinib)	For the treatment of advanced refractory metastatic non-small cell lung cancer	Genentech, OSI Pharmaceuticals
Angiotensin II receptor (3)	AT1 receptor	blocker	3	Benicar	Oral tablet for the treatment of hypertension	Forest Laboratories
				Diovan	Oral capsules and tablets for the treatment of hypertension	Novartis
				Teveten HCT (eprosartan mesylate/hydrochlorothiazide)	Tablets for the treatment of hypertension	Unimed Pharmaceuticals
Opioid receptor (2)	Kappa opioid receptor	antagonist	1	Subutex/Suboxone (buprenorphine/naloxone)	Oral tablets for the treatment of opiate dependence	Reckitt Benckiser
	Mu opioid receptor	partial agonist/ antagonist	1	Subutex/Suboxone (buprenorphine/naloxone)	Oral tablets for the treatment of opiate dependence	Reckitt Benckiser
Histamine receptor (2)	H1 receptor	antagonist	1	Clarinex	Once-daily oral tablet for the treatment of allergic rhinitis and chronic ideopathic urticaria	Schering-Plough
	H2 receptor	antagonist	1	Pepcid Complete	For use in the relief of heartburn associated with acid indigestion and sour stomach	Merck Research Laboratories
CGMP-specific 3',5'-cyclic phosphodiesterase	PDE5	inhibitor	2	Cialis (tadalafil)	Oral agent for the treatment for erectile dysfunction	Eli Lilly
				Levitra (vardenafil)	For the treatment of erectile dysfunction related to sexual activity in men	Bayer / GlaxoSmithKline
Vascular endothelial growth factor (2)		antagonist	1	Macugen (pegaptanib)	For the treatment of wet age-related macular degeneration	Pfizer/Eyetech Pharmaceuticals

		binder	1	Avastin (bevacizumab)	Colorectal Cancer	Genentech
Dihydrofolate reductase		inhibitor	2	Alimta (pemetrexed for injection)	Mesothelioma	Eli Lilly
				Malarone (atovaquone; proguanil HCl)	For the treatment and prevention of Plasmodium falciparum malaria	Glaxo Wellcome
3-oxo-5-alpha-steroid 4-dehydrogenase (1)	Steroid 5-alpha -reductase 1	inhibitor	1	dutasteride	For the treatment of symptomatic benign prostatic hyperplasia	GlaxoSmithKline
	Steroid 5-alpha -reductase 2	inhibitor	1	dutasteride	For the treatment of symptomatic benign prostatic hyperplasia	GlaxoSmithKline
4-hydroxyphenylpyruvate dioxygenase		inhibitor	1	Orfadin (nitisinone)	Capsules for the treatment of hereditary tyrosinemia type I	Orphan Pharmaceuticals
1,3-Beta-Glucan synthase		inhibitor	1	Cancidas	Intravenous infusion for the treatment of invasive aspergillosis	Merck & Co.
23S rRNA		binder	1	Ketek (telithromycin)	Respiratory Infections	Aventis Pharmaceuticals
Acetylcholinesterase		inhibitor	1	Reminyl (galantamine hydrobromide)	For the treatment of mild to moderate dementia of the Alzheimer's type	Janssen Research
Alpha-1-antitrypsin		inhibitor	1	Zemaira (alpha1-proteinase inhibitor)	For the treatment of alpha1-proteinase inhibitor deficiency (Alpha-1) and emphysema	Aventis Behring
B-lymphocyte antigen CD20		antigen	1	Bexxar	For the treatment of patients with CD20 positive, follicular, non-Hodgkin's lymphoma following chemotherapy relapse	Corixa
Calcineurin		inhibitor	1	Elidel	Topical cream for the treatment of atopic dermatitis	Novartis
Cholinesterase		inhibitor	1	Exelon (rivastigmine tartrate)	Indicated for the treatment of mild to moderate dementia of the Alzheimer's type	Novartis Pharmaceuticals Corporation
Cytochrome P450 19		inhibitor	1	Femara (letrozole)	First-line treatment of postmenopausal women with locally advanced or metastatic breast cancer	Novartis
Collagenase		inhibitor	1	Periostat (doxycycline hyclate)	Oral tablet for adjunctive treatment of adult periodontitis	Collagenex Pharmaceuticals

DNA polymerase		inhibitor	1	Hepsera (adefovir dipivoxil)	For the treatment of chronic hepatitis B in adults with evidence of active viral replication	Gilead Sciences
Endothelin receptor		antagonist	1	Tracleer (bosentan)	For the treatment of pulmonary arterial hypertension	Actelion
Estrogen receptor		antagonist	1	Faslodex (fulvestrant)	For the treatment of hormone receptor positive metastatic breast cancer	AstraZeneca
Glycinamide ribonucleotide formyltransferase		inhibitor	1	Alimta (pemetrexed for injection)	Mesothelioma	Eli Lilly
Interleukin-1		blocker	1	Kineret	Injectable therapy for the treatment of rheumatoid arthritis	Amgen
NMDA receptor		antagonist	1	Namenda (memantine HCl)	For the treatment of moderate to severe dementia of the Alzheimer	Forest Laboratories
Peroxisome proliferator activated receptor gamma		agonist	1	Avandamet (rosiglitazone maleate and metformin HCl)	For improvement of glycemic control in type 2 diabetes patients	GlaxoSmithKline
Prostaglandin F2-alpha receptor		agonist	1	Travatan (travoprost ophthalmic solution)	For the reduction of elevated intraocular pressure in patients with open-angle glaucoma or ocular hypertension	Alcon
Dopamine reuptake		blocker	1	Ritalin LA (methylphenidate HCl)	Oral capsules for the treatment Attention-Deficit/Hyperactivity Disorder (ADHD)	Novartis Pharmaceuticals
Substance-P receptor		antagonist	1	Emend (aprepitant)	For the treatment of nausea and vomiting associated with chemotherapy	Merck
Thymidylate synthase		inhibitor	1	Alimta (pemetrexed for injection)	Mesothelioma	Eli Lilly
Tumor necrosis factor		inhibitor	1	Remicade (infliximab)	For inhibiting the progression of structural damage in patients with rheumatoid arthritis; Intravenous infusion for the treatment of rheumatoid arthritis	Centocor
Tyrosine-protein kinase		inhibitor	1	Gleevec (imatinib mesylate)	Oral therapy for the treatment of chronic myeloid leukemia;For the treatment of gastrointestinal stromal tumors (GISTs)	Novartis
Ribonucleotide reductase		inhibitor	1	Clolar (clofarabine)	For the treatment of acute lymphoblastic leukemia in pediatric patients	Genzyme

Table 4-5 gives the known targets of the approved drugs by the United States FDA in 2000-2004 together with the corresponding drug name, drug action, targeted diseases and the drug inventing/marketing company. There are a total of 66 identifiable targets that are targeted by 100 distinct drugs approved during the period. Most of these targets are for antagonist/inhibitor drugs, only 17 are for agonist/partial agonist drugs. The significantly smaller number of agonist drugs is likely due in part to the higher level of difficulty in finding agonist drugs. Agonist drugs generally require more specific binding configuration than that of antagonist/inhibitor drugs. Some targets, such as 5-HT receptors and adrenoceptors, are targeted by both agonist and antagonist drugs for the treatment of different diseases.

There are 90 drugs, which constitutes 43% of the total number of approved drugs, without identifiable target described in the FDA documents. Some of these drugs are protein-based, peptide-based, or gene-therapy-based agents whose target is not specifically mentioned. Some drugs, such as trileptal (oxcarbazepine) and zonegran (zonisamide), were discovered without the knowledge of their precise molecular mechanism at the time of their filing. Trileptal is known to have blocking effects on voltage sensitive sodium and calcium channels [178]. Zonegran activates dopamine synthesis and moderately inhibits monoamine oxidase [179]. It also inhibits carbonic anhydrase, modulates GABA A receptor, and exerts blocking effects on voltage sensitive sodium and calcium channels [178]. It remains unclear how these drugs affect these proteins and which of these actions directly contribute to their therapeutic effects.

The reported mechanism of a number of drugs, such as Plavix (clopidogrel bisulfate) and Rapamune (sirolimus), was not specific enough to point to a particular target at

the time of their filing. It is noted that the mechanism of some of these drugs has since been determined. For instance, it has been reported that plavix inhibits ADP-induced platelet aggregation because one of its metabolite antagonizes platelet ADP receptor P2Y [180]. It has been found that rapamune binds to and forms a complex with cytosolic FKBP-12, which inhibits the protein kinase mTOR and thereby produces its antifungal, antiproliferative, and immunosuppressive activities [181].

Most of the 66 identifiable targets of the FDA approved drugs during the period are also targeted by drugs marketed before 2000. Given that there were ~120 known targets of marketed drugs in the previous reports [9, 37, 135], it appears that the majority of the known successful targets have been continuously explored for deriving new therapeutic agents. The targets with larger number of drugs approved during the period are 5HT receptors with 11 drugs, adrenoceptors with 8 drugs, and serotonin reuptake with 6 drugs. Moreover, coagulation factor, muscarinic acetylcholine receptor, HMG-CoA reductase, and cyclooxygenase are targeted by 5 drugs each. Dopamine receptor and noradrenergic reuptake are targeted by 4 drugs each. The other 7 targets, such as DNA topoisomerase, gonadotropin-releasing hormone, HIV-1 protease, HIV-1 reverse transcriptase, proton pump, epidermal growth factor receptor, and angiotensin II receptor are targeted by 3 drugs each. In addition, another 5 targets namely opioid receptor, histamine receptor, phosphodiesterase, vascular endothelial growth factor and dihydrofolate reductase are targeted by 2 drugs each. These targets represent highly successful targets that have been extensively explored for deriving new therapeutic agents.

There are 12 targets that are targeted by subtype-specific drugs, representing 18.2% of the total number of identifiable targets of the FDA approved drugs during the period,

which suggests that substantial efforts have been directed at the discovery of subtype specific drugs since the early 1990s and these efforts have led to the success. These targets include phosphodiesterase 5 inhibitors for the treatment of erectile dysfunction [144], cyclooxygenase 2 inhibitors for arthritis and menstrual pain [182], tumor necrosis factor alpha blockers for rheumatoid arthritis, dopamine receptor D2 agonists for schizophrenia, 5-HT2 receptor antagonists for schizophrenia, adrenoceptor alpha1 antagonist for hyperplasia, and histamine receptor H2 antagonist for heartburn [183].

According to the literature, a total of 16 innovative targets emerged since1996 [38]. Examples of these targets are receptor protein-tyrosine kinase erbB-2 (HER2/neu) with the first drug Trastuzumab approved in 1998 for HER2 positive metastatic breast cancer, *BCR-ABL* tyrosine kinase with the first drug Celecoxib approved in 2001 for chronic myeloid leukemia, vascular endothelial growth factor with the first drug Bevacizumab approved in 2004 for colorectal cancer, HBV DNA polymerase with the first drug Adefovir dipivoxil approved in 2002 for hepatitis B, HIV gp41 with the first drug Enfuvirtide approved in 2003 for HIV infection, NMDA receptor with the first drug Memantine approved in 2003 for Alzheimer's disease, platelet P2Y12 receptor with the first drug Clopidogrel approved in 1997 for stroke and heart attack, platelet glycoprotein IIb/IIIa receptor with the first two drugs Tirofiban and Eptifibatide approved in 1998 for severe chest pain and small heart attacks, endothelin receptor with the first drug Bosentan approved in 2001 for primary pulmonary hypertension, mineralocorticoid receptor with the first drug Eplerenone approved in 2002 for hypertension, phosphodiesterase 5 with the first drug Sildenafil approved in 1998 for erectile dysfunction, cyclooxygenase 2 with the first drug Celecoxib approved in1998 for arthritis, gastrointestinal lipase with the first drug Orlistat approved in 1999 for obesity, and peroxisome proliferator activated receptor with the first drug

Troglitazone approved in 1997 for type 2 diabetes mellitus.

4.2.3 Progress and difficulties of target exploration

Some of these highly explored research targets have been used for drug development well before 2000. Great progresses have been made towards the discovery and testing of agents directed at these targets. However, for some of these targets, many difficulties remain to be resolved before viable drugs can be derived. The appearance of a high number of patents associated with these targets partly reflects the intensity of efforts to find effective drug candidates against these targets.

Farnesyl-protein transferase inhibitors have been designed and tested as novel agents for the treatment of myeloid malignancies since the early 1990s [184]. Initially developed to inhibit the prenylation necessary for Ras activation, their mechanism of action seems to be more complex, involving other proteins unrelated to Ras. Preliminary results from clinical trials demonstrated inhibition of enzyme target, a favorable toxicity profile and promising efficacy [185]. This led to the initiation of phase II trials in a variety of hematologic malignancies and disease settings [186].

Phosphodiesterase-4 (PDE4) has been explored as the target of novel anti-inflammatory agents since the mid 1990s [187]. The rationale for selecting this target is, in part, from the clinical efficacy of theophylline, an orally active nonselective PDE inhibitor. It has been found that intracellular cyclic adenosine monophosphate levels regulate the function of many of the cells thought to contribute to the pathogenesis of respiratory diseases such as asthma and COPD, and these cells also selectively express PDE4 [188]. Recent clinical studies of selective PDE4 inhibitors such as cilomilast and roflumilast used for the treatment of inflammatory

lung disease showed positive results that offered some optimism, and efforts were being made to reduce the side effect of these drug candidates [188].

Matrix metalloproteinases (MMPs) have been targeted for cancer and other diseases since the early 1990s [189]. MMPs degrade the extracellular matrix, promote tumor invasion and metastasis, and regulate host defense mechanisms and normal cell function. Blocking all MMPs may not lead to a positive therapeutic outcome. So far, most clinical trials of MMP inhibitors have not yielded good results, due primarily to the lack of subtype selectivity, bioavailability and efficacy, and in some cases inappropriate study design [190]. Intensive efforts are being directed at the discovery of potent, selective, orally bioavailable MMP inhibitors for the treatment of cancer. There have been encouraging news about some inhibitors, such as ABT-518, that have entered in Phase I clinical trials in cancer patients [191].

Intensive research efforts have been directed at developing beta 3-adrenergic receptor (beta3-AR) selective agonists for the treatment of type 2 diabetes and obesity in humans since early 1990s [192]. These agonists have been observed to simultaneously increase lipolysis, fat oxidation, energy expenditure and insulin action leading to the belief that this receptor might serve as an attractive target for the treatment of diabetes and obesity. However, drug design efforts have been hindered by the obstacles in the pharmacological differences between the rodent and human beta3-AR, the lack of selectivity of leads, and unsatisfactory oral bioavailability and pharmacokinetic properties of tested agents [193]. A recent test of beta3-AR agonists directed at the human receptor showed promising results in their ability to increase energy expenditure in humans following a single dose. However, they do not appear to be able to sustain their effects when administered chronically. Further clinical testing will

be necessary, using compounds with improved oral bioavailability and potency, to help assess the physiology of the beta3-AR in humans and its attractiveness as a potential therapeutic for the treatment of type 2 diabetes and obesity [193].

Inspection of the targets reported in these patents also provides useful indication about the progress for the search of new targets. Examples of newly explored targets are 88 kDa glycoprotein growth factor for the treatment of cancer (US patent 6,670,183), anandamide amidase for pain (US patent 6,579,900), FK506-binding protein 4 for neurological disorders (US patent 6,495,549), galanin receptor type 2 for CNS disorder (US patent 6,407,136), gamma secretase for Alzheimer's disease (US patent 6,448,229), glycogen synthase kinase-3 beta for diseases characterized by an excess of Th2 cytokine (US patent 6,479,490), orexin receptor 1 for obesity (US patent 6,677,354), and tripeptidyl-peptidase II for eating disorder and obesity (US patent 6,335,360). Most of these new research targets are explored for the treatment of high impact diseases needing effective or more treatment options.

4.2.4 Targets of subtype specific drugs

There are 62 targets explored for the design of subtype-specific drugs, which represents 15.7% of the 395 identifiable targets in the US patent approved in 2000-2004. Compared with the 12 targets of FDA approved subtype-specific drugs during the same period, a significantly larger number of targets are being explored for the design of subtype-specific drugs. However, the percentage of these targets with respect to the total number of targets in the US patent is smaller than that of the FDA approved drugs during the same period, which seems to indicate the level of difficulty of finding subtype-specific agents directed at a variety of targets. For instance,

although there are 79 patents for matrix metalloproteinase (MMP), only 3 patents describe subtype-specific investigational drugs. These are MMP-9 inhibitors (US patent 6,667,388), MMP-4 inhibitors (US patent 6,544,761), and MMP-13 inhibitors (US patent 6,656,932).

The targets with a higher number of patents of subtype-specific investigational drugs are phosphodiesterase 4 with 49 patents (for the treatment of asthma, inflammation and osteoporosis), cyclooxygenase 2 with 38 patents (inflammation, cancer and others), adrenoceptor beta with 41 patents (hyperglycemia, obesity, gastrointestinal disorders and others), adrenoceptor alpha with 22 patents (hypertension, pain, gastric ulcers, vascular diseases and others), phosphodiesterase 5 with 19 patents (sexual dysfunction), cytochrome P450RAI with 15 patents (diseases responsive to retinoid treatment), 5-HT1 receptor with 17 patents (depression, eating disorders, obesity, headache and others), 5-HT2 receptor with 12 patents (irritable bowel syndrome), 5-HT3 receptor with 8 patents (blood glucose control), and 5-HT7 receptor with 4 patents (bladder disorder and urinary retention).

4.3 Characteristics of therapeutic targets

4.3.1 What constitutes a therapeutic target?

The majority of clinical drugs achieve their effect by binding to a cavity, and modifying the activity, of its protein target. Specific structural and physicochemical properties, such as the "rule-of-five[†]" [194], are required for these drugs to have a sufficient level of efficacy, bioavailability and safety, which define target sites to

[†] "Rule-of-five" was firstly introduced by Lipinski in 1997. It has become an awareness tool for discovery chemists. Compounds with two or more of the following characteristics are flagged as likely to have poor oral absorption: 1) More than 5 H-bond donors; 2) Molecular weight >500; 3) *c* log *P*>5; 4) Sum of N's and O's (a rough measure of H-bond acceptors) > 10.

which drug-like molecules can bind. In most cases, these sites exist out of functional necessity, and their structural architectures accommodate target-specific drugs that minimally interact with other functionally important but structurally similar sites. These constraints limit the types of proteins that can be bound by drug-like molecules, leading to the introduction of the concept of druggable proteins [37, 195]. Druggable proteins do not necessarily become therapeutic targets [37], only those that play key roles in diseases can be explored as potential targets. Nonetheless, analysis of the characteristics of these druggable proteins is useful for facilitating molecular dissection of the mechanism of drug targeting and for guiding new targets searching.

Certain characteristics are expected for therapeutic targets [37]. These targets play critical and preferably un-substitutable roles in disease processes. They have certain functional and structural novelty to allow for drug specificity. They are not significantly involved in other important processes in humans to limit potential side-effects. Expression of these targets is either at a constrained level or tissue selective to allow for sufficient drug efficacy. Drug-binding sites are expected to have certain structural and physicochemical properties to accommodate high-affinity site-specific binding and subsequent modification of protein activity by drug-like molecules. These characteristics likely define the sequence features, structural architectures, genomic signatures, and proteomic profiles of therapeutic targets and their roles at the pathway, cellular and physiological levels. Useful hints about some of the characteristics of therapeutic targets may be probed by analyzing their sequence properties, protein families, structural folds, biochemical classes, similarity proteins, gene locations in human genome, associated pathways. These hints may be potentially used for deriving rules and developing predictive tools for searching druggable proteins from genomic data. As part of the effort for supporting such a goal, relevant

features of 268 successful targets and 1267 research targets are described.

4.3.2 Protein families represented by therapeutic targets

The sequence and functional similarities within a protein family usually indicates general conservation of binding site architecture between family members. If a drug can specifically target one member of a family, then it is possible to design molecules with similar physicochemical properties for specific binding to some of the other members of the family, and multiple members of a family have been explored for developing drugs of different therapeutic applications [196, 197]. A recent analysis of the identifiable drug-binding domains of 399 targets (including 120 successful targets) suggested that these targets are represented by 130 protein families, nearly half of which are represented by 6 families [196], which indicate the level of extensive exploration of multiple members of specific families as therapeutic targets.

With the availability of the information of a significantly higher number of targets than that used in the recent analysis, it is of interest to re-investigate family representations of therapeutic targets. There are 173 successful targets and 906 research targets with identifiable drug-binding domain. Analysis of the Pfam [198] protein family of these domains finds that these targets are represented by 92 and 412 families respectively.

About 42% of the 173 successful targets fall into 10 families. These, in terms of Pfam family names, are 7 transmembrane receptor rhodopsin family (32 targets), nuclear hormone receptor (11 targets), protein kinase (5 targets), short chain dehydrogenase (4 targets), amino acid permease (4 targets), cytochrome P450 (4 targets), neurotransmitter-gated ion-channel l (4 targets), sodium: neurotransmitter symporter

(3 targets), reverse transcriptase (3 targets), and ion transport protein (3 targets).

About 40% of the 906 research targets fall into 26 families, which include 7 transmembrane receptor rhodopsin family (94 targets), protein kinase (87 targets), immunoglobulin (29 targets), trypsin (21 targets), nuclear hormone receptor (16 targets), receptor family ligand binding region (12 targets), papain family cysteine protease (11 targets), matrixin (10 targets), small cytokines (9 targets), 3'5'-cyclic nucleotide phosphodiesterase (8 targets), neurotransmitter-gated ion-channel (7 targets), subtilase family (7 targets), ABC transporter (7 targets), prolyl oligopeptidase family (6 targets), eukaryotic-type carbonic anhydrase (6 targets), short chain dehydrogenase (6 targets), eukaryotic aspartyl protease (6 targets), ZIP transcription factor (6 targets), TNFR/NGFR (5 targets), ion transport protein (5 targets), peptidase family (5 targets), Reprolysin (M12B) family zinc metalloprotease (5 targets), sugar transporter (5 targets), and hormone receptor (5 targets).

Overall, 40% or 436 of the 1,079 successful and research targets are distributed in 26 protein families, which include all of the 6 top target-representing families found in the recent study [196]. The remaining 60% or 643 targets are distributed in 434 families. There are 6 families both in the top 10 families of successful targets and top 26 families of the research targets. These are 7 transmembrane receptor rhodopsin families, ligand-binding domain of nuclear hormone receptor, protein kinase domain, short chain dehydrogenase, neurotransmitter-gated ion-channel ligand binding, and ion transport protein.

Two parallel lines of target exploration are indicated. One is the extensive use of successful targets and additional members of a relatively small group of protein

families. On average, 17 targets from each of the 26 heavily-used families have been explored. The other is the exploration of a diverse range of proteins in a variety of families. On average, only 1 or 2 targets from each of the other 434 protein families have been explored or are being evaluated. It is expected that more members from some of these families may be used as viable targets.

It is of interest to estimate the total number of families that represent all of the 3,000 targets that are postulated to exist. Assuming that all of the 1,535 currently explored targets are viable ones, which is doubtful but not significantly affect our estimate, there are ~1,500 un-discovered targets. If these un-discovered targets roughly follow the same pattern of protein family representation of the currently explored targets, it is expected that 40% of them are from a relatively small group of families, probably no more than a few dozen. Moreover, the bulk, say 60%, of the remaining 60% of these targets is likely from the 434 families that represent 60% of the currently explored targets. Therefore, there are no more than 24% of the un-discovered targets that are from protein families not represented by the known targets, and these targets are represented by no more than 480 families. This gives a crude estimate of no more than 940 of target-representing protein families, likely to be substantially less, for all of the therapeutic targets. The total number of protein families in Pfam database is 7677 [198]. Thus target-representing families account for less than 12% of all protein families, and 40% of the targets are expected to be represented by just a few dozen families.

4.3.3 Structural folds

A common feature of targets in a particular family is the general conservation of

binding site architecture. Binding sites of drugs are usually located within specific cavity of their target proteins, and drug binding is primarily facilitated by hydrophobic, aromatic stacking, hydrogen bonding, and van der Waals interactions [199]. Certain constrains on the architectures of drug-binding domains are expected for accommodating the binding of target-specific "rule of five" small molecules that minimally interact with other functionally important but structurally similar sites. There have been reports about specific drug-domain architecture [200-202].

Because of the distribution of therapeutic targets in a relatively small number of protein families, it is expected that these targets are represented by a relatively small number of structural folds. Examination of the structural folds of the drug-binding domains can therefore shed light on the structural characteristics of therapeutic targets. Structural folds of proteins can be obtained from the SCOP database [203], which contains 701 structural folds generated from the analysis of 1,7406 protein entries from the PDB database [204]. There are 52 successful targets that have both available 3D structure and identifiable drug binding domain. Analysis of the SCOP structural folds of these targets shows that they are represented by 29 folds, which is given in Table 4-6. All data is based on 113 successful targets that have available 3D structure.

About 60% of these targets are represented by just 8 folds. These 8 folds, given by SCOP fold names, are nuclear receptor ligand-binding domain (8 targets), TIM beta/alpha-barrel (6 targets), protein kinase-like (4 targets), 4-helical cytokines (3 targets), NAD(P)-binding Rossmann-fold domains (3 targets), trypsin-like serine proteases (3 targets), alpha/beta-hydrolases (2 targets), and galactose-binding domain-like (2 targets).

Table 4-6: Structural folds represented by successful targets. Structural folds are from the SCOP database.

SCOP Fold ID	Fold Description	Number of Targets
a.123	Nuclear receptor ligand-binding domain	8
c.1	TIM beta/alpha-barrel	6
d.144.1	Protein kinase-like (PK-like)	4
a.26.1	4-helical cytokines	3
b.47.1	Trypsin-like serine proteases	3
c.2	NAD(P)-binding Rossmann-fold domains	3
b.18.1	Galactose-binding domain-like	2
c.69	alpha/beta-Hydrolases	2
c.65.1.1	Formyltransferase	1
c.19.1	FabD/lysophospholipase-like	1
g.39.1	Glucocorticoid receptor-like (DNA-binding domain)	1
c.71.1	Dihydrofolate reductases	1
a.104.1.1	Cytochrome P450	1
b.74.1	Carbonic anhydrase	1
c.82.1	ALDH-like	1
b.68	6-bladed beta-propeller	1
d.163.1	DNA breaking-rejoining enzymes	1
d.32	Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase	1
d.68	IF3-like	1
d.174.1.1	Nitric oxide (NO) synthase oxygenase domain	1
d.6.1.1	Prion-like	1
d.110	Profilin-like	1
c.66.1	S-adenosyl-L-methionine-dependent methyltransferases	1
a.126	Serum albumin-like	1
d.179.1.1	Substrate-binding domain of HMG-CoA reductase	1
d.168.1	Succinate dehydrogenase/fumarate reductase flavoprotein, catalytic domain	1
d.117.1	Thymidylate synthase/dCMP hydroxymethylase	1
b.22	TNF-like	1
j.61.1.1	Human glutathione reductase (HGR) inhibitor	1

There are 283 research targets that have both available 3D structure and identifiable drug binding domain, which are represented by 107 folds. 60% of these targets are represented by 21 folds. These include Protein kinase-like (21 targets), 4-helical cytokines (14 targets), trypsin-like serine proteases (14 targets), P-loop containing nucleoside triphosphate hydrolases (12 targets), zincin-like (12 targets), TIM beta/alpha-barrel (11 targets), IL8-like (9 targets), cysteine proteinases (8 targets), cystine-knot cytokines (8 targets), nuclear receptor ligand-binding domain (8 targets), C-type lectin-like (7 targets), NAD(P)-binding Rossmann-fold domains (7 targets),

immunoglobulin-like beta-sandwich (6 targets), caspase-like (5 targets), flavodoxin-like (5 targets), acid proteases (4 targets), alpha/beta-hydrolases (4 targets), concanavalin A-like lectins/glucanases (4 targets), knottins (4 targets), phosphorylase/hydrolase-like (4 targets), and PLP-dependent transferases (4 targets).

4.3.4 Biochemical classes

Distribution of successful and research targets with respect to biochemical classes is given in Figure 4-2 and Figure 4-3 respectively. Biochemical classes include enzymes, receptors, nuclear receptors, channels and transporters, factors and regulators (factors, hormones, regulators, modulators, and receptor-binding proteins involved in a disease process), antigens and the remaining binding proteins not covered in other classes, structural proteins (non-receptor membrane proteins, adhesion molecules, envelop proteins, capsid proteins, motor proteins, and other structural proteins), and nucleic acids [9]. The targets unable to be assigned into any of these biochemical classes are tentatively grouped into a separate “unknown” class.

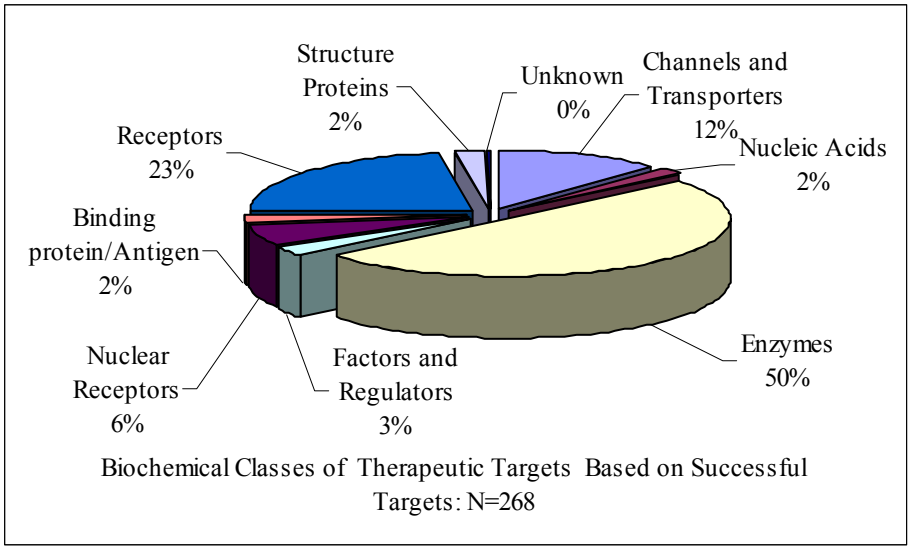


Figure 4-2: Distribution of successful targets with respect to different biochemical classes

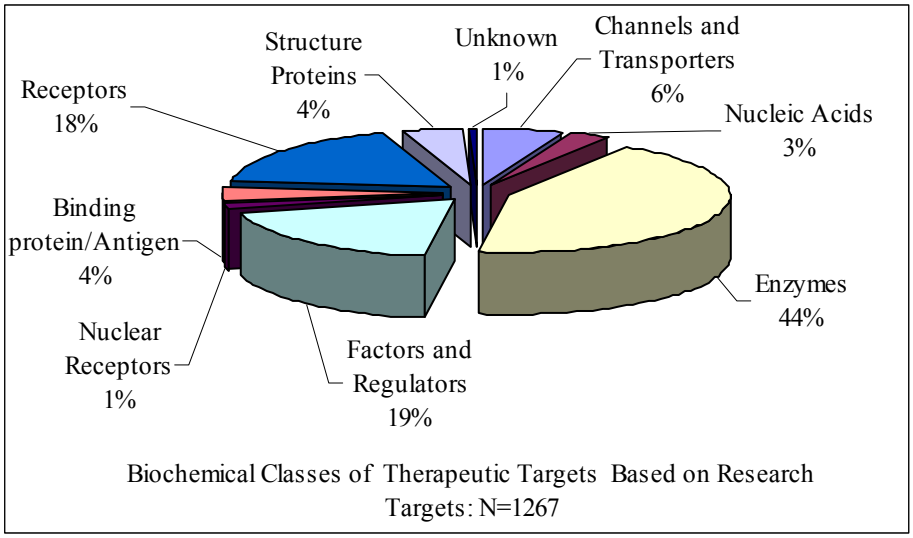


Figure 4-3: Distribution of research targets with respect to different biochemical classes

The overall distribution pattern of successful targets and research targets are roughly similar to the pattern of the 120 successful targets [37] and that of the targets with drug-like leads[9, 135]. The class with the largest number of targets is that of enzymes, which includes 134 successful and 551 research targets representing 50% and 44% of the total number of successful and research targets respectively. The second largest group of successful targets is that of receptors with 61 targets representing 23% of successful target population. The second largest group of research targets is that of factors and regulators with 242 targets representing 18% of the research target population, which is compared to the corresponding group of 8 successful targets that represents only 3% of the total successful target population. Thus there appears to be a dramatic increase in the number of factors and regulators being explored for the treatment of various diseases including cancers [205], autoimmune diseases [206], inflammation, diabetes and neurodegenerative diseases [207].

In addition, target distribution profiles of the groups with a substantial number of

successful targets are channels and transporters with 32 targets representing 12 % of the successful target population, nuclear receptors with 15 targets representing 6 % of the successful target population, and factors and regulators with 8 targets representing 3% of the successful target population. The data shows, at present, ion channels are also important targets for the treatment of pain, neurological and psychiatric disorders [208], ligand-gated channels have been used as the targets for diseases such as neuropsychiatric disorders [209], transporters are the targets of drugs like antidepressants [210], and nuclear receptors have been used as targets of cancer [211], inflammatory and immune diseases [212]. With respect to research target groups, the distribution patterns of them are receptors with 230 targets representing 18% of the research target population, channels and transporters with 75 targets representing 6% of the research target population, structural protein with 56 targets representing 4.4% of the research target population, antigens and other substrate-binding proteins with 50 targets representing 4% of the research target population, nucleic acids with 36 targets representing 3% of the research target population, and nuclear receptors with 19 targets representing 1% of the research target population.

4.3.4.1 The distribution of enzyme targets with respect to enzyme families

The biochemical class containing the largest number of successful targets is the enzyme class, which includes 134 enzymes representing 50% of the 268 successful targets collected in the TTD. This percentage is very similar to the reported figure of 47% enzyme targets among the marketed small molecule drug targets reported in 2002 [37]. There are 122 successful and 494 research enzyme targets with available enzyme classification EC number.

Substantial portion of these enzyme targets appears to be concentrated in a few enzyme families. Figure 4-4 shows the distribution of enzyme targets with respect to enzyme families. An enzyme family is represented by an enzyme classification (EC) number. Examples of therapeutically important enzyme families are EC3.4 (proteases and reverse transcriptases), EC2.7 (kinases and polymerases), EC3.1 (esterases, phosphatases, phosphodiesterases, phospholipases, and ribonucleases), EC1.1 (dehydrogenases and oxidases), EC2.3 (acyltransferases), EC2.4 (glycosyl-transferases), EC1.14 (monooxygenases and dioxygenases), and EC4.1 (carboxylases and aldolases).

The majority research enzyme targets are distributed in the family of EC2.7, EC3.4, and EC3.1, which accounts for 24%, 20%, and 11% respectively. By comparison, 14%, 11%, and 7% of successful enzyme targets are distributed in the EC2.7, EC3.4, and EC3.1 family respectively. EC2.7 contains various kinases and polymerases, EC3.4 is composed of proteases and reverse transcriptases, and EC3.1 includes esterases such as phosphodiesterases, phosphatases, phospholipases, and ribonucleases. Kinases [213], proteases [214], polymerases [215], and esterases [144, 216, 217] have been frequently explored as therapeutic targets for antiviral, antibacterial, anticancer, and cardiovascular effects because of their key roles in the regulatory, synthesis and metabolism processes essential for the progression of the relevant disease. Thus it is not surprising that these enzymes constitute the largest group of enzymatic targets.

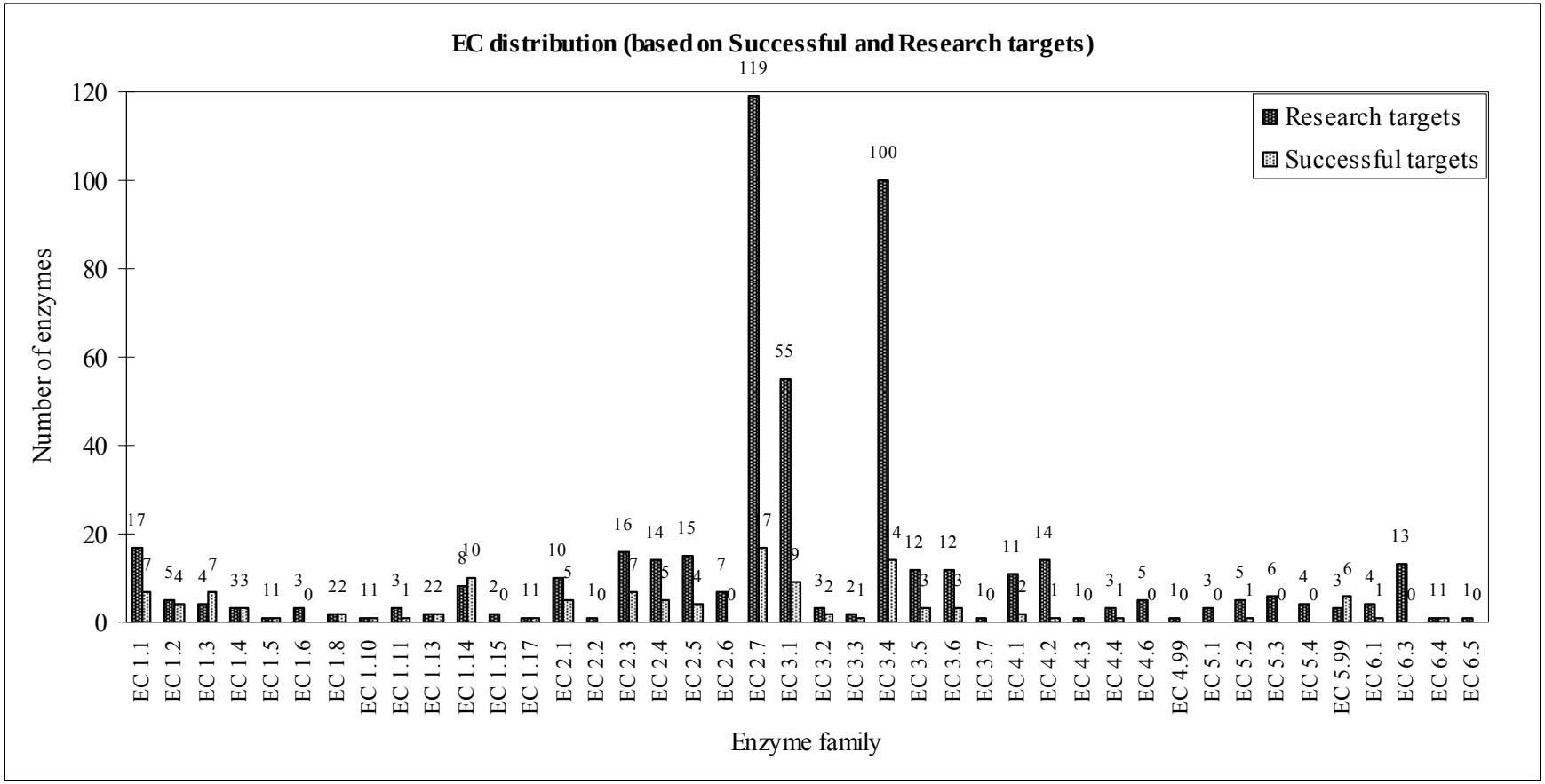


Figure 4-4: Distribution of enzyme targets with respect enzyme families

Moreover, according to the research enzyme targets, 2.8%-3.5% of them are distributed in each of the EC1.1, EC2.3, EC2.5, EC4.2, and EC2.4 family respectively. EC1.1 is composed of dehydrogenases and oxidases, EC2.3 contains acyltransferases, EC2.5 belongs to one kind of transferases, EC 4.2 is constituted by Carbon-oxygen lyases and EC2.4 includes glycosyltransferases. dehydrogenases [218], acyltransferases [219], transferases [220], lyases [221] and glycosyltransferases [222] are increasingly explored for the treatment of high impact diseases such as cancer, obesity, depression, diabetes, tumor and inflammation. This seems to indicate a trend for targeting metabolizing enzymes as a novel strategy for the treatment of these diseases.

4.3.4.2 The distribution of receptors with respect to receptor families

The biochemical class containing the second largest number of successful targets is the GPCR superfamily. This class consists of 42 GPCRs representing 16% of the successful targets collected in the TTD. In contrast, 115 GPCRs represents 9% of the research targets. The percentages are comparable to that of 30% GPCRs among the marketed small molecule drug targets reported in 2002 [37]. According to successful targets, the targets belonging to GPCRs are further divided into 37 in the rhodopsin family and 4 in the metabotropic glutamate family respectively. These receptors are the target of >50% of the current therapeutic agents in the market, including more than a quarter of the 100 top-selling drugs with benefits in the range of several billion US dollars [37, 223, 224]. They have been considered as the best drug targets because of their key roles in various signaling processes essential for such diseases as neurodegenerative diseases, epilepsy, idiopathic pain, drug addition, cardiovascular diseases, allergic inflammatory and autoimmune diseases [183].

There are 115 G-protein coupled receptors (GPCR), 22 cytokine receptors and 22 transmembrane receptor enzymes known to be explored as research therapeutic targets. These GPCR targets are further divided into 95 in the rhodopsin family, 10 in the metabotropic glutamate family, and 8 in the secretin family. Targets in the rhodopsin family account for 7.5% and those of the GPCR superfamily represent 9.1% of the total number of research targets respectively. The rhodopsin family contains a large number of targets because they have been good therapeutic targets for various diseases including neurodegenerative diseases, epilepsy, idiopathic pain, drug addition, cardiovascular diseases, allergic inflammatory and autoimmune diseases [183]. Examples of the targets in the rhodopsin family includes 5-hydroxytryptamine receptors, adenosine receptors, adrenergic receptors, bradykinin receptors, cannabinoid receptors, chemokine receptors, dopamine receptors, histamine receptors, muscarinic acetylcholine receptors, P2Y purinoceptors, and so on.

4.3.5 Human proteins similar to therapeutic targets

In the present day drug development processes, drug candidates have frequently been intentionally designed to bind to their target specifically and to avoid strong interactions with other human protein members of the same protein family to which the target belongs [6, 9, 29, 36, 135]. The successfully designed agents are thus less likely to significantly interfere with the function of human proteins of the same family, reducing the risk of some of the potential unwanted effects. However, their possible interactions with human proteins outside the family cannot intentionally avoided at the design stage, and the potential unwanted effects associated with some of these interactions can only be detected at the later testing stages. Therefore, it tends to be easier to find successful drugs for those targets that have fewer human similarity

proteins outside of their family. One can then speculate that targets with fewer human similarity proteins outside their family tend to be more likely to be explored for drug development.

Some crude estimate about the number of human similarity proteins outside the family of each individual target can be provided by conducting a sequence similarity search against the 59,618 proteins in the human genome that are currently available in protein databases. Table 4-7 summarizes the results of a BLAST search of the drug-binding domain of each of the 173 targets with identifiable drug-binding domain against available human proteins. About 57% of the targets have less than 5 human similarity proteins outside their respective family, and a further 18% of the targets have 6-10 similarity proteins. This seems to support the postulation that targets with fewer human similarity proteins outside their family tend to be more likely to be explored for drug development.

Table 4-7: Statistics of the number of human similarity proteins of successful targets that are outside the protein family of the respective target

Number of similarity proteins	Number of targets with this number of similarity proteins	Targets with this number of similarity proteins %	Examples of Targets
0 -- 5	100	57%	5-hydroxytryptamine 3 receptor, Acetylcholinesterase, Adenosine A2b receptor, ATP-sensitive K+ channel
6 -- 10	32	18%	Alpha-1D adrenergic receptor, Dopamine D1 receptor, Histamine H1 receptor, HIV-1 reverse transcriptase, Muscarinic acetylcholine receptor M1
11 -- 20	16	9%	Coagulation Factor VIIIa, Epidermal growth factor receptor, HIV-1 protease, Insulin receptor, Kappa-type opioid receptor
21 -- 40	14	8%	Androgen receptor, Estrogen receptor, Gamma-aminobutyric acid B receptor, Peroxisome proliferator activated receptor alpha
41 -- 80	6	5%	Lutropin-choriogonadotropic hormone receptor, Sulfonylurea receptor 2B , Thrombin, Urokinase-type plasminogen activator
> 80	5	3%	Human keratin, Receptor-type protein-tyrosine phosphatase S, Thyroid peroxidase, Toll-like receptor 7

However, fewer number of human similarity proteins outside the family of a target is not a necessary condition for finding successful drugs. It merely makes the tasks for finding successful drugs against these targets easier as the probability of un-wanted interactions with human proteins outside the family is reduced. For targets with a higher number of similarity proteins, it is still possible to find agents that can specifically bind to a particular target and has no significant interactions with human proteins both inside and outside of the family to which the target belongs. This is supported by the existence of several successful targets with more than 80 human proteins outside the family of the respective target.

4.3.6 Associated pathways

Targets associated with a fewer number of pathways tend to reduce the chance of un-wanted interference with other processes, and are more likely to be successfully discovered and explored for generating a higher number of clinical drugs. This can be tested by studying the 132 successful targets that have available pathway information in the KEGG database[16]. Table 4-8 gives the statistics of the number of pathways these targets are involved. There are 64 (49%), 36 (27%) and 15(11%) targets found to be associated with 1, 2, and 3 pathways respectively. Each of the remaining targets is involved in 4 to 15 pathways. Some indications about the success rate of the exploration of the targets in each group can be probed by looking at the highest number of clinical drugs directed at any single target in each group. From Table 4-8, it is found that the groups of targets associated with no more than 3 pathways have a substantially higher number of clinical drugs than those with more than 3 pathways, which seems to support the hypothesis that targets associated with a fewer number of pathways tend to be more successfully explored.

Table 4-8: Statistics of the number of pathways of successful targets

Number of pathways	Number of targets in this number of pathways	Percentage of targets in this number of pathways	Highest number of drugs for a target
1	64	49%	8
2	36	27%	8
3	15	11%	5
4	3	2%	1
5	4	3%	2
6	3	2%	3
8	4	3%	
9	1	1%	2
>10	2	2%	1

4.3.7 Tissue distribution

Some therapeutic targets have been chosen primarily because of their high and selective expression in specific tissues, despite the existence of unfavorable conditions such as high expression abundance [32]. Efforts have been made to employ more broadly tissue selective strategies [225]. This raises an interest for studying tissue distribution patterns of the successful targets to find out to what extent tissue specificity has already been used in existing therapeutics. There are 158 successful targets with available information about tissue distribution in human. Their tissue distribution patterns are given in Table 4-9. 79% of these targets are distributed in less than 6 tissues, which seem to indicate that tissue selectivity may be an important factor for the successful exploration of some of these targets.

Table 4-9: Statistics of the human tissue distribution pattern of successful targets

Number of Tissues	Number of Targets Predominantly Distributed in This Number of Tissues	Percentage of Targets Predominantly Distributed in This Number of Tissues	Examples of Targets
1	45	28%	D(3) dopamine receptor, Potassium-transporting ATPase alpha chain 1, Solute carrier family 12 member 3
2	39	25%	Lutropin-choriogonadotropic hormone receptor, Potassium voltage-gated channel subfamily H member 2, Ryanodine receptor 1
3	23	15%	Acetyl-CoA carboxylase 2, Fatty acid synthase, Pregnane X receptor
4	12	8%	Inducible Nitric oxide synthase, Peroxisome proliferator activated receptor alpha
5	5	3%	Catechol-O-methyl-transferase, Amine oxidase [flavin-containing] A
6	2	1%	Fibroblast growth factor receptor 2, Fatty-acid amide hydrolase
7	3	2%	Aldehyde oxidase, Toll-like receptor 7
8	6	4%	Peroxisome proliferator activated receptor gamma, P2Y purinoceptor 12, Insulin receptor
9	1	1%	Voltage-gated sodium channel
10	1	1%	Inhibitor of nuclear factor kappa B kinase
Many Tissue	21	12%	Adenosine deaminase, Na-K-2Cl cotransporter, Receptor-type protein-tyrosine phosphatase S

4.3.8 Chromosome locations

Members of a protein family are known to be distributed in specific clusters in genomes [226, 227]. Functionally similar but non-homologous proteins have also been found to be located at specific regions of genomes, which allow these proteins to be similarly regulated [228]. A large percentage of therapeutic targets are from multiple members of specific protein families or non-homologous proteins of similar function of other targets. It is thus of interest to study the distribution pattern of existing human targets in human genome to determine whether there are any level of clustering of these targets in specific regions of the chromosomes.

Distribution patterns of the human successful and research targets in each of the 23 chromosomes are given in Figure 4-5. These patterns are arranged from the left to

right for chromosome 1, 2, …, 22, and X respectively. For each chromosome, the pattern of successful targets is given on the left and that of research targets is given on the right. The location of each target in a chromosome is marked by a line, with a red line for a successful target and a black line for a research target. It appears that a substantial percentage of research targets are more densely distributed in or near the regions of higher concentration of successful targets. Thus, there seems to be some level of clustering of targets at specific regions where successful targets are located.

The chromosomes with larger number of targets are chromosome 1, 3, 11 and 17. Chromosomes 2, 7, 12 and 19 also contain relatively higher concentrations of targets. Distribution of targets in certain chromosomes appears to be less even than those in other chromosomes. In particular, there are specific sections of larger number of targets in chromosome 1, 3, 5, 9, 12, 17 and 19. Targets in the rest of chromosomes are relatively evenly distributed.

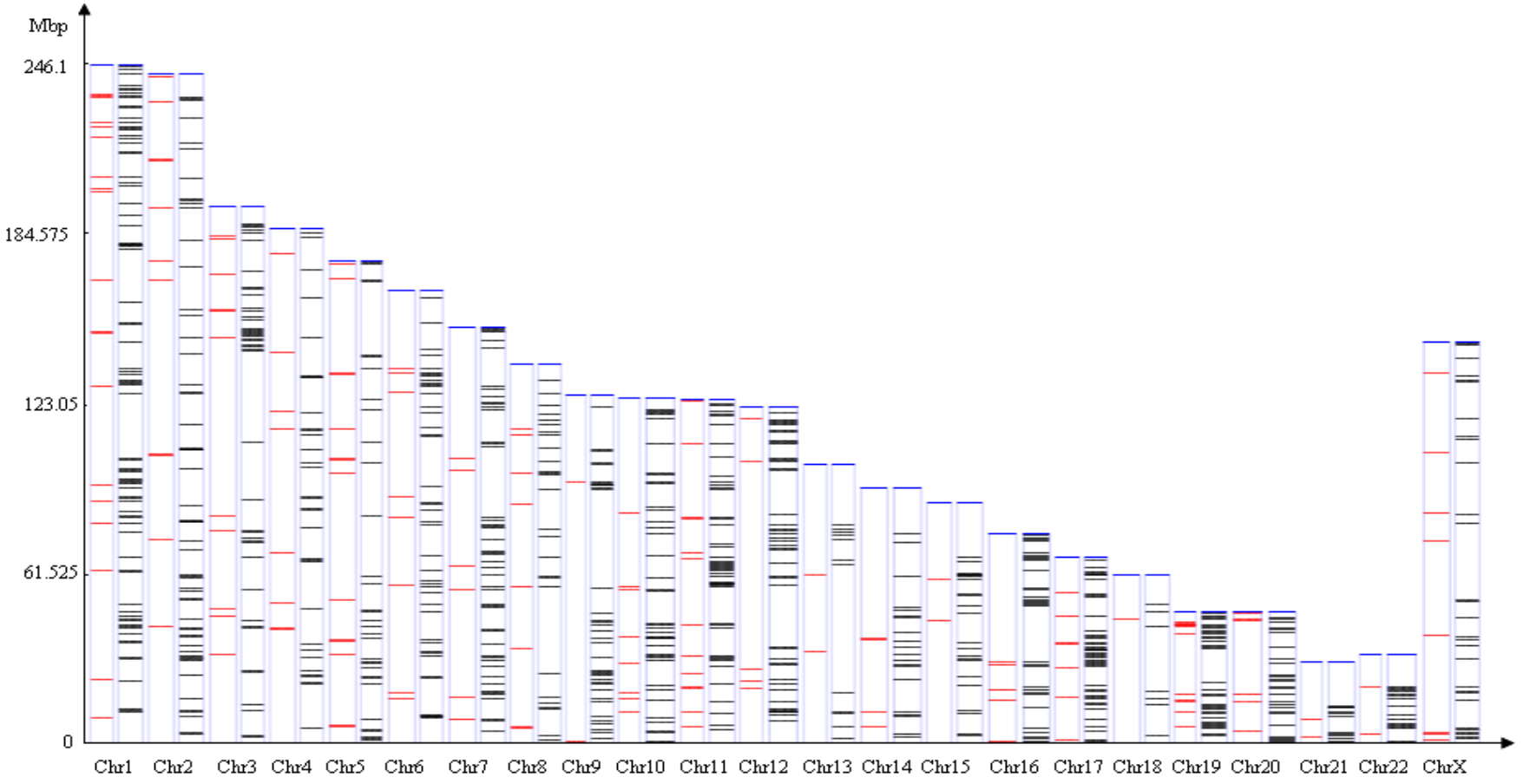


Figure 4-5: Distribution patterns of human therapeutic targets in 23 human chromosomes (For each chromosome, the pattern of successful targets is given on the left and that of research targets is given on the right.)

5 Computer prediction of druggable proteins as a step for facilitating therapeutic targets discovery

In modern drug discovery, pharmaceutical agents have been designed to exert their therapeutic effect by interacting with a pre-selected therapeutic target [9, 28, 29]. Increasing effort and considerable interest have been directed at the identification of effective targets [7, 9, 28, 29, 217]. A 1996 survey showed that, at that time, drug therapies and investigational agents were based on ~500 molecular targets [9, 28]. The reported number of identifiable targets of the marketed drugs was ~120 [9, 37, 135]. Statistical analysis of disease genes and related proteins suggested that the total number of potential targets in the human genome is 600~1,500 [37]. Investigation of the yeast genome found that antifungal targets constitute 2-5% of the genome [37]. Assuming a similar percentage of targets in disease-related microbial genomes, the number of potential targets in microbial genomes is estimated to be greater than1,000. A typical viral genome such as that of HIV-1 [74], HBV [229], and SARS coronavirus [230] contains 1-4 targets, which gives an estimated number of more than 100 potential targets in disease-related viral genomes. Therefore, the estimated total number of distinct targets is in the range of 1,700~3,000.

In this chapter, potential drug interferences with target proteins are discussed in the context of pathway and tissue distribution to provide useful hints about general trends of target exploration, current focus in drug discovery for the treatment of high impact diseases needing effective or more treatment options, and possible reasons why certain targets are easier to explore than others. Meanwhile, a computational system, Support Vector Machines (SVMs), is trained for druggable proteins prediction. How

to develop and evaluate this prediction system are also discussed. As an important item used in this section, druggable protein is elucidated firstly.

5.1 Druggable proteins and therapeutic targets

Druggable proteins represent those proteins with specific structural features that favor interactions with potent, small drug-like chemical compounds [37]. That is to say, the druggable proteins can be readily amenable to be modulated by pharmaceutical small molecules. Such kind of capability is called "druggability" [231]. Likewise, therapeutic targets here are those proteins that can be targeted and modified by drug molecules, where the modulation can change the proteins' biological functions and subsequently provide some therapeutic benefits. The definition for druggable proteins and therapeutic targets appears to be quite similar. However, they belong to different concept categories. Although they have some overlaps in the definition, druggable does not equal drug targets. Figure 5-1 illustrates how to define a potential drug target.

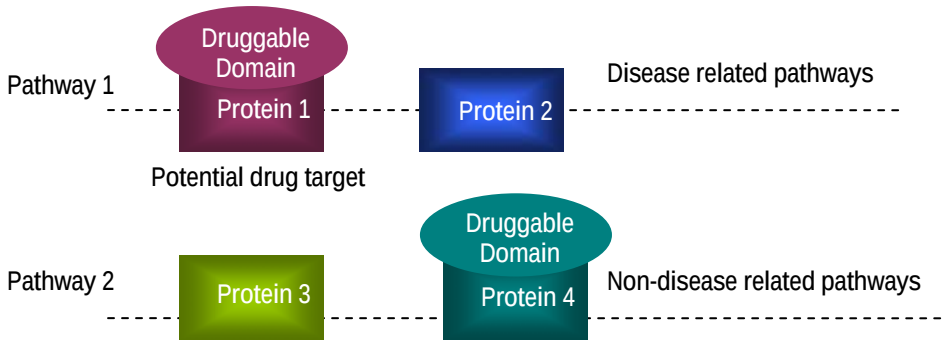


Figure 5-1: Definition of potential drug targets

In this figure, there are four different proteins and two biological pathways. These four proteins may control access to their corresponding pathways. Proteins 1 and 2 are

located in pathway 1, while proteins 3 and 4 are located in pathway 2, respectively. Among them, it is supposed that only proteins 1 and 4 have suitable druggable domains (or binding domains). According to pathways, one of them is related to disease condition. It is known that, only if the protein has appropriate druggable domain (e.g. Protein 1 and Protein 4), drug molecules can bind to the protein, modify its biological functions, and further impact cellular effects of the pathway. Thus, both proteins 1 and 4 can be considered as druggable proteins. However, only protein 1 is qualified as drug target, due to its disease relevance. Therefore, it is concluded that a protein with druggability is necessary to be considered as a potential drug target, but it is not sufficient.

In 2002, Hopkins drew a figure to explain the relationship between druggable genome and drug targets (Figure 5-2)[37]. They pointed out that "*the effective number of exploitable drug targets can be determined by the intersection of the number of genes linked to disease and the 'druggable' subset of the human genome*". In their study, they also gave some estimated number in detail.

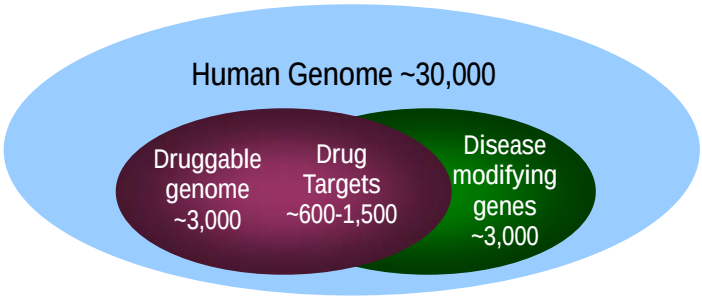


Figure 5-2: Estimated number of drug targets

As shown in Figure 5-2, the group of drug targets (~600-1,500 genes) is the intersection of the druggable genome (~3,000 genes) and disease modifying genes (~3,000 genes), which both are subsets of the human genome (~30,000 genes). With

rapid progress for Human Genome Project, an estimate of about 30,000 human genes [232, 233] has dropped to the current 22,287 genes [76]. Even now, the human gene data set is still too large to be handled. As a result, it makes the process of target development extremely complex. However, it is also noted that the number of genes in human genome is far greater than that of druggable genes. Due to this remarkable difference, we may provide a feasible and efficient way to facilitate drug target discovery (Figure 5-3).

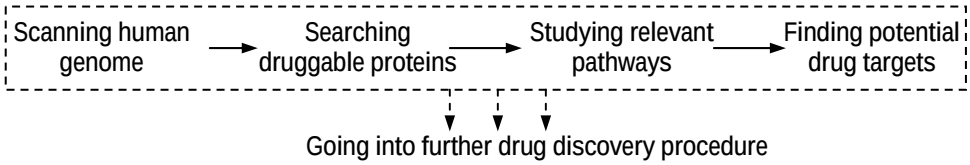


Figure 5-3: Flow chart about how to facilitate drug target discovery

Firstly, lots of druggable proteins are found by computational scanning of the whole human genome. Secondly, biological pathways relevant to the druggable proteins are studied. Then, those druggable proteins related to diseases may be picked up for further investigation. Finally, scientists will choose some of them as potential drug targets and study their various physical-chemical as well as pharmacological properties. Here, the focus will be on the first two steps and an attempt to use biological statistics and computational methods to facilitate druggable proteins searching will be discussed.

5.2 Prediction of druggable proteins from their sequence

Advances in high-throughput gene sequencing have led to rapid identification thousands of novel genes mostly with known functions. For the pharmaceutical

industry, the sequencing of the human genome and the genomes of disease species proved to be both a blessing and a curse. Where potential targets were once hard to come by, the industry is now awashed with them. This has left drug discovery communities with the difficult task of sifting through the gene data to find novel targets [32, 234]. Genomics approaches such as large-scale gene expression analysis, functional screens in model organisms, genome scans for disease susceptibility genes, and the search of new members of effective drug target classes have enabled the finding of countless candidates for many diseases [3, 235, 236]. The determination of druggable candidates still relies on experimental studies. Methods that facilitate the identification of druggable proteins from these candidates or directly from genomes are thus particularly useful for target identification.

The investigation of the features of known therapeutic targets from earlier studies [37, 195] and in the previous chapters suggests that targets have certain common characteristics, which may be used as the basis for deriving rules for the identification of druggable proteins from their sequence in a similar manner to the rule-based methods (such as “rule-of-five”) for predicting “drug-like” compounds from their structures [194, 237]. Statistical learning methods have also been successfully applied for developing tools for predicting “drug-like” molecules from their structures on the basis that they have common structural and physicochemical features [238, 239]. It is expected that these statistical learning methods are equally applicable for predicting druggable proteins from their sequences on the basis that druggable proteins share common characteristics.

5.2.1 "Rules" for guiding the search of druggable proteins

Therapeutic targets are grouped into target families, which are defined as protein domain families that contain at least one therapeutic target. These target families constitute a small percentage, 6.6%, of the 7,677 protein families in the Pfam database [198]. A study of 120 targets of clinical drugs (successful targets) and 279 targets of investigational agents (research targets) found that they are represented by 130 families [196]. In the study of 173 successful and 906 research targets in TTD [39], it was found that they are distributed in 92 and 412 families, respectively.

In addition to the distribution in a limited number of target families, certain characteristics are expected for therapeutic targets [37]. They play critical roles in disease processes, have certain level of structural novelty and physicochemical properties, and are involved minimally in other important human processes. Target expression is either at a constrained level or tissue selective to allow for drug efficacy. These characteristics are likely encoded in the sequence of targets and useful clues may be derived from comparative study of these targets against the human genome. Drug discovery has been focused on agents that bind to their targets specifically without interactions with other human proteins in the respective target family [6, 28, 29]. However, their possible interactions with human proteins outside the family are not intentionally avoided at design stages, and the corresponding unwanted-effects can only be detected at later testing stages. It tends to be easier to find viable drugs for targets having fewer number of human similarity proteins outside their family. Therefore, target comparative studies need to be conducted on the basis of separate considerations of proteins outside and inside a target family.

Targets associated with a fewer number of pathways tend to reduce the chance of undesirable interference with other processes, and are more likely to be successfully explored. Some targets have been chosen primarily because of their high and selective expression in specific tissues, despite the existence of unfavorable conditions such as high expression abundance [32]. Therefore, human pathway and tissue distribution profiles are important indicators for characterizing targets and for determining the level of difficulty of their exploration.

The profile of comparative study of therapeutic targets against the human genome is generated from the results of BLAST [240] alignment of the drug-binding domain of 173 successful targets against 59,618 proteins encoded in the human genome. Pathway distribution profile is obtained from the available human pathway information of 132 successful targets in KEGG database [16]. Tissue distribution profile is obtained from the available human tissue distribution information of 158 successful targets from the SwissProt database [25]. All these profiles are given in Table 5-1.

Table 5-1: Statistics of the characteristics of successful targets

Category	Human similarity proteins outside target family		Human similarity proteins in target family		Target participating pathways		Tissue distribution		Sub-cellular location	
Number of Targets in Statistics	173		173		132		158		153	
Item	Number of similarity proteins	Percentage of targets with this number of human similarity proteins	Number of similarity proteins	Percentage of targets with this number of human similarity proteins	Number of pathways	Percentage of targets in this number of pathways	Number of tissues	Percentage of targets primarily distributed in this number of tissues	Location	Percentage of targets primarily distributed in location
Statistical data	0-5	57%	0-5	26%	1	49%	1	28%	Membrane	60%
	6-10	18%	6-30	25%	2	27%	2	25%	Cytoplasm	16%
	11-15	3%	31-100	22%	3	11%	3	15%	Nucleus	10%
	16-20	6%	>100	29%	4	2%	4	8%	Extra-cellular and Secreted	8%
	21-40	8%			5	3%	5	3%	Mitochondrion	3%
	41-80	5%			6	2%	6	1%	Endoplasmic reticulum	2%
	>80	3%			7		7	2%	Peroxisome	1%
					8	3%	8	4%		
					9	1%	9	1%		
					>10	2%	>=10	13%		

A total of 57% of the investigated targets have less than 6, and a further 21% have 6-15 human similarity proteins outside their respective target family. In contrast, human similarity proteins inside the respective target family are evenly distributed between 1 to 40 proteins. There are 49%, 27%, and 11% of the studied targets primarily associated with 1, 2 and 3 human pathways respectively. Each remaining target is involved in 4 to 15 pathways. Targets associated with no more than 3 pathways are found to have a substantially higher number of clinical drugs than those with more than 3 pathways. A total of 79% of the investigated targets are distributed in no more than 5 human tissues. Based on the characteristics of therapeutic targets described in earlier studies [37, 195] and in the previous profiles, it seems that the following "rules" can be proposed for guiding the search of druggable proteins:

- The protein is derived from one of the target-representing protein families. The number of these families is currently estimated to be no more than 940. So far, 92 confirmed families (each containing at least one successful target) and 412 likely families (each containing at least one research target) have been found.
- Sequence variation between the drug-binding domain of a protein and those of the other human members of its protein family needs to allow sufficient degree of differential binding of a "rule-of-five" molecule to the common binding site.
- Protein preferably has less than 15 human similarity proteins outside its family (HSP). While existence of a higher number of human similarity proteins does not rule it out as a druggable protein, it generally increases the chance of undesirable interferences and thus the level of difficulty for finding viable drugs. (78% of the successful targets with identifiable drug-binding domain have less than 15 human similarity proteins).
- Protein is preferably involved in no more than 3 pathways in human (HP). While

association with a higher number of human pathways does not rule it out as a druggable protein, it generally increases the chance of undesirable interferences with other human processes and thus the level of difficulty for finding a viable target. (87% of the successful targets with pathway information are associated with no more than 3 pathways).

- For organ or tissue specific diseases, protein is preferably distributed in no more than 5 tissues in human (HT). While distribution in a higher number of tissues does not rule it out as a druggable protein, it generally increases the chance of undesirable interferences with other tissues and thus the level of difficulty for finding a viable target. (79% of the successful targets with tissue distribution information are distributed in no more than 5 tissues).
- A higher number of HSP, HP and HT doesn't preclude the protein as a potential target, it statistically increase the chance of undesirable interferences and the level of difficulty for finding viable drugs.

There are 57%, 76% and 53% of the investigated targets with the number of HSP, HP, and HT lower than those specified in rule (3), (4), and (5) respectively. Based on this result, therapeutic targets can be divided into the "easy" and "difficult" class by using a simple rule: targets with HSP≤5, HP≤2 and HT≤2 are "easy" targets, and those with a higher number are "difficult" targets. The smaller percentage of targets having a higher number of HSP, HP and HT is consistent with the notion that these targets are more difficult to explore than those with a smaller number. The suitability of using these numbers as indicators of the level of difficulty of target exploration were studied by examining the target exploration time (TET) of some of the innovative targets of the FDA approved drugs since 1994, which have no marketed drug prior to their approval [38]. TET is the number of years between the first reported compound

investigation and the first FDA approval. Table 5-2 shows that targets with a fewer number of HSP, HP and HT generally have a statistically shorter TET. There are two target difficulty levels, E represents "easy" target with a shorter expected target exploration time, and D represents "difficult" target with a longer expected target exploration time. From Table 5-2, the TET of the "easy" targets is generally shorter than 10 years and that of the "difficult" targets is generally longer than 14 years, suggesting that the level of difficulty of target exploration may be roughly estimated by using this simple rule.

Table 5-2: Profiles of some innovative targets of the FDA approved drugs since 1994

Target	Year of First Reported Compound Investigation	Year of First FDA Approval	Target Exploration Time (Years)	Number of Human Similarity Proteins Outside Target Protein Family	Number of Human Similarity Proteins in Target Protein Family	Number of Tissues Target is Primarily Distributed	Number of Pathways Target is Distributed	Predicted Target Difficulty Level	First FDA Approved Drug
Maltase-glucoamylase, intestinal	1967	1995	28	1	12	3	2	D	Acarbose
Mineralocorticoid receptor	1975	2002	27	31	101	Many	?	D	Eplerenone
Prostaglandin G/H synthase 2	1975	1998	23	33	13	4	1	D	Celecoxib
Acetyl-CoA carboxylase 2	1975	1994	19	30	0	3	5	D	metformin
Inosine-5'-monophosphate dehydrogenase 2	1979	1995	16	4	10	2	1	E	mycophenolate mofetil
Phosphodiesterase 5	1984	1998	14	3	74	5	1	D	Sildenafil
Myeloid cell surface antigen CD33	1987	2000	13	2	21	2	1	E	Gemtuzumab Ozogamicin
Type-1 angiotensin II receptor	1984	1995	11	8	388	4	2	D	Losartan Potassium
Cysteinyl leukotriene receptor 1	1986	1996	10	5	386	2	2	E	Zafirlukast
Receptor protein-tyrosine kinase erbB-2	1988	1998	10	18	482	1	4	D	Trastuzumab
FK-binding protein 12	1989	1999	10	0	30	2	?	E	Sirolimus
P2Y purinoceptor 12	1989	1997	8	3	280	2	?	E	Clopidogrel

5.2.2 Prediction of druggable proteins by a statistical learning method

New targets may not bear sequence similarity to known targets or known proteins. Consequently, a straightforward sequence similarity search against effective drug target classes [3] and known disease genes [235] may not always be useful for identification of novel targets. While targets appear to have common characteristics that are reflected in their sequences, they are from a diverse range of different families and structural folds. Thus, methods that do not rely on sequence and structure similarity are needed for facilitating the prediction of druggable proteins directly from their sequences.

5.2.2.1 Development of SVM prediction system

Statistical learning methods, such as SVMs and neural networks, have emerged in the last few years as attractive methods for the prediction of protein functional classes [82-85, 87-89] and structural classes [241, 242] without the use of sequence similarity. These classes contain proteins of diverse functions and structures. Examples of some of these classes are RNA-binding proteins, EC2.7 transferases of phosphorus-containing groups, EC3.4 peptidases, and TC1.A alpha-type channels. It appears that the prediction accuracy of these methods has reached a level sufficient for facilitating the prediction of the functional and structural classes of proteins. For instance, the overall accuracy of SVMs prediction of the functional family of 13,891 enzymes and 447 RNA-binding proteins is 86% and 98% respectively. Thus, it is of interest to investigate the feasibility of using statistical learning methods for predicting druggable proteins from their sequences.

Currently, SVM appears to be the most accurate statistical learning method for protein predictions [84, 85, 87-89, 241]. Therefore, only this method is investigated here. SVM is based on the structural risk minimization principle from statistical learning theory [92]. Known proteins are divided into druggable and non-druggable classes, each of these proteins is represented by their sequence-derived physicochemical features [85]. These features are then used by SVMs to construct a hyperplane in a higher-dimensional hyperspace that maximally separate druggable proteins and non-druggable ones. By projecting the sequence of a new protein onto this hyperspace, it can be determined if this protein is druggable from its location with respect to the hyperplane. It is a druggable protein if it is located on the side of druggable class. The accuracy of SVMs depends on the diversity of the protein samples used for finding the hyperspace and its hyperplane, the quality of the representation of protein features, and the efficiency of the SVMs algorithm. To a certain extent, no sequence and structural similarity is required *per se*. Thus SVM is an attractive approach for facilitating the prediction of classes of proteins with diverse sequences and structures, and thus the prediction of druggable proteins.

A total of 1,368 sequence entries of 1,535 successful and research targets were used to construct the druggable class, and 12,956 representative proteins from 6,856 Pfam [198] protein families (with all of the known target-representing families excluded from these families) were used to construct the non-druggable class. Multiple sequence entries of some viral protein targets were included in the druggable class because of significant sequence variations across strains. Proteins in each class were randomly divided into five subsets of approximately equal size. Four subsets were selected as the training set and the fifth as the testing set. This process was repeated five times such that every subset was selected as a testing set once.

5.2.2.2 Evaluation of prediction model

The average prediction accuracy from this 5-fold cross validation study was 69.8% for druggable proteins and 99.3% for non-druggable proteins. The accuracy for non-druggable proteins was comparable but that of druggable proteins was somehow lower than those of protein functional and structural families [84, 85, 87-89, 241], which was expected because of the significantly higher level of sequence and structural diversity of therapeutic targets. Nonetheless, these accuracies were meaningful for facilitating the prediction of druggable proteins.

To test its potential for practical applications, the constructed SVM prediction system was used to scan the human genome for identifying potential druggable proteins that are not in the training and testing sets. A total of 1,102 human proteins were predicted to be druggable, which included 153 G-protein coupled receptors (GPCR), 65 other receptors, 333 enzymes, and 56 channels. These numbers were within the estimated numbers of druggable proteins and therapeutic targets in the human genome. For instance, the total number of druggable proteins and actual targets in the human genome has been estimated to be ~3,000 and ~1,500 respectively [37], and the total number of 400 GPCRs has been suggested to be potential targets [243]. Some examples of predition results are listed in appendix A. Moreover, the yeast genome was also searched by using this SVM system to test whether the prediction results would be consistent with previous studies of this genome. The search of the yeast genome identified 353 druggable proteins, which constituted 4% of the 8,904 encoded proteins. This number was consistent with the report that antifungal targets constitute 2-5% of the yeast genome [37].

This SVM prediction system was further tested by a comparison of its predicted

druggable proteins in an HIV genome with known HIV targets. This genome was selected because it was one of the most extensively explored genomes for finding therapeutic targets, and it was highly likely that all of the potential targets in this genome have been identified [244]. The NCBI [13] HIV-1 genome entry NC_001802, with none of its encoded protein sequences used in the SVM training and testing sets, was used for this test, and the results are given in Table 5-3.

There are 4 successful and 7 research targets in HIV-1 genome. SVM was able to predict 2 successful and 6 research targets as druggable. Overall, 72% of the known successful and research targets and 100% of the non-targets were correctly predicted. This prediction accuracy was consistently similar to that of the five-fold cross validation study. These three tests seem to indicate that SVM has some potential for facilitating the identification of druggable proteins from genomic data. The prediction accuracy for druggable proteins needs to be improved. One reason for the lower accuracy of druggable proteins is the large imbalance between the number of druggable and non-druggable proteins. Such a large imbalance is known to affect the accuracy of a SVM prediction system and methods for solving these problems are being developed [89].

Table 5-3: Comparison of the known HIV-1 protein targets and the SVM predicted druggable proteins in the NCBI HIV-1 genome entry NC_001802

Protein in HIV-1 genome	NCBI protein accession number	Target status	SVM prediction status
Gag-Pol	NP_057849.4		
Gag-Pol Transframe peptide	NP_787043.1		
Pol	NP_789740.1		
protease	NP_705926.1	Successful target	Druggable
reverse transcriptase	NP_705927.1	Successful target	Druggable
reverse transcriptase p51 subunit	NP_789739.1		
integrase	NP_705928.1	Research target	Druggable
Gag	NP_057850.1	Research target	Druggable
matrix	NP_579876.2		
capsid	NP_579880.1		
p2	NP_579882.1		
nucleocapsid	NP_579881.1	Research target	Druggable
p1	NP_787042.1		
p6	NP_579883.1		
Vif	NP_057851.1	Research target	Druggable
Vpr	NP_057852.2		
Tat	NP_057853.1	Successful target	
Rev	NP_057854.1		
Vpu	NP_057855.1		
Envelope surface glycoprotein gp160	NP_057856.1	Research target	Druggable
envelope signal peptide	NP_579893.2		
Envelope surface glycoprotein gp120	NP_579894.2	Research target	
Envelope transmembrane glycoprotein gp41	NP_579895.1	Successful target	
Nef	NP_057857.2	Research target	Druggable

6 Computational analysis of drug ADME-associated proteins

Pharmacogenetic prediction and mechanistic elucidation of individual variations of drug responses is important for facilitating the design of personalized drugs and optimum dosages. One of the keys for pharmacogenetic studies is the knowledge about proteins responsible for the absorption, distribution, metabolism and excretion (ADME) of drugs. Although the original version of ADME-associated proteins (ADME-AP) database [245] have provided comprehensive information about all classes of ADME-APs described in the literature, the information about reported polymorphisms and pharmacogenetic effects need to be integrated into this database. ADME-AP database may, therefore, serve as a useful resource for understanding the known ADME-APs and molecular mechanism of drug responses and facilitating the development of personalized medicines and optimal dosages for individuals.

In previous chapters, the strategy of database development has been discussed. Similar idea was used to construct ADME-AP database. Thus, this chapter has omitted the details of database construction and simply introduces the new version of ADME-AP database. More emphasis is placed on computational analysis of ADME-APs and applications in drug discovery. In particular, a discussion on how to assess the usefulness of the relevant information for facilitating pharmacogenetic prediction of drug responses, and how to use computational methods to predict individual variations of drug responses from the polymorphisms of ADME-APs is included.

6.1 ADME-associated proteins database

Resources that provide information about ADME-APs as well as therapeutic targets and ADR-related proteins are useful for facilitating the study of pharmacogenetics [246]. To date, a number of freely accessible web-based resources have been developed and described in the literature [39, 40, 72, 245]. One of the most important web-based resources about ADME-APs is ADME-AP database, which is developed in 2002 [245]. And its updated information is introduced in more detail below.

The ADME-AP database [73, 245] provides comprehensive information about the known ADME-APs, the reported polymorphisms and pharmacogenetic effects. The updated database currently contains entries for 316 ADME-APs, 734 substrates and inhibitors, 1,337 polymorphisms in 121 proteins, and 327 reported cases of altered drug responses. The drug ADME-APs described in the literature are included in ADME-AP database. In addition, some transporter proteins and carrier proteins, not yet confirmed to play specific roles for drug ADME, are also included in this database. These proteins are capable of carrying or transporting small molecules, peptides and lipids and thus may potentially play a role in drug disposition [247-250]. Information in this database includes physiological function of each protein, site of action, tissue distributions, transport directions, driving force, substrates and inhibitors, and the potential effect on a drug in terms of ADME classes. While available, the reported polymorphisms and pharmacogenetic effects are provided. Cross-links to other databases are introduced to facilitate the access of information about the sequence, 3D structure, function, genetic disorder, nomenclature, ligand binding properties, and related literatures of each target.

Protein Name	Cytochrome P450 3A4
Synonyms	EC 1.14.14.1 Cypiiia4 Nifedipine Oxidase Nf-25 P450-Pcn1
Gene name	CYP3A4
ADME Class	M1
Species	HOMO SAPIENS (HUMAN)
Function	Involved in an nadph-dependent electron transport pathway. Known to oxidize a variety of structurally unrelated compounds, including steroids, fatty acids, and xenobiotics. Belonging to a group of heme-thiolate monooxygenases. In liver microsomes,
Substrates	(+)-(11s,12s)- And (-)-(11r,12r)-Dihydroxydibenzo[A,L]Pyrene; (Db[A,L]P-11,12-Diol); 1,2-Dihydroxy-1,2-Dihydro-5,6-Dimethylchrysene; 1,2-Dihydroxy-1,2-Dihydro-5-Methylchrysene (5-Methylchrysene-1,2-Diol); 1,4-Cineole; 1,6-Dinitropyrene; 1,8-Cineole (Eucal
Tissue distribution	Liver and intestine
Similarity	CYTOCHROME P450 FAMILY.
Variant Information	Click to see detailed infomation
Pharmacogenetic Effects	Click to see detailed infomation
Protein Properties	PROTEIN SEQUENCE / OTHER INFO (SwissProt) 3D STRUCTURE (PDB) GENETIC DISORDER (OMIM) RELATED LITERATURES (PubMed) LIGAND BINDING PROPERTIES (CLiBE) ENZYME NOMENCLATURE
Reference	Fujita K, Kamataki T.Role of human cytochrome P450 (CYP) in the metabolic activation of N-alkylnitrosamines: application of genetically engineered Salmonella typhimurium YG7108 expressing each form of CYP together with human NADPH-cytochrome P450 reductas

Figure 6-1: Web-interface of a protein entry of ADME-AP database

Detailed Information [VAR_008363]	
AC Number	P08684.
Allele(s)	Allele CYP3A4*2
Amino acid position of the variant	221
Comment	In allele CYP3A4*2; exhibits a lower intrinsic clearance towards nifedipine
Residue change	From Ser (S) to Pro (P), S221P
Status	Polymorphism

Figure 6-2: Web-interface of a polymorphism

Detailed Information [PPDHP0p7X]	
Protein	Cytochrome P450 3A4
AC Number	P08684.
Gene	CYP3A4
Drug Examples/Toxin	Carbamazepine [1]
Pharmacogenetic Effects	Polymorphisms in the DMEs influence the drug levels in the plasma and hence the toxicity toward the AEDs.

References:

1: Clues to the genetic influences of drug responsiveness in epilepsy. Epilepsia. 2003;44 Suppl 1:33-7. Review.
PubMed

Figure 6-3: The detailed information of selected ADME-associated protein

Figure 6-1 shows the database entry of an ADME-AP. Information about specific polymorphism and pharmacogenetic effect is provided in separate pages illustrated in Figure 6-2 and Figure 6-3. Each of the ADME-APs can be assigned to different ADME classes, which are defined as: Class A - proteins involved in the absorption or re-absorption of drugs into systemic system, Class D1 - transporters of chemicals across membranes of various tissue barriers from the systemic system into the target sites, Class D2 - proteins responsible for transporting drugs back into the systemic system, Class M1 - phase I drug-metabolizing enzymes, Class M2 - phase II drug-metabolizing enzymes, and Class E - proteins that enable the excretion or presystemic elimination of drugs. The distribution of the protein entries in ADME-AP database with respect to ADME classes is as follows: There are 60 proteins in class A, which is mainly distributed in the intestine. A total of 176 proteins are found in class D (103 are in D1, 20 in D2, and 17 in D3 group respectively). Moreover, there are 36 proteins that have not been reported to be involved in drug distribution by non-the-less might be potentially involved in drug distribution. These proteins are tentatively included in class D with a postfix “potential” added to their classification name. Class M includes 89 enzymes, including 45 in M1 and 44 in M2 group respectively. In addition, there are 30 proteins in class E.

Proteins in ADME-AP database appear to be diversely distributed in almost all tissues. A substantial portion of these proteins can be found in intestine (53 proteins), kidney (94 proteins), liver (93 proteins) and brain (75 proteins) where they play important roles in ADME as well as normal function. Transporters and carriers make up the majority of ADME-APs. Overall, 142 out of 316 proteins are transporters, co-transporters, transporter-like proteins, transporter-associated proteins or carriers. These transporters and carriers are mainly involved in the absorption/re-absorption of

drugs (47 proteins), uptake of drugs into cells (74 proteins), efflux of drugs out of cells (11 proteins), and drug elimination (16 proteins). Enzymes constitute another major group of ADME-APs. There are 45 M1 and 44 M2 enzymes respectively. It is noted that 39 out of 45 M1 enzymes are oxidoreductases, and 39 out of 44 M2 enzymes are transferases. No lyase or ligase is found in the database at present.

6.2 ADME-associated proteins database as a resource for facilitating pharmacogenetics research

A great number of freely accessible web-based resources provide plentiful information about studies of pharmacogenetics of drug response [251]. Up to date, many studies have explored the possibility of using polymorphisms as indicators of specific drug responses [252-256]. Computational methods have been developed for analyzing complex genetic, expression and environmental data to determine the association between drug response and the profiles of polymorphism, expression and environmental factors [257-259] and to derive pharmacogenetic predictors of individual variations of drug response [259, 260].

6.2.1 Information sources of ADME-associated proteins

Drug metabolism is associated with the interaction of a drug with specific metabolizing enzymes [261]. In certain cases, drug absorption, delivery and excretion is facilitated by drug binding to transporters and carriers [247]. Information about some of the ADME-APs can thus be obtained from specialized databases and websites focusing on specific class or group of transporters, carriers and metabolizing enzymes.

Table 6-1: Summary of web-resources of ADME-related proteins

Web-Resource and URL	Information
TP-search transporter database (http://www.tp-search.jp)	A database on drug transporters, which attract a great deal of attention in pharmacokinetics research field
49 Human ATP-Binding Cassette Transporters (http://nutrigene.4t.com/humanabc.htm)	Information on 49 human ABC transporters
ABCISSE database (http://www.pasteur.fr/recherche/unites/pmtg/abc/database.iphtml)	Sequence, structure, and evolution of ABC transporters
ABC transporters database (http://www.genome.ad.jp/kegg/ortholog/tab02010.html)	Information on ABC transporter families
ABC-Transporter Genes in HUGO Gene Nomenclature Committee (http://www.gene.ucl.ac.uk/nomenclature/genefamily/abc.html)	Nomenclature and sequence of ABC transporter genes
Human Intestinal Transport System (http://bigfoot.med.unc.edu/watkinsLab/website/hEnt.htm)	Information on human intestinal transport system proteins, substrates, and inhibitors
Human Membrane Transporter Database (HMTD) (http://lab.digibench.net/transporter/)	Information on human membrane transporters for drug transport studies and pharmacogenomics
Transporter Classification database (TCDB) (http://www.tcdb.org/)	Comprehensive info of IUBMB approved classification system of membrane transport proteins
Human Cytochrome P450 (CYP) Alleles Database (http://www.imm.ki.se/CYPalleles/)	Comprehensive info about Cytochrome P450 (CYP) Allele
Chytochrome P450 Homepage of Nelson's Lab (http://drnelson.utmem.edu/CytochromeP450.html)	Integrated info of P450 enzymes (including animals, lower eukaryotes, plants, bacteria and archaebacteria)
Cytochrome P450 family in HUGO Gene Nomenclature Committee (http://www.gene.ucl.ac.uk/nomenclature/genefamily/cyp.php)	Nomenclature and sequence of Cytochrome P450 family genes
Directory of P450-containing Systems (http://www.icgeb.trieste.it/~p450srv/)	Integrated info of P450 enzymes
Cytochrome P450 (http://www.anaesthetist.com/physiol/basics/metabol/cyp/cyp.htm)	An introduction to CYP and its importance in clinical medicine
UDP Glucuronosyltransferase database (http://som.flinders.edu.au/FUSA/ClinPharm/UGT/)	Information on UDP glucuronosyltransferases
UDP glycosyltransferases in HUGO Gene Nomenclature Committee (http://www.gene.ucl.ac.uk/nomenclature/genefamily/ugt.php)	Nomenclature and sequence of UDP glucuronosyltransferases genes
Arylamine N-Acetyltransferase (NAT) Nomenclature (http://www.louisville.edu/medschool/pharmacology/NAT.html)	Information on Arylamine N-Acyltransferase (NAT) polymorphisms and nomenclature
ADME-Associated Protein database (ADME-AP) (http://bidd.nus.edu.sg/group/admeap/admeap.asp)	Comprehensive info of ADME-associated proteins
dbSNP at NCBI (http://www.ncbi.nlm.nih.gov/SNP/)	Information on single nucleotide polymorphism in genes, including those of ADME associated proteins
PharmGKB database (http://www.pharmgkb.org/index.jsp)	Integrated data of variation in human genes and response to drugs, including those of ADME associated proteins
GeneSNPs at the Utah Genome Center (http://www.genome.utah.edu/genesnps/)	Integrated gene, sequence and polymorphism data

These databases normally provide general information about sequence, structure and biochemical characteristics of specific classes of proteins. However, the majority of them are not intended for pharmacokinetic and pharmacogenetic studies, and some of the data relevant to the pharmacogenetic studies are not provided. Examples of these data are polymorphism, variants, ligands (substrates, inducers, inhibitors, etc.), related diseases, and related drug response. Such types of data are beginning to be added in existing and newly emerging databases. Table 6-1 summaries useful freely-accessible internet resources, which are relevant to drug ADME-APs [40, 73, 245].

TP-search transporter database [262] is useful resource of drug transporters, drug-drug interactions, gender differences, and pathophysiology. There are plans to add information about genetic polymorphisms and related genetic diseases into this database. Transporter Classification database (TCDB) is another comprehensive source for IUBMB approved classification system of membrane transport proteins [263]. The Human Intestinal Transport System website (Watkins Lab, University of North Carolina at Chapel Hill) provides substrates, activators and inhibitors of P-glycoprotein, hOATP and other transporters in epithelial cells. A website of the 49 Human ATP-Binding Cassette (ABC) Transporters for the P-glycoprotein nomenclature (Allikmets R, *et al*, Frederick Cancer Research and Development Center, USA) includes 49 human ABC transporter genes with information about related genetic diseases, tissue distribution, and substrates. ABC transporter database (KEGG, Kyoto University, Japan) gives comprehensive information about ABC families of transporters. Other useful websites are ABC-Transporter Genes (University College London, UK) which gives the nomenclature and sequence of ABC transporter genes, and ABCISSE database (Institute Pasteur, France) which provides information about the sequence, structure, and evolution of ABC

transporters.

The Cytochrome P450 Homepage of Nelson's Lab (Nelson D, University of Tennessee, USA) includes 2,383 P450s from different species. It provides sequences, phylogenetic trees, and hyperlinks to other databases. The Directory of P450-Containing Systems (Degtyarenko KN, et al; International Centre for Genetic Engineering and Biotechnology, Italy) provides access to internet resources of P450 proteins, P450-containing systems, steroid ligands known to bind to P450 and cross-links to a number of sequence, structure and function databases. UDP Glucuronosyltransferase home page (Committee for naming UDP Glucuronosyltransferases, Flinders University, Australia) gives detailed information about the sequence, multiple alignments, neighbor joining tree, and human alleles of UDP Glucuronosyltransferase.

Information about polymorphism of ADME-APs and possible links to variations of drug responses can be obtained from general genomics databases and those specializing in pharmacokinetics- and ADME-APs. Home page of the Human Cytochrome P450 (CYP) Allele Nomenclature Committee (http://www.imm.ki.se/CYPalleles/default.htm) provides comprehensive information about the genetic polymorphisms of 22 CYP alleles. The dbSNP database (NCBI, USA) provides comprehensive information about single nucleotide polymorphism (SNP) data for 29 organisms including human. It currently contains over 20 million SNP entries and over 8.9 million of these have been validated. GeneSNPs (University of Utah, USA) is a web resource that integrates gene, sequence and polymorphism data into individually annotated gene models. The human genes included are related to DNA repair, cell cycle control, cell signaling, cell division, homeostasis and metabolism,

and are thought to play a role in susceptibility to environmental exposure. PharmGKB database [264] is a central repository for genetic and clinical information about people who have participated in research studies at various medical centers in the PGRN. In addition, genomic data, molecular and cellular phenotype data, and clinical phenotype data are accepted from the scientific community at large. These data are organized and the relationships between genes and drugs are categorized into the categories of clinical outcome, pharmacodynamics and drug responses, pharmacokinetics, and molecular and cellular functional assays.

As indicated in previous sections, ADME-AP database [40, 73, 245] is also a useful resource not only for providing comprehensive information about the known ADME-APs, but also for obtaining pharmacogenetic data which currently contains information about 1,337 polymorphisms in 121 ADME- APs and 327 reported cases of altered drug responses.

6.2.2 Reported polymorphisms of ADME-associated proteins

Current progress in investigating pharmacogenomic polymorphisms of pharmacokinetic origin can be revealed from the analysis of the literature-reported polymorphisms of ADME-APs. A comprehensive search of the abstracts of Medline database [77] identified 1,337 SNPs in the coding regions, and a total of 13189 SNPs in all of the coding, non-coding and regulatory regions of 121 ADME-APs reported in the literature, some of which are given in Table 6-2.

Table 6-2: Examples of ADME-associated proteins with reported polymorphisms

ADME-associated Protein	Gene Name	ADME class	Protein Function	Number of reported SNPs in coding region	Total Number of reported SNPs*
Retinal-specific ATP-binding cassette transporter	ABCA4	D1	Transporter of retinoids	219	593
cAMP-dependent chloride channel	CFTR	D2	Transport of chloride ions	175	503
ATP-binding cassette sub-family D member 1	ABCD1	D	A possible transporter	125	843
Copper-transporting ATPase 2	ATP7B	E	Export of copper out of the cells	125	196
Serum albumin	ALB	D3	Plasma protein binding to water, ions, fatty acids, hormones, bilirubin and drugs	64	68
Multidrug resistance-associated protein 6	ABCC6	E	Transporter of glutathione conjugates and drug efflux	37	430
ATP-binding cassette, sub-family A, member 1	ABCA1	D2	Camp-dependent and sulfonylurea-sensitive transporter of anions	34	709
Sulfonylurea receptor 1	ABCC8	D2	Regulator of ATP-sensitive k+ channels and insulin release	28	344
Cytochrome P450 2D6	CYP2D6	M1	Metabolizing enzyme for structurally unrelated compounds (steroids, fatty acids, and xenobiotics) and drugs (antiarrhythmics, antidepressants and beta-blockers)	18	51
Cytochrome P450 3A4	CYP3A4	M1	Metabolizing enzyme for structurally unrelated compounds (steroids, fatty acids, and xenobiotics) and over 50% of drugs	17	129
UDP-glucuronosyl-transferase 1A1	UGT1A1	M2	Enzyme responsible for conjugation and subsequent elimination of xenobiotics and endogenous compounds	17	65
Copper-transporting ATPase 1	ATP7A	D2	Transporter of copper to copper-requiring proteins	17	98
Solute carrier family 21 member 6	SLCO1B1	A	Sodium-independent transporter of cystine and neutral and dibasic amino acids	15	418
Neutral and basic amino acid transport protein rBAT	SLC3A1	A; D1	Na(+)-independent transporter of organic anions (pravastatin, estrone sulfate, prostaglandin e2, thromboxane b2, leukotriene c3, thyroxine)	15	145
Cytochrome P450 1B1	CYP1B1	M1	Metabolizing enzyme for structurally unrelated compounds (steroids, fatty acids, and xenobiotics) and an unknown molecule in eye development	14	85
Dimethylaniline monooxygenase	FMO3	D1	Metabolizing enzyme of various xenobiotics such as drugs and pesticides	13	112
Thiazide-sensitive sodium-chloridecotransporter	SLC12A3	A	Transporter mediating sodium and chloride reabsorption	13	231
Sodium-dependent noradrenaline transporter	SLC6A2	D1	Sodium-dependent reuptake of noradrenaline	13	219
Antigen peptide transporter 1	TAP1	D2	Transporter of antigens from cytoplasm to a membrane-bound compartment	10	66
Multidrug resistance-associated protein 1	ABCC1	D2; E	Energy-dependent efflux pump transporting drugs into subcellular organelles	9	807

Bile salt export pump (ATP-binding cassette, sub-family B, member 11)	ABCB11	E	ATP-dependent secretion of bile salts into the canaliculus of hepatocytes	9	464
Cytochrome P450 1A1	CYP1A1	M1	Metabolizing enzyme for structurally unrelated compounds (steroids, fatty acids, and xenobiotics)	9	38
Cytochrome P450 2C8	CYP2C8	M1	Metabolizing enzyme for structurally unrelated compounds (steroids, fatty acids, and xenobiotics)	9	208
Arylamine N-acetyltransferase 1	NAT1	M2	Enzyme catalyzing the n- or o-acetylation of various arylamine and heterocyclic amine substrates	9	384
B(0,+)-type amino acid transporter 1	SLC7A9	A	Transporter of cystine, and neutral and dibasic amino acids.	9	158
Antigen peptide transporter 2	TAP2	D2	Transporter of antigens from the cytoplasm to a membrane-bound compartment	8	210
Arylamine N-acetyltransferase 2	NAT2	M2	Enzyme catalyzing the n- or o-acetylation of various arylamine and heterocyclic amine substrates	8	176
Dimethylaniline monooxygenase [N-oxide forming] 2	FMO2	M1	Enzyme catalyzing the n-oxidation of certain primary alkylamines to their oximes	8	214
Sodium/iodide cotransporter	SLC5A5	D1	Iodide uptake in the thyroid gland	8	58
Solute carrier family 2 (Glucose transporter type 1, erythrocyte/ brain)	SLC2A1	D1	Basal and growth factor-stimulated transporter of glucose and aldoses	8	167
Multidrug resistance-associated protein 2	ABCC2	D2; E	Hepatobiliary excretion of numerous organic anions	8	239
Steroidogenic acute regulatory protein	STAR	D1	Protein enhancing the metabolism of cholesterol into pregnenolone, involved in transport of cholesterol	8	212
Multidrug resistance protein 1	ABCB1	D2; E	Energy-dependent efflux pump transporting drugs into subcellular organelles	8	445
Cytochrome P450 2B6	CYP2B6	M1	Metabolizing enzyme for structurally unrelated compounds (steroids, fatty acids, and xenobiotics)	7	162
Dihydropyrimidine dehydrogenase	DPYD	M1	Enzyme involved in the reduction of uracil and thymine	7	1983
Prostacyclin synthase	PTGIS	M1	Enzyme catalyzing the isomerization of prostaglandin h2 to prostacyclin	7	276
Thiopurine S-methyltransferase	TPMT	M2	Enzyme catalyzing the s-methylation of thiopurine drugs	7	102
Mitochondrial ornithine transporter 1	SLC25A15	D	Ornithine transporter	7	88
Organic cation/carnitine transporter 2	SLC22A5	A	Transporter of organic cations	6	123
Epoxide hydrolase 1	EPHX1	M1	Enzyme catalyzing the hydrolysis of arene and aliphatic epoxides	6	111
Cytochrome P450 2C9	CYP2C9	M1	Metabolizing enzyme for structurally unrelated compounds (steroids, fatty acids, and xenobiotics)	6	248
Prostaglandin G/H synthase 2	PTGS2	M1	Likely a major mediator of inflammation and/or a role for prostanoid signaling	6	150
Cytochrome P450 2A6	CYP2A6	M1	Metabolizing enzyme for structurally unrelated compounds (steroids, fatty acids, and xenobiotics) and anti-cancer drugs (cyclophosphamide and ifosphamide)	5	91
Cytochrome P450 4F2	CYP4F2	M1	Metabolizing enzyme for structurally unrelated compounds (steroids, fatty acids, and xenobiotics)	5	208
Potential phospholipid-transporting ATPase	ATP8B1	A	Transporter of aminophospholipids	5	307
Sulfate transporter	SLC26A2	A; D1	Transporter of sulfate	5	62
Cytochrome P450 2J2	CYP2J2	M1	Metabolizing enzyme for arachidonic acid	5	193

* Including SNPs in coding, non-coding, and regulatory regions

Mutations arising from coding regions may lead to altered protein structure, and polymorphisms in non-coding and regulatory regions such as promoters may influence the level of expression, inducibility, and post-transcription processing, thereby affecting the functional roles of these ADME-APs. Most of these proteins are important drug transporters such as multidrug resistance-associated proteins [265] and metabolizing enzymes such as cytochrome P450s [66, 266] and UDP-glucuronosyl-transferases [267].

Examples of proteins containing a higher number of reported coding region SNPs are retinal-specific ATP-binding cassette transporter with 219 variants, cAMP-dependent chloride channel with 175 variants, adrenoleukodystrophy protein with 125 variants, copper-transporting ATPase 2 with 125 variants, serum albumin with 64 variants, multidrug resistance-associated protein 6 with 37 variants, ATP-binding cassette sub-family A member 1 protein with 34 variants, cytochrome P450 2D6 with 18 variants, cytochrome P450 3A4 with 17 variants, multidrug resistance-associated protein 1 with 17 variants, UDP-glucuronosyltransferase A with 17 variants, copper-transporting ATPase 1 with 17 variants, solute carrier family 21 member 6 protein with 15 variants, neutral and basic amino acid transport protein rBAT with 15 variants, and cytochrome P450 1B1 with 14 variants.

Examples of proteins containing a higher number of reported SNPs in their respective coding, non-coding and regulatory regions are dihydropyrimidine dehydrogenase with 1983 SNPs, ATP-binding cassette sub-family D member 1 with 843 SNPs, multidrug resistance-associated protein 1 with 807 SNPs, ATP-binding cassette, sub-family A member 1 with 709 SNPs, retinal-specific ATP-binding cassette transporter with 593 SNPs, cAMP-dependent chloride channel with 503 SNPs, bile salt export pump

(ATP-binding cassette, sub-family B, member 11) with 464 SNPs, multidrug resistance protein 1 with 445 SNPs, multidrug resistance-associated protein 6 with 430 SNPs, and solute carrier family 21 member 6 with 418 SNPs.

6.2.3 ADME-associated proteins linked to reported drug response variations

The role of specific ADME-APs in pharmacogenetics can be probed from its relationship with the reported drug response variations. Table 6-3 gives 35 ADME-APs that have been linked to the reported variations in drug response, many of which are drug metabolizing enzymes [66] and multidrug resistance-associated proteins [265]. The altered drug responses include both altered pharmacological effect and altered kinetics.

Examples of proteins linked to the reported variations in drug response are cytochrome P450 2D6 which are associated with variations for 61 drugs, multidrug resistance protein 1 for 25 drugs, cytochrome P450 2C19 for 22 drugs, cytochrome P450 2C9 for 22 drugs, arylamine N-acetyltransferase 2 for 17 drugs, cytochrome P450 3A4 for 18 drugs, sodium-dependent serotonin transporter for 9 drugs, cytochrome P450 1A2 for 13 drugs, cytochrome P450 2E1 for 12 drugs, thiopurine S-methyltransferase for 4 drugs, and UDP-glucuronosyltransferase 1A1 for 7 drugs.

Table 6-3: Examples of ADME-associated proteins linked to reported cases of individual variations in drug response

ADME-associated Protein	ADME class	Protein Function	Drugs with altered response* linked to protein	
			Number of drugs	List of Drugs
Cytochrome P450 2D6	M1	Metabolizing enzyme for structurally unrelated compounds (steroids, fatty acids, and xenobiotics) and drugs (antiarrhythmics, antidepressants and beta-blockers)	61	Alprenolol, Amiodarone, Amitriptyline, Bufuralol, Carbamazepine, Carvedilol, Chlorpromazine, Clarithromycin, Clomipramine, Codeine, Debrisoquin, Debrisoquine, Desipramine, Dextromethorphan, Diltiazem, Dihydrocodeine, Encainide, Ethylmorphine, Flecainide, Fluoxetine, Fluvoxamine, Guanoxan, Haloperidol, Hydrocodone, Imipramine, Losartan, Maprotiline, Maprotyline, Methoxyamphetamine, Metoprolol, Mexiletine, Mianserin, Mianserine, Nefazodon, Nortriptyline, N-propylajmaline, Ondasetron, Oxycodone, Perhexiline, Perphenazine, Phenacetin, Phenformin, Phenformine, Phenytoin, Propafenone, Propranolol, Resperidone, Risperidone, Ritinovir, Simvastatin, S-Mianserin, Sparteine, Tamoxifen, Theophylline, Thioridazine, Timolol, Tramadol, Trazodon, Tropisetron, Venlafaxine, Venlafazine
Multidrug resistance protein 1 (ABCB1 or MDR1)	D2; E	Energy-dependent efflux pump transporting drugs into subcellular organelles	25	Amiodarone, Cefazolin, Cefotetan, Cis-flupenthixol, Cyclosporin A, Cyclosporine, Digoxin, Diltiazem, Efavirenz, Fexofenadine, Indinavir, Irinotecan, Mitoxantrone, Morphine, Nelfinavir, Nicardipine, Nortriptyline, Ondansetron, Phenytoin, Quinidine, Tacrolimus, Tamoxifen, Topotecan, Trifluperazine, Verapamil
Cytochrome P450 2C19	M1	Metabolizing enzyme for structurally unrelated compounds (steroids, fatty acids, and xenobiotics) and anticonvulsant drugs	22	Citalopram, Clarithromycin, Diazepam, Difebarbamate, Febarbamate, Fluoxetine, Hexobarbital, Imipramine, Isoniazid, MePhenytoin, Mephobarbital, Nortriptyline, Omeprazole, Phenobarbital, Phenytoin, Proguanil, Propranolol, Rifampin, Sertraline, Valproate, Warfarin, Zonisamide
Cytochrome P450 2C9	M1	Matabolizing enzyme for structurally unrelated compounds (steroids, fatty acids, and xenobiotics) and many polar drugs (ibuprofen, naproxen, diclofenac and sulphaphenazole)	22	Diclofenac, Fluoxetine, Glibenclamide, Glimepiride, Glipazide, Glipizide, Glyburide, Ibuprofen, Imipramine, Irbesartan, Isoniazid, Lornoxicam , Losartan, Naproxen, Nateglinide , Phenytoin, Piroxicam, Rifampin, Tenoxicam, Tolbutamide, Verapamil, Warfarin

Cytochrome P450 3A4	M1	Metabolizing enzyme for structurally unrelated compounds (steroids, fatty acids, and xenobiotics) and over 50% of drugs.	18	Carbamazepine, Cisapride, Clonazepam, Cyclophosphamide, Cyclosporin A, Ethosuximide, Etoposide, Ganaxolone, Ifosphamide, Midazolam, Paclitaxel, Phenytoin, Tacrolimus, Teniposide, Tiagabine, Trimethadione, Vincas, Zonisamide
Arylamine N-acetyltransferase 2 (NAT2)	M2	Enzyme catalyzing the n- or o-acetylation of various arylamine and heterocyclic amine substrates.	17	Amonafide, Amonifide, Amrinone, Caffeine, Dapson, Dapsone, DiHydralazine, Hydralazine, Isoniazid, Paraminosalicylic acid, Phenelzine, Procainamide, Sufasalazine, Sulfamethazine, Sulfasalazine, Sulphamethoxazole, Sulphonamide Hydralysine
Cytochrome P450 1A2	M1	Metabolizing enzyme for structurally unrelated compounds (steroids, fatty acids, and xenobiotics)	13	Amonafide, Carbamazepine, Diltiazem, Erythromycin, Fluoxetine, Imipramine, Isoniazid, Naproxen, Nortriptyline hydrochloride, Phenytoin, Rifampin, Theophylline, Verapamil
Cytochrome P450 2E1	M1	Metabolizing enzyme for structurally unrelated compounds (steroids, fatty acids, and xenobiotics), and certain precarcinogens, drugs and solvents.	12	Alcohol, Diepoxybutane, Ethanol, Felbamate, Fluoxetine, Isoniazid, Phenobarbital, Phenytoin, Theophylline, Trimethadione, Valproate, Verapamil
Sodium-dependent serotonin transporter (5-HTT)	D1	Sodium-dependent reuptake of serotonine into presynaptic terminals.	9	Citalopram, Clomipramine, Fenfluramine, Fluoxetine, Fluvoxamine, Lithium, Nortriptyline, Paroxetine, Sertraline
UDP-glucuronosyl-transferase 1A1 (UGT1A1)	M2	Enzyme responsible for conjugation and elimination of xenobiotics and endogenous compounds.	7	Bilirubin (endogenous), Irinotecan, Estradiol, Tranilast, Etoposide, Atazanavir, Indinavir
Glutathione S-transferase Mu 1 (GSTM1)	M2	Enzyme responsible for conjugation of reduced glutathione to exogenous and endogenous hydrophobic electrophiles.	6	5-fluorouracil, Cyclophosphamide, Doxorubicin, D-penicillamine, Platinum, Tacrine
Glutathione S-transferase theta 1 (GSTT1)	M2	Enzyme responsible for conjugation of reduced glutathione to exogenous and endogenous hydrophobic electrophiles.	6	5-fluorouracil, Cyclophosphamide, Doxorubicin, Diepoxybutane, Platinum, Tacrine
Cytochrome P450 2A6	M1	Metabolizing enzyme for structurally unrelated compounds (steroids, fatty acids, and xenobiotics) and anti-cancer drugs (cyclophosphamide and ifosphamide)	6	Carbamazepine, Coumarin, Losigamone, Halothane, Nicotine, Valproate,
Catechol O-methyl-transferase (COMT)	M2	Enzyme for o-methylating and inactivating catecholamine neurotransmitters and catechol hormones	5	Ascorbic acid, Levodopa, Isoetharine, Isoprenaline, Methyldopa (Alpha-methyldopa)
Cytochrome P450 2C18	M1	Metabolizing enzyme for structurally unrelated compounds (steroids, fatty acids, and xenobiotics)	5	Fluoxetine, Imipramine, Phenytoin, Piroxicam, Rifampin
Thiopurine S-methyl-transferase (TPMT)	M2	Enzyme catalyzing the s-methylation of thiopurine drugs such as 6-mercaptopurine.	4	Azathioprine, Azathiopurin, Thioguanine (6-Thioguanine), Mercaptopurine (6-Mercaptopurine)
Cytochrome P450 2C8	M1	Metabolizing enzyme for structurally unrelated compounds (steroids, fatty acids, and xenobiotics), arachidonic acid, and anti-cancer drug paclitaxel	4	Carbamazepine, Paclitaxel, Phenytoin, Trimethadione

		(taxol)]		
Cytochrome P450 3A5	M1	Metabolizing enzyme for structurally unrelated compounds (steroids, fatty acids, and xenobiotics)	4	Cyclosporine, Tacrolimus, Phenytoin, Zonisamide
Cytochrome P450 2B6	M1	Metabolizing enzyme for structurally unrelated compounds (steroids, fatty acids, and xenobiotics)	4	Cyclophosphamide, Mephobarbital, Phenytoin, Valproate
ATP-binding cassette sub-family G member 2 (ABCG2 or BCRP)	D2; E	Xenobiotic transporter involved in the multidrug resistance phenotype of a specific mcf-7 breast cancer cell line	4	Diflomotecan, Irinotecan, mitoxantrone, topotecan
UDP-glucuronosyl-transferase 2B7 (UGT2B7)	M2	Enzyme responsible for conjugation and elimination of xenobiotics and endogenous compounds.	2	Epirubicin, Irinotecan
Amine oxidase [flavin-containing] A (MAOA)	M1	Enzyme catalyzing the oxidative deamination of biogenic and xenobiotic amines	2	Fluvoxamine, Moclobemide
Arylamine N-acetyl-transferase 1 (NAT1)	M2	Enzyme catalyzing the n- or o-acetylation of various arylamine and heterocyclic amine substrates	2	P-aminobenzoic acid, Para-substituted arylamine
Dihydropyrimidine dehydrogenase (DPD)	M1	Enzyme catalyzing the reduction of uracil and thymine	1	Fluorouracil (5-Fluorouracil)
Sulfonylurea receptor 1	D2	Regulator of ATP-sensitive k+ channels and insulin release.	1	Tolbutamide
Cytochrome P450 3A7	M1	Metabolizing enzyme for structurally unrelated compounds (steroids, fatty acids, and xenobiotics)	1	Phenytoin
ATP-binding cassette, sub-family A, member 1 (ABCA1)	D2	Camp-dependent and sulfonylurea-sensitive transportor of anions.	1	Fluvastatin
Epoxide hydrolase 1 (EH)	M1	Enzyme catalyzing the hydrolysis of arene and aliphatic epoxides to less reactive and more water soluble dihydrodiols.	1	Diepoxybutane
Sodium-dependent dopamine transporter (DAT1)	D1	Sodium-dependent reuptake of dopamine into presynaptic terminals.	1	Cocaine
Multidrug resistance-associated protein 1 (ABCC1 or MRP1)	D2; E	Energy-dependent efflux pump transporting drugs into subcellular organelles	1	Doxorubicin
Liver carboxylesterase	M1	Enzyme hydrolyzing aromatic and aliphatic esters	1	SN-38 (from the prodrug irinotecan)
NAD(P)H dehydrogenase [quinone] 1 (NQO1 or DT-diaphorase)	M1	Enzyme responsible for conjugation reactions of hydroquinons, a quinone reductase	1	Menadione
Carbonyl reductase [NADPH] 1 (Carbonyl reductase 1)	M1	Enzyme catalyzing the reduction of various carbonyl compounds	1	Doxorubicin
Multidrug resistance-associated protein 4 (ABCC4 or MRP4)	D2; E	Transporter acting as an organic anion pump	1	Azidothymidine
Excitatory amino acid transporter 2 (EAAT2)	D1	Transporter of l-glutamate and also l- and d-aspartate	1	3-Nitropropionic acid

Cytochrome P450 2D6 is responsible for the metabolism of most psychoactive drugs and it accounts for 20-30% of drugs metabolized by all cytochrome P450 enzymes [266]. It metabolizes drugs for several diseases including depression, psychosis, cancer, and pain. Changes in the metabolism of these drugs are expected to have a significant impact on the level of toxic effects as well as therapeutic effects induced by these drugs. Thus it is not surprising that this enzyme affects the response of a large number of drugs. Examples of other cytochrome P450 enzymes affecting a wide spectrum of drugs are CYP2C9 [268], which metabolizes 10% drugs and affects drugs for depression, cardiovascular, and epilepsy, and CYP2C19, which metabolizes 5% drugs and affects drugs for depression and ulcer [268]. Although CYP3A4 is known to metabolize 40-45% of drugs, there has been insufficient study about the clinical effects of the polymorphisms of this enzyme [266].

Multidrug resistance protein 1, which affects the response to the second largest number of drugs, is an energy-dependent cellular efflux protein responsible for the efflux of a wide spectrum of drugs including bilirubin, some anticancer agents, cardiac glycosides, immunosuppressive agents, glucocorticoids, HIV-1 protease inhibitors [269-271]. It plays important roles in the excretion of xenobiotics and metabolites into urine, bile, and intestine lumen [272, 273]. It also limits the accumulation of many drugs in the brain including digoxin, ivermectin, vinblastine, dexamethasone, cyclosporine, domperidone, and loperamide [272-274].

6.2.4 Development of rule-based prediction system

Established links between polymorphisms of ADME-APs and individual drug responses have been used in combination with genetic studies as indicators for

predicting individual variations of drug response [252-256]. Based on the analysis of clinical samples of the variation of drug response and the results of genetic analysis of the participating patients, simple rules can be derived for the prediction of individual variations of drug response from the polymorphism of specific protein [252, 253, 255] or combination of polymorphisms of multiple proteins [275, 276]. Experimental techniques capable of differentiating between a single wild-type sequence and mutant sequences can then be used to detect these polymorphisms and predict drug response.

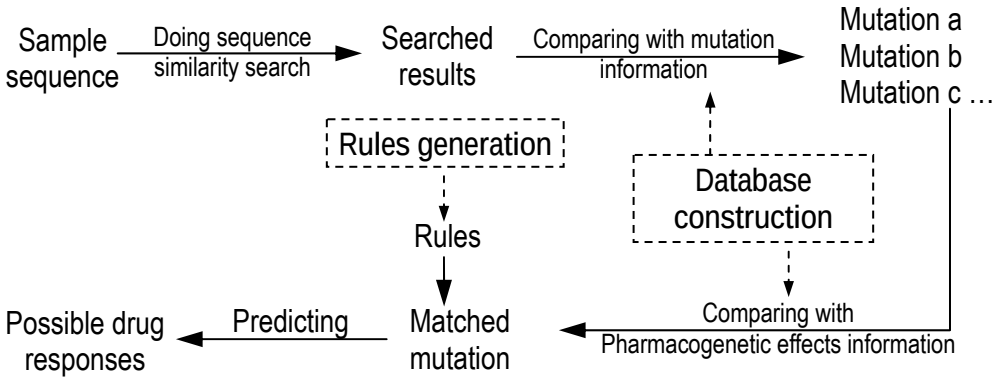


Figure 6-4: The flow chart of development of rule-based prediction system

Figure 6-4 describes the process of developing rule-based prediction system. Firstly, a sample sequence is conducted to do sequence similarity search by using BLAST [240]. The searched results are then introduced to compare with mutation information collected in ADME-APs database. Several mutations occurring in sample sequence are confirmed and they are searched in pharmacogenetic effects library integrated in ADME-APs database. Based on the rules generated in advance, the matched mutation can be used to predict possible drug responses. In this flow chart, there are two key components in this process. One is about database construction, which has been discussed in previous sections. The other is about rules generation. Generally, the rules can be simple rules which list several possibilities of the prediction results. The

rules can also be some numeric profiles generated by computational methods and interpreted by computer.

6.2.4.1 Rule-based prediction of drug responses from the polymorphisms of ADME-associated proteins

Established links between polymorphisms of ADME-APs and individual drug responses have been used in combination with genetic studies as indicators for predicting individual variations of drug response [252-256]. Based on the analysis of clinical samples of the variation of drug response and the results of genetic analysis of the participating patients, simple rules can be derived for the prediction of individual variations of drug response from the polymorphism of specific protein [252, 253, 255] or combination of polymorphisms of multiple proteins [275, 276]. Experimental techniques capable of differentiating between a single wild-type sequence and mutant sequences can then be used to detect these polymorphisms and predict drug response. The simple rules generated and applied in these studies may be collected and used for developing a computer prediction system in a similar fashion like that of the HIV drug resistant genotype interpretation systems [277]. Table 6-4 gives examples of the ADME-APs with a known pharmacogenetic polymorphism and a reasonably accurate rule for predicting responses to a specific drug or drug group reported in the literature.

Table 6-4: Prediction of specific drug responses from the polymorphisms of ADME associated proteins by using simple rules

Protein	Drugs and Treatment/Action	Drug Responses	Polymorphism Rules and Year of Report	Number of Patients with Polymorphism	Prediction Accuracy
Cytochrome P450 1A2	Antipsychotic agents for schizophrenia patients	Tardive dyskinesia	Bsp120I (C→A) polymorphism in CYP1A2 gene, 2000 [278]	85	69%
Cytochrome P450 2C9	Anticoagulant agents for the initial phase of phenprocoumon treatment	Severe over-anticoagulation	CYP2C9*2 genotype, 2004 [279]	61	26%
			CYP2C9*3 genotype, 2004 [279]	37	22%
Cytochrome P450 2D6	Neuroleptic agents for chronic schizophrenic patients	Tardive dyskinesia	CYP2D6*4 genotype, 1998 [280]	13	81%
Cytochrome P450 2D6	Psychochopic drugs for psychiartric illness	Extrapyramidal drug side effects	CYP2D6 PM phenotypes, 1999 [281]	22	45%
Cytochrome P450 2D6	CYP2D6-dependent antidepressants	Drug non-response	CYP2D6 EM phenotypes, 2004 [282]	16	75%
UDP-glucuronysl-transferase	Capecitabine/irinotecan for the treatment of metastatic colorectal cancer	Greater antitumor response with low toxicity	UGT1A7*2/*2 genotype, 2005 [256]	6	100%
			UGT1A7*3/*3 genotype, 2005 [256]	7	100%
UDP-glucuronysl-transferase I	Tranilast for the prevention of restenosis following coronary revascularization	Hyper-bilirubinemia	Homozygosity for a (TA)7-repeat element within the promotor region of UGT1A1 gene, 2004 [252]	146	40%
N-acetyl-transferase 2	Trimethoprim-sulfamethoxazole for the treatment of infections in infants	Idiosyncratic reactions such as fever, skin rash and multiorgan toxicity	NAT2*5A allele, 1997 [283]	18	89%
			NAT2*5C allele, 1997 [283]	5	80%
			NAT2*7B, 1997 [283]	3	67%
N-acetyl-transferase 2	Aromatic amine carcinogens in tobacco smoke	Hepatitis B related hepatocellular carcinoma	NAT2*4 allele, 2000 [284]	76	53%
N-acetyl-transferase 2	Isonaiazid for the prophylaxis and treatment of tuberculosis	ADRs such as peripheral neuritis, fever and hepatic toxicity	SA type (NAT2*6/*6, NAT2*6/*7, and NAT2*7/*7), 2002 [254]	6	83%
Tryptophan hydroxylase	Fluvoxamine for the treatment of depression	Antidepressant response	A218C A/C phenotypes, 2001 [285]	107	76%
			A218C C/C phenotypes, 2001 [285]	70	81%
			A218C A/A phenotypes, 2001 [285]	40	65%
Nor-epinephrine transporter	Milnacipran for the treatment of depression	Antidepressant response	T allele of the NET T182C polymorphism, 2004 [255]	50	72%
Seratonin transporter	Serotonin reuptake inhibitors for the	Antidepressant response	s/s genotype of serotonin transporter gene promoter	11-72	54% at 6th week

	treatment of depression		region, 2000-2004 [59, 60, 282, 286]		
			s/l genotype of serotonin transporter gene promoter region, 2000-2004 [59, 60, 282, 286]	20-47	55% at 6th week
			l/l genotype of serotonin transporter gene promoter region, 2000-2004 [59, 60, 282, 286]	4-16	48% at 6th week
Multidrug resistance-associated protein 1	Epileptic drugs for the treatment of epilepsy	Drug resistant epilepsy	ABCB1 C3435T C/C genotype, 2003 [287]	73	75%
			ABCB1 C3435T C/T genotype, 2003 [287]	169	63%
			ABCB1 C3435T T/T genotype, 2003 [287]	73	53%
Multidrug resistance-associated protein 1	Combination therapy of nelfinavir, efavirenz, and nucleoside reverse transcriptase inhibitors for HIV-1 infected children	Virologic response by week 8	MDR1 C3435T C/C genotype, 2005 [288]	31	59%
			MDR1 C3435T C/T genotype, 2005 [288]	33	91%

The reported prediction accuracy for the patients with specific polymorphism reported in the literature is also given. Based on the test of the patients described in these reports, most of these rules are capable of predicting drug responses at accuracies of 50%~100%, which are not too much lower than and in many cases comparable to the accuracies of 81%~97% for predicting HIV drug resistance mutations from the HIV resistant genotype interpretation systems [277]. This suggests that these simple rules have certain level of capacity for facilitating pharmacogenetic prediction of drug response and they may be used as the basis for developing more sophisticated interpretation systems like those of HIV resistant genotype interpretation systems [277].

Variation of response to some drugs is known to be associated with interactions between genetic polymorphisms in more than one protein [275, 276]. For instance, specific polymorphisms in cytochrome P450 7A1 (CYP7A1) and ATP-binding

cassette transporters G5 and G6 (ABCG5/G6) are known to affect LDL cholesterol-lowering response to atorvastatin. The combination of the polymorphisms in CYP7A1 and ABCG5/G6 explained a greater percentage of response variations (8.5%) than the single polymorphism in each of these proteins (4.2% in CYP7A1 and 3.0% in ABCG5/G6) [276]. Therefore, in such cases, simple rules based on single polymorphism in one protein are insufficient for predicting individual variations of drug responses. Rules that take into consideration of complex interaction of polymorphisms in multiple proteins [275, 276], gene expression patterns [258], and environmental factors [259] likely give more accurate prediction ranges and accuracy.

Some of the pharmacogenetic studies have been based on a limited number of samples and the derived data may show various degrees of deviations. For instance, in a systematic review of the literature on the influence of polymorphisms in the serotonin transporter gene on SSRI response, it was found that both the investigation methodologies and research outcomes showed large heterogeneity, which led to the conclusion that the current information is insufficiently reliable as a basis for implementing pharmacogenetic testing of depressive patients [59, 60, 282, 286]. This is not surprising when the neurochemistry of the drugs is considered. Increased synaptic availability of serotonin is known to stimulate a large number of post-synaptic receptors yet down-regulate others. Therefore, the relevant data may need to be interpreted cautiously [289] particularly in applying them for pharmacogenetic prediction of individual variation of drug response.

6.2.4.2 Computational methods for analysis and prediction of pharmacogenetics of drug responses from the polymorphisms of ADME-associated proteins

The complex pharmacogenetic interactions of proteins [275, 276], complicated microarray-based gene expression profiles [258], and multitude of patient data (physical conditions, medications, food consumptions, outdoor activities etc.) [259] used in pharmacogenetic analysis and prediction of drug responses require the application of more sophisticated statistical analysis and statistical learning methods than those of simple rule-based and linear methods [257-260]. Table 6-5 summarizes the computational methods recently explored for pharmacogenetic prediction of drug responses. These methods include discriminant analysis (DA) [259], unconditional logistic regression [284], random regression model [290], conditional logistic regression, 2004 [260], artificial neural networks (ANN) [257, 259], and maximum likelihood context model from haplotype structure provided by HapMap [291].

Table 6-5: Statistical analysis and statistical learning methods used for pharmacogenetic prediction of drug responses

Method and Year of Report	Protein Polymorphisms	Drugs and Treatment/ Action	Drug Responses	Number of Patients	Computed percentage of drug response
Unconditional logistic regression, 2000 [284]	NAT2*4 allele of N-acetyl-transferase 2	Aromatic amine carcinogens in tobacco smoke	Hepatitis B related hepatocellular carcinoma	76	53%
Random regression model, 2001 [290]	Serotonin transporter gene-linked functional polymorphisms	Fluvoxamine for antidepressant activity	Variation of response to antidepressan t activity	155	70%~87%
Discriminant analysis, 2003 [259]	Four polymorphisms in cytochrome P450c17, E-cadherin, urokinase and VEGF	Various foodstuffs and drinks	Calcium oxalate stone disease	151	74%
Artificial neural networks, 2003 [259]	Four polymorphisms in cytochrome P450c17, E-cadherin, urokinase and VEGF	Various foodstuffs and drinks	Calcium oxalate stone disease	151	89%
Artificial neural	Polymorphisms in the	Antidepressant	Variation of	121	78% for

networks, 2004 [257]	transcriptional control region upstream of serotonin receptor coding sequence and in tryptophan hydroxylase	agents for the treatment of mood disorders	drug response		responders, 51% for non-responders
Logistic regression, 2004 [260]	Polymorphisms in MDR1 (C3435T) and IL-10 (-1082G/A, -592A/C)	Corticosteroid for immuno-suppression in pediatric heart transplant patients	Steroid dependency	47 with MDR1 C3435T CT/TT; 15, 28 and 26 with IL-10 high, intermediate and low producer genotype	62% for MDR1 C3435T CT/TT; 43% ~ 87 % for IL-10 genotypes
Maximum likelihood context model from haplotype structure provided by HapMap, 2005 [291]	2AR allele Gly16 at codon 16 and allele Glu27 at codon 27	Dobutamine for the treatment of cardiovascular disease	Variation of drug response	107	14% of the total observed variation

DA determines a linear combination of input feature variables and forms a linear discriminate function which could provide the maximum degree of distinction among the different drug response groups [259]. RRM attempts to explain the relationship between the drug responses and their pharmacogenetic origins by constructing a statistical model that fits to the multi-variable data [290]. Logistic regression (conditional and unconditional) produces a prediction equation by determining regression coefficients which measure the predictive capability of the input independent variables [260, 284]. It predicts the occurrence possibility of an event which could be interpreted as the ratio of the probability of the occurrence of a particular pharmacogenetic event to that without the event.

ANN trains a hidden-layer-containing network and uses its outcomes for pattern recognition and classification of the input feature vectors [292, 293], with each vector representing various data of a patient. A classifier for ANN is $y = g\sum_j w_{0j} h_j$, where

w_{0j} is the output weight of a hidden node j to an output node, g is the output function, h_j is the value of a hidden layer node: $h_j = \delta(\sum_j w_{ji} x_j + w_j)$, w_{ji} is the input weight from an input node i to a hidden node j, w_j is the threshold weight from an input node of value 1 to a hidden node j, and δ is a sigmoid function. Known resistance and non-resistance samples are used for training an ANN such that all the weights are determined, and the resulting classifier can be used for determining whether or not a new input data of a patient responds to a drug.

The haplotype structures of HapMap reveal variation patterns in DNA sequences, from which a statistical model can be developed to directly characterize specific DNA sequence variants responsible for drug response [291]. One such model has been developed in the maximum likelihood context, which is represented by clinically meaningful mathematical functions modeling drug response and is implemented by an integrative EM algorithm.

The application and performance of statistical analysis and statistical learning methods depends on several factors including knowledge of related proteins, availability of sequence and polymorphism data, establishment of quantitative relationship between polymorphism and drug response from sufficient number of patients, and appropriate representation of genetic polymorphisms and other properties such as expression profiles and environmental factors. For instance, a sufficiently diverse set of response and non-response samples is needed for training a sophisticated statistical learning system such as ANN and SVMs which have been successfully applied for predicting drug resistance mutations directly from protein sequence [251]. Thus these methods are not applicable for proteins and drugs with

little or no polymorphism and drug response data. Mining of polymorphism and drug response data from the literature [294-296] and other sources [73, 245, 264], is a key to more extensive exploration of statistical learning methods as well as rule-based methods for pharmacogenetic prediction of individual variations of drug responses.

6.3 Conclusion

Knowledge about ADME-APs, polymorphisms and drug responses appears to have reached a meaningful level to facilitate pharmacogenetic prediction of various types of individual variations of drug responses. Internet sites such as the ADME-AP database and PharmGKB database serve as convenient resources for obtaining the relevant information. With the rapid development of genomics [32], pharmacokinetics [297-300], and pharmacogenomics [64, 66, 67], more information about ADME-APs, polymorphisms and variations of drug responses are expected to become available. Moreover, progress in the study of proteomics [31] and pathways [301] related to drug ADME-APs is expected to further facilitate our understanding of the mechanism of drug disposition and their possible contribution to individual variations in drug response.

Both rule-based methods and statistical learning methods have consistently shown a promising capability for predicting individual variations of drug responses from polymorphisms of ADME-APs as well as those of therapeutic targets and ADR-related proteins. The availability of more comprehensive information about ADME-APs, polymorphisms and variations of drug responses will further extend the range of the application of these methods. It is expected that rules and methods that predict individual variations of drug responses on the basis of complex pharmacogenetic interactions will be more extensively explored. Methods that

improve the prediction accuracy in cases of imbalanced datasets, such as those with too small number of drug respondents, are being developed [302] and these may be applied to further improve the accuracy of drug responses.

7 Conclusion

In the post-genomic age, multi-disciplinary bioinformatics approaches are widely used to advance drug discovery. In this regard, the objective of this study were to develop and update three related pharmainformatics databases, namely, TTD, TRMP database, and ADME-AP database.

In the new version of TTD, the number of reported targets was increased to 1,535. Not only much additional relevant information has been included in the database to provide more comprehensive knowledge about the therapeutic targets, but also the data structure has been rearranged and the web interface has been rewritten to facilitate the better search of targets and corresponding drug/ligand, and disease information. Likewise, TRMP database has been developed to understand comprehensively the relationship between different targets of the same disease and facilitate mechanistic study of drug actions. It contains 11 entries of multiple pathways, 97 entries of individual pathways, 120 targets covering 72 disease conditions together with 120 sets of drugs directed at each of these targets. Also, information about 1,337 polymorphisms in 121 proteins, and 327 drugs with altered responses linked to ADME-APs has been added into the new version of ADME-AP database. By studying pharmacogenetic data, we find it could be feasible to do pharmacogenetic prediction of drug responses and individual variations from the polymorphisms of ADME-APs.

Consequently, these databases provide comprehensive information of known therapeutic targets, pathways, and ADME-APs and can serve as platforms to the scientific understanding of therapeutically relevant events. Particularly, knowledge of targets is helpful for molecular dissection of the mechanism of action of drugs, and for

predicting features that guide new drug design and the search for new targets. Based on therapeutic targets relevant profiles and their characteristics described in earlier investigations [37, 195], the following simple rules were derived for characterizing druggable proteins:

- The druggable protein is from one of the target-representing protein families.
- Sequence variation between the drug-binding domain of a druggable protein and those of the other human members of its protein family needs to allow sufficient degree of differential binding of a "rule-of-five" molecule to the common binding site.
- The druggable protein is preferable to have less than 15 human similarity proteins outside its family (HSP).
- The druggable protein is preferable to be involved in no more than 3 pathways in human (HP).
- For organ or tissue specific diseases, the druggable protein is preferable to be distributed in no more than 5 tissues in human (HT).
- A higher number of HSP, HP and HT does not preclude the protein as a potential target, it statistically increase the chance of unwanted interferences and the level of difficulty for finding viable drugs.

Furthermore, a SVM prediction system, which was developed by using 1,174 targets and 12,956 non-druggable proteins from 6,856 non-target families, was constructed to predict possible therapeutic targets. Its estimated prediction accuracy was 69.8% and 99.3% for druggable and non-druggable proteins respectively, based on a 5-fold cross validation study. In addition, to test its potential for practical applications, the constructed SVM prediction system was used to scan the human, yeast, and HIV genomes to identify potential druggable proteins that were not in the training and

testing sets. The results suggested that statistical learning methods such as SVM would be potentially useful for facilitating genome search for druggable proteins.

In conclusion, this study provided knowledge platforms to facilitate pharmaceutical research by developing several useful databases, studied the possibilities of predicting pharmacogenetic effects from the polymorphisms of ADME-APs, and constructed a feasible prediction system to search potential candidates for therapeutic targets. Also, some limitations about the research are discussed here. According to the generated "rules" in the previous section, several limitations are listed as follows:

- The rules are generated from the statistics of the 1,535 currently known targets.
- The number of known successful targets is limited.
- The research targets need to be proved as successful targets in further clinical experiments.
- The annotation of many protein targets needs to be completed.

Therefore, the "rules" listed here are rough rules, which can be considered as a flexible profile to facilitate the search of druggable proteins. The feasibility of these rules is still waiting to be proven. Furthermore, with the development of modern biological technologies, more and more targets discovered by experiments will be added into the database, which would require generation of more elaborate rules from more comprehensive data. Eventually, the rules will play an important role in shortening the procedure of target discovery and speeding up the whole drug discovery. In addition, the prediction accuracy for druggable proteins needs to be improved. One reason for the lower accuracy of druggable proteins is the large imbalance between the number of druggable and non-druggable proteins. Such a large imbalance is known to affect the accuracy of a SVM prediction system.

In future research, there are several aspects that can be studied. Firstly, reliable data is the key to successful prediction. Thus, targets and non-targets should be chosen more strictly from the clinical experiment reports according to their different pharmacological properties, physical and chemical properties. Besides, since drug discovery is a fast growing area with many different communities internationally. The standards of America for drug designs and trials, or drug patents can be rather different from those of Europe and Asia. As a result, future work should be extended to study those marketed drug approved by other communities. Secondly, knowledge about therapeutic pathways as well as that of drug targets and ADME-APs should be used in analysis of mechanism of drug action and disease relevant events, especially on the aspect of relationships between drugs, targets and diseases. As a result, future work should pay more attention to understand the therapeutic targets and ADME-APs in overall views. That is to say, it would be better put them into a specific pathologic context for study, rather than consider them individually. Thirdly, regarding database development, an open architecture with the databases should be added to facilitate public to submit entries missed by our databases. The new submitted data can be manually checked in further and filled into the databases. Moreover, effective text mining technique also needs to be explored to facilitate information collection. Fourthly, the kernel function plays an important role in SVM prediction. Therefore, in order to effectively improve the prediction accuracy, the SVM kernel function, kernel KBF, should be further modified to address specified problem, druggable proteins prediction. In addition, other kernels, such as kernel PCA, kernel ICA, or introducing text kernel, should be explored in future research. Finally, although Support Vector Machine (SVM) methods have many advantages in huge data classification, they have a few disadvantages inherently, such as inability to handle the imbalance data properly,

inability to distinguish the predominant features, and working as a black box. As a consequence, other effective prediction systems (neural networks, consensus model, QSAR, etc.) should be explored as complements to SVM for classifying imbalance data.

REFERENCES

1. Drews, J., *In Human disease - from genetic causes to biochemical effects*, J. Drews and S. Ryser, Editors. 1997, Blackwell: Berlin. p. 5-9.
2. WHO, *The world health report 2004 – changing history*. 2004, World Health Organization.
3. Sanseau, P., *Impact of human genome sequencing for in silico target discovery.* Drug Discov Today, 2001. **6**(6): p. 316-323.
4. Duckworth, D.M. and P. Sanseau, *In silico identification of novel therapeutic targets.* Drug Discov Today, 2002. **7**(11 Suppl): p. S64-9.
5. Walke, D.W., et al., *In vivo drug target discovery: identifying the best targets from the genome.* Curr Opin Biotechnol, 2001. **12**(6): p. 626-31.
6. Terstappen, G.C. and A. Reggiani, *In silico research in drug discovery.* Trends Pharmacol Sci, 2001. **22**(1): p. 23-6.
7. Swindells, M.B. and J.P. Overington, *Prioritizing the proteome: identifying pharmaceutically relevant targets.* Drug Discov Today, 2002. **7**(9): p. 516-21.
8. Lindsay, M.A., *Target discovery.* Nat Rev Drug Discov, 2003. **2**(10): p. 831-8.
9. Drews, J., *Drug discovery: a historical perspective.* Science, 2000. **287**(5460): p. 1960-4.
10. Wong, A.H., Gottesman, II, and A. Petronis, *Phenotypic differences in genetically identical organisms: the epigenetic perspective.* Hum Mol Genet, 2005. **14 Spec No 1**: p. R11-8.
11. NIH, *Working Definition of Bioinformatics and Computational Biology.* 2000.
12. Altman, R.B., *A curriculum for bioinformatics: the time is ripe.* Bioinformatics, 1998. **14**(7): p. 549-50.
13. Wheeler, D.L., et al., *Database resources of the National Center for Biotechnology Information: update.* Nucleic Acids Res, 2004. **32 Database issue**: p. D35-40.
14. Brooksbank, C., G. Cameron, and J. Thornton, *The European Bioinformatics Institute's data resources: towards systems biology.* Nucleic Acids Res, 2005. **33**(Database issue): p. D46-53.
15. Tateno, Y., et al., *DDBJ in collaboration with mass-sequencing teams on annotation.* Nucleic Acids Res, 2005. **33**(Database issue): p. D25-8.
16. Kanehisa, M., *The KEGG database.* Novartis Found Symp, 2002. **247**: p. 91-101; discussion 101-3, 119-28, 244-52.
17. Gibas, C. and P. Jambek, *Developing Bioinformatics Computer Skills*. 2001: O'Reilly & Associates. 427.
18. Sanger, F., *Chemistry of insulin; determination of the structure of insulin opens the way to greater understanding of life processes.* Science, 1959. **129**(3359): p. 1340-4.
19. Holley, R.W., et al., *Structure Of A Ribonucleic Acid.* Science, 1965. **147**: p. 1462-5.
20. Dayhoff, M.O., et al., *Atlas of Protein Sequence and Structure.* National Biomedical Research Foundation, 1965.
21. Dayhoff, M.O., *Atlas of Protein Sequence and Structure.* National Biomedical Research Foundation, 1972. **5**.
22. Dayhoff, M.O., *Atlas of Protein Sequence and Structure.* National Biomedical Research Foundation, 1979. **5**(Suppl. 3).
23. Barker, W.C., et al., *The PIR-International Protein Sequence Database.*

Nucleic Acids Res, 1999. **27**(1): p. 39-43.
24. King, J.N., *Notes and News: Protein Data Bank.* Acta crystallographica. Section B: Structural crystallography and crystal chemistry., 1973. **B29**: p. 1764.
25. Boeckmann, B., et al., *The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.* Nucleic Acids Res, 2003. **31**(1): p. 365-70.
26. Galperin, M.Y., *The Molecular Biology Database Collection: 2005 update.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D5-24.
27. Frishman, D., et al., *Comprehensive, comprehensible, distributed and intelligent databases: current status.* Bioinformatics, 1998. **14**(7): p. 551-61.
28. Drews, J., *Strategic choices facing the pharmaceutical industry: a case for innovation.* Drug Discov. Today., 1997. **2**: p. 72-78.
29. Ohlstein, E.H., R.R. Ruffolo, Jr., and J.D. Elliott, *Drug discovery in the next millennium.* Annu Rev Pharmacol Toxicol, 2000. **40**: p. 177-91.
30. Sali, A., *100,000 protein structures for the biologist.* Nat Struct Biol, 1998. **5**(12): p. 1029-32.
31. Dove, A., *Proteomics: translating genomics into products?* Nat Biotechnol, 1999. **17**(3): p. 233-6.
32. Debouck, C. and B. Metcalf, *The impact of genomics on drug discovery.* Annu Rev Pharmacol Toxicol, 2000. **40**: p. 193-207.
33. Peltonen, L. and V.A. McKusick, *Genomics and medicine. Dissecting human disease in the postgenomic era.* Science, 2001. **291**(5507): p. 1224-9.
34. Baker, M.D. and J.N. Wood, *Involvement of Na+ channels in pain pathways.* Trends Pharmacol Sci, 2001. **22**(1): p. 27-31.
35. Macdonald, I.A., *Obesity: are we any closer to identifying causes and effective treatments?* Trends Pharmacol Sci, 2000. **21**(9): p. 334-6.
36. Drews, J. and S. Ryser, *The role of innovation in drug development.* Nat Biotechnol, 1997. **15**(13): p. 1318-9.
37. Hopkins, A.L. and C.R. Groom, *The druggable genome.* Nat Rev Drug Discov, 2002. **1**(9): p. 727-30.
38. Zambrowicz, B.P. and A.T. Sands, *Knockouts model the 100 best-selling drugs--will they model the next 100?* Nat Rev Drug Discov, 2003. **2**(1): p. 38-51.
39. Chen, X., Z.L. Ji, and Y.Z. Chen, *TTD: Therapeutic Target Database.* Nucleic Acids Res, 2002. **30**(1): p. 412-5.
40. Ji, Z.L., et al., *Internet resources for proteins associated with drug therapeutic effects, adverse reactions and ADME.* Drug Discov Today, 2003. **8**(12): p. 526-9.
41. Ignar-Trowbridge, D.M., et al., *Coupling of dual signaling pathways: epidermal growth factor action involves the estrogen receptor.* Proc Natl Acad Sci U S A, 1992. **89**(10): p. 4658-62.
42. Abdel-Latif, A.A., *Cross talk between cyclic AMP and the polyphosphoinositide signaling cascade in iris sphincter and other nonvascular smooth muscle.* Proc Soc Exp Biol Med, 1996. **211**(2): p. 163-77.
43. Zemel, M.B., *Agouti/melanocortin interactions with leptin pathways in obesity.* Nutr Rev, 1998. **56**(9): p. 271-4.
44. Adcock, I.M. and G. Caramori, *Cross-talk between pro-inflammatory transcription factors and glucocorticoids.* Immunol Cell Biol, 2001. **79**(4): p. 376-84.
45. Minatoguchi, S., et al., *Cross-talk among noradrenaline, adenosine and*

protein kinase C in the mechanisms of ischemic preconditioning in rabbits. J Cardiovasc Pharmacol, 2003. **41 Suppl 1**: p. S39-47.
46. Takayanagi, H., et al., *T-cell-mediated regulation of osteoclastogenesis by signalling cross-talk between RANKL and IFN-gamma.* Nature, 2000. **408**(6812): p. 600-5.
47. Kumar-Sinha, C., et al., *Molecular cross-talk between the TRAIL and interferon signaling pathways.* J Biol Chem, 2002. **277**(1): p. 575-85.
48. Alaoui-Jamali, M.A. and H. Qiang, *The interface between ErbB and non-ErbB receptors in tumor invasion: clinical implications and opportunities for target discovery.* Drug Resist Updat, 2003. **6**(2): p. 95-107.
49. Grodesky, M., et al., *Combination therapy with indinavir, ritonavir, and delavirdine and nucleoside reverse transcriptase inhibitors in patients with HIV/AIDS who have failed multiple antiretroviral combinations.* HIV Clin Trials, 2001. **2**(3): p. 193-9.
50. El-Rayes, B.F. and P.A. Philip, *Systemic therapy for advanced pancreatic cancer.* Expert Rev Anticancer Ther, 2002. **2**(4): p. 426-36.
51. Sagae, S., et al., *Combination therapy with granisetron, methylprednisolone and droperidol as an antiemetic prophylaxis in CDDP-induced delayed emesis for gynecologic cancer.* Oncology, 2003. **64**(1): p. 46-53.
52. Tariot, P.N. and H.J. Federoff, *Current treatment for Alzheimer disease and future prospects.* Alzheimer Dis Assoc Disord, 2003. **17 Suppl 4**: p. S105-13.
53. Kriz, J., G. Gowing, and J.P. Julien, *Efficient three-drug cocktail for disease induced by mutant superoxide dismutase.* Ann Neurol, 2003. **53**(4): p. 429-36.
54. Bays, H. and E.A. Stein, *Pharmacotherapy for dyslipidaemia--current therapies and future agents.* Expert Opin Pharmacother, 2003. **4**(11): p. 1901-38.
55. Huang, S., *Rational drug discovery: what can we learn from regulatory networks?* Drug Discov Today, 2002. **7**(20 Suppl): p. S163-9.
56. Lewis, D.A., *Antiretroviral combination therapy for HIV infection.* Dent Update, 2003. **30**(5): p. 242-7.
57. Yin, M.J., Y. Yamamoto, and R.B. Gaynor, *The anti-inflammatory agents aspirin and salicylate inhibit the activity of I(kappa)B kinase-beta.* Nature, 1998. **396**(6706): p. 77-80.
58. Alpert, D. and J. Vilcek, *Inhibition of IkappaB kinase activity by sodium salicylate in vitro does not reflect its inhibitory mechanism in intact cells.* J Biol Chem, 2000. **275**(15): p. 10925-9.
59. Evans, W.E. and M.V. Relling, *Moving towards individualized medicine with pharmacogenomics.* Nature, 2004. **429**(6990): p. 464-8.
60. Lindpaintner, K., *Pharmacogenetics and the future of medical practice.* J Mol Med, 2003. **81**(3): p. 141-53.
61. Marzolini, C., et al., *Polymorphisms in human MDR1 (P-glycoprotein): recent advances and clinical relevance.* Clin Pharmacol Ther, 2004. **75**(1): p. 13-33.
62. Evans, W.E., et al., *Preponderance of thiopurine S-methyltransferase deficiency and heterozygosity among patients intolerant to mercaptopurine or azathioprine.* J Clin Oncol, 2001. **19**(8): p. 2293-301.
63. Marshall, A., *Laying the foundations for personalized medicines.* Nat Biotechnol, 1997. **15**(10): p. 954-7.
64. Altman, R.B. and T.E. Klein, *Challenges for biomedical informatics and pharmacogenomics.* Annu Rev Pharmacol Toxicol, 2002. **42**: p. 113-33.
65. Kalow, W., *Pharmacogenetics, pharmacogenomics, and pharmacobiology.*

Clin Pharmacol Ther, 2001. **70**(1): p. 1-4.
66. Marshall, A., *Getting the right drug into the right patient.* Nat Biotechnol, 1997. **15**(12): p. 1249-52.
67. Nicholls, H., *Improving drug response with pharmacogenomics.* Drug Discov Today, 2003. **8**(7): p. 281-2.
68. Marth, G.T., et al., *A general approach to single-nucleotide polymorphism discovery.* Nat Genet, 1999. **23**(4): p. 452-6.
69. Sachidanandam, R., et al., *A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms.* Nature, 2001. **409**(6822): p. 928-33.
70. Evans, W.E. and J.A. Johnson, *Pharmacogenomics: the inherited basis for interindividual differences in drug response.* Annu Rev Genomics Hum Genet, 2001. **2**: p. 9-39.
71. Yates, C.R., et al., *Molecular diagnosis of thiopurine S-methyltransferase deficiency: genetic basis for azathioprine and mercaptopurine intolerance.* Ann Intern Med, 1997. **126**(8): p. 608-14.
72. Ji, Z.L., et al., *Drug Adverse Reaction Target Database (DART): proteins related to adverse drug reactions.* Drug Saf, 2003. **26**(10): p. 685-90.
73. Sun, L.Z., et al., *Absorption, distribution, metabolism, and excretion-associated protein database.* Clin Pharmacol Ther, 2002. **71**(5): p. 405.
74. Miller, M.D. and D.J. Hazuda, *New antiretroviral agents: looking beyond protease and reverse transcriptase.* Curr Opin Microbiol, 2001. **4**(5): p. 535-9.
75. Wen, Y.M., X. Lin, and Z.M. Ma, *Exploiting new potential targets for anti-hepatitis B virus drugs.* Curr Drug Targets Infect Disord, 2003. **3**(3): p. 241-6.
76. Consortium, I.H.G.S., *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**(7011): p. 931-45.
77. Wheeler, D.L., et al., *Database resources of the National Center for Biotechnology.* Nucleic Acids Res, 2003. **31**(1): p. 28-33.
78. Connolly, T. and C. Begg, in *Database Solutions: A step-by-step guide to building databases*. 2003, Addison Wesley: Boston. p. 528.
79. Rao, J., *Reasoning about probabilistic parallel programs.* ACM Transactions on Programming Languages and Systems, 1994. **16**(3): p. 798-842.
80. Miaou, S.P. and J.J. Song, *Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion, and spatial dependence.* Accid Anal Prev, 2005. **37**(4): p. 699-720.
81. Kretschmann, E., W. Fleischmann, and R. Apweiler, *Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT.* Bioinformatics, 2001. **17**(10): p. 920-6.
82. des Jardins, M., et al., *Prediction of enzyme classification from protein sequence without the use of sequence similarity.* Proc Int Conf Intell Syst Mol Biol, 1997. **5**: p. 92-9.
83. Jensen, L.J., et al., *Prediction of human protein function from post-translational modifications and localization features.* J Mol Biol, 2002. **319**(5): p. 1257-65.
84. Karchin, R., K. Karplus, and D. Haussler, *Classifying G-protein coupled receptors with support vector machines.* Bioinformatics, 2002. **18**(1): p. 147-59.
85. Cai, C.Z., et al., *SVM-Prot: Web-based support vector machine software for*

functional classification of a protein from its primary sequence. Nucleic Acids Res, 2003. **31**(13): p. 3692-7.
86. Cai, Y.D. and S.L. Lin, *Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence.* Biochim Biophys Acta, 2003. **1648**(1-2): p. 127-33.
87. Cai, C.Z., et al., *Enzyme family classification by support vector machines.* Proteins, 2004. **55**(1): p. 66-76.
88. Han, L.Y., et al., *Prediction of RNA-binding proteins from primary sequence by a support vector machine approach.* Rna, 2004. **10**(3): p. 355-68.
89. Bhasin, M. and G.P. Raghava, *Classification of nuclear receptors based on amino acid composition and dipeptide composition.* J Biol Chem, 2004. **279**(22): p. 23262-6.
90. Vapnik, V., *Estimation of Dependences Based on Empirical Data [in Russian]. [English translation: Springer Verlag, New York, 1982]*. 1979.
91. Vapnik, V., *The Nature of Statistical Learning Theory*. 1995, New York: Springer.
92. Burges, C., *A tutorial on Support Vector Machine for pattern recognition.* Data Min. Knowl. Disc., 1998. **2**: p. 121-167.
93. Kim, K.I., et al., *Support vector machine-based text detection in digital video.* Pattern Recognition, 2001. **34**(2): p. 527-529.
94. Drucker, H., D.H. Wu, and V.N. Vapnik, *Support Vector Machines for Spam Categorization.* IEEE Transactions on Neural Networks, 1999. **10**: p. 1048-1054.
95. de Vel, O., et al., *Mining e-mail content for author identification forensics.* Sigmod Record, 2001. **30**(4): p. 55-64.
96. Thubthong, N. and B. Kijsirikul, *Support vector machines for Thai phoneme recognition.* International Journal of Uncertainty Fuzziness and Knowledge-Based Systems, 2001. **9**(6): p. 803-813.
97. Ben-Yacoub, S., Y. Abdeljaoued, and E. Mayoraz, *Fusion Face and Speech Data for Person Identity Verification.* IEEE Transactions on Neural Networks, 1999. **10**: p. 1065-1074.
98. Karlsen, R.E., D.J. Gorsich, and G.R. Gerhart, *Target classification via support vector machines.* Optical Engineering, 2000. **39**(3): p. 704-711.
99. Papageorgiou, C. and T. Poggio, *A trainable system for object detection.* International Journal of Computer Vision, 2000. **38**(1): p. 15-33.
100. Huang, C., L.S. Davis, and J.R.G. Townshend, *An assessment of support vector machines for land cover classification.* International Journal of Remote Sensing, 2002. **23**(4): p. 725-749.
101. Liong, S.Y. and C. Sivapragasam, *Flood stage forecasting with support vector machines.* Journal of the American Water Resources Association, 2002. **38**(1): p. 173-186.
102. Rasmussen, M. and L. Bjorck, *Unique regulation of SclB - a novel collagen-like surface protein of Streptococcus pyogenes.* Mol Microbiol, 2001. **40**(6): p. 1427-38.
103. Furey, T.S., et al., *Support vector machine classification and validation of cancer tissue samples using microarray expression data.* Bioinformatics, 2000. **16**(10): p. 906-914.
104. Fritsche, H.A., *Tumor Markers and Pattern Recognition Analysis: A New Diagnostic Tool for Cancer.* J. Clin. Ligand Assay, 2002. **25**: p. 11-15.
105. Brown, M.P., et al., *Knowledge-based analysis of microarray gene expression*

data by using support vector machines. Proc Natl Acad Sci U S A, 2000. **97**(1): p. 262-7.
106. Ding, C.H. and I. Dubchak, *Multi-class protein fold recognition using support vector machines and neural networks.* Bioinformatics, 2001. **17**(4): p. 349-58.
107. Hua, S. and Z. Sun, *A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach.* J Mol Biol, 2001. **308**(2): p. 397-407.
108. Bock, J.R. and D.A. Gough, *Predicting protein--protein interactions from primary structure.* Bioinformatics, 2001. **17**(5): p. 455-60.
109. Cristianini, N. and J. Shawe-Taylor, *An introduction to Support Vector Machines: and other kernel-based learning methods*. 2000, New York: Cambridge University Press.
110. Han, L.Y., et al., *Prediction of functional class of novel viral proteins by a statistical learning method irrespective of sequence similarity.* Virology, 2005. **331**(1): p. 136-43.
111. Yuan, Z., K. Burrage, and J.S. Mattick, *Prediction of protein solvent accessibility using support vector machines.* Proteins, 2002. **48**(3): p. 566-70.
112. Roulston, J.E., *Screening with tumor markers: critical issues.* Mol Biotechnol, 2002. **20**(2): p. 153-62.
113. Baldi, P., et al., *Assessing the accuracy of prediction algorithms for classification: an overview.* Bioinformatics, 2000. **16**(5): p. 412-24.
114. Han, L.Y., et al., *Prediction of RNA-binding proteins from primary sequence by a support vector machine approach.* RNA, 2004. **10**(3): p. 355-368.
115. Burges, C.J.C., *A tutorial on Support Vector Machine for pattern recognition.* Data Min. Knowl. Disc., 1998. **2**: p. 121-167.
116. Provost, F., T. Fawcett, and R. Kohavi, *The case against accuracy estimation for comparing induction algorithms*, in *Proc. 15th International Conf. on Machine Learning*. 1998, Morgan Kaufmann: San Francisco, CA. p. 445-453.
117. World Health Organization., *International statistical classification of diseases and related health problems*. 10th revision. ed. 1992, Geneva: World Health Organization.
118. Gasteiger, E., et al., *ExPASy: the proteomics server for in-depth protein knowledge and analysis.* Nucleic Acids Res, 2003. **31**(13): p. 3784-8.
119. Kanehisa, M., et al., *The KEGG databases at GenomeNet.* Nucleic Acids Res, 2002. **30**(1): p. 42-6.
120. Miller., K.J., *Metabolic Pathways of Biochemistry.* http://www.gwu.edu/~mpb/. 1998.
121. Higashi-ku., H., *SPAD: Signaling Pathway database.* http://www.grt.kyushu-u.ac.jp/eny-doc/spad.html. 1998.
122. Takai-Igarashi, T., Y. Nadaoka, and T. Kaminuma, *A database for cell signaling networks.* J Comput Biol, 1998. **5**(4): p. 747-54.
123. Selkov, E., et al., *The metabolic pathway collection from EMP: the enzymes and metabolic pathways database.* Nucleic Acids Res, 1996. **24**(1): p. 26-8.
124. Schilkey, F., *PathDB: a pathway database.* http://www.ncgr.org/pathdb. 2003.
125. Karp, P.D., et al., *The EcoCyc Database.* Nucleic Acids Res, 2002. **30**(1): p. 56-8.
126. The BioCarta team, *Biocarta: Charting pathways of life.* http://www.biocarta.com. 2003.
127. Ellis, L.B., et al., *The University of Minnesota Biocatalysis/Biodegradation Database: post-genomic data mining.* Nucleic Acids Res, 2003. **31**(1): p.

262-5.
128. Shoemaker., R., *SoyBase: Soybean metabolic pathways.* *http://cgsc.biology.yale.edu/metab.html.* 2003.
129. Nicholson, D.E., *IUBMB-Nicholson metabolic pathways charts.* Biochem. Mol. Biol. Educ., 2001. **29**: p. 42-44.
130. Xenarios, I., et al., *DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions.* Nucleic Acids Res, 2002. **30**(1): p. 303-5.
131. Bader, G.D., D. Betel, and C.W. Hogue, *BIND: the Biomolecular Interaction Network Database.* Nucleic Acids Res, 2003. **31**(1): p. 248-50.
132. Krull, M., et al., *TRANSPATH: an integrated database on signal transduction and a tool for array analysis.* Nucleic Acids Res, 2003. **31**(1): p. 97-100.
133. Gough, N.R., *Signal transduction pathways as targets for therapeutics.* Sci STKE, 2001. **2001**(76): p. PE1.
134. Schwartz, M.W., et al., *Central nervous system control of food intake.* Nature, 2000. **404**(6778): p. 661-71.
135. Drews, J., *In Human disease - from genetic causes to biochemical effects.* 1997: p. 5-9.
136. Chiesi, M., C. Huppertz, and K.G. Hofbauer, *Pharmacotherapy of obesity: targets and perspectives.* Trends Pharmacol Sci, 2001. **22**(5): p. 247-54.
137. Kumar, S., S.M. Blake, and J.G. Emery, *Intracellular signaling pathways as a target for the treatment of rheumatoid arthritis.* Curr Opin Pharmacol, 2001. **1**(3): p. 307-13.
138. Matter, A., *Tumor angiogenesis as a therapeutic target.* Drug Discov Today, 2001. **6**(19): p. 1005-1024.
139. Helmuth, L., *New therapies. New Alzheimer's treatments that may ease the mind.* Science, 2002. **297**(5585): p. 1260-2.
140. Greenfeder, S. and J.C. Anthes, *New asthma targets: recent clinical and preclinical advances.* Curr Opin Chem Biol, 2002. **6**(4): p. 526-33.
141. Ilag, L.L., et al., *Emerging high-throughput drug target validation technologies.* Drug Discov Today, 2002. **7**(18 Suppl): p. S136-42.
142. Chen, Y.Z. and D.G. Zhi, *Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule.* Proteins, 2001. **43**(2): p. 217-26.
143. Vane, J.R., Y.S. Bakhle, and R.M. Botting, *Cyclooxygenases 1 and 2.* Annu Rev Pharmacol Toxicol, 1998. **38**: p. 97-120.
144. Torphy, T.J. and C. Page, *Phosphodiesterases: the journey towards therapeutics.* Trends Pharmacol Sci, 2000. **21**(5): p. 157-9.
145. O'Neill, E.A., *A new target for aspirin.* Nature, 1998. **396**(6706): p. 15, 17.
146. Razinkov, V., et al., *RFI-641 inhibits entry of respiratory syncytial virus via interactions with fusion protein.* Chem Biol, 2001. **8**(7): p. 645-59.
147. Luchner, A. and H. Schunkert, *Interactions between the sympathetic nervous system and the cardiac natriuretic peptide system.* Cardiovasc Res, 2004. **63**(3): p. 443-9.
148. Widdicombe, J. and L.Y. Lee, *Airway reflexes, autonomic function, and cardiovascular responses.* Environ Health Perspect, 2001. **109 Suppl 4**: p. 579-84.
149. Toda, N., *Vasodilating beta-adrenoceptor blockers as cardiovascular therapeutics.* Pharmacol Ther, 2003. **100**(3): p. 215-34.
150. Turner, P., *Therapeutic uses of beta-adrenoceptor blocking drugs in the*

central nervous system in man. Postgrad Med J, 1989. **65**(759): p. 1-6.
151. Middlemiss, D.N., D.A. Buxton, and D.T. Greenwood, *Beta-adrenoceptor antagonists in psychiatry and neurology.* Pharmacol Ther, 1981. **12**(2): p. 419-37.
152. Emilien, G. and J.M. Maloteaux, *Current therapeutic uses and potential of beta-adrenoceptor agonists and antagonists.* Eur J Clin Pharmacol, 1998. **53**(6): p. 389-404.
153. Ducruet, A.P., et al., *Dual Specificity Protein Phosphatases: Therapeutic Targets for Cancer and Alzheimer's Disease.* Annu Rev Pharmacol Toxicol, 2004.
154. Lyon, M.A., et al., *Dual-specificity phosphatases as targets for antineoplastic agents.* Nat Rev Drug Discov, 2002. **1**(12): p. 961-76.
155. Buolamwini, J.K., *Novel anticancer drug discovery.* Curr Opin Chem Biol, 1999. **3**(4): p. 500-9.
156. Dubowchik, G.M. and M.A. Walker, *Receptor-mediated and enzyme-dependent targeting of cytotoxic anticancer drugs.* Pharmacol Ther, 1999. **83**(2): p. 67-123.
157. Elsayed, Y.A. and E.A. Sausville, *Selected novel anticancer treatments targeting cell signaling proteins.* Oncologist, 2001. **6**(6): p. 517-37.
158. Persidis, A., *Cardiovascular disease drug discovery.* Nat Biotechnol, 1999. **17**(9): p. 930-1.
159. Bicknell, K.A., E.L. Surry, and G. Brooks, *Targeting the cell cycle machinery for the treatment of cardiovascular disease.* J Pharm Pharmacol, 2003. **55**(5): p. 571-91.
160. Lewis, A.J. and A.M. Manning, *New targets for anti-inflammatory drugs.* Curr Opin Chem Biol, 1999. **3**(4): p. 489-94.
161. Bray, G.A. and L.A. Tartaglia, *Medicinal strategies in the treatment of obesity.* Nature, 2000. **404**(6778): p. 672-7.
162. Campfield, L.A., F.J. Smith, and P. Burn, *Strategies and potential molecular targets for obesity treatment.* Science, 1998. **280**(5368): p. 1383-7.
163. Ahima, R.S. and S.Y. Osei, *Molecular regulation of eating behavior: new insights and prospects for therapeutic strategies.* Trends Mol Med, 2001. **7**(5): p. 205-13.
164. Clapham, J.C., J.R. Arch, and M. Tadayyon, *Anti-obesity drugs: a critical review of current therapies and future opportunities.* Pharmacol Ther, 2001. **89**(1): p. 81-121.
165. Best, J.D. and A.J. Jenkins, *Novel agents for managing dyslipidaemia.* Expert Opin Investig Drugs, 2001. **10**(11): p. 1901-11.
166. Chong, P.H. and B.S. Bachenheimer, *Current, new and future treatments in dyslipidaemia and atherosclerosis.* Drugs, 2000. **60**(1): p. 55-93.
167. Scheinfeld, N.S., et al., *The preauricular sinus: a review of its clinical presentation, treatment, and associations.* Pediatr Dermatol, 2004. **21**(3): p. 191-6.
168. Lin, P.C., et al., *Female genital anomalies affecting reproduction.* Fertil Steril, 2002. **78**(5): p. 899-915.
169. Kobayashi, H. and M.D. Stringer, *Biliary atresia.* Semin Neonatol, 2003. **8**(5): p. 383-91.
170. Wagman, A.S. and J.M. Nuss, *Current therapies and emerging targets for the treatment of diabetes.* Curr Pharm Des, 2001. **7**(6): p. 417-50.
171. Blake, S.M. and B.A. Swift, *What next for rheumatoid arthritis therapy?* Curr

Opin Pharmacol, 2004. **4**(3): p. 276-80.
172. Windisch, M., B. Hutter-Paier, and E. Schreiner, *Current drugs and future hopes in the treatment of Alzheimer's disease.* J Neural Transm Suppl, 2002(62): p. 149-64.
173. Irizarry, M.C. and B.T. Hyman, *Alzheimer disease therapeutics.* J Neuropathol Exp Neurol, 2001. **60**(10): p. 923-8.
174. Bush, K. and M. Macielag, *New approaches in the treatment of bacterial infections.* Curr Opin Chem Biol, 2000. **4**(4): p. 433-9.
175. Hossain, M.A. and M.A. Ghannoum, *New investigational antifungal agents for treating invasive fungal infections.* Expert Opin Investig Drugs, 2000. **9**(8): p. 1797-813.
176. De Clercq, E., *2001 ASPET Otto Krayer Award Lecture. Molecular targets for antiviral agents.* J Pharmacol Exp Ther, 2001. **297**(1): p. 1-10.
177. Olliaro, P.L. and Y. Yuthavong, *An overview of chemotherapeutic targets for antimalarial drug discovery.* Pharmacol Ther, 1999. **81**(2): p. 91-110.
178. White, H., *Mechanism of action of newer anticonvulsants.* Journal of Clinical Psychiatry, 2003. **64 Suppl 8**: p. 5-8.
179. Murata, M., *Novel therapeutic effects of the anti-convulsant, zonisamide, on Parkinson's disease.* Curr Pharm Des, 2004. **10**(6): p. 687-693.
180. Storey, F., *The P2Y12 receptor as a therapeutic target in cardiovascular disease.* Platelets, 2001. **12**(4): p. 197-209.
181. Sehgal, S., *Sirolimus: its discovery, biological properties, and mechanism of action.* Transplantation Proceedings, 2003. **35**(3 Suppl): p. 7S-14S.
182. Turini, M. and R. DuBois, *Cyclooxygenase-2: a therapeutic target.* Annual Review of Medicine, 2002. **53**: p. 35-57.
183. Chantry, D., *G protein-coupled receptors: from ligand identification to drug targets.* Expert Opin. Emerg. Drugs, 2003. **8**(1): p. 273-276.
184. Gibbs, J.B., et al., *Selective inhibition of farnesyl-protein transferase blocks ras processing in vivo.* J Biol Chem, 1993. **268**(11): p. 7617-20.
185. Jabbour, E., H. Kantarjian, and J. Cortes, *Clinical activity of farnesyl transferase inhibitors in hematologic malignancies: possible mechanisms of action.* Leuk Lymphoma, 2004. **45**(11): p. 2187-95.
186. Karp, J.E. and J.E. Lancet, *Farnesyltransferase inhibitors (FTIs) in myeloid malignancies.* Ann Hematol, 2004. **83 Suppl 1**: p. S87-8.
187. Barnette, M.S., et al., *Association of the anti-inflammatory activity of phosphodiesterase 4 (PDE4) inhibitors with either inhibition of PDE4 catalytic activity or competition for [3H]rolipram binding.* Biochem Pharmacol, 1996. **51**(7): p. 949-56.
188. Spina, D., *Phosphodiesterase-4 inhibitors in the treatment of inflammatory lung disease.* Drugs, 2003. **63**(23): p. 2575-94.
189. Docherty, A.J., et al., *The matrix metalloproteinases and their natural inhibitors: prospects for treating degenerative tissue diseases.* Trends Biotechnol, 1992. **10**(6): p. 200-7.
190. Ramnath, N. and P.J. Creaven, *Matrix metalloproteinase inhibitors.* Curr Oncol Rep, 2004. **6**(2): p. 96-102.
191. Wada, C.K., *The evolution of the matrix metalloproteinase inhibitor drug discovery program at abbott laboratories.* Curr Top Med Chem, 2004. **4**(12): p. 1255-67.
192. Howe, R., et al., *Selective beta 3-adrenergic agonists of brown adipose tissue and thermogenesis. 1.*

[4-[2-[(2-Hydroxy-3-phenoxypropyl)amino]ethoxy]phenoxy]acetates. J Med Chem, 1992. **35**(10): p. 1751-9.
193. de Souza, C.J. and B.F. Burkey, *Beta 3-adrenoceptor agonists as anti-diabetic and anti-obesity drugs in humans.* Curr Pharm Des, 2001. **7**(14): p. 1433-49.
194. Lipinski, C.A., et al., *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.* Adv Drug Deliv Rev, 2001. **46**(1-3): p. 3-26.
195. Hardy, L.W. and N.P. Peet, *The multiple orthogonal tools approach to define molecular causation in the validation of druggable targets.* Drug Discov Today, 2004. **9**(3): p. 117-26.
196. Chantry, D., *G protein-coupled receptors: from ligand identification to drug targets. 14-16 October 2002, San Diego, CA, USA.* Expert Opin Emerg Drugs, 2003. **8**(1): p. 273-6.
197. Gronemeyer, H., J.A. Gustafsson, and V. Laudet, *Principles for modulation of the nuclear receptor superfamily.* Nat Rev Drug Discov, 2004. **3**(11): p. 950-64.
198. Bateman, A., et al., *The Pfam protein families database.* Nucleic Acids Res, 2004. **32 Database issue**: p. D138-41.
199. Yu, E.W., et al., *Structural basis of multiple drug-binding capacity of the AcrB multidrug efflux pump.* Science, 2003. **300**(5621): p. 976-80.
200. Striessnig, J., et al., *Structural basis of drug binding to L Ca2+ channels.* Trends Pharmacol Sci, 1998. **19**(3): p. 108-15.
201. Benke, D., C. Michel, and H. Mohler, *GABA(A) receptors containing the alpha4-subunit: prevalence, distribution, pharmacology, and subunit architecture in situ.* J Neurochem, 1997. **69**(2): p. 806-14.
202. Poulos, T.L., *Cytochrome P450: molecular architecture, mechanism, and prospects for rational inhibitor design.* Pharm Res, 1988. **5**(2): p. 67-75.
203. Andreeva, A., et al., *SCOP database in 2004: refinements integrate structure and sequence family data.* Nucleic Acids Res, 2004. **32 Database issue**: p. D226-9.
204. Sussman, J.L., et al., *Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules.* Acta Crystallogr D Biol Crystallogr, 1998. **54**(Pt 6 Pt 1): p. 1078-84.
205. Darnell, J.E., Jr., *Transcription factors as targets for cancer therapy.* Nat Rev Cancer, 2002. **2**(10): p. 740-9.
206. Eggert, M., et al., *Transcription factors in autoimmune diseases.* Curr Pharm Des, 2004. **10**(23): p. 2787-96.
207. Collins, J.L., *Therapeutic opportunities for liver X receptor modulators.* Curr Opin Drug Discov Devel, 2004. **7**(5): p. 692-702.
208. Faber, E. and P. Sah, *Calcium-activated potassium channels: multiple contributions to neuronal function.* Neuroscientist, 2003. **9**(3): p. 181-194.
209. Mihailescu, S. and R. Drucker-Colin, *Nicotine, brain nicotinic receptors, and neuropsychiatric disorders.* Archives of Medical Research, 2000. **31**(2): p. 131-144.
210. Pacher, P. and V. Kecskemeti, *Trends in the development of new antidepressants. Is there a light at the end of the tunnel?* Current Medicinal Chemistry, 2004. **11**(7): p. 925-943.
211. Jordan, V., *Selective estrogen receptor modulation: concept and consequences in cancer.* Cancer Cells, 2004. **5**(3): p. 207-213.
212. Adcock, I., *Glucocorticoids: new mechanisms and future agents.* Curr. Allergy

Asthma Rep., 2003. **3**(3): p. 249-257.
213. Orchard, S., *Kinases as targets: prospects for chronic therapy.* Curr. Opin. Drug Discov. Devel., 2002. **5**(5): p. 713-717.
214. Docherty, A., et al., *Proteases as drug targets.* Biochemical Society Symposia, 2003(70): p. 147-161.
215. Lai, M., *RNA polymerase as an antiviral target of hepatitis C virus.* Antiviral Chemistry and Chemotherapy, 2001. **12 Suppl 1**: p. 143-147.
216. Loverix, S. and J. Steyaert, *Ribonucleases: from prototypes to therapeutic targets?* Current Medicinal Chemistry, 2003. **10**(9): p. 779-785.
217. van Huijsduijnen, R.H., A. Bombrun, and D. Swinnen, *Selecting protein tyrosine phosphatases as drug targets.* Drug Discov Today, 2002. **7**(19): p. 1013-9.
218. Ristimaki, A., *Cyclooxygenase 2: from inflammation to carcinogenesis.* Novartis Foundation Symposium, 2004. **256**: p. 215-221.
219. Subauste, A. and C. Burant, *DGAT: novel therapeutic target for obesity and type 2 diabetes mellitus.* Curr. Drug Targets Immune Endocr. Metabol. Disord., 2003. **3**(4): p. 263-270.
220. Baranczyk-Kuzma, A., et al., *Glutathione S-transferase pi as a target for tricyclic antidepressants in human brain.* Acta Biochim Pol, 2004. **51**(1): p. 207-12.
221. Vullo, D., et al., *Carbonic anhydrase inhibitors. Inhibition of the transmembrane isozyme XII with sulfonamides-a new target for the design of antitumor and antiglaucoma drugs?* Bioorg Med Chem Lett, 2005. **15**(4): p. 963-9.
222. Oikonomakos, N., *Glycogen phosphorylase as a molecular target for type 2 diabetes therapy.* Curr. Protein Pept. Sci., 2002. **3**(6): p. 561-586.
223. Gudermann, T., B. Nurnberg, and G. Schultz, *Receptors and G-proteins as primary components of transmembrane signal transduction.* Journal of Molecular Medicine, 1995. **73**: p. 51-63.
224. Flower, D.R., *Modelling G-protein-coupled receptors for drug design.* Biochim Biophys Acta, 1999. **1422**(3): p. 207-34.
225. Blagosklonny, M.V., *Tissue-selective therapy of cancer.* Br J Cancer, 2003. **89**(7): p. 1147-51.
226. Zhang, Z., et al., *Genomic analysis of the nuclear receptor family: new insights into structure, regulation, and evolution from the rat genome.* Genome Res, 2004. **14**(4): p. 580-90.
227. Yanase, H., H. Sugino, and T. Yagi, *Genomic sequence and organization of the family of CNR/Pcdhalpha genes in rat.* Genomics, 2004. **83**(4): p. 717-26.
228. Feldman, M. and G. Segal, *A specific genomic location within the icm/dot pathogenesis region of different Legionella species encodes functionally similar but nonhomologous virulence proteins.* Infect Immun, 2004. **72**(8): p. 4503-11.
229. Wen, Y., X. Lin, and Z. Ma, *Exploiting new potential targets for anti-hepatitis B virus drugs.* Curr. Drug Targets Infect. Disord., 2003. **3**(3): p. 241-246.
230. Davidson, A. and S. Siddell, *Potential for antiviral treatment of severe acute respiratory syndrome.* Curr. Opin. Infect. Dis., 2003. **2003**(16): p. 6.
231. Hopkins, A.L. and C.R. Groom, *Target analysis: a priori assessment of druggability.* Ernst Schering Res Found Workshop, 2003(42): p. 11-7.
232. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

233. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
234. Smith, C., *Drug target validation: Hitting the target.* Nature, 2003. **422**(6929): p. 341, 343, 345 passim.
235. Desany, B. and Z. Zhang, *Bioinformatics and cancer target discovery.* Drug Discov Today, 2004. **9**(18): p. 795-802.
236. Dohrmann, C.E., *Target discovery in metabolic disease.* Drug Discov Today, 2004. **9**(18): p. 785-94.
237. Baurin, N., et al., *Drug-like annotation and duplicate analysis of a 23-supplier chemical database totalling 2.7 million compounds.* J Chem Inf Comput Sci, 2004. **44**(2): p. 643-51.
238. Zernov, V.V., et al., *Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions.* J Chem Inf Comput Sci, 2003. **43**(6): p. 2048-56.
239. Byvatov, E., et al., *Comparison of support vector machine and artificial neural network systems for drug/nondrug classification.* J Chem Inf Comput Sci, 2003. **43**(6): p. 1882-9.
240. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
241. Cai, Y.D., et al., *Support vector machines for prediction of protein domain structural class.* J Theor Biol, 2003. **221**(1): p. 115-20.
242. Zhou, G.P. and N. Assa-Munt, *Some insights into protein structural class prediction.* Proteins, 2001. **44**(1): p. 57-9.
243. Wise, A., K. Gearing, and S. Rees, *Target validation of G-protein coupled receptors.* Drug Discov Today, 2002. **7**(4): p. 235-46.
244. Turpin, J.A., *The next generation of HIV/AIDS drugs: novel and developmental antiHIV drugs and targets.* Expert Rev Anti Infect Ther, 2003. **1**(1): p. 97-128.
245. Sun, L.Z., et al., *ADME-AP: a database of ADME associated proteins.* Bioinformatics, 2002. **18**(12): p. 1699-700.
246. Zheng, C.J., et al., *Drug ADME-associated protein database as a resource for facilitating pharmacogenomics research.* Drug Development Research, 2004. **62**(2): p. 134-142.
247. Zhang, L., C.M. Brett, and K.M. Giacomini, *Role of organic cation transporters in drug absorption and elimination.* Annu Rev Pharmacol Toxicol, 1998. **38**: p. 431-60.
248. Tamai, I. and A. Tsuji, *Transporter-mediated permeation of drugs across the blood-brain barrier.* J Pharm Sci, 2000. **89**(11): p. 1371-88.
249. Vizi, E.S., *Role of high-affinity receptors and membrane transporters in nonsynaptic communication and drug action in the central nervous system.* Pharmacol Rev, 2000. **52**(1): p. 63-89.
250. de Wolf, F.A. and G.M. Brett, *Ligand-binding proteins: their potential for application in systems for controlled delivery and uptake of ligands.* Pharmacol Rev, 2000. **52**(2): p. 207-36.
251. Zheng, C.J., et al., *Drug ADME-Associated Protein Database as a Resource for Facilitating Pharmacogenomics Research.* Drug Development Research, 2004. **62**(2): p. 134-142.
252. Hosford, D.A., et al., *Pharmacogenetics to Predict Drug-Related Adverse Events.* Toxicol Pathol, 2004. **32**(Supplement 1): p. 9-12.

253. Ranganathan, P., et al., *Will pharmacogenetics allow better prediction of methotrexate toxicity and efficacy in patients with rheumatoid arthritis?* Ann Rheum Dis, 2003. **62**(1): p. 4-9.
254. Hiratsuka, M., et al., *Genotyping of the N-acetyltransferase2 polymorphism in the prediction of adverse drug reactions to isoniazid in Japanese patients.* Drug Metab Pharmacokinet, 2002. **17**(4): p. 357-62.
255. Yoshida, K., et al., *Prediction of antidepressant response to milnacipran by norepinephrine transporter gene polymorphisms.* Am J Psychiatry, 2004. **161**(9): p. 1575-80.
256. Carlini, L.E., et al., *UGT1A7 and UGT1A9 polymorphisms predict response and toxicity in colorectal cancer patients treated with capecitabine/irinotecan.* Clin Cancer Res, 2005. **11**(3): p. 1226-36.
257. Serretti, A. and E. Smeraldi, *Neural network analysis in pharmacogenetics of mood disorders.* BMC Med Genet, 2004. **5**(1): p. 27.
258. Gunther, E.C., et al., *Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles in vitro.* Proc Natl Acad Sci U S A, 2003. **100**(16): p. 9608-13.
259. Chiang, D., et al., *Prediction of stone disease by discriminant analysis and artificial neural networks in genetic polymorphisms: a new method.* BJU Int, 2003. **91**(7): p. 661-6.
260. Zheng, H.X., et al., *The impact of pharmacogenomic factors on steroid dependency in pediatric heart transplant patients using logistic regression analysis.* Pediatr Transplant, 2004. **8**(6): p. 551-7.
261. Green, M.D., E.M. Oturu, and T.R. Tephly, *Stable expression of a human liver UDP-glucuronosyltransferase (UGT2B15) with activity toward steroid and xenobiotic substrates.* Drug Metab Dispos, 1994. **22**(5): p. 799-805.
262. Ozawa, N., et al., *Transporter database, TP-Search: a web-accessible comprehensive database for research in pharmacokinetics of drugs.* Pharm Res, 2004. **21**(11): p. 2133-4.
263. Busch, W. and M.H. Saier, Jr., *The IUBMB-endorsed transporter classification system.* Mol Biotechnol, 2004. **27**(3): p. 253-62.
264. Klein, T.E., et al., *Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base.* Pharmacogenomics J, 2001. **1**(3): p. 167-70.
265. Sparreboom, A., et al., *Pharmacogenomics of ABC transporters and its role in cancer chemotherapy.* Drug Resist Updat, 2003. **6**(2): p. 71-84.
266. Ingelman-Sundberg, M., *Pharmacogenetics of cytochrome P450 and its applications in drug therapy: the past, present and future.* Trends Pharmacol Sci, 2004. **25**(4): p. 193-200.
267. Guillemette, C., *Pharmacogenomics of human UDP-glucuronosyltransferase enzymes.* Pharmacogenomics J, 2003. **3**(3): p. 136-58.
268. Rogers, J.F., A.N. Nafziger, and J.S. Bertino, Jr., *Pharmacogenetics affects dosing, efficacy, and toxicity of cytochrome P450-metabolized drugs.* Am J Med, 2002. **113**(9): p. 746-50.
269. Borst, P., et al., *A family of drug transporters: the multidrug resistance-associated proteins.* J Natl Cancer Inst, 2000. **92**(16): p. 1295-302.
270. Brinkmann, U., I. Roots, and M. Eichelbaum, *Pharmacogenetics of the human drug-transporter gene MDR1: impact of polymorphisms on pharmacotherapy.* Drug Discov Today, 2001. **6**(16): p. 835-839.
271. Weinshilboum, R., *Inheritance and drug response.* N Engl J Med, 2003.

348(6): p. 529-37.
272. Rao, V.V., et al., *Choroid plexus epithelial expression of MDR1 P glycoprotein and multidrug resistance-associated protein contribute to the blood-cerebrospinal-fluid drug-permeability barrier.* Proc Natl Acad Sci U S A, 1999. **96**(7): p. 3900-5.
273. Thiebaut, F., et al., *Cellular localization of the multidrug-resistance gene product P-glycoprotein in normal human tissues.* Proc Natl Acad Sci U S A, 1987. **84**(21): p. 7735-8.
274. Schinkel, A.H., et al., *P-glycoprotein in the blood-brain barrier of mice influences the brain penetration and pharmacological activity of many drugs.* J Clin Invest, 1996. **97**(11): p. 2517-24.
275. Anttila, S., et al., *Interaction between NOTCH4 and catechol-O-methyltransferase genotypes in schizophrenia patients with poor response to typical neuroleptics.* Pharmacogenetics, 2004. **14**(5): p. 303-7.
276. Kajinami, K., et al., *Interactions between common genetic polymorphisms in ABCG5/G8 and CYP7A1 on LDL cholesterol-lowering response to atorvastatin.* Atherosclerosis, 2004. **175**(2): p. 287-93.
277. Zazzi, M., et al., *Comparative evaluation of three computerized algorithms for prediction of antiretroviral susceptibility from HIV type 1 genotype.* J Antimicrob Chemother, 2004. **53**(2): p. 356-60.
278. Basile, V.S., et al., *A functional polymorphism of the cytochrome P450 1A2 (CYP1A2) gene: association with tardive dyskinesia in schizophrenia.* Mol Psychiatry, 2000. **5**(4): p. 410-7.
279. Schalekamp, T., et al., *Effects of cytochrome P450 2C9 polymorphisms on phenprocoumon anticoagulation status.* Clin Pharmacol Ther, 2004. **76**(5): p. 409-17.
280. Kapitany, T., et al., *Genetic polymorphisms for drug metabolism (CYP2D6) and tardive dyskinesia in schizophrenia.* Schizophr Res, 1998. **32**(2): p. 101-6.
281. Vandel, P., et al., *Drug extrapyramidal side effects. CYP2D6 genotypes and phenotypes.* Eur J Clin Pharmacol, 1999. **55**(9): p. 659-65.
282. Rau, T., et al., *CYP2D6 genotype: impact on adverse effects and nonresponse during treatment with antidepressants-a pilot study.* Clin Pharmacol Ther, 2004. **75**(5): p. 386-93.
283. Zielinska, E., et al., *Genotyping of the arylamine N-acetyltransferase polymorphism in the prediction of idiosyncratic reactions to trimethoprim-sulfamethoxazole in infants.* Pharm World Sci, 1998. **20**(3): p. 123-30.
284. Yu, M.W., et al., *Role of N-acetyltransferase polymorphisms in hepatitis B related hepatocellular carcinoma: impact of smoking on risk.* Gut, 2000. **47**(5): p. 703-9.
285. Serretti, A., et al., *Influence of tryptophan hydroxylase and serotonin transporter genes on fluvoxamine antidepressant activity.* Mol Psychiatry, 2001. **6**(5): p. 586-92.
286. Smits, K.M., et al., *Influence of SERTPR and STin2 in the serotonin transporter gene on the effect of selective serotonin reuptake inhibitors in depression: a systematic review.* Mol Psychiatry, 2004. **9**(5): p. 433-41.
287. Siddiqui, A., et al., *Association of multidrug resistance in epilepsy with a polymorphism in the drug-transporter gene ABCB1.* N Engl J Med, 2003. **348**(15): p. 1442-8.
288. Saitoh, A., et al., *An MDR1-3435 variant is associated with higher plasma*

nelfinavir levels and more rapid virologic response in HIV-1 infected children. Aids, 2005. **19**(4): p. 371-80.
289. Peters, E.J., et al., *Investigation of serotonin-related genes in antidepressant response.* Mol Psychiatry, 2004. **9**(9): p. 879-89.
290. Zanardi, R., et al., *Factors affecting fluvoxamine antidepressant activity: influence of pindolol and 5-HTTLPR in delusional and nondelusional depression.* Biol Psychiatry, 2001. **50**(5): p. 323-30.
291. Lin, M., et al., *Sequencing drug response with HapMap.* Pharmacogenomics J, 2005. **5**(3): p. 149-56.
292. Wang, D. and B. Larder, *Enhanced prediction of lopinavir resistance from genotype by use of artificial neural networks.* J Infect Dis, 2003. **188**(5): p. 653-60.
293. Draghici, S. and R.B. Potter, *Predicting HIV drug resistance with neural networks.* Bioinformatics, 2003. **19**(1): p. 98-107.
294. Meyer, U.A., *Pharmacogenetics and adverse drug reactions.* Lancet, 2000. **356**(9242): p. 1667-71.
295. Roses, A.D., *Pharmacogenetics and future drug development and delivery.* Lancet, 2000. **355**(9212): p. 1358-61.
296. Evans, W.E. and H.L. McLeod, *Pharmacogenomics--drug disposition, drug targets, and side effects.* N Engl J Med, 2003. **348**(6): p. 538-49.
297. Eddershaw, P.J., A.P. Beresford, and M.K. Bayliss, *ADME/PK as part of a rational approach to drug discovery.* Drug Discov Today, 2000. **5**(9): p. 409-414.
298. Li, A.P., *Screening for human ADME/Tox drug properties in drug discovery.* Drug Discov Today, 2001. **6**(7): p. 357-366.
299. Lin, J.H. and A.Y. Lu, *Role of pharmacokinetics and metabolism in drug discovery and development.* Pharmacol Rev, 1997. **49**(4): p. 403-49.
300. Mobley, C. and G. Hochhaus, *Methods used to assess pulmonary deposition and absorption of drugs.* Drug Discov Today, 2001. **6**(7): p. 367-375.
301. Scharpe, S. and I. De Meester, *Peptide truncation by dipeptidyl peptidase IV: a new pathway for drug discovery?* Verh K Acad Geneeskd Belg, 2001. **63**(1): p. 5-32; discussion 32-3.
302. Kim, H. and H. Park, *Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor.* Proteins, 2004. **54**(3): p. 557-62.

APPENDIX A

Some examples of druggable proteins selected by human genome screening:

2',3'-cyclic nucleotide 3'-phosphodiesterase	Immunoglobulin alpha Fc receptor
40S ribosomal protein S12	Intercellular adhesion molecule-1
40 kDa peptidyl-prolyl cis-trans isomerase	Inositol polyphosphate 1-phosphatase
5-hydroxytryptamine 5A receptor	Inositol-1(or 4)-monophosphatase
5-hydroxytryptamine 6 receptor	Insulin receptor substrate-1
69 kDa islet cell autoantigen	Insulin-like growth factor binding protein 1
72 kDa type IV collagenase	Integrin-linked protein kinase 1
92 kDa type IV collagenase	Interferon regulatory factor 1
Acetyl-CoA carboxylase 1	Interleukin-1 beta convertase
Aconitate hydratase, mitochondrial	Junction plakoglobin
Acylamino-acid-releasing enzyme	Kallikrein 6
Adenosine kinase	Kallikrein 7
Adenosylhomocysteinase	Keratin, type I cytoskeletal 19
Adenylate cyclase, type II	Kinesin-like protein KIF11
Adipocyte-derived leucine aminopeptidase	Kininogen
Adiponectin	Kynureninase
Adrenocorticotropic hormone receptor	Lactadherin
Adenosine A3 receptor	Lactase-phlorizin hydrolase
Bile acid receptor	Lactosylceramide alpha-2,3-sialyltransferase
B1 bradykinin receptor	Melanoma-associated antigen 4
B2 bradykinin receptor	Membrane copper amine oxidase
Baculoviral IAP repeat-containing protein 4	Metabotropic glutamate receptor 1
Baculoviral IAP repeat-containing protein 5	Myeloperoxidase
Bax inhibitor-1	Myotubularin-related protein 1
Beta crystallin B1	NAD(P)H dehydrogenase [quinone] 1
Beta platelet-derived growth factor receptor	Neprilysin
Beta-3 adrenergic receptor	Neural-cadherin
Beta-adrenergic receptor kinase 1	Neuroendocrine convertase 1
Beta-catenin	Neuroendocrine convertase 2
Calgranulin D	Neurogenic locus notch homolog protein 1
Calmodulin	Nigral tachykinin NK(3) receptor
cAMP response element binding protein	Oncostatin M
Cell division protein kinase 9	Orexin
cathepsin B	Orexin receptor type 1
multidrug resistance-associated protein 2	Ornithine aminotransferase, mitochondrial
Cannabinoid receptor 1	Orphan nuclear receptor DAX-1
Delta-aminolevulinic acid dehydratase	P2Y purinoceptor 6
Delta-type opioid receptor	Paired box protein Pax-5
Deoxyhypusine synthase	Presenilin 2
Early activation antigen CD69	Prostacyclin receptor
Ets-domain protein elk-3	Proto-oncogene C-crk
Endothelin-converting enzyme 1	Peripheral-type benzodiazepine receptor
Elastase 1	Phosphatidylethanolamine N-methyltransferase
Elav-like protein 1	Prostaglandin E synthase
Elongation factor 2	Placenta growth factor
Endoglin	Plasminogen activator inhibitor-1
Endothelin-1	Platelet endothelial cell adhesion molecule

Eosinophil peroxidase	Renin, renal
Ephrin type-A receptor 2	Reticulon 4 receptor
Fanconi anemia group F protein	Retinoic acid receptor alpha
Farnesyl-diphosphate farnesyltransferase	Rhodopsin
Fascin	Rhombotin-2
Ferrochelatase	Ribosomal protein S6 kinase
Fibroblast growth factor receptor 3	Ryanodine receptor 2
fibroleukin	Somatostatin receptor type 1
Filamin A	Steroid hormone receptor ERR1
FL cytokine receptor	Small inducible cytokine A2
Folate receptor alpha	Serum paraoxonase/arylesterase 1
G2/mitotic-specific cyclin B1	Seprase
Galanin receptor type 1	Serine protease hepsin
Gamma-synuclein	Suppressor of tumorigenicity 14
Gap junction alpha-1 protein	Synaptosomal-associated protein 25
Gastric inhibitory polypeptide	T-box transcription factor TBX21
Gastrin/cholecystokinin type B receptor	T-cell-specific surface glycoprotein CD28
Gastrin-releasing peptide receptor	Telomerase reverse transcriptase
Glucagon receptor	Tenascin
Glucagon-like peptide 1 receptor	Thioredoxin
Glucose-6-phosphatase	Ubiquitin-protein ligase E3A
Heat shock 27 kDa protein	UDP-glucose 4-epimerase
Heat shock factor protein 1	Urotensin II receptor
Guanylyl cyclase C	vascular endothelial growth factor B
Heme oxygenase 1	Vascular endothelial growth factor receptor 2
Heparin cofactor II	Vascular endothelial-cadherin
Heparin-binding growth factor 1	Vasoactive intestinal polypeptide receptor 1
Heparin-binding growth factor 2	Vasopressin V1a receptor
Hepatocyte growth factor	Wilms' tumor protein
Hepatocyte growth factor receptor	Xaa-Pro dipeptidase
Hepatocyte nuclear factor 1-alpha	Zinc finger protein OZF

APPENDIX B

Selected publications:

1. C.J. Zheng, L.Y. Han, C. W. Yap, Z. L. Ji, Z. W. Cao and Y. Z. Chen. Therapeutic Targets: Progress of Their Exploration and Investigation of Their Characteristics. *Pharmacological Reviews*, 58:259-279. (2006).

2. C.J. Zheng, L.Y. Han, C. W. Yap, B. Xie, and Y. Z. Chen. Progress and Difficulties in the Exploration of Therapeutic Targets. *Drug Discovery Today*, **11**(9-10):412-420 (2006).

3. C.J. Zheng, L.Y. Han, X. Chen, Z.W. Cao, J. Cui, H.H. Lin, H.L. Zhang, H. Li and Y. Z. Chen. Information of ADME-associated proteins and potential application for pharmacogenetic prediction of drug responses. *Current Pharmacogenomics*, **4**(17):87-103 (2006).

4. C.J. Zheng, L.Y. Han, C. W. Yap, B. Xie, and Y. Z. Chen. Trends in Exploration of Therapeutic Targets. *Drug News & Perspectives*, **18**(2): 109-127. (2005).

5. C.J. Zheng, L. Z. Sun, L.Y. Han, Z. L. Ji, X. Chen, and Y. Z. Chen. Drug ADME-Associated Protein Database as a Resource for Facilitating Pharmacogenomics Research. *Drug Development Research*, 62, 134–142 (2004).

6. C.J. Zheng, H. Zhou, B. Xie, L.Y. Han, C.W. Yap, and Y. Z. Chen. TRMP: A Database of Therapeutically Relevant Multiple-Pathways. *Bioinformatics*, **20**(14), 2236- 2241. (2004).