COMPUTER AIDED DRUG DESIGN:

DRUG TARGET DIRECTED *IN SILICO* APPROACHES

CHEN XIN

NATIONAL UNIVERSITY OF SINGAPORE
2003

Founded 1905

COMPUTER AIDED DRUG DESIGN:
DRUG TARGET DIRECTED IN SILICO APPROACHES

BY

CHEN XIN

(B.Sc. (Biotech. & Comp. Sci.), SJTU)

A THESIS SUBMITTED
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
DEPARTMENT OF COMPUTATIONAL SCIENCE
NATIONAL UNIVERSITY OF SINGAPORE
2003

Acknowledgements

First and foremost, I would like to express my sincerest appreciation to my supervisor, Associate Professor Chen Yuzong for his patient guidance, supervision, invaluable advices and suggestions throughout my whole research progress.

Sincere gratitude is also expressed to Dr. Cai Congzhong, Dr. Li Zherong, Dr. Xue Ying for their helpful suggestions and co-operations; to Lixia, Zhiliang, Zhiwei, Lianyi, Chanjuan, Jifeng, and Chunwei, who are lab-mates as well as friends, for being ever so willing to share with me their valuable ideas, as well as my joy and sorrow at all times.

I would also like to thank Ms Lindah, Ms. Hwee sim, Ms.Lucee, Ms Elaine and Ms.Wei har, for their kind and timely assistances.

Last but not the least, I am eternally grateful to my parents for encouraging me throughout my life.

Chen Xin

September 2003

Table of Contents

Acronyms V

Synopsis VII

1. Introduction

1.1 Introduction to drug discovery 1

1.1.1 History of drug discovery 2

1.1.2 Modern drug discovery 3

1.1.2.1 Combinatory chemistry based approaches 5

1.1.2.2 Receptor structure based drug design 5

1.1.2.3 Chemical structure activity relationship based drug design 7

1.2 Therapeutic targets and drug discovery 9

1.2.1 Information resources of therapeutic targets 9

1.2.2 Discovery of novel therapeutic targets 10

1.2.3 Study of novel therapeutic mechanisms 12

1.3 Thesis outline 13

2. Therapeutic target database development

2.1 Introduction 15

2.2 Collection of therapeutic target information 20

2.3 Therapeutic target database development 24

2.3.1 Requirement analysis 24

2.3.1.1 Databases development approaches 25

2.3.1.2 Selection of RDBMS 29

2.3.2 Database design & implementation 31

2.3.2.1 Conceptual design 32

2.3.2.2 Logical design 34

2.3.2.2.1 ERD derived database structure 35

2.3.2.2.2 Revised database structure 40

2.3.2.2.3 Further analysis of the revised database structure 43

2.3.2.3 Physical design 46

2.3.3 Implementation 47

2.4 Preliminary analysis of TTD 52

2.5 Extension of the TTD database schema and interface 52

2.6 Summary 55

3. Prediction of drug-target like proteins

3.1 Introduction 56

3.2 Statistical learning 59

3.2.1 Classification algorithms 59

3.2.1.1 Decision tree 60

3.2.1.2 K-nearest neighbor 66

3.2.1.3 Support vector machine 67

3.2.2 Pre-processing for classification 74

3.2.2.1 Scaling 74

3.2.2.2 Principal component analysis 75

3.2.2.3 Independent component analysis 77

3.3 Problem definition 82

3.3.1 Description of data 83

3.3.2 Measurements of prediction accuracy 87

3.4 Prediction of drug-target like proteins 90

3.4.1 Decision tree prediction 91

3.4.2 K-nearest neighbor prediction 92

3.4.3 Support vector machine prediction 100

3.5 Prediction results and analysis 106

3.6 Summary 112

4. ***In silico* study of the mechanisms of action of active ingredients from medicinal plants**

4.1 Introduction 113

4.2 *In silico* methods for target identification of MP ingredients 115

4.3 A closer examination of an *in silico* method – INVDOCK 118

4.3.1 Feasibility 118

4.3.2 Algorithm 119

4.3.3 Validation studies on synthetic chemicals 123

4.4 *In silico* prediction of therapeutic targets of MP ingredients 128

4.4.1 Genistein 130

4.4.2 Ginsenoside Rg1 135

4.4.3 Quercetin 137

4.4.4 Acronycine 141

4.4.5 Baicalin 143

4.4.6 Emodin 145

4.4.7 Allicin 147

4.4.8 Catechin 149

4.4.9 Camptothecin 153

4.5 Limitations and suggested improvement of INVDOCK 155

4.6 Summary 158

5. Summary 160

References 164

Acronyms

ADME-AP	Absorption, distribution, metabolism, excretion associated protein
ADO	ActiveX data objects
AI	Artificial intelligence
ANN	Artificial neural network
ANSI	American national standards institute
ASP	Active server pages
CADD	Computer aided drug design
CAS RN	Chemical abstract service registration number
CGI	Common gateway interface
DART	Drug Adverse Reaction Target
DBI	Database interface
DBMS	Database management system
DNA	Deoxyribonucleic acid
ERD	Entity relationships diagram
FDA	Food and drug administration, USA
GA	Genetic algorithm
GPCR	G-protein coupled receptor
HMM	Hidden markov model
HTML	Hypertext markup language
ICA	Independent component analysis
IEM	Information engineering methodology
IUPAC	International union of pure and applied chemistry
JSP	Java server pages
kNN	K-nearest neighbor
MBDD	Mechanism base drug design
MP	Medicinal plant

NCBI	National center for biotechnology information
NF	Normal form
NMR	Nuclear magnetic resonance
ODBC	Open database connectivity
OLE-DB	Object linking and embedding database
OOP	Object oriented programming
OSH	Optimal separating hyperplane
PCA	Principal component analysis
PDB	Protein data bank
Perl	Practical extraction and reporting language
PHP	Personal home page
PLS	Partial least squares
QSAR	Quantitative structure activity relationship
R&D	Research and development
RDBMS	Relational database management system
RNA	Ribonucleic acid
SAR	Structure activity relationship
SQL	Structured query language
SRM	Structural risk minimization
SVM	Support vector machine
TTD	Therapeutic target database

Synopsis

In modern drug discovery practices, drug leads are screened / designed against a pre-selected drug target. As a prerequisite step, target identification directs further research and developments. It has become increasingly important and received more and more attention from researchers.

This work begins with the development of the Therapeutic Target Database (TTD), which provides a comprehensive information source of known therapeutic targets and serves as a basis for the development of other *in silico* tools. A relational data model was designed specifically for this database which aims to maximize the ability to accommodate future extensions and facilitate the integration of information.

Rapid discovery of new therapeutic targets is also very important as it may not only introduce more efficient therapeutic targets for certain diseases, but also increase the flexibility in designing of novel therapeutic intervention strategies by exploiting the synergies between known and newly discovered targets. With this database, statistical learning approaches are explored in rapid drug target discovery. Our results showed that support vector machine, a novel statistical learning approach, may be useful in the prediction of drug-target like proteins in human genome.

Besides more effective therapeutic targets, delicate therapeutic mechanisms involving multiple cooperating targets may also help to improve the treatment effectiveness. Novel therapeutic mechanisms discovered from studies of herbal

medicines have routinely been used in new drug discovery. However, the insufficient mechanistic understanding of Medicinal Plants (MPs) hinders the efforts of developing new drugs based on the novel therapeutic mechanisms of MP ingredients. With known drug target information, virtual screening technologies are explored in the rapid analysis of the therapeutic mechanisms of effective herbal medicines. While a number of methods bear the potential in this application, our testing results on an extended docking method, the inverse docking approach, suggests its usefulness in facilitating the rapid analysis of the therapeutic mechanisms of effective herbal medicines.

Currently, computer aided drug design approaches mainly focus on the structure properties of a drug target and its possible binder to find or design a chemical that could bind the target tightly. However, these approaches based on the “lock and key” principle neglect the important processes prior to and after drug–receptor interactions. Therefore, the success rate of new drug candidates is still low. Introducing the consideration of mechanisms of drug action into the early stages of drug design process becomes a popular idea among drug design experts. In this regard, the drug target directed *in silico* approaches discussed in this work can be regarded as part of the efforts toward therapeutic mechanism based drug design. Novel approaches introducing the consideration of ADME profile, potential toxicity effects and other important factors into the early stages of drug discovery process would be interesting topics that follow this work.

Chapter 1

Introduction

This thesis is submitted to the Faculty of Science in partial fulfillment of the requirements of the degree of Doctor of Philosophy.

1.1 Introduction to drug discovery

The search for new, effective and safe drugs has become increasingly sophisticated. Two pronounced characteristics marked the modern age of the pharmaceutical industry: "competitiveness" and "high cost". Driven by the high exclusive marketing profit, competition between pharmaceutical companies is much more intensive than before. Moreover, it is a competition by innovation [1], as highlighted by the title of an article in a research management journal: "'Innovate or die' is the first rule of international industrial competition" [2].

Besides the profit, the cost of discovering a new drug is also very high. Recent statistics shows that it would take 10-12 years, 200-350 million U.S. dollars to discover a new drug [3]. And this cost has been growing at a rate of 20% per year [3]. To alleviate this problem, efforts have been directed to reduce the cost and time span needed for the discovery of a new drug. In consideration of the current patent protection period of 20 years for new drugs, any advance in getting a drug out more quickly is desirable. In addition to its great contribution to the improvement of our life qualities, it is enormously profitable. If the research and development (R&D) stage took 10 years, the exclusive marketing period would only have 10 years left. If the R&D time were to be shortened for 2 or 3 years, not only a big amount of R&D funding could be saved, but also a longer precious exclusive marketing period would be rewarded.

More and more computer approaches are now being developed to reduce the cost and cycle time for discovering a new drug. In order to appreciate the drug target directed *in silico* approaches in drug discovery and development, the background of drug discovery is necessary to be introduced first.

1.1.1 History of drug discovery

Around the period from 1872 to 1874, as a medical student in the laboratory of the anatomist Wilhelm Waldeyer at the University of Strasbourg in Germany, Paul Ehrlich observed that certain dyes showed selective affinity to biological tissues. This observation led Ehrlich to postulate the "chemoreceptor hypothesis" [4]. This hypothesis argued that certain chemoreceptors on parasites, micro-organisms, and cancer cells would be different from analogous structures in host tissues, and that these differences could be exploited therapeutically. This idea gave rise to the birth of chemotherapy, laid the ground in immunology and pharmacology, and subsequently led to the drug discovery practices.

In the late 19th and early 20th century, the development of analytical chemistry methodologies such as chromatography, mass spectrometry, Nuclear Magnetic

Resonance (NMR) spectrometry [5,6] and purification techniques used in organic chemistry [7-9] had been proved fruitful in the purification and characterization of active ingredients form medicinal plants. For instance, morphine [10] was first isolated from opium extract in 1815 and papaverin [11] in 1848. Another prominent example is the discovery of penicillin [12] as an antibiotic by Alexander Fleming in 1929 from a penicillium mold. The discovery of penicillin had opened a door for other scientists to search for other chemically related derivatives as well as new antibiotics. Since then, many drug companies established their own research units to search for drugs that exerted other pharmacological or chemotherapeutic properties.

The advances in biochemistry [13] also influenced drug discovery significantly. Many drugs were found to exert their effects by interacting with biological macromolecules such as enzymes, DNA (deoxyribonucleic acid) or RNA (ribonucleic acid), glycoproteins, hormones, receptors and transcription factors, which are regarded as drug targets. It is also well understood that in most of the cases, drugs exert their functions by interacting with their targets mainly by non-covalent bonds such as van der Waals interactions, the same hydrogen bond interactions, and electrostatic interactions [14]. Only in few instances are covalent interactions formed [15].

1.1.2 Modern drug discovery

After more then 150 years of development, the discovery and development of a new drug is still a long and expensive process while it has become much more competitive. At present, new agents discovered not only need to show the desired therapeutic effects, but also need to be demonstrably better than existing drugs in terms of less side effects and higher efficacy. The development and improvement of drug discovery technologies is indispensable in order to win the competition of innovation [1] in the modern pharmaceutical industry.

As illustrated in Fig 1.1, a typical new drug discovery process starts from target identification, which is followed by the search for drug leads and then clinical trials.

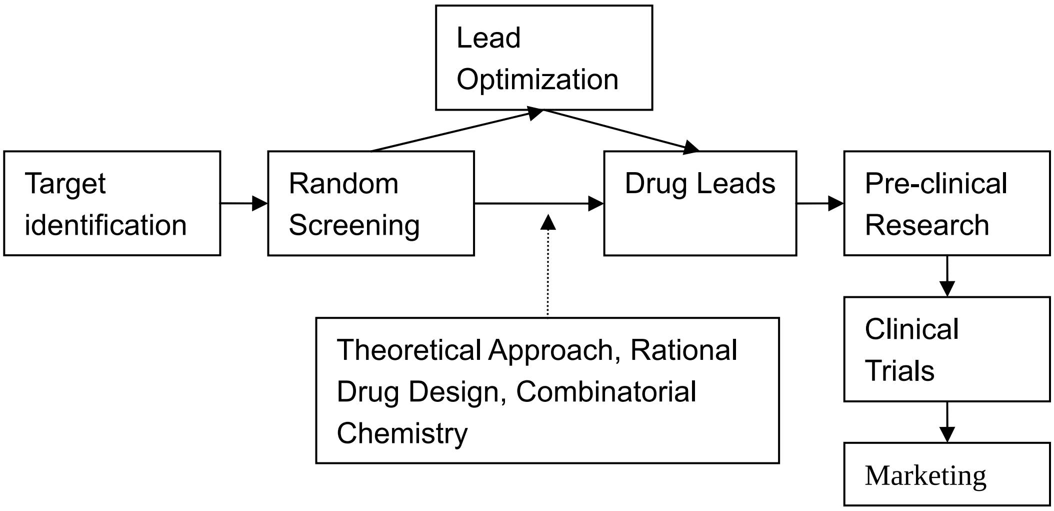


Figure 1.1 Stages of the new drug discovery process

The step of lead discovery is considered a bottle-neck of the drug discovery process [16,17]. In the past, leads were mainly discovered by random screening of a large chemical library. The sources of chemicals can be diverse such as active ingredients of natural products, derivatives of existing drugs, or even random synthesized chemicals. Most large pharmaceutical companies have their own corporate libraries, which contain the chemicals accumulated from years of efforts. It was reported that only one potential lead can be identified by random screening of

10-20 thousand of chemicals [3]. Therefore, the efficiency of mere random screening is very low.

The increasingly better understanding of the drug-target interaction mechanism and rapid advances in biochemistry and organic chemistry lead to the advent of computer aided drug design (CADD) [18-24], which aims to help the rapid and efficient discovery of drug leads. These approaches can be grouped into three categories according to their different strategies.

1.1.2.1 Combinatorial chemistry based approaches

One way to improve the efficiency of lead discovery is to reduce the average time and cost required for individual target-chemical binding affinity assay. This idea is fulfilled by the emergence of combinatorial chemistry [25] in the 1990s. Combinatorial chemistry provides a tool to do systematic screening of a large number of small chemicals. Building blocks are first designed by computer software using molecular modeling techniques. A combinatorial chemical library is then synthesized or virtually synthesized maximizing the molecular diversity [26,27]. With the help of high-throughput screening technologies, the average time and cost for screening an individual compound in a large chemical library are significantly reduced [28]. Combinatorial chemistry is mainly based on wet-lab experiments and is not within the scope of this work. Therefore, it will not be covered in detail here.

1.1.2.2 Receptor structure based drug design

Another way to improve the efficiency of lead discovery is to focus on those chemicals that are more possible to be drug leads, which is fulfilled by rational drug design approaches.

In case that a specific drug target and its 3D structure are known, receptor structure based drug design can be conducted. With the progressing of molecular biology, X-ray crystallography and NMR techniques, the structures of many drug targets have been determined [29]. More structures of drug targets can be modeled using homology-based methods [30]. Based on the 3D structure of the macromolecule receptor, molecular modeling techniques [23] are first applied to infer the mechanism of interaction between the target and its ligands. The essential structural features of the target are then summarized from the mechanism, such as electrostatic interaction areas, hydrophobic interaction areas, hydrogen bond donors and acceptors. Base on these features, rational drug design methods can then be used to obtain possible starting structures for leads optimization. There are two kinds of such methods, namely the “whole-molecule method” and the “connection method”.

The “whole-molecule method” mainly relies on the molecular docking technique [31-38]. It searches an entire 3D structure database of small molecules to find putative drug leads for a specific therapeutic target. In this course, docking single or multiple small molecules in single or multiple conformations to the receptor binding sites of the target is attempted, in order to find the best putative ligand-receptor complex conformation. Testing results on a number of flexible docking algorithms have shown that these algorithms are capable of finding binding conformations close to experimentally determined ones [39-41]. Based on geometric and chemical complementarities, a score is given to each putative ligand-receptor complex to reflect the “expected” binding affinity. Chemicals are considered as potential drug leads if their scores pass certain threshold.

Connection methods work progressively like building a house by bricks. Functional groups that best interact with important receptor sites are first placed on the receptor, and then they gradually “grow” to a full molecule. This is like the greedy search method often used in mathematical optimizations. Many drug design tools have been developed implementing this idea, such as CLIX [42], LUDI [43], CAVEAT [44], LEGEND [45], and MCDNLG [46].

The receptor structure based drug design strategy has showed more and more significance in new drug discovery [47-54]. There are many successful examples, one of which can be found in Inviraser [51], approved as an anti-HIV drug by FDA (Food and Drug Administration, USA) in 1995. This drug was developed by Hoffmann La Roche co. Ltd. It was the first HIV protease inhibitor approved by FDA.

1.1.2.3 Chemical structure activity relationship based drug design

In the case when some effective drugs / ligands of a target are known, Structure Activity Relationship (SAR) based drug design can be performed. Usually, by studying a series of small chemicals that have similar pharmacological effects through the same mechanism, Quantitative Structure Activity Relationship (QSAR) / 3D-QSAR models [55-59] are constructed to reflect the relationship between their activities and their quantitative structure properties. Then the QSAR / 3D-QSAR models can be used to screen a chemical library for potential drug leads, as well as provide theoretical guidance on lead structure optimization. Furthermore, by means of conformation analysis and molecular modeling, 3D pharmacophore models [60-62] can be inferred from the SAR models. Based on the pharmacophore models, 3D chemical structure database queries [63] can be performed to obtain possible drug leads. It is also possible to optimize lead structures according to the 3D pharmacophore models [64,65].

The key step in this strategy is the derivation of QSAR/3D-QSAR models. In the year of 1868, Crum-Brown and Fraser published the first equation in the field of QSAR (Equation 1.1), which set forth the idea that the biological activity of a compound Φ is a function of its structure properties C [66].

$$\Phi = f(C) \qquad \text{Equation 1.1}$$

Nearly one century later, Hansch and Fujita [67,68] discovered the extra thermodynamic approach (also called Hansch approach, Equation 1.2), which says that the activity of a drug is related to, in a linear model, three descriptors, namely the hydrophobicity parameter π or $\lg P$, the electrostatic parameter σ, and the stereo parameter E_s.

$$\lg {}^{1}\!/_{C} = a \lg P + b\sigma + cE_s + ... + Const \quad (a, b, c, Const \in R) \qquad \text{Equation 1.2}$$

Modern QSAR / 3D QSAR studies use much more complicated descriptors to capture the structure features of small chemicals, such as hydrophobicity parameters [69,70], electrostatic parameters (such as Hammett parameter σ [71], field parameter F and resonance parameter R [72]), stereo parameters (such as Taft constant [73], STERIMOL parameters[74]), indicator variables (such as molecular topological index [75,76]) and computed theoretical parameters (such as electron structure parameters, force field parameters and free energy related parameters). Also, much more complicated statistical learning algorithms have been explored in QSAR studies to construct better models, which include partial least squares (PLS) [77,78], principal component analysis (PCA) [79], genetic algorithm (GA) [80,81], and artificial neural network (ANN) [82-86]. The competition for the best descriptors and the best models are still far from the end.

Small molecule structure activity relationship based drug design is one of the most “classical” approaches used in drug design. One successful example of the classic Hansch-Fujita QSAR method can be found in the development of the anti-cancer drug asulacrine (CT921) [87]. In the QSAR research, Denny et. al. focused not only on the DNA-binding ability of the chemical, but also tried to optimize the solubility and pKa. So far, asulacrine had entered phase II clinical trial and possessed a good prospect in the treatment of breast cancer [88].

1.2 Therapeutics target and drug discovery

The above mentioned technologies are powerful tools in new drug discovery. However, their successes are built on an appropriate selection of therapeutic intervention strategy and therapeutic targets. As the initial step in the chain process of drug discovery, this step shall be paid full attention.

1.2.1 Information resources of therapeutic targets

A comprehensive knowledge database on therapeutic targets summarizing known drug target information will undoubtedly help the selection of therapeutic targets and the design of therapeutic intervention strategies that explore the synergies between known targets [89]. However, the information about known drug targets is still scattered among the millions of available references. Work needs to be done in order to collect and sort the drug target information. We therefore directed our effort in developing a database of known therapeutic targets with the aim to facilitate convenient access of the relevant information and knowledge discovery [90].

All the information in the Therapeutic Target Database (TTD) was manually collected from available literature data with the help of a few simple automated text retrieval programs. A relational data model [91] was designed specifically for this database with deliberate effort to maximize the ability to accommodate future extensions and facilitate the integration of information. The database was finally implemented on an Oracle 9i DBMS (DataBase Management System) [92] and a public accessible web interface was built using the Active Server Page (ASP) technology [93,94]. The database schema and web interface of TTD has been extended to develop two other databases -- Drug Adverse Reaction Target (DART) database and drug Absorption, Distribution, Metabolism, Excretion Associated Protein (ADME-AP) database.

1.2.2 Discovery of novel therapeutic targets

Besides a central information source for known targets, rapid discovery of new therapeutic targets is also very important. It may not only introduce more efficient therapeutic targets, but also increase the flexibility in designing novel therapeutic intervention strategies by exploiting the synergies between known and newly discovered targets. The discovery of new targets that are sufficiently robust to yield marketable therapeutics is an enormous challenge [95,96]. The completion of human genome project [97] brought a new opportunity for target discovery by the way of systematic genome scale screening.

Conventional approaches of target discovery are mainly disease-dependent, such as screening of disease-derived cell lines, analysis of crucial elements of disease-affected pathways, examination of gene transcript levels and protein expression levels of cells in disease status [95]. These methods involve heavy wet-lab experiments as well as domain expertise in respective diseases and therefore are difficult to be applied in the genome scale target identification. Hence, rapid *in silico* disease-independent target discovery methods are desired.

The search for novel targets is, to a certain degree, similar to the search for novel drug leads in rational drug design. For example, the ligands of a certain protein share some common structural features. In a typical QSAR study, a statistical model is first constructed to learn the common features represented by a proper set of descriptors, and then used to predict new ligands of this protein according to their descriptors. Proteins targeted by drugs are belonging to a unique group among all others [89]. An appropriate set of descriptors may also reflect some common features they share, which might be used to identify new potential drug targets. This leads to the study on the prediction of drug-target like proteins by statistical learning methods described in Chapter 3.

With the known drug targets as examples, we explored the usefulness of statistical learning methods [98-103] in the prediction of drug-target like proteins based on protein sequences, which may have the potential to be applied in genome scale drug target screening. Specifically, our studies on one statistical learning method, support vector machine [104], showed that it is able to train a statistical model reasonably well to facilitate the identification of potential new drug targets in the human genome. Its overall prediction accuracy is nearly 90% high and the prediction accuracy may be further improved by new developments in learning algorithms, descriptors, and pre-processing techniques.

1.2.3 Study of novel therapeutic mechanisms

Proven efficient therapeutic intervention strategies are of great value to the designing of new therapeutic intervention strategies. Medicinal plants serve as a good repository for clinical effective drug mechanisms [105] as they have been explored therapeutically in traditional medicines for hundreds of years and have already been used as an important source for potential drug leads in modern drug discovery [106-108]. It was known that 1/3 of the currently available drugs were developed from herbal ingredients [108]. However, there are lots of effective herbal medicines that do not have their therapeutic mechanisms understood yet.

Insufficient knowledge about the molecular mechanism of these medicinal plants limits the scope of their application and hinders the effort to design new drugs using the therapeutic principles of herbal medicines. This problem can be partially alleviated if efficient methods for rapid identification of protein targets of herbal ingredients can be introduced.

Efforts have been directed at developing efficient computer methods facilitating the target identification for small molecules. The rational drug design technologies developed for searching drug leads for a certain target [41,58,109,110] may also be inversely used for the identification of therapeutic targets of effective herbal medicines with unknown mechanisms of action. For example, the virtual binding test, originally designed to search for protein binders, shows a good potential to be extended to analyze novel therapeutic mechanisms of herbal medicines. One computer program, INVDOCK [111], has been developed to search the therapeutic target database for therapeutic targets of active herbal ingredients. We selected nine herbal ingredients to evaluate usefulness of INVDOCK in the identification of therapeutic targets of medicinal herbal ingredients [112]. The results showed that the majority of INVDOCK identified therapeutic targets and their associated therapeutic effects have been confirmed or implicated by previous studies, which suggests the potentiality of *in silico* methods in facilitating the study of molecular mechanisms of medicinal plants.

1.3 Thesis outline

As introduced above, although the problems addressed in this thesis are focused on drug targets, the techniques used in this work span several relatively independent areas, namely information technology, statistical learning and molecular modeling.

As a multi-disciplinary work, two distinct audiences are addressed, one of specialists in pharmacology, the other of specialists in computer science or bioinformatics. Despite the fact that either group may find certain sections of this work elementary, such sections are included to cover backgrounds for the benefit of individuals from outside of the given field.

The multi-disciplinary nature of this work requires a slightly different thesis organization. Because the approaches used in different chapters are virtually dissimilar and independent, these methods and their backgrounds are discussed in their respective chapters to maintain the best coherency. This thesis is divided into five chapters. Chapter 1 introduces the general background of this work. Chapter 2, therapeutic target database development, describes the effort to establish a public accessible information source of known therapeutic targets. The attempt to construct a statistical model for the prediction of drug-target like proteins is detailed in Chapter 3. The study of the molecular mechanisms of medicinal plants by an *in silico* approach is documented in chapter 4. And finally, a summary of this work is presented in Chapter 5.

Chapter 2

Therapeutic target database development

This chapter describes our work in developing a publicly accessible drug target database, Therapeutic Target Database (TTD), which provides information about the known protein and nucleic acid therapeutic targets together with the targeted diseases / conditions, their pathway information and those corresponding drugs / ligands directed at each of these targets. An ontology-like database structure is devised to manage the drug target information as well as maintaining the maximum flexibility to accommodate new interests in drug mechanisms. Web interfaces built on this database structure inherits this flexibility. The work of TTD has been extended to the construction of two other drug mechanism information databases, namely Drug Adverse Reaction Database (DART) and drug Absorption Distribution Metabolism and Excretion Associated Protein database (ADME-AP).

2.1 Introduction

Pharmaceutical agents generally exert their therapeutic effects by binding to some particular protein or nucleic acid targets [89,113]. So far, hundreds of proteins and nucleic acids have been explored as therapeutic targets [89]. Rapid advances in genetic [114,115], structural [29,30] and functional [116] understandings of disease related genes and proteins not only raise strong interest in the search for new therapeutic targets, but also promote the study of various aspects of known targets including the molecular mechanisms of their binding agents, related adverse effects [117], and pharmacogenetic implications [118], etc. The knowledge gained from such studies is important in facilitating the design of more potent, less toxic, and personalized drugs. Development of advanced computational methods for bioinformatics [119], molecular modeling [120], drug designing and pharmacokinetics analysis [54,56,111] increasingly uses known therapeutic targets and drugs to refine and test algorithms and parameters. Therefore, a database that provides comprehensive information about therapeutic targets will be helpful in catering to the needs and interests of the relevant communities in general and those unfamiliar with a specific therapeutic target in particular.

Database development is one of the major concerns in the field of bioinformatics [121]. The motivation for design and development of bioinformatics databases comes from the challenge of bridging the gap between knowledge and their efficient management (storage, retrieval and processing) in biomedical sciences. It is said that in the post-genomic area, the annotation of sequences would be a major direction of bioinformatics [121-123]. The development of specialized domain knowledge databases such as TTD can be regarded as part of this effort.

In order to provide a background for readers who are not familiar with biological databases, a brief history of bioinformatics with the focus on publicly accessible databases is briefly introduced below.

Although the term bioinformatics was first coined in the 1980s, the idea of using computers to store and manage biological data was actually initiated by X-ray protein crystallographers in the 1960's [124]. Their early work led to the establishment of the first bioinformatics database in 1971: Brookhaven National Laboratory's Protein Data Bank (PDB), a database of 3D protein structures [125].

However, the advent of what we call bioinformatics today was mainly driven not by X-ray crystallographers but by the development of improved automatic DNA sequencing technology [126,127]. Prior to these Nobel-prize winning developments, it would take a laboratory at least two months to sequence just 150 nucleotides. By the end of 1970s, it was possible to sequence around 200 bases per day. Owing to the introduction of fluorescence labeling technology [128] and multiplexed capillary electrophoresis [129-131], fully automated DNA sequencers soon appeared. Now with instruments such as the ABI 3700 or the Pharmacia Megabase 500, it is possible to sequence 500,000 bases per day on a single machine. Today, companies such as Celera, Incyte, Monsanto and others are capable of sequencing up to 100 million bases a day.

Because of these new technologies, DNA sequencing activities became heavily dependent on computer software for assembling, storing and managing DNA sequence data [132-135]. The rapid accumulation of DNA sequence data also stimulated much interest in the development of statistical methods and computer programs for analyzing DNA and protein sequences [136-139]. The need for computational tools was especially amplified with the launch of the Human Genome

Project in 1990 [140]. Beginning as a 15 year effort coordinated by the U.S. Department of Energy and the National Institutes of Health, its ultimate goal is to map sequence and identify all 30,000+ genes in the human genome. The first draft of human genome was completed on June 25, 2000 and released publicly on Feb 15, 2001 [141-143]. It is expected that a finished version of human genome will be released very soon. By then it is also expected that the genomes of many other organisms will have been sequenced. Not only has bioinformatics played a key role in handling, sorting and storing this genomic information, it is also expected to help with the new challenges ahead in inferring gene and protein functionality [122,123], which is critical in the advances of biomedical and pharmaceutical sciences.

Another stimulus for the rapid advances of bioinformatics has been the spectacular growth in computer technology [144]. It is uncannily predicted by Gordon Moore in 1965: “The processing speed of a microchip will double about every 18 months”. Today, this trend still holds true and it is know as the Moore’s Law [145]. Such rapid rate of computer hardware development has led to the creation of a thriving computer industry that delivers very high performance machines at relatively low prices. This in turn has led to the ubiquitous distribution of desktop computers, allowing easy access to computational tools among biologists and drug designers. Also, there is another significant reason for the rapid growth in computer usage among biologists and pharmaceutical researchers, which is the emergence of the “Information Superhighway” – Internet [146-148]. Originally developed in 1969 by the U.S. Department of Defense for research in communication networkings, ARPANET

[149] (as it was called then) grew from a text-only messaging system to a graphical, interactive communication medium, enabling rapid information exchange [150,151]. By 1993, Internet uses exploded with the introduction of browsers such as Mosaic and Netscape. These web browsers and their special communication language, HTML (Hypertext Markup Language) [152], have greatly facilitated the access and communication between individuals, research labs, universities and other large research organizations. Taking advantages of the information highway, centralized biological databases have been established. Dedicated bioinformatics web servers such as EXPASY (http://www.expasy.org) [153] and National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov) [154] heralded the establishment of the Internet as the primary means of communication among biological and pharmaceutical researchers, causing the field of bioinformatics to truly take off. So far, there are more than 500 published biological databases and this number keeps growing annually [155-157].

While many databases have been built for knowledge exchange and discovery with different focuses, a public database focusing on therapeutic target information has not yet been established. Although probably all the proteins and nucleic acids targeted by drugs are in other databases, they are not specified as therapeutic targets and it is troublesome for researchers to search for information across multiple databases. Moreover, because of the different focuses of these databases, the annotations they provide may not be so relevant to drug discovery. A database focusing on therapeutic targets are therefore needed as a basic tool in pharmaceutical sciences to support the multi-disciplinary effort of modern drug discovery. For example, with comprehensive drug target information, *in silico* approaches may be applied to facilitate the discovery of novel therapeutic targets and therapeutic mechanisms, which are discussed later in the next two chapters.

2.2 Collection of therapeutic target information

A survey of modern drug design approaches reveals that the information on three types of molecules is of great interest to relevant communities: drug targets (proteins or nucleotides), drugs / chemicals that bind these targets and natural ligands of these targets [48,50,54,56,74,113,158].

Drug targets are the primary focus of the database. Important properties of a target include its synonyms, related diseases and pathways. Unlike small chemicals, there is no systematic naming protocol designed for macromolecules. Contemporary naming for proteins and genes are not unified. For example, prostaglandin H2 synthase, a well known therapeutic target for inflammation, is also known as cyclooxygenase [159], while the two names bear no obvious morphological similarity. The prevailing heterogeneous naming in literature makes it necessary to enforce a standardized or systematic nomenclature for drug targets. Therefore, a unique identifier needs to be assigned to each target, which is also the solution adopted by major sequence databases such as SWISS-PROT and NCBI. In TTD, the most popularly used target name is chosen for each target and other names of the protein are stored as its synonyms. The therapeutic effects achieved through the regulation of target activity are no doubt the most important feature of a therapeutic target. To understand the therapeutic effects, this regulation of target activity shall be examined in the complex pathways in the host organisms [160]. The pathway information is therefore very useful to a variety of applications such as finding alternative therapeutic targets, designing a therapeutic intervention strategy which involves multiple co-operating targets, and analyzing potential drug-drug interactions. As introduced in Chapter 1, receptor 3D structure based approaches require the 3D structures of target molecules. In case that the 3D structure of a target has not been resolved, the primary sequence of the target may be used to derive its 3D model. Therefore, cross reference to PDB, the protein 3D structure database, and SWISS-PROT, a major protein sequence database with wealthy annotation, shall be established whenever possible.

A new drug discovery process can also start from the structural information of the small molecules that bind a certain target. In this case, a series of known binders of a target are analyzed to derive a structure activity relationship model. Information on drugs, investigational drugs, and other chemicals that have activities on a certain target is therefore very important. A target may have multiple binding sites [161-164]. Different drugs may bind to different binding sites of a target and exert different regulatory effects on the target activity. Therefore, drugs of different types may have different binding sites and shall be differentiated as their structure activity relationship may be different.

Drug binding is competitive in nature [165-169]. This binding competitiveness is an important factor in drug design. Natural ligands of drug targets are prevailing competitors. A drug is less likely to be effective if it binds to its receptor non-competitively against the natural ligands of the receptor. Thus the information on natural ligands that can bind to the known drug targets is also desirable.

Drugs and natural ligands are generally small chemicals. Although there is a standard IUPAC (International Union of Pure and Applied Chemistry) name for every small chemical, they are not the most widely used ones. Therefore, a unique identifier is assigned for drugs and natural ligands in the database. With the help of IUPAC names, it is much easier to identify the synonyms of the same chemical. In this database, the molecular formula, molecular weight, CAS RN (Chemical Abstract Service Registration Number, an identification number given to each registered chmeical), and chemical classification of a drug or natural ligand are provided.

Because the information collected is mostly reported in experiments using different methodologies, equipment and reagents, the heterogeneous quality of data in this database requires that the references to the original information sources be provided. The citation of the literature is therefore provided wherever applicable.

With the rapid advances in new drug discovery, more and more information about explored drug targets and new drug targets are being generated. An automated literature information extraction system, if available, is desired with this consideration. However, biological literature are unstructured materials, which are considered very difficult for automated information extraction [170-172]. A survey of current literature information extraction technology showed that there are some major obstacles in this application besides those inherited from natural language processing. First, the molecule names are difficult to recognize [173]. They are often composed of several words that also have their own meanings respectively. This makes it difficult to determine the boundary of protein names. Also, various abbreviations are used for proteins and nucleotides. The same abbreviation may or may not mean the same thing in different contexts. Previous successful attempts in automated biomedical literature information retrieval usually work in small domains and use name dictionaries to avoid the problem of recognizing protein and nucleic acid names [174]. This is not a feasible solution to our application as the work needed to construct a complete name dictionary of human proteins and nucleic acids is too heavy to be afforded. Second, pronouns are extensively used when describing complex relationships between molecules. The determination of the objects indicated becomes particularly difficult especially when there are more than one pronouns in the same sentence [175]. The third difficulty lies in the understanding of generalized terms and narrowed terms. For example, rhodopsin is a kind of G-Protein Coupled Receptor (GPCR). The description of the common characteristics of all GPCRs will also apply to rhodopsin. To address this problem, it needs the biological domain knowledge to be “hard-coded” into the information extraction system [176,177], which is an extremely difficult work. Therefore, fully automated literature information extraction methods may not be ready for this application until the above mentioned problems are sufficiently addressed. However, simple automated text retrieval programs based on key word searching are developed to facilitate our search for therapeutic target information, which are proved to be helpful in reducing the burden of data collators.

With the help of a few automated text retrieval programs developed in PERL, we downloaded all the literature in NCBI that contains the phrase “therapeutic target” in their abstracts. Efforts have been made to manually extract information from available literature. Only those proteins and nucleic acids that had already been explored by current therapies or had been suggested explicitly by the author as potential therapeutic targets were included in the database. A total number of 433 targets were found in the literature. It has been reported that approximately 500 therapeutic targets have been exploited in the currently available medical treatments [89]. The search for therapeutic targets aims to collect as many known targets as possible. However, descriptions of some of the targets in the literature were not specific enough to point to any particular protein or nucleic acid as the targets. Hence these targets were not included in this database.

2.3 Therapeutic target database development

Before undertaking a discussion of therapeutic target database development, it is important to determine the expected system functionalities and guidelines for further designing processes. The technology platform and software tools suitable for this project shall be selected accordingly.

2.3.1 Requirement analysis

The database system is expected to store and manage the information about known therapeutic targets. The interested types of information have been discussed in the previous section. In consideration of the fast enriching data about known drug targets or potential drug targets, a convenient updating mechanism is needed for this database. Also, a good system should not be designed to satisfy only the current needs; it should also take full consideration of the possible change and extension in future. Rational drug design is pacing fast nowadays. Different designing approaches take interests in different facets of drug targets. In the foreseeable future, more types of information will be needed by new or improved rational drug design approaches and therefore needed to be added into this database. Small changes or a complete overhaul of the database structure may be needed with this regard. No database or other software can be suitable for use forever; however, a flexible database structure that can be extended to incorporate these new interests with minimum necessary changes is desired. Moreover, the data collection work of two other drug mechanism information databases was in progress parallel to the development of TTD. As an augmented goal, it would be better that the design and implementation codes of TTD can be re-used in the development of these two databases.

Before the actual database development, it is also to be determined which technology platform and software tool are to be used to establish this database.

2.3.1.1 Databases development approaches

There are several approaches to establish a database in past bioinformatics practices. The common ones include the flat-file approach, the relational approach, and the manual approach.

In the flat-file approach, data is organized in text files where individual records of data are represented by a set of lines in strict order with symbols that allow the computer to find and retrieve specific pieces of information [178,179]. SWISS-PROT, EMBL, GenBank and OMIM are examples of the databases using this approach.

In the relational approach, a set of tables (also known as relations) are created by a database developer to reflect the inter-relationship between the data stored in the database [180-185]. Typically database management software such as SQL Server, Access, IBM DB2, or Oracle is used to manage the querying, updating and re-structuring of the database. The relational approach is so far the most widely utilized and dominant mainstream approach to data management.

In the manual approach, the information is manually coded into static web pages. Data are organized hierarchically and can be navigated following the hyper-links from the portal page. Usually no software helping the search and management of the data is used and limited search facility is provided. This approach mainly serves databases with very small scale and highly specific scope, such as the protease inhibitor database (http://www.yorvic.york.ac.uk/~proteinase/).

Manual approach is the easiest way to create an information repository without the requirements of any specific software. However, its limitations in search facility and data maintenance are severe drawbacks for middle and large scale databases. The scale of TTD rules out the manual approach as a good choice.

The flat-file approach comes from the earliest way of exchanging biological information -- distribution of copies of flat-files to researchers. Since then, a large number of biological repositories have emerged, and the availability of the Internet has made it possible for researchers to access them without having to install and manage local copies of the data. Due in parts to the flat-file based origins, many of the major present day biological data repositories, e.g. SWISS-PROT, are established using flat-file indexing systems. They are mainly efficient searching engines built on the concept of indexing. Information retrieval is performed by keywords, or by conjunctive or disjunctive combination of a set of keywords. Numerous interfaces [186,187] has been built to allow one to search for desired information in a collection of heterogeneous databases.

The advantage of such flat-file based systems is self-explanatory content that is optimized for human readability. The incorporation of hyperlinks into such records further allows for extensive cross-referencing. However, the flat-file approach is relying on mere text-matching indexing. It shows significant drawbacks in comparison to the more comprehensive relational approach, which provides a number of desired capabilities for complex queries and data maintenance with the support of a relational database management system (RDBMS):

1. Complex query support is limited. Present flat-file based repositories offer HTML forms that accept search terms as input. Search engines parse indexes of keyword to find matches, and retrieve matching records. A more elaborate search form allows the specification of field specific terms and Boolean combination of different search terms. It is important to note that, in the majority of cases, queries

are limited to the text matching approach. There is no support for complex queries with analytical requirements or nested queries such as “what are the targets unique to liver?” This kind of comprehensive listing of organ-wise unique targets has the potential to minimize drug side effects. In the contemporary relational databases, Structured Query Language (SQL) [188,189], an industrial standard, is supported almost unanimously, which could be used to construct virtually all kinds query logic in order to get comprehensive and specific results. For instance, the above mentioned question may be answered by a SQL query like the following:

```
SELECT DISTINCT [Target Name]
FROM [Target Table]
WHERE [Organ] = ‘liver’
      AND [Target Name] NOT IN
                 (SELECT DISTINCT [Target Name] FROM [Target Table]
                   WHERE [Organ] <> ‘liver’);
```

This SQL statement contains nested query which is not supported by indexing engines such as the one in SWISS-PROT.

2. Data maintenance is tedious and difficult. A flat-file indexing system only provides the functionality of a search engine with no support for data maintenance. For example, when updating the content of a flat file, there is no constraint that can be enforced in normal text-editors to check whether reasonable data have been entered in the correct format. For example, one might enter “30-Feb-1997” in a text editor as a date by mistake. This obvious error could only be identified by a thorough check of the data later. However, A RDBMS will refuse to accept this kind of “illegal”

data if appropriate constraints have been enforced. This will significantly reduce the error rate in data preparation. Also, any updates in the data files will not automatically take effect in the flat-file indexing system. It usually requires a re-indexing process to make the changes effective. For big repositories like SWISS-PROT and TREMBLE, a complete re-indexing would take more than 10 hours. However in RDBMS, efficient algorithms are implanted to keep indices up to date on the fly while you are modifying the data.

3. Views cannot be decoupled from underlying data. In other words, records are always retrieved in entirety. This will reduce the query performance due to the operations wasted in retrieving irrelevant data. In RDBMS, SQL is able to control which information is needed by defining appropriate views upon the tables. An RDBMS is able to carry out the queries optimally according to the user defined SQL.

In view of these advantages of the relational approach over the others, it is considered to be the best one for the development of TTD

2.3.1.2 Selection of RDBMS

The relational approach requires the support of an RDBMS. The RDBMS enables users to create and maintain a relational database. They are designed to control data redundancy, restrict unauthorized access, provide persistent storage for program objects and data structures, permit inference and actions using rules, provide multiple user interfaces, represent complex relationships among data, enforce integrity constraints, and provide backup and recovery supports [92,190].

Although the direct objective is to develop a therapeutic target database, the selection of the RDBMS should take further considerations for the subsequent database projects and the data analysis requirements by other projects of our group. Specifically, a desired RDBMS should be able not only to support several small databases, but also to hold local copies of existing major public biological databases and provide integrated analytical query ability. Therefore, a high performance large scale RDBMS is needed.

As illustrated in Figure 2.1 (May 2001 IDC report on 2000 RDBMSs), the market share of RDBMSs is: Oracle (Oracle 9i) 46%, IBM DB2 (DB2 UDB 7.2) 24%, Microsoft SQL Server (SQL Server 2000) 7% and others 23%. For better compatibility with other existing applications, we decided to choose a RDBMS from the best selling ones, namely Oracle 9i, DB2 UDB 7.2 and SQL Server 2000.

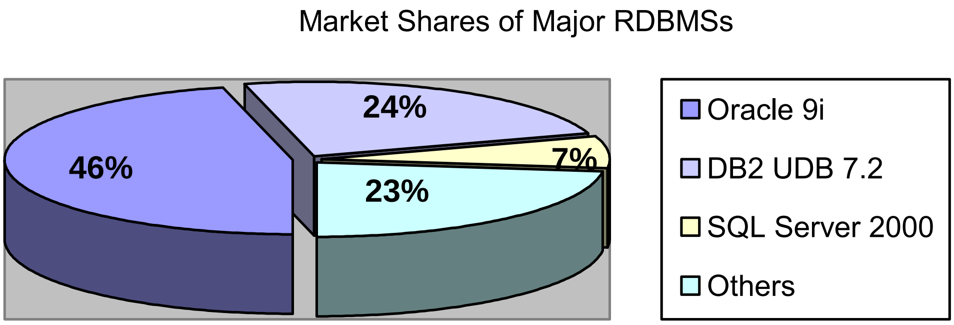


Figure 2.1: Market shares of major RDBMSs. Based on May 2001, IDC report on 2000 RDBMSs

The factors affecting the selection of RDBMS include: platform and system requirements, supported data types, program language supporting, application development features, manageability features, security features, analysis ability, internet ability, price and performance.

After a careful technical assessment of the major RDBMSs, Oracle 9i is selected. It is clear that SQL Server 2000 was the least expensive of the three. However, it could only run on the Windows platform, which limited its performance. DB2 UDB 7 had just moved from its main frame to Client/Server based database market. Similar to Oracle 9i, it could run on many operating systems, and its data types were best compatible with ANSI (American National Standards Institute) SQL definitions. However, Oracle 9i had the most variety of modules and development tools, including modules for data mining and online analytical process, which is essential for high-end data analysis purposes. For many years, it had led the way in indexing and query optimization technologies while it is not worse than its competitors in other important aspects. Also, Oracle 9i is a fully object-oriented database, which conforms to the trend toward Object Oriented Programming (OOP). Oracle had kept the biggest market share for years. According to the 2001 statistics, over half of the fortune 100 corporations used Oracle as their database servers. And finally, it was said that the price for a full featured Oracle RDBMS was comparable to that of a full featured DB2 RDBMS. Therefore, Oracle 9i was selected to be the platform for our database projects and other data analysis tasks.

After the determination of approach and software platform, the actual database design begins.

2.3.2 Database design

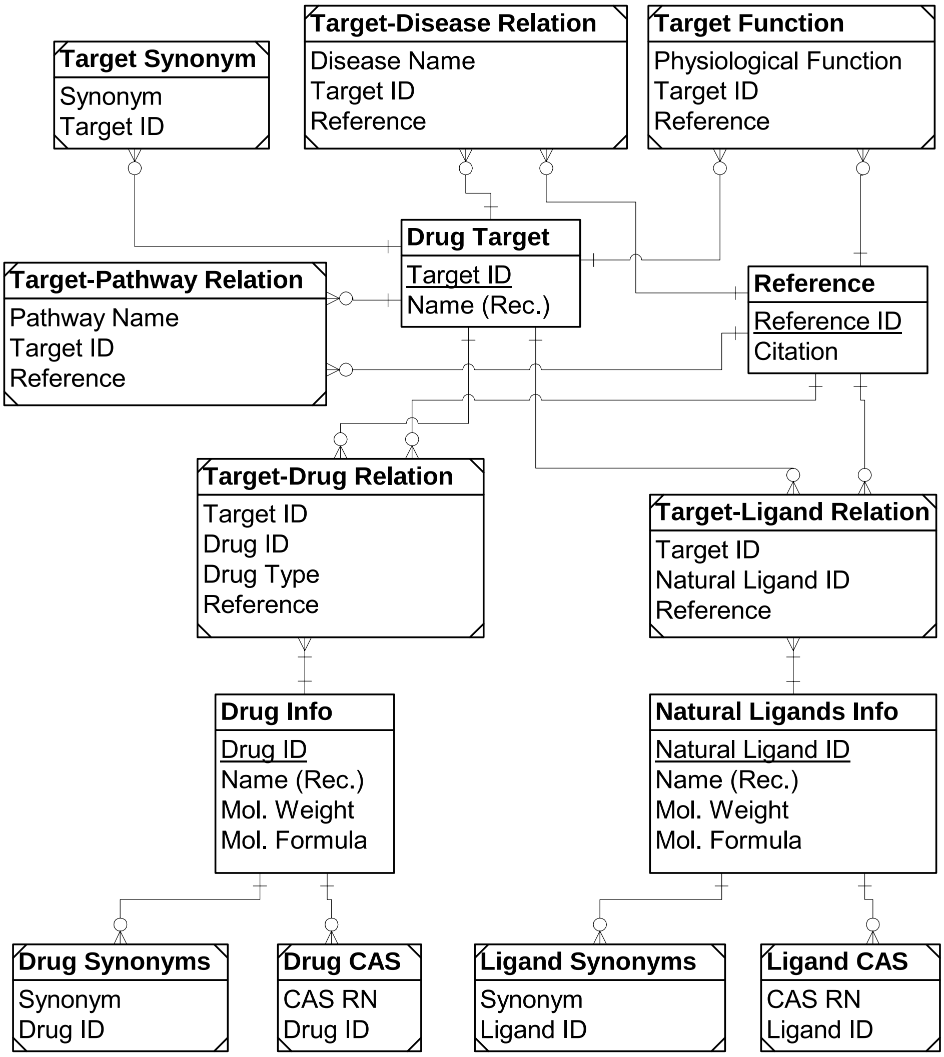
The database designing process can and should be divided into three phases: conceptual design, logical structure design and physical design, as given in the widely accepted Information Engineering Methodology (IEM) [191,192]. The conceptual design phase consists of defining the types of information to be stored in the database and documenting them. The logical design phase consists of putting the conceptual design into practice in the software of choice by creating data tables and the relationships between them. Physical design phase allows the designer to determine how the data is to be stored on the magnetic media of a computer.

2.3.2.1 Conceptual design

The conceptual design phase is a “high-level” phase of design and is independent of the choice of database management systems. The result of this designing phase is a set of documentation diagrams, whose purpose is to create discussion and understanding of the database design before the implementation work begins.

The design documentation that depicts the semantic relationship between data is called the entity relationships diagram (ERD) [193,194]. This data modeling technique breaks data types down into entity types, attributes and relationship types. An entity type is a collection of entities that share common properties or characteristics. The properties or characteristics of an entity type or relationship type that are of interest to the organization are called attributes. Relationship types are meaningful associations between / among entity types. There are three categories of relationship types: one-one relationships, one-many relationships and many-many relationships.

A natural way of drawing the entity relationship diagram is to take target molecules, drugs, ligands, and references as entity types and their respective properties as their attributes while relationships are established accordingly. For instance, the entity type target molecule will have attributes including a unique ID, synonyms, cross references, related diseases, functions, pathways involved. And they are linked to natural ligands and different types of their drugs by many-many relationships. However, this simple method is not applicable because it may result in potential multiple-valued attributes which violates the first norm in database design and will cause problem in query and data maintenance [195,196]. For instance, the attribute “synonyms” may have multiple values for one target. It would be hard for applications and SQL queries to distinguish different synonyms within the same data item. To address this problem, these “attributes” that might have multiple values are “promoted” to weak entity types. A weak entity is an entity of which its existence depends on the relation with another entity (the identifying entity). For example, the identifying entity type for a weak entity of “target synonym” is an entity of “target molecule”. The existence of “target synonyms” depends on their relationship to a “target molecule”. A weak entity has no key attribute because it cannot exist without the relation it has to its identifying entity. The resulted first draft of ERD (using "Crow's Foot" notation, which is the accepted IEM convention) is shown in Figure 2.2. To have a clearer view of the attributes of relationships, in this ERD, any relationship with attributes is drawn as weak entities identified by both sides of the relationship.



Therapeutic Target DB	Edit Date: 4/26/2001 2:20:57 PM	
Description: Direct modeling of real-world entity relationships. Four main entity types: therapeutic target, drugs, natrual ligands and references.		
Target DB: Oracle 9i	Rev: 0	Creator: Chen Xin
Filename: ERD0		Company: BIDD Group

Figure 2.2: The first Entity Relationship Diagram of TTD.

2.3.2.2 Logical design

In the logical structure design phase, the structure of the database (also known as database schema) is created, which is a set of data tables and their connections. The data tables and the connections between them can be directly derived from the ERD. However, this approach does not, by itself, assure a good relational database for every purpose. The first design of TTD schema was derived from the above ERD, which was evaluated on a small set of sample data and disclosed several problems about this design. Accordingly, a series of modifications on the first schema is carried to introduce more unification on the data representation and flexibility to accommodate extensions. The latest revised schema resembles a semantic network in ontology research very much. These designing processes are detailed as follows.

2.3.2.2.1 ERD derived database structure

The algorithm for translating a sound conceptual design into a relational data structure is given in [197,198] as the following four steps:

First, construct a table for each entity type, containing all the attributes of the entity type and having a primary key or a unique identifier filed.

Second, construct a table for each many-many relationship type containing the unique identifier for each side of the relationship along with the attributes of the relationship.

Third, for each one-many relationship type, add the unique identifier from the “one” side to the table corresponding to the entity on the “many” side, along with all the attributes of the relationship.

Finally, for each one-one relationship type, add the unique identifier from either side to the table for the other side, along with the attributes of the relationship.

Using this algorithm, an initial version of therapeutic target database can be constructed, which consists of sixteen tables, as shown in Table 2.1. One table is created for the target molecule entity type, which includes two columns -- the unique identifier assigned to each target and its recommended name. Those types of information with one-many relationships to a target, such as synonyms, related diseases / conditions, functions, pathways involved, are stored in their respective tables with the unique identifiers of their identifying targets. The many-many relationship between target molecules and their different types of drugs are stored in one table, which contains the unique identifiers of targets, the unique identifiers of drugs that bind the targets, and the types of the drugs. The many-many relationship between target molecules and their natural ligands are also stored in one table structurally similar to the table of target-drug relationship. To provide the information sources for data quality assessments, the tables storing information collected from literature, such as the tables for target functions, target pathways, different types of drugs and ligands, all contain one column storing the unique identifier of corresponding references. For each drug, a unique identifier was assigned and it was stored in one table with the recommended name of that drug, its molecular weight and its molecular formula. Those types of information with one-many relationships to a drug, such as drug synonyms, CAS RN, and chemical classification, are stored in their respective tables with the unique identifier of their corresponding drugs. In this design, CAS RN is treated as an entity type with one-many relationship to a drug, which is different from most of the current available chemical databases, such as ACX and the Merck index, which usually provide only one CAS RN for each chemical. This is because while in most of the cases, a chemical only has one CAS RN and vice versa, a small number of exceptions do exist. For example, GTP (guanosine triphosphate) has two CAS RNs: 86-01-1 and 56001-37-7. Drug and natural ligands are all small chemicals that have the same types of information. Therefore, the data tables created for natural ligands are similar to those for drugs. The table storing information about references is relatively simple. It only contains two columns of unique identifier and reference citation.

Table 2.1: The data tables created in the first design of TTD.

Table Name	Columns	Table Relationships
TTDTG: Drug targets	TID: Target unique identifier NM: Recommended target name	
TTDDG: Drugs	DID: Drug unique identifier NM: Recommended drug name MW: Molecular weight MF: Molecular formula	
TTDLG: Natural Ligands	LID: Natural Ligand unique identifier NM: Recommended ligand name MW: Molecular weight MF: Molecular formula	
TTDRF: References	RFID: Reference unique identifier RF: Reference citation	

TTDTS Target synonyms	TID: Target unique identifier SN: Target synonym	TID ∈ TTDTG.TID
TTDTP: Target involved pathways	TID: Target unique identifier PW: Pathway name RF: Reference unique identifier	TID ∈ TTDTG.TID RF ∈ TTDRF.RFID
TTDTD: Target related diseases	TID: Target unique identifier DN: Disease name RF: Reference unique identifier	TID ∈ TTDTG.TID RF ∈ TTDRF.RFID
TTDTF: Target functions	TID: Target unique identifier FN: Target function RF: Reference unique identifier	TID ∈ TTDTG.TID RF ∈ TTDRF.RFID
TTDTDG: Drugs that bind targets	TID: Target unique identifier DG: Drug ID TP: Drug category RF: Reference unique identifier	TID ∈ TTDTG.TID RF ∈ TTDRF.RFID DG ∈ TTDDG.DID
TTDTLG: Natural ligands that bind targets	TID: Target unique identifier LG: Natural ligand ID RF: Reference unique identifier	TID ∈ TTDTG.TID RF ∈ TTDRF.RFID LG ∈ TTDLG.LID
TTDDS: Drug synonyms	DID: Drug unique identifier SN: Drug synonym	DID ∈ TTDDG.DID
TTDDR: Drug CAS RN	DID: Drug unique identifier CAS: CAS Registration Number	DID ∈ TTDDG.DID
TTDDC: Drug chemical classification	DID: Drug unique identifier DC: Drug classification	DID ∈ TTDDG.DID
TTDLS: Ligands synonyms	LID: Natural ligand unique identifier SN: Natural ligand synonym	LID ∈ TTDLG.LID
TTDLR: Ligand CAS RN	LID: Natural ligand unique identifier CAS: CAS Registration Number	LID ∈ TTDLG.LID

TTDLC: Ligand chemical classification	LID: Natural ligand unique identifier LC: Natural ligand classification	LID $\in$ TTDLG.LID

After the tables and their relationship (constraints) were created, a web interface of the database was sketchily implemented and the system was analyzed on a small set of testing data. The above data structures showed several weaknesses in the test run. They can be summarized as below:

First, the information retrieval of a single target involves many tables. It is not a big issue when the tables are queried or analyzed by SQL statements. However, when building the web interface for the database, each table needs a distinct record set object in the web server, which will lead to inefficient use of web server resources. Also, the big number of objects requires a corresponding size of codes to manipulate them, which makes the debug and maintenance of these codes troublesome.

Second, the interface codes developed on this structure are fairly rigid to accommodate extension of the database. For example, in the later designing stages, it was suggested that it would be better to classify the therapeutic targets into different categories according to its related diseases / conditions, in order to facilitate systematical studies. Accordingly, the classification information for therapeutic targets should be added into the database. With the rapid progress in rational drug design approaches, it is foreseeable that more types of information will be needed by novel or improved methods. The ability to accommodate these extensions with minimal changes in the database structure and all its application codes is desired. In this design, in order to incorporate the classification information, a new table storing the unique identifier of targets and their classes needed to be added into the database. Big changes on the interfacing codes were required to use the new table. Actually, not only the web interface but also all the database applications written in programming languages requiring extensive coding (such as C/C++) will suffer from this problem. The expandability and life cycle of this database are therefore considerably compromised. Also the database structure and interface codes, developed in this manner, are unable to be re-used in the development of other databases.

2.3.2.2.2 Revised database structure

A careful analysis of the first design revealed possible directions to address the above problems. First, in the above data structure, most of the tables were quite similar in terms of the data types stored in them. Many of them consisted of three columns: a column for the unique identifier of an entity type (target, drug or ligand), a column for a property of the entity type (i.e. natural ligands of targets or synonyms of drugs) and a column for the corresponding unique identifier of a reference citation. This similarity provides the possibility of reducing the number of tables by merging similar ones. Second, in order to minimize the effort needed when new types of information were to be added into the database, the applications shall be able to "notice" the changes in the database and adapt to the changes automatically. One possible solution would be providing the "database structure" information in one of the tables.

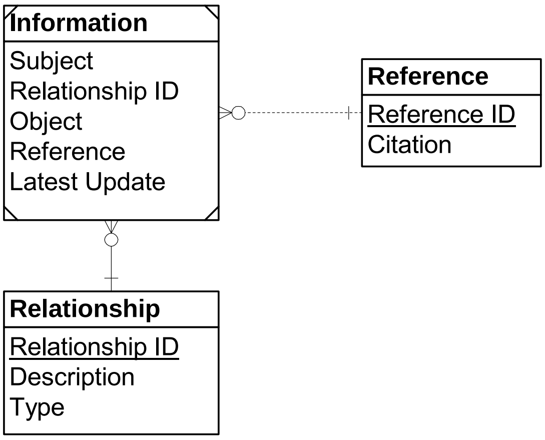
The above considerations led to the revised ERD shown in Figure 2.3. Each kind of relationship type, regardless of its type (one-one, one-many or many-many), are given a unique identifier. The unique identifiers of different relationships and their descriptions are stored in one table so that applications will be able to access the types of relationships stored in the database. The relationships are divided into two categories. One category of relationships links an entity to its attributes, regardless of single-valued attributes or multi-valued attributes. The other category of relationships links two entities together, where the other entity is represented by its unique identifier. All the information in this database can therefore be represented in a quadruple notation and stored uniformly. The quadruple notation includes the unique identifier of an entity, the unique identifier of a relationship, the right hand side of the relationship (either a property value or a unique identifier of another entity, according to the relationship category) and the unique identifier of the reference where this piece of information comes from. For example, “prostaglandin H2 synthase is a therapeutic target for treating inflammation” will be stored in the database in the quadruple form:

(“TTT0000600”, “D105”, “Inflammation”, “10878289”)

where “TTT0000600” is the unique identifier for the target prostaglandin H2 synthase; “D105” is the unique identifier for the relationship between targets and their related diseases; “10878289” is the unique identifier of the reference where this information is extracted. Similarly, “prostaglandin H2 synthase has an inhibitor, aspirin” will be stored in the database in the quadruple form:

(“TTT0000600”, “D141”, “TTD0000354”, “8728890”)

where “D141” is the unique identifier for the relationship between targets and their inhibitors, “TTD0000354” is the unique identifier of the drug aspirin and “8728890” is the unique identifier of the reference where this information is extracted. In this revised schema, there are only three tables and applications can access the “logical structure” of the database by reading the information in the relationships table. Also, to support the continuous work of data maintenance, a housekeeping field storing the latest update time of each piece of information is added. This revised schema is shown in Table 2.2



Therapeutic Target DB	Edit Date: 6/17/2001 12:05:05 PM	
Description: Modified based on Rev. 5. A unified format is used to store all the data. The definintion of logical relationships is given in another table.		
Target DB: Oracle 9i	Rev: 7	Creator: Chen Xin
Filename: ERD8		Company: BIDD Group

Figure 2.3: Revised Entity Relationship Digram of TTD.

Table 2.2: The data tables created in the revised design of TTD.

Table Name	Columns	Table Relationships
TTDINFO: Information	ID: Unique identifier of an entity TP: Unique identifier of a relationship CT: Property value RF: Unique identifier of the reference UPD: Latest update time	TP $\in$ TTDTP.TP RF $\in$ TTDRF.RFID
TTDRF: References	RFID: Reference unique identifier RF: Reference citation	
TTDTP: Relationships	TP: Unique identifier of a relationship NM: Relationship name MD: Relationship category.	

2.3.2.2.3 Further analysis of the revised database structure

This quadruple notation is applicable not only to this database but also to the other drug mechanism databases that were developed in parallel, which enabled the reuse of the TTD database schema and interface codes in the development of these databases. This universal applicability of this schema may be partly explained by its resemblance to semantic networks that are used in ontology research.

Ontologies are frameworks developed principally by the AI (Artificial Intelligence) community in the 1970s and 1980s to represent the key concepts in any research field and their inter-relationships [199]. Many bioinformatics specialists believe that they are necessary not only to make database annotations accessible to analysis tools, but also to facilitate information retrieval. For example, searching for all G-protein-coupled receptors in a database would be easier with software that “knows” that these proteins might be variously annotated as “transmembrane protein", "7TM protein", "GPCR" or "opsin". In the 1960s and 1970s, the artificial intelligence (AI) community developed several systems to embody this sort of complexity. One such system, the semantic networks, represents concepts as nodes in a graph, which are joined by arcs that specify their relationships [200-202]. Semantic network is the approach taken by many bioinformatics ontologies. It is hoped that by building these lattices of semantic associations and by hooking database entries to the appropriate points, ontologies can be used to resolve the problem of database integration.

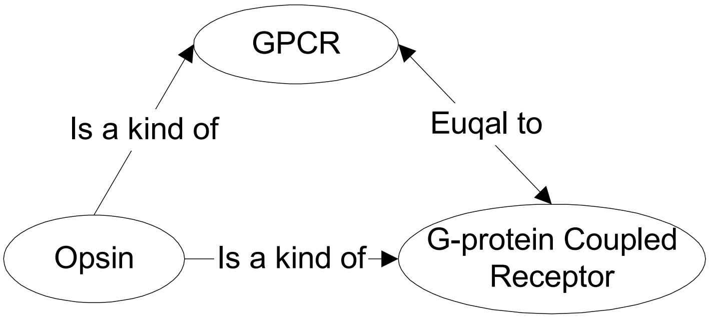


Figure 2.4: An example of semantic networks.

When a semantic network notation is devised, it is necessary to specify not only the types of the node and arcs, as well as the ways they can be combined, but also their meanings. In the above example, there are three issues. First, synonyms are used to describe the same entity (for example, "GPCR" and "G-protein-coupled receptor"). Second, relationships are required to group related concepts together (for example, "opsin" is a kind of "GPCR"). Third, there are subtle differences between terms (for example, "transmembrane" describes the location of a protein that might or might not be a GPCR). Unless the meanings of the relationships are specified precisely, the semantic network is meaningless [203]. An example of a semantic network is shown in Figure 2.4.

The revised data structure could be regarded as an extended semantic network. In this sense, the quadruple notation could be interpreted as if two nodes were linked by an arc with the reference. The exact meanings of the arcs (relationships) are stored in the relationships table. Ontologies are built for the integration of databases. Therefore, this ontology-like data structure is also expected to enable easy integration with other databases and maximizes the re-usability of the application codes developed on this data structure.

Normalization is also a very important issue in database design [195,198]. The following normal forms (NFs) have been defined: 1NF, 2NF, 3NF, BCNF, 4NF, and PJ/NF(5NF) [204]. Normalization theory is simply a formalization of the “one fact in one place” principle of good design. If a database schema satisfies a specific set of rules, it is said to be in some normal form. Normal forms are thus specific sets of rules. Each higher normal form includes all the rules of all lower normal forms. For example, a table in the third normal form satisfies all requirements for 1NF, 2NF and 3NF. The normal form requirements up to 3NF are listed below:

1NF: This normal form specified the granularities of data. It requires no relations within attributes, no composite attributes, no multi-valued attributes and no nested relations.

2NF: In this normal form, all the requirements of 1NF shall be satisfied besides every non-prime attribute is fully functionally dependent on the key.

3NF: In this normal form, all the requirements of 2NF shall be satisfied besides no non-prime attribute can be determined by another non-prime attribute.

Normalization may be an abused principle. Excessive normalization will lead to a poor performance of the database. Thus normalization should be carried out cautiously [204]. In general, the third normal form (3NF) is regarded sufficient for therapeutic target database and the revised schema conforms to the third normal form.

One problem that the revised data structure might face is that the referential integrity could no longer be enforced by simply adding foreign key constraints on relevant columns. However, this problem could be solved by setting up triggers to check the referential integrity in Oracle 9i.

The revised schema and the first ERD derived schema are mutually convertible. In a sense, the revised data structure could be viewed as each table in the first schema was attached by a tag of its identity and stored together. To convert data from the revised schema to the first ERD derived schema, one only needs to group the records according to their relationship type column (TP) and store different groups of records in different tables. To convert data from the first ERD derived schema to the revised schema, one just needs to add the relationship type information into the each table and then store them centrally. During the development of the database, a script was written to do these conversions automatically when needed.

2.3.2.3 Physical design

After the data structure was fixed, how the data would be stored on the magnetic media of the computer needed to be determined. As most of the modern database management systems, Oracle 9i can take care of this problem for the user. This designing phase is mentioned here as a distinct section only for the sake of completeness.

After all the above designing activities, the database can be then implemented and the maintenance work will keep on going all across the life cycle of the database.

2.3.3 Implementation

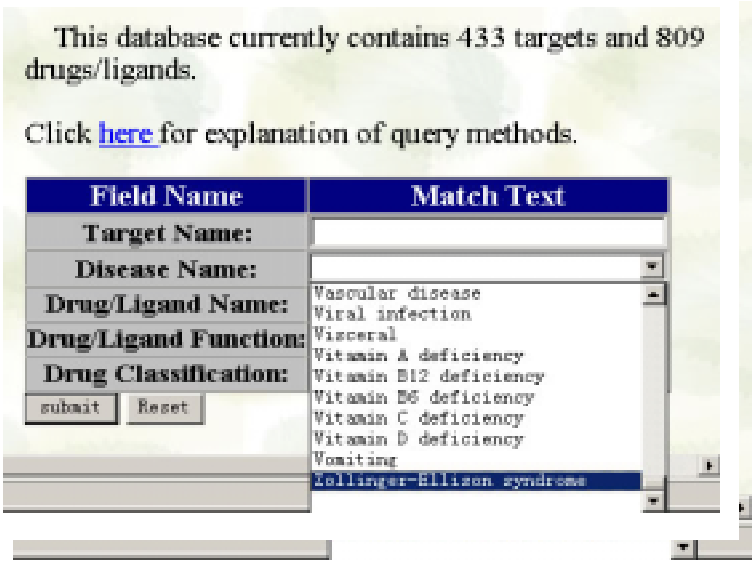
There are two parts of work needed to be done to establish a publicly accessible database. One is on the RDBMS side and the other is on the web interface side.

According to the above revised design, tables were set up in the RDBMS and the referential integrity was enforced by setting up triggers. A package was created to collect all the PL/SQL programs written for this database, including scripts to load data, check referential integrity, convert data formats, do housekeeping work as well as support the functionalities of the web interface [92,190].

Various technologies can be used to build the web interface which creates dynamic web presentations according to a viewer’s interest. Common techniques include JSP (Java Server Pages) [205], PHP (Personal Home Page) [206], ASP [207], and Perl (Practical Extraction and Reporting Language) DBI (DataBase Interface) [208] based applications. Among them, ASP has a big advantage on its ease of use. It supports ADO (ActiveX Data Objects) objects [209], ODBC (Open

DataBase Connectivity) [210] and OLE-DB (Object Linking and Embedding DataBase) [211] which make the development of web interface for databases much easier. Although in terms of run-time resource usage, ASP is not as good as JSP, it is still better than traditional CGI (Common Gateway Interface) [212] based approaches such as Perl DBI based applications. It is said that ASP technology is very good for middle-range applications, which fits TTD very well.

Using the ASP technology, the interface of TTD was developed, which has a URL at http://xin.cz3.nus.edu.sg/Group/ttd/ttd.asp. The portal page is shown in Figure 2.5. TTD is searchable by target name or drug/ligand name. It can also be accessed by selection of disease name, drug/ligand function, or drug therapeutic classification from the list provided in the corresponding selection field. Searches involving any combination of these five search or selection fields are also supported. Each search or selection field in this page will match one or more types of information in the database. For example, if any of the "recommended name" or "synonyms" of a target matches the term specified in the "target name" filed, this target is considered a hit.



This database currently contains 433 targets and 809 drugs/ligands.

Click here for explanation of query methods.

Field Name	Match Text
Target Name:	
Disease Name:	
Drug/Ligand Name:	
Drug/Ligand Function:	
Drug Classification:	

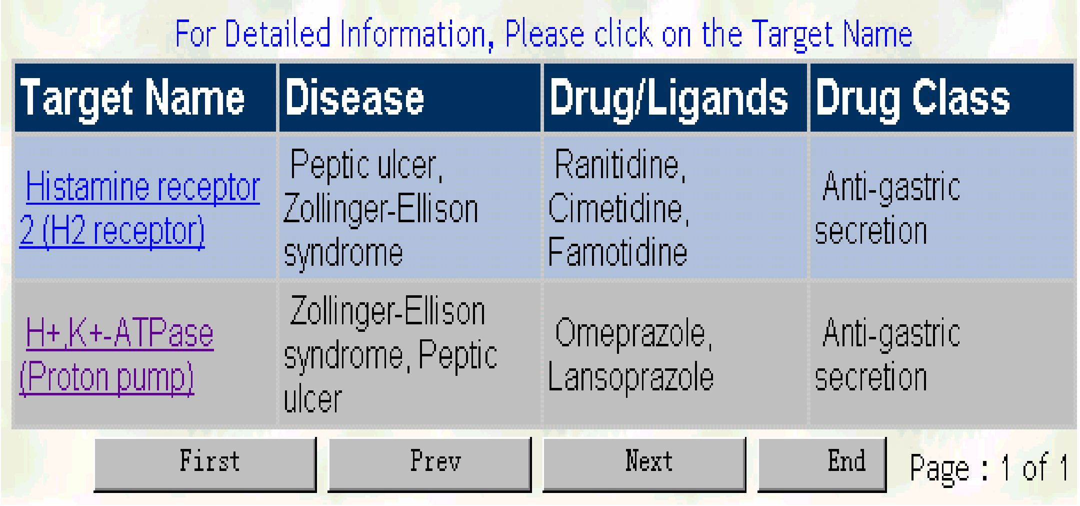
Vascular disease
Viral infection
Visceral
Vitamin A deficiency
Vitamin B12 deficiency
Vitamin B6 deficiency
Vitamin C deficiency
Vitamin D deficiency
Vomiting
Zollinger-Ellison syndrome

Figure 2.5: The portal page of TTD web interface

The search is case insensitive. In a query, a user can specify full name or any part of the name in a text field, or choose one item from a selection field. Wild characters of “*” and “?” are supported in the text field. “?” represents any one character and “*” represents a string of characters of any length. For example, input of “phosphatase” in the target name field finds entries containing “phosphatase” in their name, such as Cdc25A phosphatase or tyrosine phosphatase. On the other hand, input of “Cdc25? phosphatase” finds entries with names like Cdc25A phosphatase, Cdc25B phosphatase and Cdc25C phosphatase. Likewise, input of “Cdc* phosphatase” finds the same entries as above. In this case, “*” represents “25A”, “25B” or “25C”. “*” and “?” are not the wild characters used in SQL, therefore, all the terms are pre-processed so that they can be correctly interpreted.

The query conditions are persevered throughout a query session by cookies. The result of a typical search is illustrated in Figure 2.6. SQL query statements were dynamically constructed to pick out the summary information of the targets satisfying the criteria specified in the first page. In this page, all the therapeutic targets found are listed along with the disease conditions to be treated, drugs or ligands that bind the target, and its classification. This summary information is generated automatically by a PL/SQL scripts with parameters specifying which types of information shall be included. ASP codes for this page were written in a manner that they are able to automatically adapt to different types of summary information. Detailed information of a target can be obtained by clicking the corresponding target name.

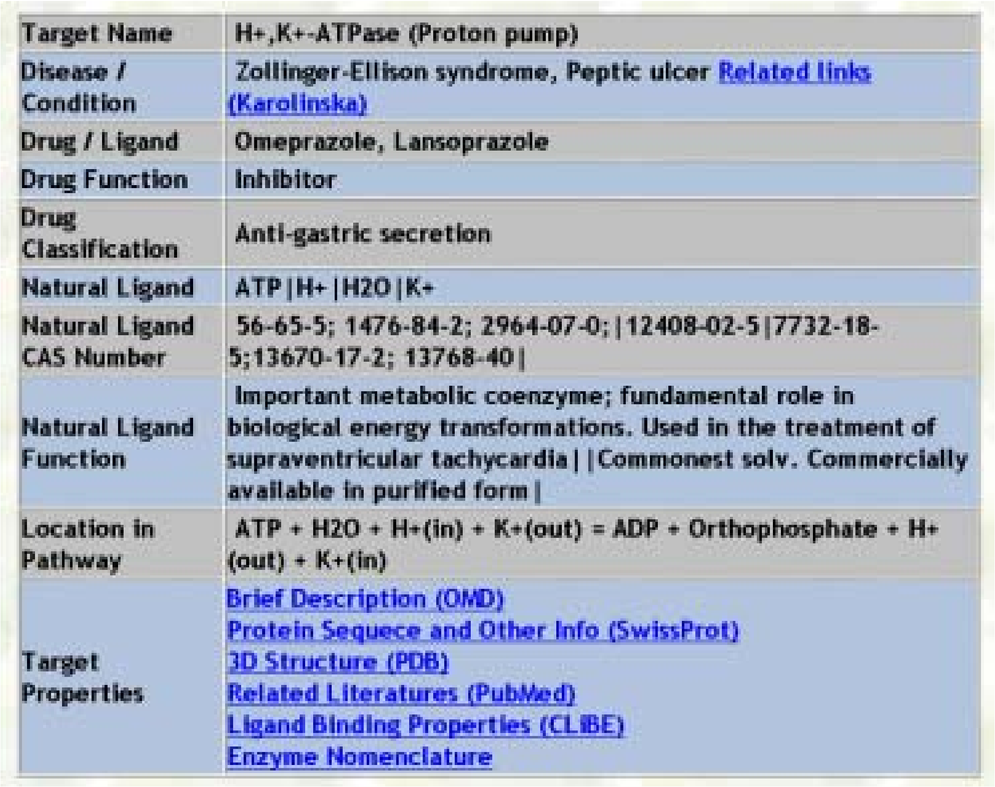
The interface displaying the detailed information of a target is shown in Figure 2.7. ASP codes for the generation of this page read the relationships type table TTDTP in the database and display all types of information about the target currently available. From the page shown, one finds target name, corresponding disease condition and cross-link to Karolinska disease database (http://www.kib.ki.se/), target function, pathway, corresponding natural ligand, known drugs or ligands directed at the target, drug type (such as inhibitor, antagonist, and blocker etc.), drug therapeutic classification, and additional cross-links to other databases that provide useful information about the target.



For Detailed Information, Please click on the Target Name

Target Name	Disease	Drug/Ligands	Drug Class
Histamine receptor 2 (H2 receptor)	Peptic ulcer, Zollinger-Ellison syndrome	Ranitidine, Cimetidine, Famotidine	Anti-gastric secretion
H+,K+-ATPase (Proton pump)	Zollinger-Ellison syndrome, Peptic ulcer	Omeprazole, Lansoprazole	Anti-gastric secretion

Figure 2.6: The TTD web interface of a search result



Target Name	H+,K+-ATPase (Proton pump)
Disease / Condition	Zollinger-Ellison syndrome, Peptic ulcer Related links (Karolinska)
Drug / Ligand	Omeprazole, Lansoprazole
Drug Function	Inhibitor
Drug Classification	Anti-gastric secretion
Natural Ligand	ATP|H+|H2O|K+
Natural Ligand CAS Number	56-65-5; 1476-84-2; 2964-07-0;|12408-02-5|7732-18-5;13670-17-2; 13768-40|
Natural Ligand Function	Important metabolic coenzyme; fundamental role in biological energy transformations. Used in the treatment of supraventricular tachycardia||Commonest solv. Commercially available in purified form|
Location in Pathway	ATP + H2O + H+(in) + K+(out) = ADP + Orthophosphate + H+(out) + K+(in)
Target Properties	Brief Description (OMD) Protein Sequece and Other Info (SwissProt) 3D Structure (PDB) Related Literatures (PubMed) Ligand Binding Properties (CLiBE) Enzyme Nomenclature

Figure 2.7: The TTD web interface of the detailed information of a target

2.4 Preliminary analysis of TTD

A total number of 433 protein and nucleic acid targets were collected in TTD. As shown in Figure 2.8, two major classes of molecules contribute to more then three quarters of the total therapeutic targets, which are enzymes (44%) and receptors (33%). Other significant classes of therapeutic targets include hormones and factors (10%), ion channels (4%) and nucleotides (3%). This composition is generally in agreement with that reported in 1997[213]. These targets cover 125 different diseases / conditions, and 809 distinct drugs / ligands directed at these targets are collected in this database.

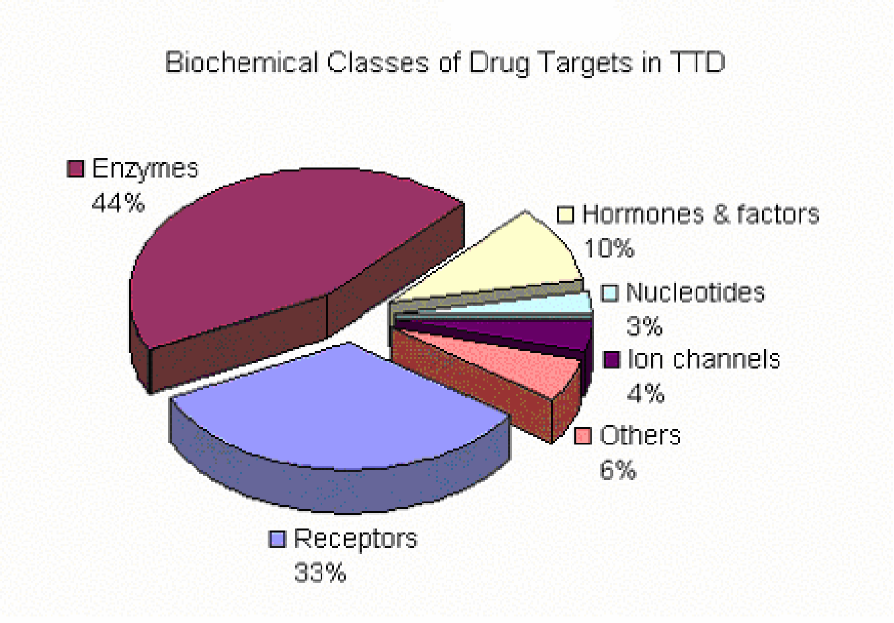


Figure 2.8: Biochemical classes of drug targets in TTD

2.5 Extension of the TTD database schema and interface

The information on drug adverse reaction and drug Absorption Distribution

Metabolism and Excretion (ADME) associated proteins are collected in parallel to the development of TTD. These are also very important aspects that affect the success of a drug.

While developing the database schema and web interface for the therapeutic target database, attentions have been paid to develop re-usable modules. By reading the relationship table, the ontology like database schema of TTD and its interface codes can virtually adapt to any predefined sets of information types, including those needed by drug adverse reaction information and drug ADME associated proteins information. Therefore, the work of TTD is readily extendable to the development of these two drug mechanism databases, namely the Drug Adverse Reaction Database (DART) [214] and drug ADME associated protein database (ADME-AP) [215].

Drug adverse reaction is often induced by the interaction of a drug or its metabolites with specific protein targets related to toxicity or side effects [117,216-219]. Knowledge about these targets is both important in facilitating the study of the mechanism of drug adverse reaction and in new drug discovery. It is also useful in the development and testing of rational drug design and safety evaluation tools [220-223]. The Drug Adverse Reaction Database (DART) is intended to provide comprehensive information about toxicity and side effect targets to the relevant communities. DART contains information about known toxicity and side effect related proteins described in the literature together with physiological function of each target, related diseases, corresponding agonists / antagonists /

activators / inhibitors, IC50 values of the inhibitors, and the toxic effect or side effect resulting from the binding of a drug. Cross-links to other databases are also introduced to facilitate the access of information about the sequence, 3D structure, function, and nomenclature of each target along with drug/ligand binding properties, and related literature. Each entry can be retrieved through multiple methods including target name, target physiological function, toxicity or side effect, ligand name, and biological pathways. This database can be accessed at http://xin.cz3.nus.edu.sg/group/dart/dart.asp.

Drug absorption, distribution, metabolism and excretion are the processes prior to and after drug-target interaction. It often involves interaction of a drug with specific proteins [224-229]. Knowledge about these ADME-related protein targets is important in facilitating the study of the mechanism of drug transportation, disposition as well as therapeutic action. It is also useful in the development and testing of rational drug design and pharmacokinetics prediction tools [225,230-235]. The ADME associated protein database ADME-AP is intended to provide information about proteins acting as ADME targets described in the literature. This database gives description about physiological function of each target, membrane location and tissue distribution, transport direction, driving force, substrates that bind to a target, pharmacokinetic effect in terms of ADME classification, synonyms and gene name. Cross-links to other databases are also provided to facilitate the access of information about the sequence, 3D structure, function, genetic disorder and nomenclature of each target along with drug/ligand binding properties, and related literature. Each entry can be retrieved through multiple methods including target name, ADME class, ligand/substrate name, and target physiological function. This database can be accessed at http://xin.cz3.nus.edu.sg/group/admet/admet.asp.

2.6 Summary

Therapeutic target database is developed from information in available literature, which is a result of collective and persistent effort over the years. It integrates the general information of therapeutic targets such as physiological functions and their therapeutic related aspects. With the rapid development of proteomics [95,236] and pathway analysis [237], the relevant information can be incorporated or the corresponding databases can be cross-linked to TTD to provide more comprehensive information about the drug targets and their relationship to other biomolecules and cellular processes.

An ontology-like database schema is designed for TTD which can easily incorporate new interests in therapeutic targets. Interface codes developed on this schema are also highly flexible. The work in TTD has been extended to develop two other drug mechanism information databases – DART and ADME-AP.

The completion of TTD not only provides a convenient way of looking up therapeutic target information, but also brings new research opportunities, such as the study of novel approaches in discovery of new therapeutic targets and new therapeutic intervention strategies, which are discussed next.

Chapter 3

Prediction of Drug-target like proteins

In this chapter, we explore the use of statistical learning approaches to predict drug-target like proteins from their primary sequences in order to facilitate the rapid discovery of new potential therapeutic targets from the large quantity of sequences in human genome. A number of statistical learning methods and pre-processing techniques are explored. It was found that the Support Vector Machine (SVM) algorithm with a fine-tuned Gaussian kernel was able to make reasonably accurate prediction, which showed its potential to be used in the genome scale rapid drug target discovery, as a novel *in silico* approach supplementary to the conventional experimental approaches.

3.1 Introduction

Target discovery constitutes one of the main components of today's early stage pharmaceutical research [3,74]. The aim of target discovery is to identify and validate suitable drug targets (i.e. proteins or nucleotides to which drug binding produces therapeutic effects) for therapeutic interventions. Only a small fraction of proteins are actually targeted by today's drugs. Indeed, a review article published in 2000 estimated that current therapies explored less than 500 distinct targets [89]. The total number of tractable targets remains difficult to estimate given the uncertainty surrounding the total number of human genes. However, it has been estimated that the number of drug targets is probably in the range of 5,000-10,000 [89]. This number is 10-20 times greater than that of the currently explored drug targets.

The discovery of targets that are sufficiently robust to yield marketable therapeutics is an enormous challenge. Through the years, many approaches have been used with varying degrees of success. Most of them are disease dependent [95], for example, target-independent screening of tumor-derived cell lines, reductionist approaches to identifying crucial elements in disease-affected pathways, “global” examination of gene transcript levels, and global examination of protein expression levels. These are mainly wet-lab based approaches which require the consumption of large amount of money and time. Disease-independent approaches were also reported, such as screening of homologs of previously drugged targets.

The fruits of the Human Genome Project will undoubtedly change how and where we look for new drugs and how we assess drug targets [238,239]. With the exception of these infectious disease targets, which are proteins or nucleotides essential to viral replication or bacterial metabolism in the infectious organisms, most of the drug targets are human proteins. Many more targets responsible for debilitating human diseases are waiting to be uncovered from the large number of genes composing the human genome. It is expected that the search engines and powerful analytical techniques developed in bioinformatics and rational drug design may help a lot in future target discovery. Rapid genome scale *in silico* disease-independent target discovery methods are desired in this endeavor.

Drug targets are a unique group of proteins bearing certain common characteristics. For example, a good target must possess substantial regulatory effect on a pathogenic pathway and the effect should be limited to that pathway so that normal processes of human body will not be disturbed. Therefore not surprisingly, a big portion of explored targets are receptors whose functions are highly important and specific to certain pathways. According to 2000 statistics, these receptors contribute to 45% of all current targets. Also, enzymes, whose activities are usually highly specific, make up for 28% targets [89]. It is easy to see that some protein classes are, obviously, more “successful” or exhibit better tractability in the drug discovery process. This shows that the drug-target likeness of a protein is related to its classes of structure and function, which are ultimately determined by its sequence. Statistical learning approaches have been applied to find the relationship between protein sequences and their functions [240-244], which lead to the hypothesis that the statistical learning methods may be equally applicable in prediction of drug-target like proteins, which is an efficient approach to pick out candidate targets from the huge number of proteins in the human genome.

The establishment of therapeutic target database has provided a useful resource for statistical model training. Various statistical approaches are evaluated in this work to examine whether statistical learning approaches can show satisfactory capacity in recognizing novel potential therapeutic targets by analysis of the sequences of explored therapeutic targets.

3.2 Statistical learning

With regard to this application, a supervised binary classification method is needed [245]. Specifically, examples are represented by a vector of a fixed number of attributes (denoted by $A_1, A_2 ... A_N$), also known as features, describing different characteristics of the examples. A given set of class labels (denoted by C), for example $\{+1,-1\}$ in binary classification, labels all the examples, where the examples labeled "+1" are called positive examples, and the examples labeled "-1" are called negative examples. Supervised learning methods, also known as classification methods, are to build a set of models that can correctly predict the class labels of a set of different examples (testing examples) based on the knowledge represented by a set of examples with known class labels (training examples).

3.2.1 Classification algorithms

Over the years, a variety of different classification algorithms have been developed by the machine learning community. Examples of such algorithms are decision tree bases [246-248], rule-based [249,250], probabilistic [251], neural network [252,253], genetic [254], instance based [255,256], and support vector machine [104,245]. Depending on the characteristics of the data sets being classified, certain algorithms tend to perform better than others. In recent years, algorithms based on the support vector machine and the k-nearest neighbors have been shown to produce reasonably good results for problems in which features are continuous. For this reason, we are mainly interested in these two algorithms. We also include the decision tree algorithm, which is the “classical” benchmark for classification algorithms and can be applied universally. These algorithms are described briefly below.

3.2.1.1 Decision tree

Decision trees are powerful and popular tools for classification and prediction [246-248]. The attractiveness of decision trees is due to the fact that, in contrast to neural networks and support vector machines, decision trees represent rules. Rules can readily be expressed so that human can understand them. They can even be directly used in a database access language like SQL so that records falling into a particular category may be retrieved.

There are a variety of algorithms for building decision trees that share the desirable quality of interpretability. A well known and frequently used decision tree algorithm over the years is ID3 (or its improved version C4.5 and its commercial counterpart See5/C5.0).

Decision tree is a classifier in the form of a tree structure (Figure 3.1), where each node is either a leaf node, which indicates the class labels of examples, or a decision node, which specifies some test to be carried out on a single attribute value and has one branch and sub-tree for each possible outcome of the test. A decision tree can be used to classify an example by starting at the root of the tree and moving through it until a leaf node, which provides the class label of the example.

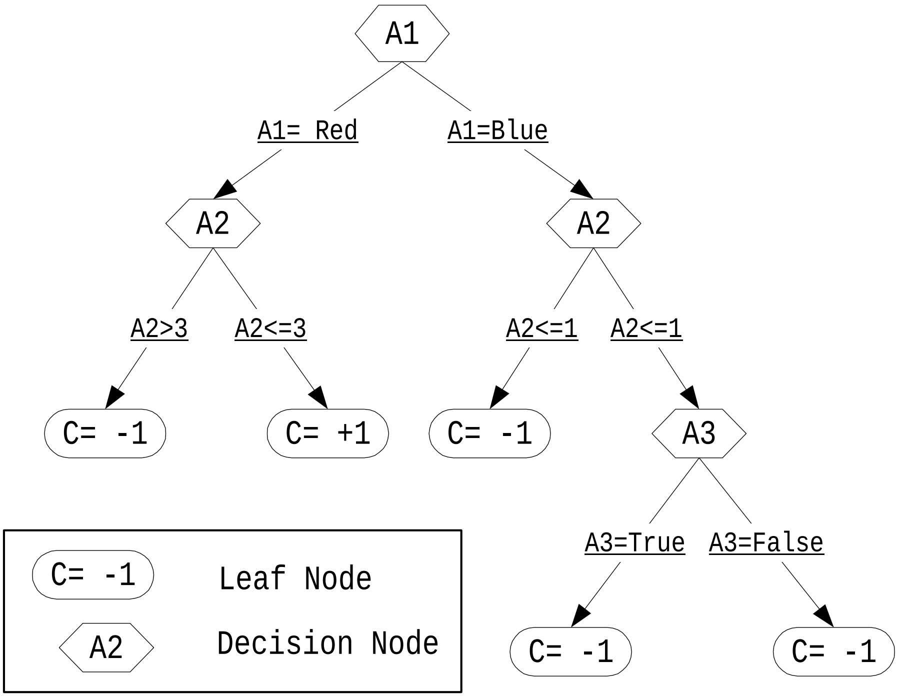


Figure 3.1: An example of decision trees.

Decision tree induction is a typical inductive approach to learn knowledge on classification. Most algorithms that have been developed for learning decision trees are variations on a core algorithm that employs a top-down, greedy search through the space of possible decision trees. ID3 is a widely used decision tree learning algorithm [257]. It uses fixed sets of attributes, and creates a decision tree to classify an example into a fixed set of class labels. At every step, if the remaining examples are all of the same class, it predicts that class, otherwise, it chooses the attribute with the highest "information gain" and creates a decision node based on that attribute to split the remaining training examples into one subset per discrete value of that attribute. It recursively does this until each leaf node contains only examples of one class, or all the attributes are used up.

The primary focus of the decision tree growing algorithm is to select which attribute to test at each decision node in the tree. The quantitative measure of the worth of an attribute used in ID3 is a statistical property called "information gain" that measures how well a given attribute separates the training examples according to their class labels.

In order to define information gain precisely, we need to define a measure commonly used in information theory, called entropy (denoted by E), which characterizes the impurity of an arbitrary collection of examples. Given a set S , in the binary classification setting, the entropy of set S is defined as:

$$E(S) = -p\log_2 p - n\log_2 n \qquad \text{Equation 3.6}$$

where p is the proportion of positive examples in S and n is the proportion of negative examples in S . In all calculations involving entropy we define $0\log_2 0$ to be 0. Notice that the entropy is 0 if all members of S belong to the same class, which is the stop criteria of tree splitting; and the entropy is 1 (at its maximum) when the collection contains equal numbers of positive and negative examples. If the collection contains unequal numbers of positive and negative examples, the entropy is between 0 and 1.

One interpretation of entropy from information theory is that it specifies the minimum number of bits of information needed to encode the classification of an arbitrary member of S (i.e., a member of S drawn at random with uniform probability) [258]. For example, if p is 1, a receiver knows the drawn example will be positive, so no message need be sent, and the entropy is 0. On the other hand, if p is 0.5, one bit is required to indicate whether the drawn example is positive or negative. If p is 0.8, then a collection of messages can be encoded using less than 1 bit (on average) per message by assigning shorter codes to collections of positive examples and longer codes to less likely negative examples.

Given entropy as a measure of the impurity in a collection of training examples, we can now define a measure of the effectiveness of an attribute in classifying the training data. The measure we will use, called information gain, is simply the expected reduction in entropy caused by partitioning the examples according to this attribute. More precisely, the information gain $G(S, A_i)$ of an attribute A_i, relative to a collection of examples S, is defined as

$$G(S, A_i) = E(S) - \sum_{v \in V(A_i)} \frac{|S_{A_i=v}|}{|S|} E(S_{A_i=v}) \qquad \text{Equation 3.7}$$

where $V(A_i)$ is the set of all possible discrete values for attribute A_i, and $S_{A_i=v}$ is the subset of S in which attribute A_i has the value v. Note the first term in the equation of $G(S, A_i)$ is just the entropy of the original collection S and the second term is the expected value of the entropy after S is partitioned using attribute A_i. The expected entropy described by this second term is simply the sum of the entropies of each subset $S_{A_i=v}$, weighted by the fraction of examples $\frac{|S_{A_i=v}|}{|S|}$ that belong to $S_{A_i=v}$. $G(S, A_i)$ is therefore the expected reduction in entropy caused by knowing the value of attribute A_i. Put another way, $G(S, A_i)$ is the information provided about the class labels, given the values of some attribute A_i. The value of $G(S, A_i)$ is the number of bits saved when encoding the class labels of an arbitrary member of S, by knowing the value of attribute A_i.

The measure G tends to favor those attributes with more possible discrete values. For example, a decision tree can be established to predict the disease of a patient using only one attribute: the case serial number. However, this decision tree would probably fail when a new patient with a new case serial number comes. Another measurement, Gain Ratio (R) is defined to avoid this bias, which can be calculated as follows.

$$R(S,A_i)=G(S,A_i)/IV(S,A_i) \text{ where} \qquad \text{Equation 3.8}$$

$$IV(S,A_i)=-\sum_{v\in V(A_i)}\frac{|S_{A_i=v}|}{|S|}\log\frac{|S_{A_i=v}|}{|S|} \qquad \text{Equation 3.9}$$

where $V(A_i)$ is the set of all possible discrete values for attribute A_i; $S_{A_i=v}$ is the subset of S for which attribute A_i has value v and the norm of a set $|\cdot|$ is defined as the number of elements in the set. $IV(S,A_i)$ can be interpreted as the total information content of the attribute A_i. Therefore, Gain ratio is the ratio of information gained that is pertinent to classification by branching on A_i.

The initial definition of decision tree is restricted to attributes that take on a discrete set of values. In other words, the attributes tested in the decision nodes of the tree must also be discretely valued. This restriction can easily be removed so that continuous-valued decision attributes can be incorporated into the learned tree [259]. This can be accomplished by dynamically defining new discrete-valued attributes that partition the continuous attribute value into a discrete set of intervals. For instance, for an attribute A_i that is continuous-valued, the ID3 algorithm can dynamically create a new Boolean attribute A'_i that is true if $A_i < c$ and false otherwise. The only question is how to select the best value for the threshold c.

Clearly, we would like to pick a threshold that produces the greatest information gain. By sorting the examples according to the continuous attribute A_i, then identifying adjacent examples that differ in their class labels, we can generate a set of candidate thresholds midway between the corresponding values of A_i. It can be proven that the value of c that maximizes information gain must always lie at such a boundary. These candidate thresholds can then be evaluated by computing the information gain associated. This dynamically created Boolean attribute can then compete with the other discrete-valued candidate attributes available for growing the decision tree.

Normally, the process of selecting a new attribute and partitioning the training examples is repeated for each non-terminal descendant node, using only the training examples associated with that node. This process continues for each new leaf node until either of two conditions is met:

Every attribute has already been included along this path through the tree, or

The training examples associated with this leaf node all have the same class labels (i.e., their entropy is zero).

In principle, the above decision tree algorithm can be used to grow each branch of the tree just deeply enough to perfectly classify the training examples. While this is sometimes a reasonable strategy, in many cases it can lead to difficulties when there is noise in the data, or when the number of training examples is too small to give a representative sample of the reality. In either of these cases, this simple algorithm will produce trees that over-fit the training examples. Therefore, after the tree is constructed, a pruning process is applied to gradually remove decision nodes that give the least improvements on accuracy and assign to these nodes the class label of the majority of remaining examples [260,261]. In this case, the prune level will be a free parameter to be optimized in the decision tree induction, which controls the complexity of the tree.

3.2.1.2 K-nearest neighbor

K-nearest neighbor (kNN) is a well known and widely used instance-based classification algorithm due to its conceptual simplicity, general applicability and efficiency [255,262-264]. It can be used as an initial tool to analyze a data set before proceeding to more sophisticated methods. It is also used to benchmark results of other classification methods. K-nearest neighbor is an algorithm that uses all available examples and classifies new instances based on a similarity measure.

The basic idea behind this classification paradigm is first to compute the similarity between a test example and all the examples in the training set, then to select the k most similar training examples, and finally to determine the class label of the test example based on the class labels of these k nearest neighbors.

One of the advantages of k-nearest neighbor method is that it is well suited for composite classes (classes consists of examples whose features have different characteristics for different subsets, or sub-classes) as its classification decision is based on a small neighborhood of similar examples.

Two steps are critical to the performance of the k-nearest neighbor. The first is the method used to determine the similarity between a test example and the examples in the training set and the second is the method used to determine the class label of the test example based on the class labels of the nearest neighbors. For data sets for which the objects are represented by multi-dimensional vectors, like our application, the approach that is commonly used to compute the similarity is using the Euclidean distance or any other norm based distance. We use the Euclidean distance as the similarity measurement in our experiments.

The other step to determine the class of the test example based on the classes of its k-nearest neighbors is to assign it to the majority class, i.e., the class to which most of the k-nearest neighbors belong. This decision function can be illustrated by Equation 3.10.

$$C = sign\left(\sum_{i=1}^{k} c(V_i) \right) \qquad \text{Equation 3.10}$$

where $c(V_i)$ is the class label of the i-th nearest neighbor V_i.

In the k-nearest neighbor classification, the value of k is needed. It has been found that $k < \sqrt{N}$ is a general criterion that should be met for good results, where N is the total number of training examples [255]. Therefore, the number of effective nearest neighbors k will be a free parameter in the k-nearest neighbor algorithm to be optimized according to test results.

3.2.1.3 Support vector machine

Although the basis of support vector machine had been laid in the 1960s, the idea of support vector machine was only officially proposed in 1995 by Vapnik and his co-workers [104,245]. Then, the research on its theoretical aspects and application aspects soared up because of the strong predictive power that this statistical learning algorithm had shown. It has been applied in a wide range of problems including text categorization [265,266], hand-written digit recognition [104], image classification and object detection [267,268], flood stage forecasting [269], micro-array gene expression data analysis [270], drug design [158], prediction of protein solvent accessibility [271], protein fold recognition [272], protein secondary structure prediction [273], prediction of protein-protein interaction [274]. These studies have demonstrated that SVM is consistently superior to other supervised learning methods [158,270,275].

Support vector machine (SVM) separates a given set of labeled training examples in a multi-dimensional space via a hyper-plane optimally positioned between the positive samples and negative samples. The test examples are then placed onto this multi-dimensional space to recognize which are positive and which are negative based on their relative positions to the hyper-plane. For most of real-world problems, the dataset can not be separated by this linear method. Special “kernels” are introduced in SVM to automatically conduct nonlinear mapping from the input space onto a high-dimensional feature space in which the training examples can be linearly separated. The optimal hyper-plane thus determined in the feature space corresponds to a nonlinear decision boundary in the input space.

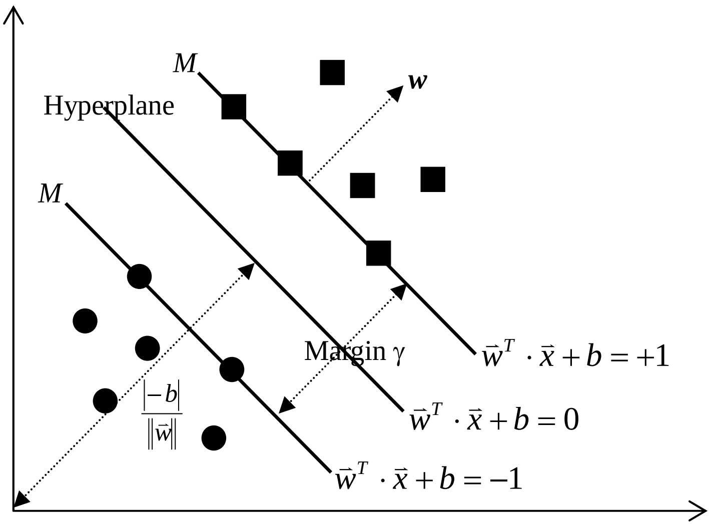

M
Hyperplane
M
w
Margin γ
$\frac{|-b|}{\|\vec{w}\|}$
$\vec{w}^T \cdot \vec{x} + b = +1$
$\vec{w}^T \cdot \vec{x} + b = 0$
$\vec{w}^T \cdot \vec{x} + b = -1$

Figure 3.2: Definition of Hyperplane and Margin. The circular dots and square dots represent samples of class -1 and class +1, respectively.

Let the training data of two separable classes, which contains *n* samples, be represented by $(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_n, y_n)$ $i = 1,2,\ldots,n.$ where $x_i \in R^N$ is a vector in an N dimensional space, and $y_i \in \{-1,+1\}$ indicates class label. Given a weight vector w and a bias b (figure 3.2), it is assumed that these examples can be separated by a hyperplane with a margin of 1:

$$w \cdot x_i + b \geq +1, \quad \text{for } y_i = +1 \qquad \text{Equation 3.11}$$

$$w \cdot x_i + b \leq -1, \quad \text{for } y_i = -1 \qquad \text{Equation 3.12}$$

where $w = (w_1, w_2, \ldots, w_n)^T$ is a vector of n elements.

Equation 4.11 and Equation 4.12 can be combined into a single inequality:

$$y_i(w \cdot x_i + b) \geq 1, \quad \text{for } i = 1,2,\ldots,n \qquad \text{Equation 3.13}$$

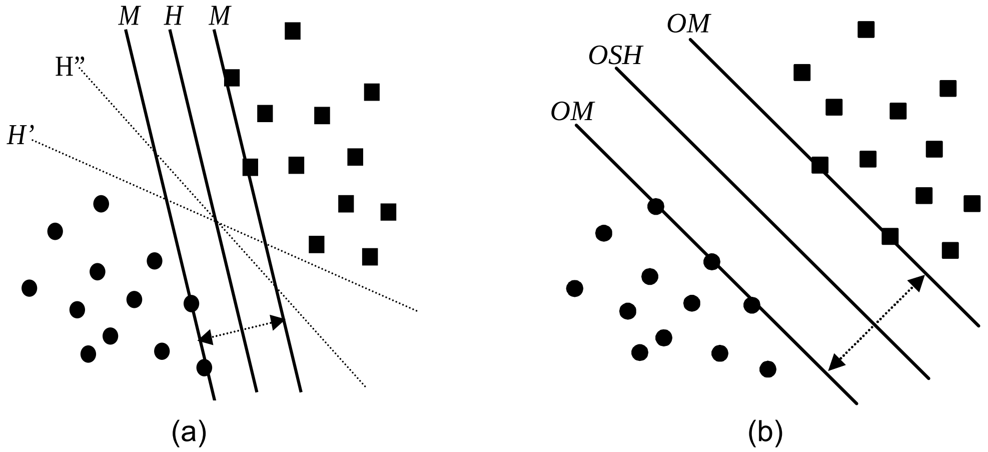


Figure 3.3: Available separating hyperplanes and Optimal Separating Hyperplane
(a) Available Hyperplanes H, H', H",…
(b) Unique Optimal Separation Hyperplane

As shown in Figure 3.3(a), there exist a number of separating hyperplanes for an identical group of training data. The objective of SVM is to determine the optimal weight w_o and optimal bias b_o such that the corresponding hyperplane separates the positive and negative training data with maximum margin, which is expected to produce the best generalization performance. This hyperplane (Figure 3.3(b)) is called the Optimal Separating Hyperplane (OSH).

The equation for an arbitrary hyperplane can be written as

$$w^T \cdot x_i + b = 0, \qquad \text{Equation 3.14}$$

and the width of the two corresponding margins is

$$\gamma(w,b) = \min_{\{x|y=+1\}} \frac{w^T \cdot x}{\|w\|} - \max_{\{x|y=-1\}} \frac{w^T \cdot x}{\|w\|} \qquad \text{Equation 3.15}$$

Given the constraint Equation 3.13, one obtains

$$\gamma_{\max} = \gamma(w_o, b_o) = \frac{2}{\|w\|} \qquad \text{Equation 3.16}$$

The OSH can therefore be obtained by minimizing the norm of $\|w\|$ under the inequality constraint Equation 3.13. The saddle point of the following Lagrangian gives solutions to the above optimization problem:

$$L(w,b,\alpha)=\frac{1}{2}w^{T}\cdot w-\sum_{i=1}^{n}\alpha_i[y_i(w^{T}\cdot x_i+b)-1] \qquad \text{Equation 3.17}$$

where $\alpha_i \geq 0$ are Lagrange multipliers. The solution to this Quadratic Programming (QP) problem requires that the gradient of $L(w,b,\alpha)$ with respect to w and b vanish, i.e., $\left.\frac{\partial L}{\partial w}\right|_{w=w_o}=0$ and $\left.\frac{\partial L}{\partial b}\right|_{b=b_o}=0$, which gives rise to the following conditions:

$$w_o=\sum_{i=1}^{n}\alpha_i y_i x_i \qquad \text{Equation 3.18}$$

$$\sum_{i=1}^{n}\alpha_i y_i=0 \qquad \text{Equation 3.19}$$

By substituting Equation 3.18 and Equation 3.19 into Equation 3.17, the QP problem becomes the maximization of the following expression:

$$L(\alpha)=\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}a_i\alpha_j y_i y_j(x_i\cdot x_j) \qquad \text{Equation 3.20}$$

under the constraints $\sum_{i=1}^{n}\alpha_i y_i=0$ and $\alpha_i\geq 0$, $i=1,2,...,n.$

The points located on the two optimal margins will have non-zero coefficients α_i among the solutions to Equation 3.20, and are called Support Vectors (SVs). The bias b_o can be calculated as follows:

$$b_o=-\frac{1}{2}\left\{\min_{\{x_i|y=+1\}}(w_o{}^{T}\cdot x_i)+\max_{\{x_i|y=-1\}}(w_o{}^{T}\cdot x_i)\right\} \qquad \text{Equation 3.21}$$

After determination of support vectors and bias, the decision function that separates the two classes can be written as

$$f(X) = sign[\sum_{i=1}^{n} \alpha_i y_i (x_i^T \cdot x) + b_o] = sign[\sum_{SV} \alpha_i y_i (x_i^T \cdot x) + b_o] \qquad \text{Equation 3.22}$$

Real-world problem are usually nonlinear in nature. The linear classification scheme described above is thus inapplicable to these problems. A nonlinear classification scheme can be introduced such that the original training data x in the input space X is projected into a high-dimensional feature space F via a Mercer kernel operator K [276,277] followed by the construction of OSH in the feature space (Figure 3.4).

Mathematically, the above set of equations is transformed into the following form by substituting the inner product in input space $(x_i \cdot x)$ to the inner product in feature space $K(x_i, x)$, where K is a symmetric positive definite function that satisfies Mercer's conditions:

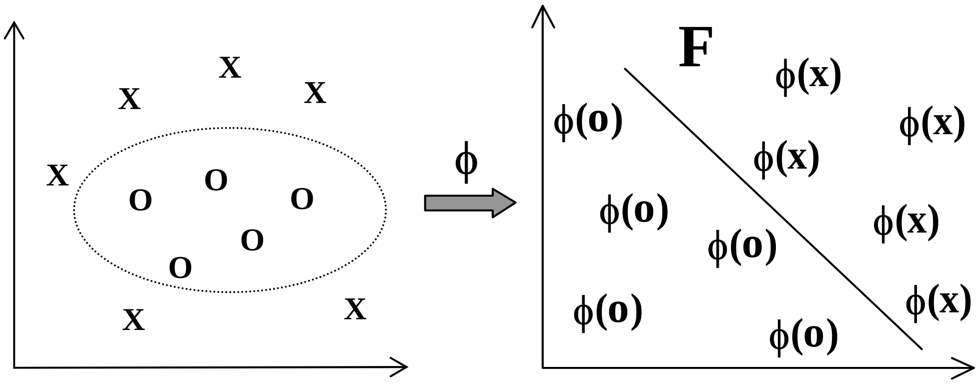


Figure 3.4: Project the training data nonlinearly into a higher-dimensional feature space and construct a hyperplane to separate positive and negative datasets with maximum margin there.

$$K(x,y)=\sum_{m=1}^{\infty}\alpha_m\Phi(x)\cdot\Phi(y),\qquad \alpha_m\ge 0,\qquad \text{Equation 3.24}$$

$$\iint K(x,y)g(x)g(y)dxdy>0,\qquad \int g^2(x)dx<\infty,\qquad \text{Equation 3.25}$$

In this case, the Kernel function can represent a legitimate inner product in a feature space:

$$K(x,y)=(\Phi(x)\cdot\Phi(y))\qquad \text{Equation 3.26}$$

where Φ is an implicit mapping function from the input space to the feature space F .

In F , the dual Lagrangian, given in Equation 3.20, becomes

$$L(\alpha)=\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}a_i\alpha_j y_i y_j K(x_i,x_j)\,,$$

s.t. $\sum_{i=1}^{n}\alpha_i y_i=0$ and $\alpha_i\ge 0$, $i=1,2,...,n.$ Equation 3.27

Thus the decision function changes to be

$$f(X)=sign[\sum_{SV}\alpha_i y_i K(x_i\,,x)+b_o]\qquad \text{Equation 3.28}$$

$$b_o=-\frac{1}{2}\left\{\min_{\{x_i|y=+1\}}(\sum_{SV}\alpha_i y_i K(x_i,x))+\max_{\{x_i|y=-1\}}(\sum_{SV}\alpha_i y_i K(x_i,x))\right\}\qquad \text{Equation 3.29}$$

Linear classification can also be integrated in the non-Linear classification framework. By defining $K(x,y)=(x^T\cdot y)$, the equations for non-linear classification immediately become equations for linear classification.

Usually, many candidate kernel functions can be used in a SVM, such as Polynomial kernel $K(x,y)=(1-x^T\cdot y)^{-d}$, Gaussian kernel $K(x,y)=\exp(-\frac{\|x-y\|^2}{2\sigma^2})$, and others [278], as well as their combinations such as the sum or tensor products of kernels. Among them, Gaussian kernel is the most popular one and we use Gaussian kernel in our classification. Usually, there are some parameters to be optimized in kernel function, such as the parameter σ in a Gaussian kernel.

3.2.2 Pre-processing for classification

While neglected by many machine learning research in the area of bioinformatics, pre-processing is regarded as a necessary step for serious real world data mining in the machine learning community. Here, several popular pre-processing techniques are explored.

3.2.2.1 Scaling

One of these widely used pre-processing techniques is normalization. Empirically, normalization will help to improve the prediction accuracy. In this work, Equation 3.30 is used for normalization, which is the approach adopted by LIBSVM [279], a support vector machine classification toolbox.

$$A_i^{new} = \frac{2*\left(A_i - \overline{A_i}\right)}{\max(A_i) - \min(A_i)} \qquad \text{Equation 3.30}$$

where A_i is the i-th feature, $\overline{A_i}$ is the average value of A_i among all the examples. After this process, all the features will be in the region of [-1,+1].

Another very important issue of pre-processing is dimensionality reduction. Bellman (1961) [280] first proposed the term “curse of dimensionality”, which refers to the exponential growth of hyper-volume as a function of dimensionality. Most statistical learning models can be thought of mappings from an input space to an output space. Thus, loosely speaking, a statistical learning model needs to somehow "monitor", cover or represent every part of its input space in order to know how that part of the space should be mapped. Covering the input space takes resources, and, in the most general case, the amount of resources needed is proportional to the hyper-volume of the input space. The exact formulation of "resources" and "part of the input space" depends on the type of the model and should probably be based on the concepts of information theory and differential geometry. The curse of dimensionality causes a model with lots of irrelevant inputs to behave relatively badly. When the dimension of the input space is high, the model uses almost all its resources to represent irrelevant portions of the space. Even if a statistical learning algorithm is able to focus on important portions of the input space, the higher the dimensionality of the input space is, the more examples are needed to make a reasonable sampling.

Dimensionality reduction has been the focus of pre-processing research for quite some time [281-284]. Conventional approaches have been developed to select the “best” subset of the original features that can best describe the problem. Recent advances in dimensionality reduction incorporate methods that will construct a new set of features from the original features to minimize the information loss when discarding any of the original features. Such methods are represented by Principal Component Analysis (PCA) and Independent Component Analysis (ICA).

3.2.2.2 Principal component analysis

Principal Component Analysis (PCA) is widely used in signal processing, statistics, and neural computing [285,286]. In some application areas, this is also called the (discrete) Karhunen-Loève transform, or the Hotelling transform.

The basic idea in PCA is to find the components $s_1, s_2 ......s_n$ so that they explain the maximum amount of variances in the input space by n linearly transformed components. PCA can be defined in an intuitive way using a recursive formulation. Define the direction of the first principal component, say w_1, by

$$w_1 = \arg\max_{\|w\|=1} E\left\{\left(w^T \cdot x\right)^2\right\} \qquad \text{Equation 3.31}$$

where w_1 is of the same dimension as the example vector x. Thus the first principal component is the projection on the direction in which the variance of the projection is maximized. Having determined the first $k-1$ principal components, the k-th principal component is determined as the principal component of the residual:

$$w_1 = \arg\max_{\|w\|=1} E\left\{\left[w^T\left(x - \sum_{i=1}^{k-1} w_i w_i^T x\right)\right]^2\right\} \qquad \text{Equation 3.32}$$

The principal components are then given by Equation 3.33.

$$s_i = w_i^T x \qquad \text{Equation 3.33}$$

In practice, the computation of the w_i can be simply accomplished using the (example) covariance matrix $E\{x^T x\} = C$. The w_i are the eigenvectors of C that correspond to the k largest eigenvalues of C.

By choosing the n first components, PCA is used to reduce the dimensionality of the input data. One usually chooses $n < N$ (N is the dimension of the original feature vector). It can be proven that the representation given by PCA is an optimal linear dimension reduction technique in the mean-square sense [285]. By this means, noise may be reduced, as the data not contained in the n first components may be mostly due to noise. A simple illustration of PCA is found in Fig. 3.5, in which the first principal component of a two-dimensional data set is shown.

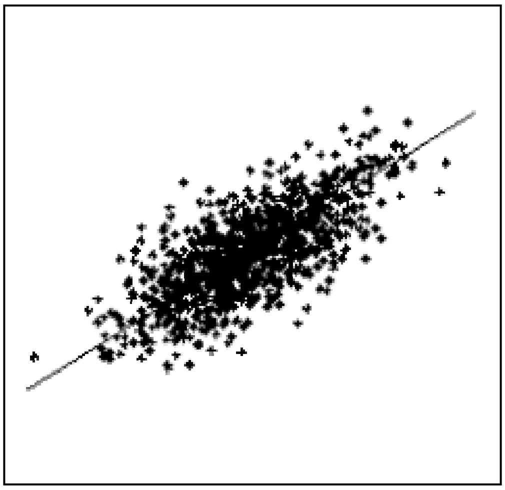


Figure 3.5: Principal component analysis of a two-dimensional data set. The line shown is the direction of the first principal component, which gives an optimal (in the mean-square sense) linear reduction of dimension from 2 to 1 dimension.

3.2.2.3 Independent component analysis

Independent component analysis (ICA) [287,288] is a statistical and computational technique for revealing hidden factors that underlie sets of random variables, measurements, or signals.

Assume that we observe N linear mixtures $x\,(A_1, A_2 ... A_N)$ (the features of an example) of N independent components $s_1, s_2 ... s_N$,

$$A_j = m_{j1}s_1 + m_{j2}s_2 + ... + m_{jN}s_N \qquad \text{Equation 3.34}$$

In the ICA model, it is assumed that each mixture A_j as well as each independent component s_k is a random variable. The observed values $A_j\,(x)$, e.g., the j-th feature of all the examples, are then a sample of this random variable.

Without loss of generality, it is assumed that both the mixture variables and the independent components have zero mean: If this is not true, then the observable variables A_j can always be centered by subtracting the sample mean, which makes the model zero-mean.

It is convenient to use vector-matrix notation instead of the sums like in the previous equation. The above mixing model is written as

$$x = Ms \qquad \text{Equation 3.35}$$

where x is the random vector whose elements are the mixtures $A_1, A_2 ... A_N$; s is the source vector with elements $s_1, s_2 ... s_N$; and M is the mixing matrix with elements m_{ij}.

The statistical model in Equation 3.36 is called independent component analysis, or ICA model. The ICA model is a generative model, which means that it describes how the observed data are generated by a process of mixing the components s_j. The independent components are latent variables, meaning that they cannot be directly observed. Also the mixing matrix is assumed to be unknown. All we observe is the random vector x, and we must estimate both M and s using it. This must be done under assumptions as general as possible.

$$s = Wx \qquad \text{Equation 3.36}$$

The starting point for ICA is the very simple assumption that the components s_j are statistically independent. It is assumed that the independent component must have non-gaussian distributions. However, in the basic model we do not assume that these distributions are known. For simplicity, it is also assumed that the unknown mixing matrix M is square. Then, after estimating the matrix M, its inverse W can be computed.

The non-gaussianity can be measured in different ways. Suppose y is a random variable, classical measure of non-gaussianity is kurtosis which is defined by Equation 3.37:

$$kurt(y) = E\{y^4\} - 3(E\{y^2\})^2 \qquad \text{Equation 3.37}$$

A second very important measure of non-gaussianity is given by negentropy. Negentropy is based on the information-theoretic quantity of (differential) entropy. Entropy E is defined for a discrete random variable y as

$$E(y) = -\sum_i P(y = a_i)\log_2 P(y = a_i) \qquad \text{Equation 3.38}$$

where the a_i are the possible values of y. This very well-known definition can be generalized for continuous-valued random variables, in which case it is often called differential entropy. The differential entropy E of a random variable y with density $f(y)$ is defined as [289]:

$$E(y) = -\int f(y)\log_2 f(y)dy \qquad \text{Equation 3.39}$$

A fundamental fact in information theory is that a gaussian variable has the largest entropy among all random variables of equal variance. To obtain a measure of non-gaussianity that is zero for a gaussian variable and always nonnegative, one often uses a slightly modified version of the definition of differential entropy, called negentropy. Negentropy J is defined as follows:

$$J = E(y_{Gauss}) - E(y) \qquad \text{Equation 3.40}$$

where y_{Gauss} is a Gaussian random variable sharing the same covariance matrix as

y. However, negentropy is computationally very difficult. Estimating negentropy using the definition would require an estimate (possibly nonparametric) of the probability density function. Therefore, simpler approximations of negentropy are very useful, for example Equation 3.41 [290].

$$J \propto \sum_{i=1}^{p} k_i \left[E\{G_i(y)\} - E\{G_i(v)\}\right]^2 \qquad \text{Equation 3.41}$$

where k_i are some positive constants, and v is a Gaussian variable of zero mean and unit variance. The variable y is assumed to be of zero mean and unit variance, and the functions G_i are some non-quadratic functions [290]. The following choices of G have been proved useful:

$$G(y) = \frac{1}{a} \log \cosh ay \quad \text{where} \quad 1 \le a \le 2 \quad \text{and} \qquad \text{Equation 3.42}$$

$$G(y) = -\exp(-y^2 / 2) \qquad \text{Equation 3.43}$$

By choosing an appropriate G the classical kurtosis measure of non-gaussianity can be unified in this framework, i.e. $G(y) = y^4$.

Equation 3.41 is used as the contrast function (non-gaussianity measure) to be maximized in order to find the first independent component:

$$w_1 = \arg \max_{\|w\|=1} J(w^T \cdot x) \qquad \text{Equation 3.44}$$

Thus the first independent component is the projection on the direction in which the non-gaussianity (measured by an approximation of negentropy) of the projection is maximized. To estimate several independent components, several units with weight vectors $w_1, w_2 ......w_N$ need to maximized together using Equation 3.44. To prevent different vectors from converging to the same maxima, the outputs are decorrelated after every iteration of the optimization process. The independent components are then given by Equation 3.45.

$$s_i = w_i^T x \qquad \text{Equation 3.45}$$

ICA can be seen as an extension to principal component analysis and factor analysis. ICA is a powerful technique capable of finding the underlying factors or sources. It is interesting to note that ICA makes explicit connection to projection pursuit. Projection pursuit is a technique developed in statistics for finding "interesting" projections of multidimensional data [291-294]. Such projections can then be used for decision making, and for such purposes as density estimation and regression. In basic (one dimensional) projection pursuit, we try to find a direction such that the projections of the data in this directions have an interesting distribution, i.e., display some structure. It has been argued by Huber [293] and by Jones and Sibson [294] that the Gaussian distribution is the least interesting one, and that the most interesting directions are those that show the least Gaussian distribution. This is exactly how to estimate the ICA model. The usefulness of finding such projections can be seen in Fig. 4.6, where the projection on the projection pursuit direction, which is horizontal, clearly shows the clustered structure of the data. The projection on the first principal component (vertical), on the other hand, fails to show this structure.

So far, there are no reported indicators that can show whether a statistical learning algorithm or a pre-processing procedure is suitable for a certain application. The performances of different classifiers and the effect of scaling, PCA and ICA on them are evaluated and compared below.

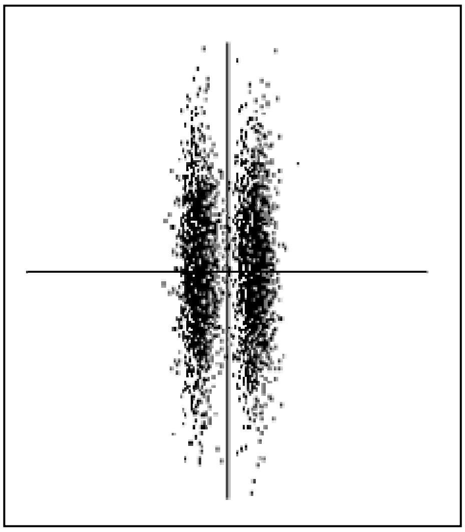


Figure 3.6: An illustration of projection pursuit and the "interestingness" of non-gaussian projections.

3.3 Problem definition

The input data to the above statistical learning methods are a set of examples (training set) and the class labels attached to these examples (training labels). The accuracy of the output models can be evaluated by comparing the model predictions on a set of different examples (testing set) with prior knowledge of the class labels of these examples (testing labels). In the majority of cases, there are some free parameters in statistical learning algorithms which control the generation of models, such as the prune level in decision tree induction and the kernel parameter in support vector machine. The "best" model is usually selected from a series of models generated using different parameter sets according to certain model accuracy measurements. This leads to the problem that the selection of the "best" model is not independent of the testing set. Therefore, the best model may "over-fit" to a particular testing set. In order to have an unbiased estimation of the prediction accuracy of the selected models, an independent evaluation data set with known class labels, in addition to the training and testing data sets, is needed. The unbiased estimation of the prediction accuracy of the selected model can therefore be measured on the independent evaluation data set. The procedures used to construct the data sets and the measurements used to evaluate the model accuracy are detailed below.

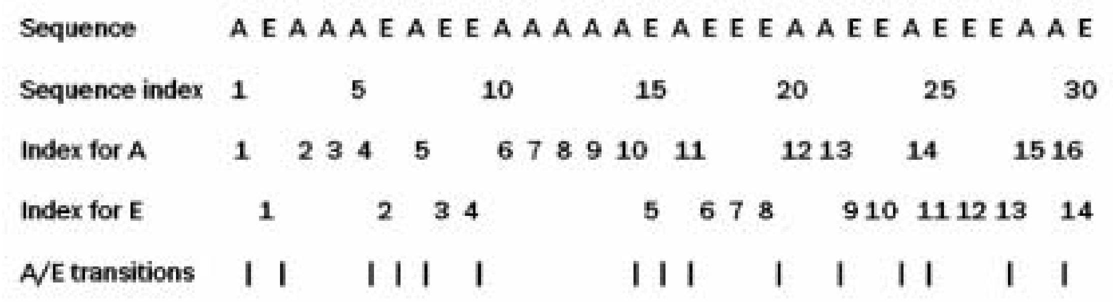
3.3.1 Description of data

All drug-target protein sequences used in this study are retrieved from the SWISS-PROT database release 40 [295]. A total of 339 human target proteins are obtained by an automated sequence retrieval program linked to therapeutic target database. These proteins are labeled “+1”. The non-drug-target proteins are composed of 1620 other proteins randomly selected. These proteins are given the label “-1”. As shown in Table 3.1, the training set is comprised of 235 positive examples and 1131 negative examples; the testing set is comprised of 64 positive examples and 301 negative examples; and the independent evaluation set is comprised of 40 positive examples and 188 negative examples. A perl program is written to randomly distribute these proteins while maximizing the protein family diversity [296] in each set. The proportions of positive examples in all the three data sets are close to 17%, which is the expected proportion of targets in human genome (5,000 / 30,000) [89].

Table 3.1: Composition of training, testing and independent evaluation data sets.

Data Set	No. of positive examples	No. of negative examples
Training set	235	1131
Testing set	64	301
Independent evaluation set	40	188

Training and testing of the statistical learning model requires each example be represented as a feature vector consisting of a fixed number of real valued components. These feature vectors are assembled from the encoded representations of tabulated residue properties including amino acid composition, hydrophobicity, normalized Van der Waals volume, polarity, polarizability, charge, surface tension, secondary structure and solvent accessibility for each residue in the sequence [271-274,297]. Three types of descriptors, composition (C), transition (T) and distribution (D), are used to describe global composition of each of these properties[298]. C is the number of amino acids of a particular property (such as hydrophobicity) divided by the total number of amino acids. T characterizes the percent frequency with which amino acids of a particular property is followed by amino acids of a different property. D measures the chain length within which the first, 25, 50, 75 and 100% of the amino acids of a particular property is located respectively.



Sequence	A	E	A	A	A	E	A	E	E	A	A	A	A	A	E	A	E	E	E	A	A	E	E	A	E	E	E	A	A	E
Sequence index	1				5					10					15					20					25					30
Index for A	1		2	3	4		5			6	7	8	9	10		11				12	13			14				15	16	
Index for E		1				2		3	4						5		6	7	8			9	10		11	12	13			14
A/E transitions		|	|			|	|	|		|					|	|	|			|		|		|	|			|		|

Figure 3.7: A hypothetical sequence for illustration of derivation of the feature vector of a protein.

A hypothetical protein sequence AEAAAEAEEAAAAAEAEEEAAEEAEEEAAE, as shown in Figure 3, has 16 alanines ($n_1 = 16$) and 14 glutamic acids ($n_2 = 20$). The composition for these two amino acids are $n_1 / (n_1 + n_2) = 53.33\%$ and $n_2 / (n_1 + n_2) = 46.67\%$ respectively. There are 15 transitions from A to E or from E to A in this sequence and the percent frequency of these transitions is (15/29)=51.72%. The first, 25, 50, 75 and 100% of "A" are located within the first 1, 5, 12, 20 and 29 residues, respectively. The D descriptor for "A" is thus 1/30=3.33%, 5/30=16.67%, 12/30=40.0%, 20/30=66.67%, 29/30=96.67%. Likewise, the D descriptor for "E" is 6.67%, 26.67%, 60.0%, 76.67%, 100.0%. Overall, the amino acid composition descriptors for this sequence are C=(53.33, 46.67), T=(51.72) and D=(3.33, 16.67, 40.0, 66.67, 96.67, 6.67, 26.67, 60.0, 76.67, 100.0), respectively.

Descriptors for other properties can be computed by a similar procedure and all the descriptors are combined to form the feature vector. In most studies, amino acids are divided into three classes for each property and thus the three types of descriptors for each property consist of 21 elements: three for C, three for T and 15 for D. Taking hydrophobicity as an example, the steps of feature construction are as follows: first, 20 aminoacid residues can be devided into three groups according to their hydrophobicity (denoted as H, M, L). The original protein sequence can thus be translated to a pseduosequence of H, M and L according this grouping. Then, the composition of H, M and L (3 elements), transitions of H/M, H/L and M/L (3 elements), and distributions of H, M, and L (15 elements) can be computed. They add up to 21 elements. The physicochemical properties of amino acid residues are can be found in KEGG database. The secondary structure and solvent accessibility properties of each residue are predicted by the PHD program. Details of the formula can be found in the respective publications and references therein[271-274,297,298]. These properties and their respective dimensionality are given in Table 3.2. All the features are continuous and the total dimensionality of the vector is 188 (all the vectors are column vectors by default in this work).

Table 3.2: Feature vector composition.

Feature description	No. of dimensions
Amino acids composition	20
Hydrophobicity	21
Normalized Van der Waals volume	21
Polarity	21
Polarizability	21
Charge	21
Surface tension	21
Secondary structure	21
Solvent accessibility	21

Total	188

3.3.2 Measurements of prediction accuracy

Before discussing the result of the drug-target like protein prediction, it is important to explain the measurements used to evaluate the effectiveness of a classification algorithm. First, confusion Matrix is the most simple and informative way to analyze the behavior of a classifier. It contains information about both actual and predicted classifications. Table 3.3 shows a common confusion matrix for binary classification, in which “a” is the number of correctly classified positive examples, called true positive (TP); “d” is the number of correctly classified negative examples, called true negative (TN); “b” is the number of incorrect predictions of the positive class, called false positive (FP); “c” is the number of incorrect predictions of the negative class, called false negative (FN).

Table 3.3: An example of confusion matrix in binary classification

		Actual	
		Positive	Negative
Predicted	Positive	a (TP)	b (FP)
	Negative	c (FN)	d (TN)

Based on the confusion matrix, accuracy measurements of a classifier can be calculated by the following quantities.

1. Overall accuracy, A for short, is the proportion of the total number of

predictions that are correct. It is calculated using Equation 3.1.

$$A = (a+d)/(a+b+c+d) \qquad \text{Equation 3.1}$$

2. Precision, P for short, is the proportion of the predicted positive examples that are correct. It is calculated using Equation 3.2.

$$P = a/(a+b) \qquad \text{Equation 3.2}$$

3. Recall, R for short, is the proportion of the real positive examples that are correctly predicted. It is calculated using Equation 3.3.

$$R = a/(a+c) \qquad \text{Equation 3.3}$$

4. F value, F for short, is designed to have a balanced evaluation on both precision and recall. It is the geometric mean of R and P, as in Equation 3.4.

$$F = 2*P*R/(P+R) \qquad \text{Equation 3.4}$$

The most commonly used measurement for evaluating a classifier is the overall accuracy. However, it is usually found that positive examples are more interesting. For example, in our case, finding a drug-target like protein is of more interest to pharmaceutical companies than finding a protein that is unlikely to be a drug target. Under this situation, the precision and recall, which reflects the prediction accuracy on positive examples, are important measurements. Precision gives the idea about how much confidence we have when the model gives a positive prediction and recall measures how many real positive examples are predicted by this model. The geometric average of precision and recall, F value, is a balanced view of these two factors.

The data sets in this application are not balanced. The number of negative examples is around five times bigger than that of positive examples. Therefore, the prediction accuracy on negative examples will have five times more influence on the overall accuracy. Inspired by the F value, we devised a new measure “balanced precision” in this work, which is the geometric mean of the precision on positive examples and the precision on negative examples. Balanced precision is denoted by G and can by calculated by Equation 3.5:

$$G = 2 * RN * R/(RN + R) \quad \text{where} \quad RN = d/(d + b) \qquad \text{Equation 3.5}$$

In this work, the statistical models are selected according to three criteria, A, F and G, respectively.

As discussed above, the free parameters controlling the generation of the statistical models needs to be optimized. In this work, a series of models are generated using different parameters. We define the term “best A” of a learning algorithm as the A measured on an independent evaluation set using the same parameter set that maximizes A on the testing data set. The “best F” and “best G” are defined similarly as the measurements made on the independent evaluation set using the same parameter sets that maximize them on the testing data set respectively. Note, under this definition, the “best” measurements might not be the highest values achieved on the independent evaluation set.

Furthermore, we consider one algorithm to be better than another if and only if this algorithm wins two or more times when its best A, best F and best G are compared to those of the other algorithm respectively. Also, we say one model is “better” than another if and only if two or more of its three measurements (A, F

and G) are better than those of the other.

3.4 Prediction of drug-target like proteins

In order to evaluate different classification and pre-processing techniques, an efficient tool to implement different algorithms is needed. The matrix operation support provided by MatLab [299] makes the representation of numerical data and implementation of the different algorithms much easier. Therefore, we choose MatLab as our platform of computation.

With the help of standard MatLab matrix functions and standard toolboxes, i.e. statistics toolbox and optimization toolbox, we implemented the algorithms for scaling, PCA, decision tree, k-nearest neighbor, and support vector machine. An ICA package for MatLab, FastICA 2.1, was used for ICA analysis, which is developed by Jarmo Hurri et.al. [300] and can be downloaded freely from http://www.cis.hut.fi/projects/ica/fastica/.

The decision tree algorithm was implemented with the information gain branching criterion. This is because the gain ratio criterion is mainly designed to deal with attributes with many discrete values. In our case, all the attributes are continuous where a threshold is chosen to bisect all the possible values, which is equivalent to all the attributes having two discrete values. Therefore, it is not necessary to use the gain ratio branching criterion. And our later study also confirmed that the gain ratio criterion performed no better than the information gain criterion.

The support vector machine algorithm was implemented with a Gaussian kernel, $K(x,y)=\exp(-\frac{\left\|x-y\right\|^2}{2\sigma^2})$. This is because the Gaussian kernel always performs better than others in our previous study of protein function classification [275,301].

In the ICA analysis, the nonlinearity function used in the non-gaussianity measurement was $G(u)=u^3$, which is the default choice of the FastICA package.

All the programs in this work were developed using MatLab R13 licensed from NUS and executed on a Dell Optiplex GX240 computer with one 2.4GHz Intel Pentium IV CPU and 512M memory.

3.4.1 Decision tree prediction

After training with the training data set without any pre-processing, the decision tree that perfectly classifies the training set is pruned from level 0 (original tree) to the maximal level (only the root is kept). By the definitions in section 3.3.2, the best A, F, and G achieved by the decision trees pruned in different levels are 85.09%, 54.05% and 68.40% respectively, which are shown in Figure 3.7.

The pre-process of scaling does not affect the tree induction because the decision tree induction algorithm only cares about the relative order of the attribute values when selecting the thresholds to bisect the continuous attributes. Therefore, the scaled data sets give the same tree and the same performance.

In order to evaluate the effect of the PCA dimensionality reduction, the first n principal components of the original data sets are selected to construct a new set of training, testing and independent evaluation data and the decision tree algorithm is evaluated on the new set of data with reduced dimensionality. Here, n is scanned from 1 to the maximal number of principle components with an interval of 10. The best A, best F, and best G achieved using different numbers of principal components are plotted in Figure 3.8. Overall, The best A, best F, and best G achieved, as shown in Table 3.5, are 79.39%, 43.37% and 46.62% respectively which is worse than those of the original or scaled data sets.

Similar to PCA, the evaluation of the effect of ICA dimensionality reduction on decision tree induction is conducted. The first n independent components of the original data sets were first estimated to construct a new set of training, testing and independent evaluation data. The decision tree algorithm is then evaluated on the new set of data with reduced dimensionality. The best A, best F, and best G achieved using different numbers of independent components are plotted in Figure 4.9. As shown in Table 4.6, the best A, best F, and best G achieved are 81.58%, 22.78% and 35.25% respectively, which are also inferior to those of the original or scaled data sets.

In summary, the decision tree algorithm works better on original or scaled data sets. The best A, best F, and best G achieved are 85.09%, 54.05% and 68.40% respectively.

3.5.2 K-nearest neighbor prediction

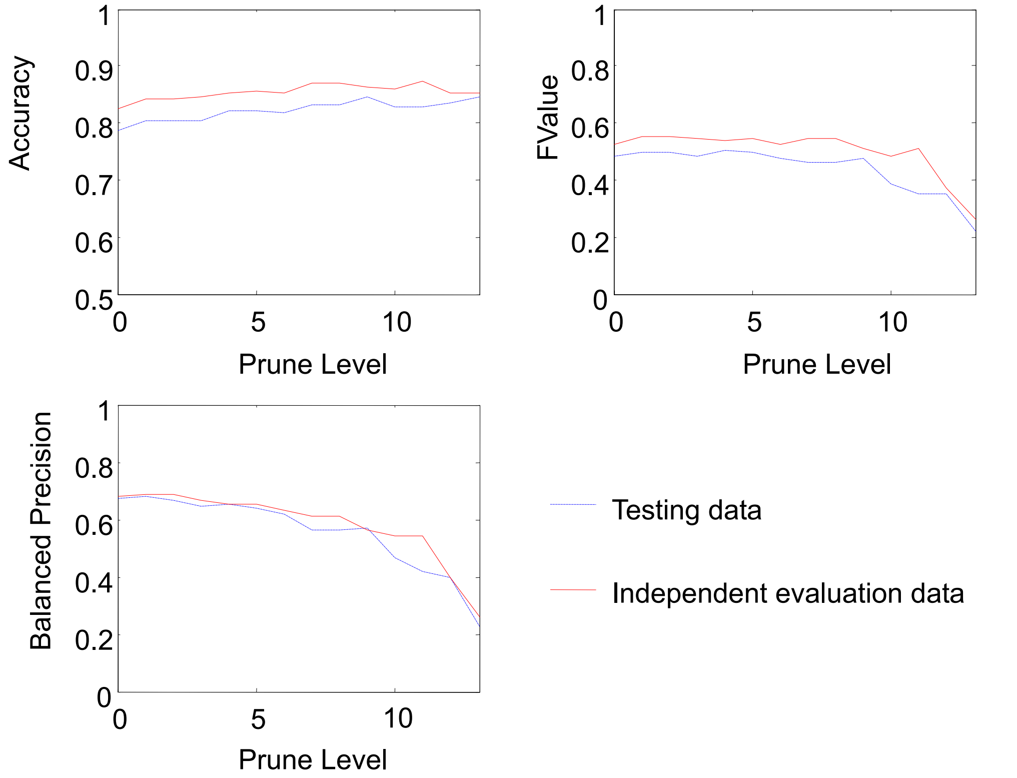


Figure 3.7: Decision tree prediction of drug-target like proteins on original datasets.

Table 3.4: Summary of the decision tree performance on original or scaled data sets. The maximum prune level of the complete tree is 13.

Measure Optimized	Prune Level	Overall Accuracy	F Value	Balanced Precision
Overall Accuracy	13	0.8509	0.2609	0.2609
F Value	4	0.8509	0.5405	0.6493
Balanced Precision	1	0.8421	0.5500	0.6840

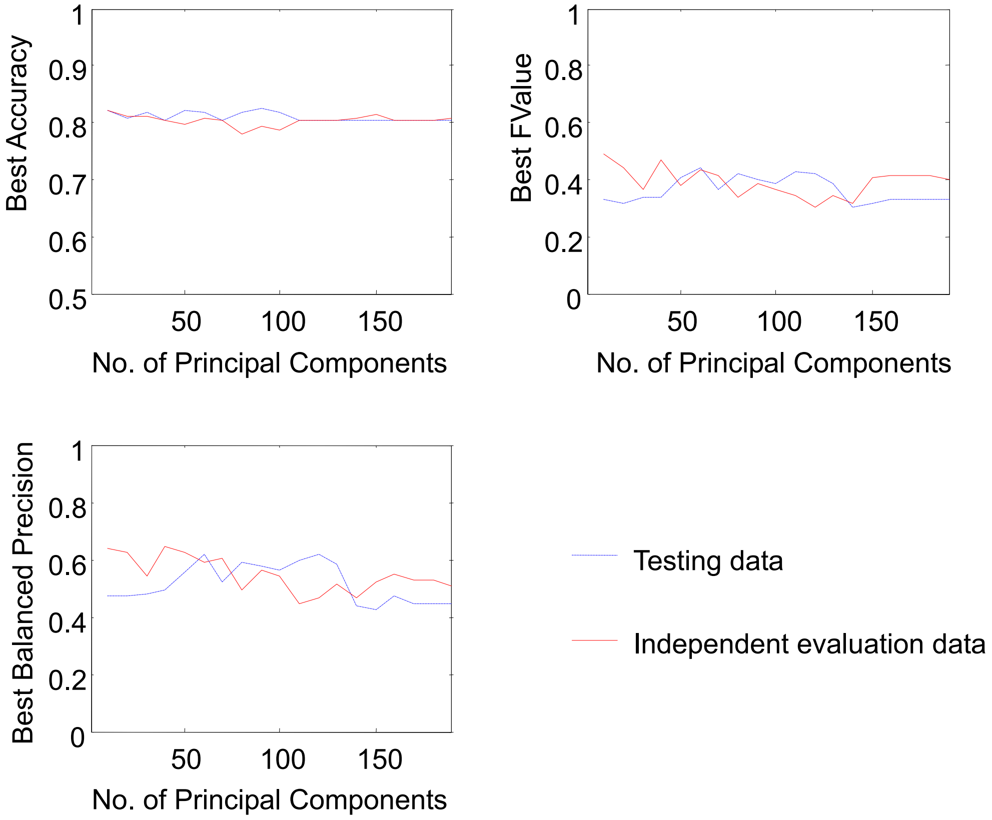


Figure 3.8: Decision tree prediction of drug-target like proteins after PCA dimensionality reduction.

Table 3.5: Summary of the decision tree performance on PCA processed data sets.

Measure Optimized	No. of Principal Components	Overall Accuracy	F Value	Balanced Precision
Overall Accuracy	90	0.7939	0.3864	0.5633
F Value	60	0.8070	0.4337	0.5925
Balanced Precision	120	0.3023	0.3291	0.4662

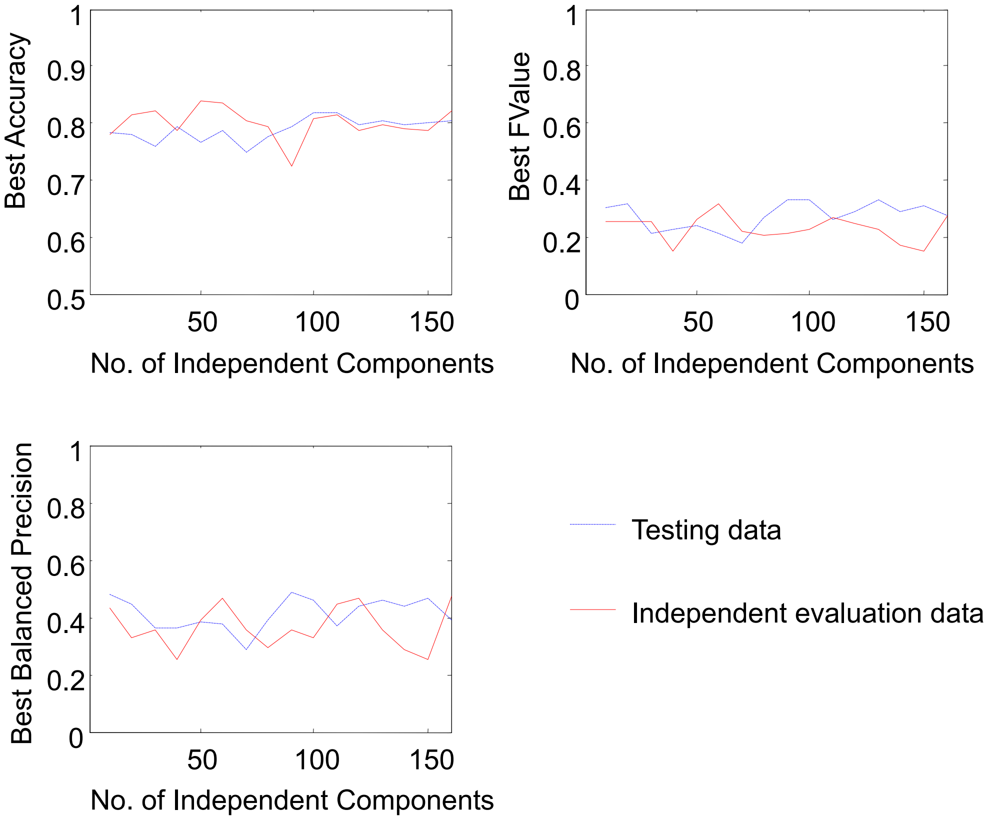


Figure 3.9: Decision tree prediction of drug-target like proteins after ICA dimensionality reduction.

Table 3.6: Summary of the decision tree performance on ICA processed data sets.

Measure Optimized	No. of Independent Components	Overall Accuracy	F Value	Balanced Precision
Overall Accuracy	110	0.8158	0.2667	0.4422
F Value	130	0.7982	0.2278	0.3550
Balanced Precision	90	0.7237	0.2143	0.3525

On the original and scaled data sets, the parameter k used in kNN is scanned in the range of 1…37. The number 37 is calculated as the square root of the number of training examples. The kNN algorithm performs slightly better on original data sets. The best A, best F, and best G on original data sets are 83.77%, 50.00% and 69.64% respectively, which is comparable to those of decision trees. The classification results on original data sets are plotted in Figure 3.10 and summarized in Table 3.7.

The effect of PCA dimensionality reduction on kNN classification is evaluated similarly as in the last section. The first n principle components of the original data sets are estimated to construct a new set of training, testing and independent evaluation data and the kNN algorithm is evaluated on this new set of data with reduced dimensionality. The number of principal components n is scanned from 1 to the maximum with an interval of 10. The best A, best F, and best G achieved using different numbers of principal components are 83.77%, 56.84% and 75.30% respectively. These values are better than those of the original data sets. These results are plotted in Figure 3.11 and summarized in Table 3.8. Also, it is interesting to see the best models are built when only a small number of principal components, i.e. less than one third of the total dimensions, are used in the training, testing and independent evaluation data sets.

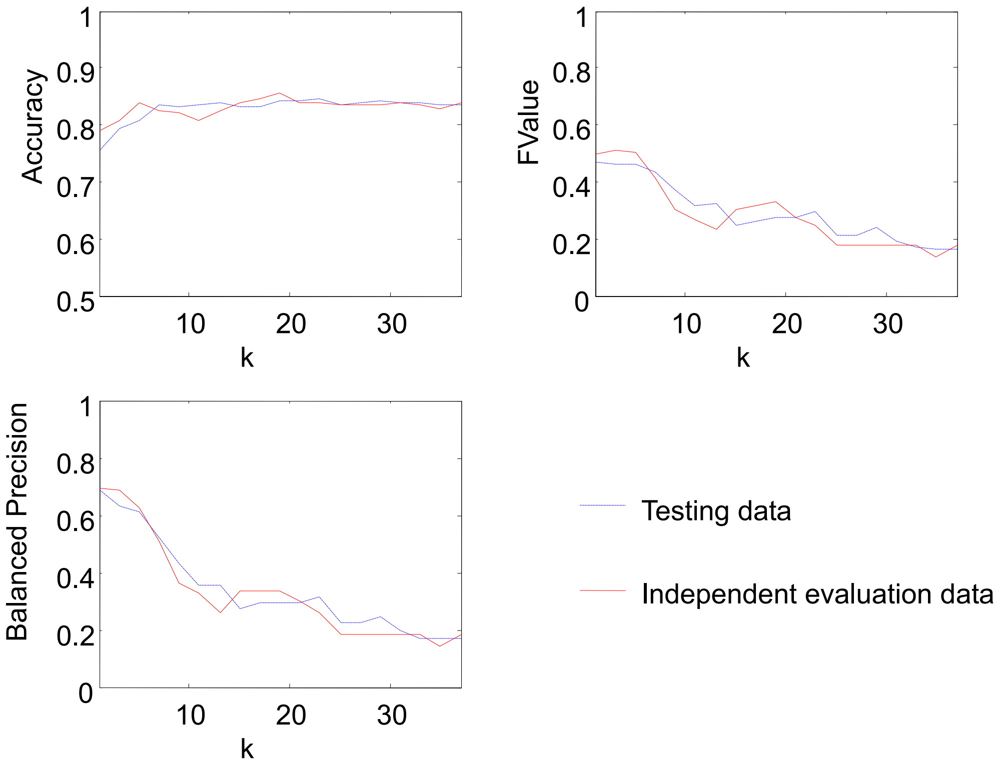


Figure 3.10: K-nearest neighbor prediction of drug-target like proteins using original data sets.

Table 3.7: Summary of the kNN performance on original data sets.

Measure Optimized	k	Overall Accuracy	F Value	Balanced Precision
Overall Accuracy	23	0.8377	0.2449	0.2603
F Value	1	0.7895	0.5000	0.6964
Balanced Precision	1	0.7895	0.5000	0.6964

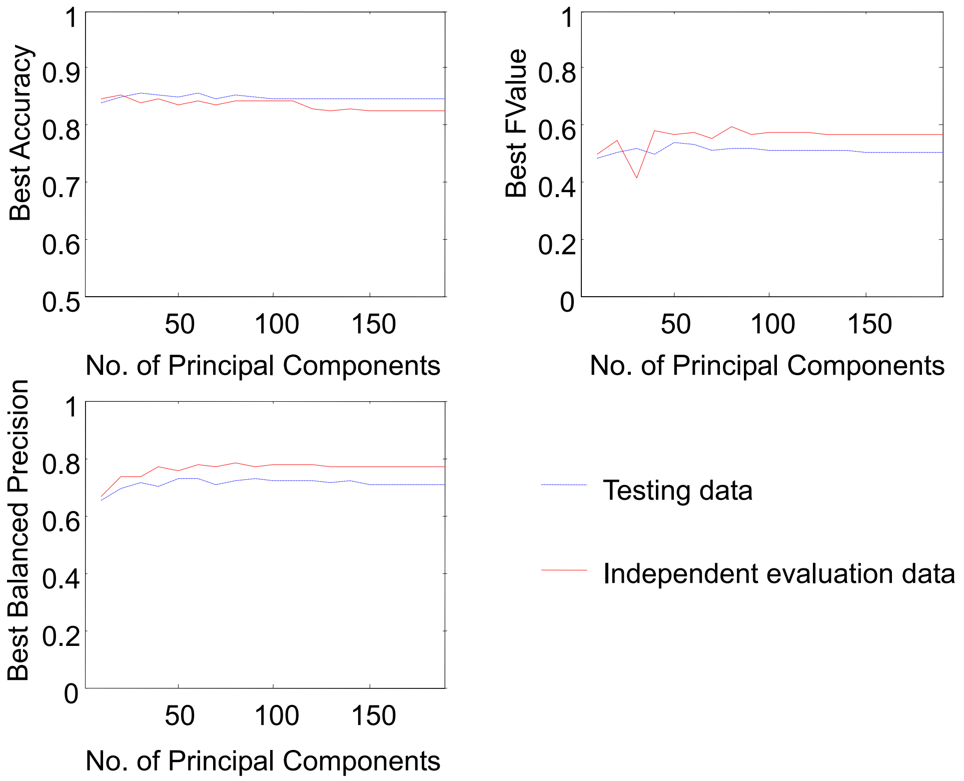


Figure 3.11: K-nearest neighbor prediction of drug-target like proteins after PCA dimensionality reduction.

Table 3.8: Summary of the kNN performance on PCA processed data sets.

Measure Optimized	No. of Principal Components	Overall Accuracy	F Value	Balanced Precision
Overall Accuracy	30	0.8377	0.4127	0.7330
F Value	50	0.8333	0.5684	0.7429
Balanced Precision	60	0.8421	0.5743	0.7530

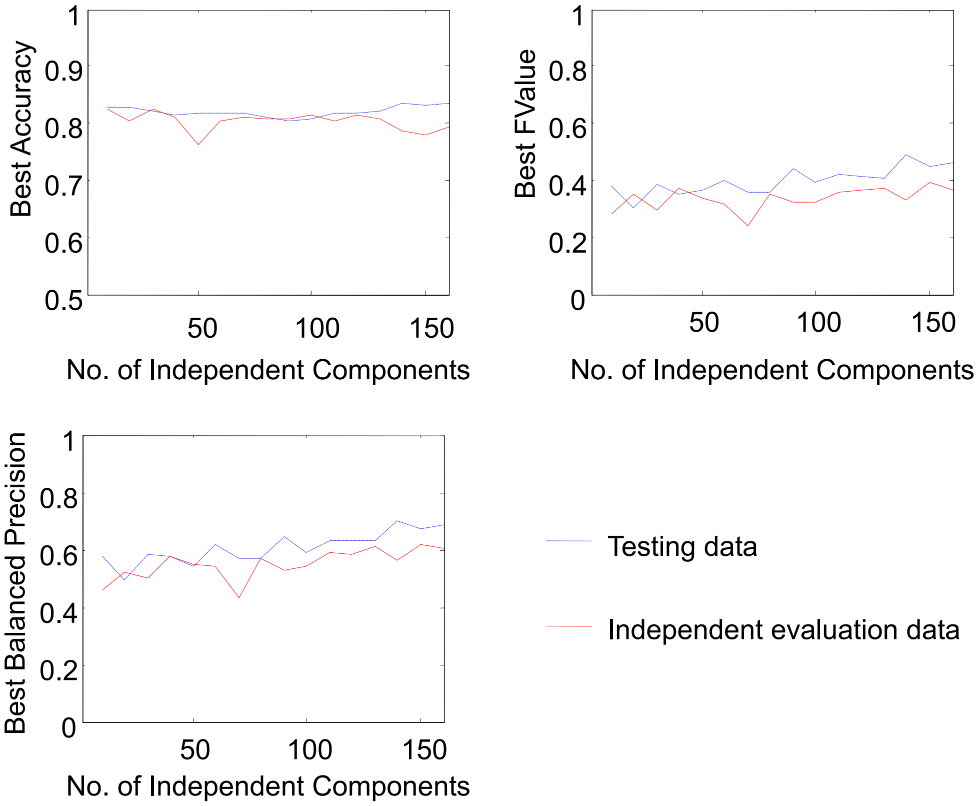


Figure 3.12: K-nearest neighbor prediction of drug-target like proteins after ICA dimensionality reduction.

Table 3.9: Summary of the kNN performance on ICA processed data sets.

Measure Optimized	No. of Independent Components	Overall Accuracy	F Value	Balanced Precision
Overall Accuracy	160	0.7939	0.3621	0.6027
F Value	140	0.7851	0.3276	0.5649
Balanced Precision	140	0.7851	0.3276	0.5649

The ICA pre-processing is also evaluated similarly. The best A, best F, and best G are plotted in Figure 3.12 against the number of independent components used. As shown in Table 3.9, the best A, best F, and best G achieved using different number of independent components are 79.39%, 32.76%, and 56.49% respectively, which is also better than the original results.

In summary, the k-nearest neighbor algorithm works best on PCA pre-processed data sets. The best A, best F, and best G achieved are 83.77%, 56.84% and 75.30% respectively. These results are slightly better than those of decision trees.

3.5.3 Support vector machine prediction

On original data sets, SVMs are trained with the kernel parameter σ scanned in the range of [1...75] with an interval of 1, which is the range that empirically gives optimal classification results in protein function classification [275]. The measurements concerned, A, F and G, are plotted against σ in Figure 3.13. The best A, best F, and best G, as summarized in Table 3.10, are 87.28%, 56.72%, and 72.47% respectively.

On scaled data sets, the kernel parameter σ is scanned in the range of [0.04..3] with an interval of 0.04. The A, F, and G obtained with different σ are plotted in Figure 3.14. The best A, best F, and best G are found in a single SVM model with $\sigma = 1.28$, which are 89.91%, 68.49%, and 75.63% respectively. These results are better than those of the original data sets.

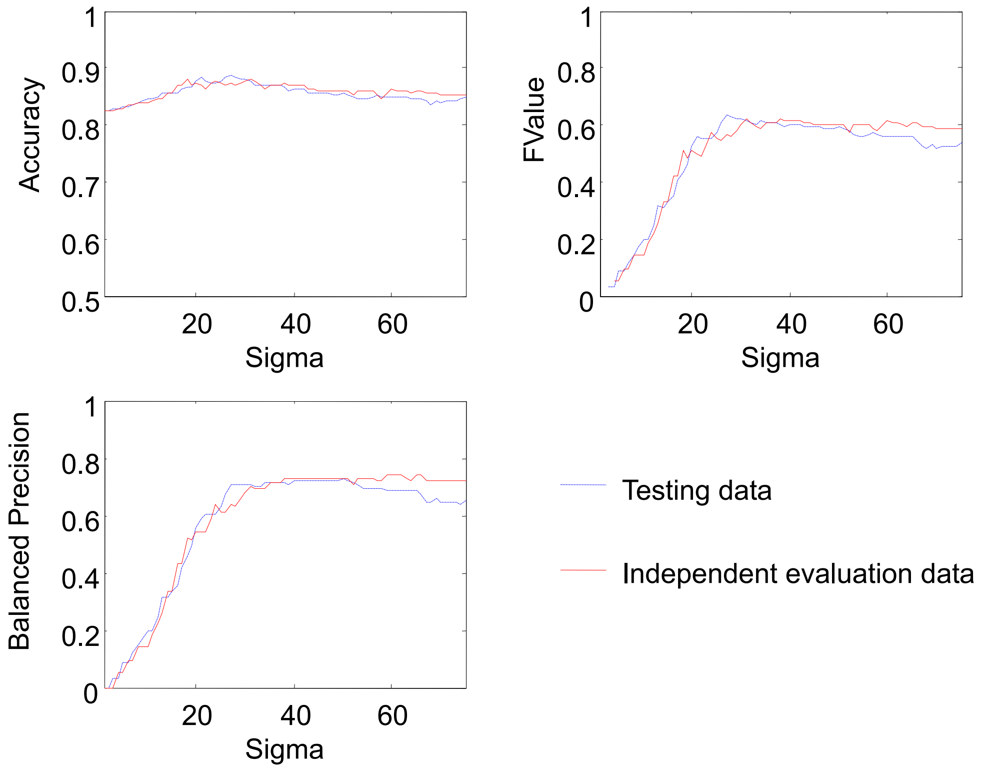


Figure 3.13: Support vector machine prediction of drug-target like proteins on original data sets.

Table 3.10: Summary of the SVM performance on original data sets..

Measure Optimized	Kernel Parameter σ	Overall Accuracy	F Value	Balanced Precision
Overall Accuracy	27	0.8728	0.5672	0.6350
F Value	27	0.8728	0.5672	0.6350
Balanced Precision	50	0.8596	0.6000	0.7247

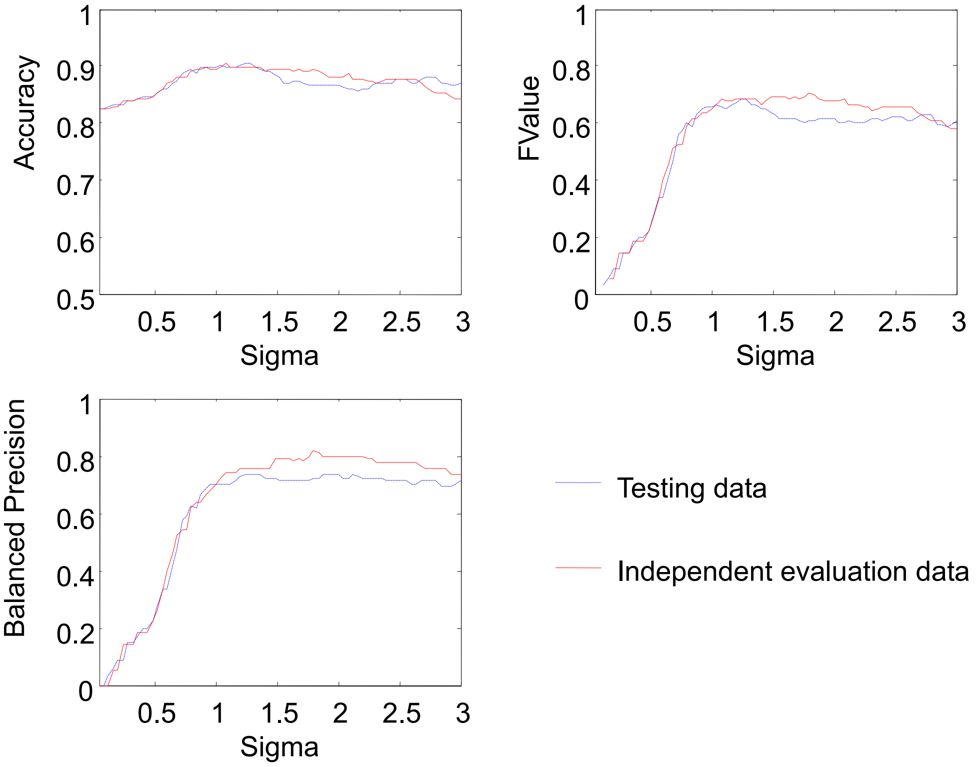


Figure 3.14: Support vector machine prediction of drug-target like proteins on scaled data sets.

Table 3.11: Summary of the SVM performance on scaled data sets.

Measure Optimized	Kernel Parameter σ	Overall Accuracy	F Value	Balanced Precision
Overall Accuracy	1.2800	0.8991	0.6849	0.7563
F Value	1.2800	0.8991	0.6849	0.7563
Balanced Precision	1.2800	0.8991	0.6849	0.7563

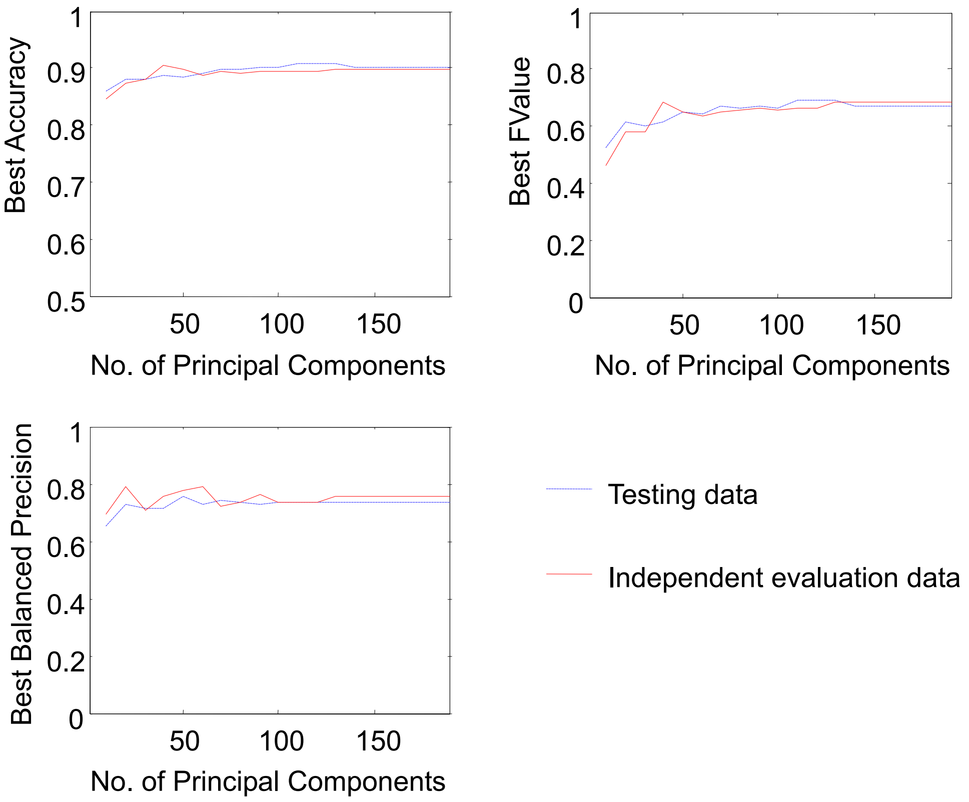


Figure 3.15: Support vector machine prediction of drug-target like proteins after PCA dimensionality reduction.

Table 3.12: Summary of the SVM performance on PCA pre-processed data sets.

Measure Optimized	No. of Principal Components	Overall Accuracy	F Value	Balanced Precision
Best Overall Accuracy	130	0.8991	0.6849	0.7563
Best F Value	130	0.8991	0.6849	0.7563
Best Balanced Precision	50	0.8991	0.6512	0.7730

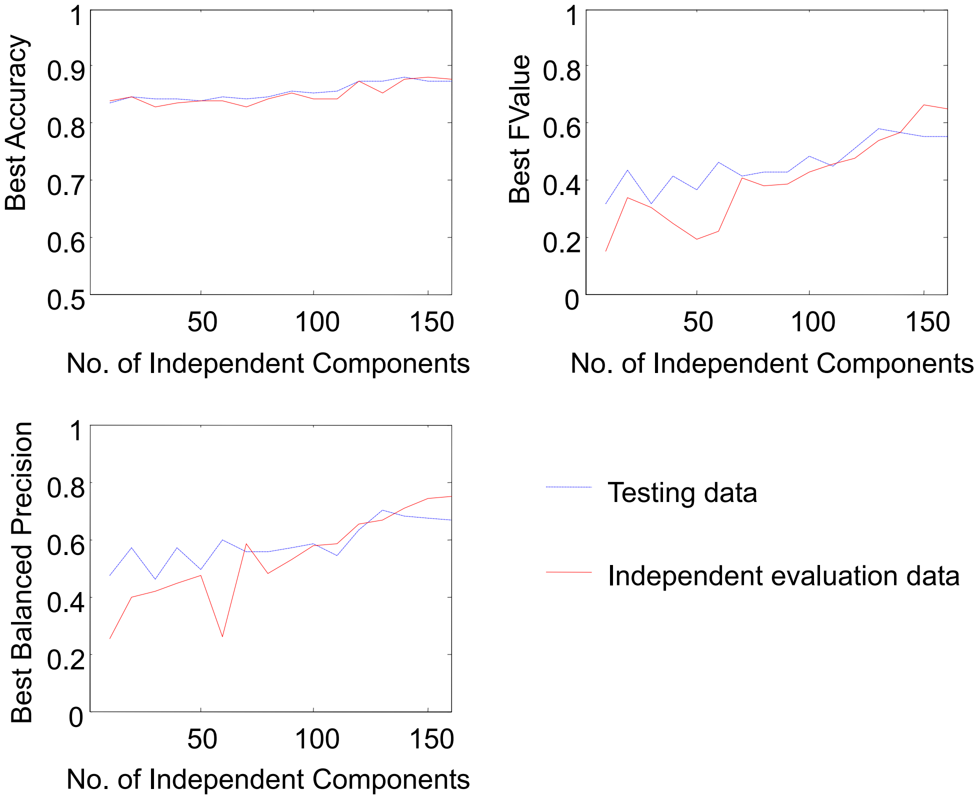


Figure 3.16: Support vector machine prediction of drug-target like proteins after ICA dimensionality reduction.

Table 3.13: Summary of the SVM performance on ICA pre-processed data sets.

Measure Optimized	No. of Independent Components	Overall Accuracy	F Value	Balanced Precision
Best Overall Accuracy	140	0.8772	0.5634	0.7046
Best F Value	130	0.8509	0.5385	0.6657
Best Balanced Precision	130	0.8509	0.5385	0.6657

The effect of PCA dimensionality reduction on SVM prediction is evaluated in a similar way as that discussed in the previous sections. The number of principal components estimated is scanned from 1 to the maximum with an interval of 10 and the kernel parameter σ is scanned in different ranges that empirically give optimal results under different conditions. The best A, best F, and best G achieved using different numbers of principal components are plotted in Figure 3.15. Overall, The best A, best F, and best G achieved with different numbers of principal components, as shown in Table 3.12, are 89.91%, 68.49%, and 77.30% respectively, which are comparable (slightly better in terms of balanced precision) to those on the scaled data sets.

The effect of the ICA dimensionality reduction on SVM is also evaluated. The number of independent components estimated is scanned from 1 to the maximum with an interval of 10 and the kernel parameter σ is scanned in different ranges which empirically give optimal solutions while different numbers of independent components are used. The best A, best F, and best G achieved are plotted against the number of independent components in Figure 3.16. As shown in Table 3.13, the best A, best F, and best G achieved, are 87.72%, 53.85%, and 66.57% respectively, which are not as good as those of the original data sets.

In summary, SVM performance can be improved by the pre-processing of scaling and PCA. Among all the three classification techniques explored, SVM classifies our data best. The best A, best F, and best G, if optimized individually, can reach 89.91%, 68.49% and 77.30% respectively.

3.6 Prediction results and analysis

Table 3.14 summarizes the comparison between different statistical learning methods evaluated in this work. Overall, SVM gives the best results with the best A, best F, and best G reaching 89.91%, 68.49% and 77.30% in different SVM models. The accuracy of SVM prediction, if successfully generalized in real-world application, is reasonably good to provide valuable information for genome scale target discovery.

Table 3.14: Performance comparison between different statistical methods

Measurement Optimized	Decision Tree	K-nearest Neighbor	Support Vector Machine
Best Overall Accuracy	85.09%	83.77%	89.91%
Best F Value	54.05%	56.84%	68.49%
Best Balanced Precision	68.40%	75.30%	77.30%

Errors in statistical learning arise for a number of reasons. It is not expected that exhaustive experiments have been done to verify whether each known protein is a target or not. Also, the therapeutic targets collected in TTD are not complete. This may result in that, with a small possibility, some drug targets are included in the negative examples. Although, most of the statistical learning methods are able to deal with a certain level of noise, these approaches are generally based on a large number of observations (examples). Here in our application, the number of positive examples used for training is only 235, which accounts for less then 5% of the expected population of drug-target like proteins. Therefore, these wrongly assigned class labels, if any, may considerably confuse the learning algorithm and compromise the accuracy of the constructed classification models.

Also, the training data are not balanced. The number of negative examples is significantly higher than that of positive examples. It is expected that the negative examples represent a better sample of the problem space and provide more comprehensive information about its classification. Our results indeed show significantly better performance on negative examples than on positive examples. For example, while one SVM model has an overall accuracy of 89.91%, its F value is only 68.49%.

Anticipated rapid progress in pharmaceutical sciences is expected to provide larger number of and more accurate training examples. Knowledge from study of protein functions also facilitates the selection of training examples for prediction of potential drug-target like proteins.

According to the principal of Occam's razor, the simpler a model is, the better it will generalize. However, too simple models may not catch all the essential characteristics of a problem. The following inequality was found to describe this relationship during the development of the support vector machine algorithm. With the probability of $1-\eta$ $(0 \le \eta \le 1)$, the following bound holds for all models α that are generated using statistical learning approaches,

$$R(\alpha) \le R_{emp}(\alpha) + \sqrt{\frac{1}{l}\left(h\left(1+\ln\frac{2l}{h}\right)+\ln\frac{4}{\eta}\right)} + \frac{1}{l} \qquad \text{Equation 4.46}$$

where h is a measure of the model complexity called the Vapnik-Chervonenkis (VC) dimension, l is the number of training examples and $R_{emp}(\alpha)$ is the empirical risk calculated as Equation 4.47.

$$R_{emp}(\alpha) = \frac{1}{2l}\sum_{i=1}^{l}\left|y_i - f(x_i,\alpha)\right| \qquad \text{Equation 4.47}$$

where $f(x,\alpha)$ is the classification function used for a model α, x_i is the i-th training example and y_i the corresponding class label. Therefore, a good statistical learning method shall strike a balance between the model complexity and the empirical risk.

The better performance displayed by support vector machine might be partly explained with the above theorem. By implementing a scheme called Structural Risk Minimization (SRM), SVM finds the optimal separation hyper-plane, which proves to be the model with lowest complexity when the empirical risk is fixed.

In comparison, the decision tree induction process does not guarantee the result to be the simplest tree, whether in terms of tree height or number of decision nodes. The decision tree induction process always branches on the candidate attribute that gives the maximal information gain. Therefore, it can be regarded as a greedy search algorithm to build a tree as simple as possible. As we know, the greedy search algorithm usually does not guarantee the finding of the global minimal. On the other hand, although the information gain criterion used in selecting branching attributes well reflects the nature of classification – find the class information that is hidden in a number of attributes, the binary discretization method used to deal with continuous attribute may not well reflect the class information hidden in an attribute. This binary discretization will work well when the attribute values of each class have only one center. If the attribute values of each class have multiple clusters (sub-classes), this binary discretization method will overlook much useful information by combining those multiple clusters into one. A good discretization process needs to separate the attribute into multiple segments that generate maximum gain ratio. However, it would be computationally too expensive to afford as the search for the best number of segments and their boundaries will have to evaluate $2^{n-1}-1$ possible solutions, where n is the number of places where adjacent examples belong to different classes.

In comparison to the decision tree algorithm, the algorithm of kNN works well with attribute values that have multiple clusters. However, this advantage can only be fully displayed when the problem space is well sampled and every sub-cluster (or sub-class) of each class is well represented by training examples. However, this is not likely to be our case – the examples of the two classes display no obvious structure on any of the dimensions and 1366 training examples can by no means sample a space of 188 dimensions well. Even if the input space is reduced to only 50 dimensions, the training examples are still far from enough to sample the input space. This kind of insufficient sampling will carry more information on a particular sample set (training set) besides the information on the attribute-class relationship. An algorithm that is less selective may therefore tend to use more irrelevant information to build the classification model. In other words, the constructed model may easily "over-fit" to that particular sampling of training examples and its expected prediction accuracy will be compromised. While SVM implements SRM to avoid over-fitting, the distance measure used in kNN is too simple to effectively avoid this over-fitting, which may be one of the reasons that explains the better performance of SVM over kNN. Future advances in small sample statistics might provide better learning algorithms to derive attribute-class relationship from a small number of samples in a high-dimensional space.

The performance of classification may also be improved with the advances in the formulation of protein sequence descriptors. Although our test results demonstrated this set of descriptors are useful, they may not be perfect.

Obviously, the descriptors used here do not have any direct logical connection to the drug-target likeness. For example, descriptors that reflect the sequence uniqueness and function relationship may give better information on the drug-target likeness. It is desirable for such quantitative descriptors to be devised.

Also, the descriptors are not independent, which means the same information is given in more than one descriptor, which gratuitously increases the complexity of input space and consumes more "resources" of a statistical model. As shown in Figure 3.17, about 50 principal components are able to represent 90% of the total variances in all 188 dimensions, and 140 principal components can represent 99.9% of the total variances in all dimensions. However, in this application, mere PCA or

ICA dimensionality reduction approach seems not effective in improving the predictive accuracy. This might be explained by the unsupervised nature of these dimensionality reduction approaches. PCA and ICA do not use the class information when they are making linear combinations of attributes. Therefore, the constructed principal components or independent components may not be so relevant to the problem of classification. When the attributes-class relationship is intricate and the input space is insufficiently sampled, such as in our application, dimensionality reductions by mere PCA or ICA analysis may not preserve enough information relevant to classification. In this regard, dimensionality reduction processes that take consideration of the class distribution of examples and preserve useful information are desired.

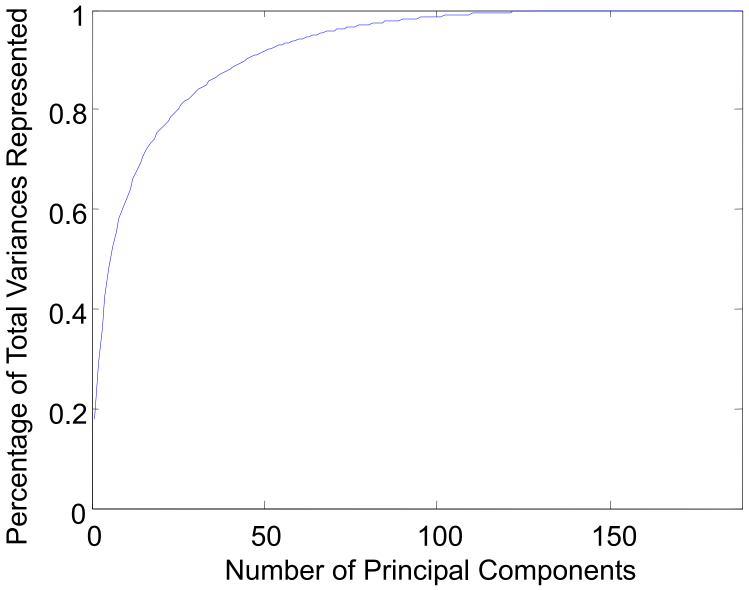


Figure 3.17: Number of principal components and the percentage of total variance they can represent.

3.7 Summary

A number of statistical learning methods and pre-processing techniques are investigated for the application of drug-target like protein prediction, which includes the learning algorithms of decision tree, k-nearest neighbor and support vector machine and the pre-processing techniques of scaling, PCA and ICA dimensionality reduction. The support vector machine approach gives the best classification results. Scaling and PCA help to improve the performance of SVM. The highest A, F and G achieved in different SVM models reach 89.91%, 68.49% and 77.30% respectively. This accuracy seems to be reasonably good to facilitate the genome scale drug target discovery.

Performance and applicability of the statistical learning methods may be further improved by incorporation of new information from advances in pharmaceutical sciences, proteomics, and protein function. Efficiency and accuracy of statistical learning methods in prediction of drug-target like proteins can also be enhanced from new progress in learning algorithms, descriptors, and pre-processing techniques.

Apart from discovery of new drug targets, discovery of efficient therapeutic intervention strategies that explore the synergies between existing targets is also critical in facilitating the combat against diseases. With the help of the therapeutic target database, it may also be possible to explore the unknown therapeutic mechanisms of effective herbal medicines, which is discussed in the next chapter.

Chapter 4

In silico study of the mechanisms of action of active ingredients from medicinal plants

Therapeutic mechanisms of effective herbal medicines are very useful in designing novel therapeutic intervention strategies. So far, medicinal plants still serve as a major source of novel therapeutic mechanisms. The molecular mechanism related to therapeutic effects of a medicine can be probed if its therapeutic targets can be identified. *In silico* approaches to study mechanisms of the therapeutic action of herbal medicines are developed in this endeavor. So far, one extended docking approach, INVDOCK, has been developed to facilitate such studies. Our results on nine selected active ingredients derived from herbs showed that over half of the INVDOCK identified potential therapeutic targets of the selected active herbal ingredients have relevant experimental findings, and about 70% of their therapeutic implications have been reported to occur in cultivated cells, animal models or clinical trails. These results suggest the usefulness of *in silico* tools in facilitating the discovery of novel therapeutic mechanisms of effective herbal active ingredients.

4.1 Introduction

Novel therapeutic mechanisms discovered from studies of MP ingredients have routinely been used to derive new therapeutic intervention approaches [105]. However, so far, studies on the therapeutic mechanism of herbal medicines are still very limited.

Medicinal plants (MPs) have been widely explored in traditional medicines [302-304] and in drug discovery [106-108]. Approximately one third of the top-selling drugs currently in the market were derived from plants [108]. Because of their broad structural diversity and wide range of known pharmacological activities, MPs have served and are still used as a valuable source for new drug discovery [106-108].

As part of the new drug discovery effort, a large number of ingredients have been extracted from MPs and other natural sources, and they have been studied for their potential therapeutic effects[305-309]. However, the basic and clinical pharmacology is known for only a portion of these ingredients [305,309]. Thus the insufficient mechanistic understanding of MPs hinders the efforts of developing new drugs based on the novel therapeutic mechanisms of MP ingredients. It also limits the scope of therapeutic exploration of MPs used in traditional medicines and other herbal medicines. Hence much more research remains to be done in order to probe the mechanisms of action of MP active ingredients. Systematic study of the mechanisms of a large number of MP ingredients by means of traditional assay based methods is a costly and time-consuming process due to the difficulties in extraction, synthesis, and activity test of herbal ingredients. Therefore alternative approaches for low-cost and rapid analysis of the mechanisms of MP ingredients are useful for the study and exploration of MPs.

4.2 *In silico* methods for target identification of MP ingredients

The molecular mechanism related to the therapeutic effects of an MP ingredient can be probed if its therapeutic targets can be identified. Target identification is also an important step in rational drug design [238,310]. Thus varieties of target search strategies have been explored or are under consideration.

A popular approach is to derive target information from the variation of micro-array expression data between normal cells and cells in disease states [109]. Micro-array technology relies on the hybridization properties of nucleic acids to monitor DNA or RNA abundance on a genomic scale. By arraying different immobilized probe oligo-nucleotides on a chip, micro-arrays have revolutionized the study of genomes by allowing researchers to study the expression of thousands of genes simultaneously. The ability to simultaneously study thousands of genes under a host of differing conditions presents an immense challenge in the fields of computational science and data mining. Conventionally, t-test is used to identify those genes that have a significantly different expression level in normal and disease states. Now various computational methods have been used to facilitate the analysis of micro-array expression data to find the genes that act differently and the patterns of gene expression that strongly relate to certain kinds of disease conditions. Such approaches include regularized t-test based on a Bayesian statistical framework, neural networks, principal component analysis, Bayes belief networks, clustering, mixture models and expectation minimization, gene shaving, support vector machine, hidden markov model (HMM), and other statistical learning methods [311,312]. For instance, statistical methods were exploited to find schizophrenia associated genes [238] and diffuse large B-cell lymphoma related genes [313] based on microarray expression data. These comparative methods may be potentially used for facilitating the search of the targets of MP ingredients by analysis of the change of expression profiles induced by the addition of specific MP ingredients into disease-state cells.

High-throughput screening has been routinely used for the identification of small molecule drugs of a specific target [110]. The same technology can be readily applied to the search of targets of MP ingredients as well as synthetic chemicals. A comprehensive library of potential targets can be built and used for the search of targets [90]. It has been reported that protein chips [314] are capable of large-scale, high-throughput analysis of protein-small-molecule interactions. Similar to DNA micro-arrays, in protein chips, proteins are prepared, densely arrayed on the surface of the chip in their active conformations. This technology enables the large-scale parallel analysis of the binding affinity between a pool of proteins and a certain chemical, which makes it feasible to screen a target library for those proteins targeted by MP ingredients. Production, segregation, purification and immobilization of proteins required for the fabrication of protein chips is still a complex and costly procedure. Therefore, *in silico* virtual screening is expected to be a potentially good choice for low-cost and efficient search of targets of MP ingredients.

One of the widely used virtual screening approaches in rational drug design is quantitative structure activity relationship [58,315]. As discussed in previous sections, QSAR is based on the statistical analysis of the relationship between biological activity of a chemical and its quantitative and structural properties. The derived statistical model could then be used to predict the activity of an unknown chemical by its quantitative attributes usually calculated from its structure. By generating QSAR models of a sufficient pool of potential targets, this method can be extended to facilitate the search of the protein targets of an MP ingredient as well as a synthetic chemical. Specific targets can be identified from this pool if the molecule is predicted to have sufficient activities.

QSAR is a target specific approach which requires the construction of a unique model for each potential protein target, which makes it a complex matter to screen a large number of potential targets. An interesting alternative method has recently emerged which overcomes the difficulty of using QSAR. It has been reported that support vector machine, an relatively new statistical learning algorithm, may be used to construct a universal model for prediction of the binding affinity between a protein and a compound using the quantitative descriptors constructed from both the protein sequence and the chemical structure properties [316]. This method may be easily extended to screen a target library to identify potential protein targets of MP ingredients. However, this approach requires sufficiently diverse and accurate known ligand-protein binding affinities in order to satisfactorily train the statistical learning model. A database of theoretically computed ligand-receptor binding energies [317] has been used as the training examples in the previous work, which limits the reliability of the models generated. Hopefully, further developments of this method will enable it to become a useful tool for target identification.

Ligand-protein docking is also a popular virtual screening technology in rational drug design. In this approach, a chemical is structurally and chemically placed into the binding site of a protein based on the 3D structure of both molecules and the computed interaction energy between them [39-41]. Testing results on a number chemicals have consistently indicated that they are capable of finding the protein-small-molecule binding conformations at a receptor site close to experimentally determined structures[39-41]. Thus the ligand-protein docking method can be readily extended to identify therapeutic targets of MP ingredients as well as synthetic chemicals based on the structural and molecular mechanical analysis of the bindings between the molecules and the therapeutic targets collected in TTD [318]. So far, an extended ligand-protein docking method, INVDOCK, has been specifically used for automated drug target identification of several small molecules [111].

4.3 A closer examination of an *in silico* method - INVDOCK

INVDOCK is the only *in silico* method specifically applied to identification of protein targets of small chemicals so far. It is worthy of a closer examination of the principles and algorithms as well as the performance of this method.

4.3.1 Feasibility

INVDOCK is based on a ligand-protein inverse docking strategy such that a compound is attempted to dock to known ligand-binding pockets of each of the proteins in a protein 3D structural database. A protein is considered as a candidate potential target of a compound if that compound can be docked into the protein and the binding satisfies a molecular-mechanics based criterion for chemical complementarity [111]. Because of their capability in identifying potential ligands and binding conformations, docking algorithms are expected to be equally applicable to the inverse docking procedure for finding multiple protein targets to which a compound can bind or weakly bind [111,223].

The inverse docking algorithm requires a sufficient number of proteins of known 3D structure to cover a diverse range of potential therapeutic effects. At present there are 19860 protein entries in the Protein Data Bank (PDB) and the number increases at a rate of well over 100 per month[319]. About 17% of these have unique sequence[320]. Introduction of high-throughput methods is expected to enable structural determination of 10,000 proteins with unique sequence within five years[29]. Thus the number of proteins is approaching a meaningful level to cover a diverse set of potential therapeutic targets.

4.3.2 Algorithm

The flowchart of the inverse docking algorithm is shown in Figure 4.1. A small molecule is attempted to dock to proteins with known 3D structures. By evaluating a molecular-mechanics based criterion for chemical complementarity, the docking filter generate a putative protein / nucleotide targets list of the small molecule. This list is further filtered to find the putative therapeutic targets of that small molecule.

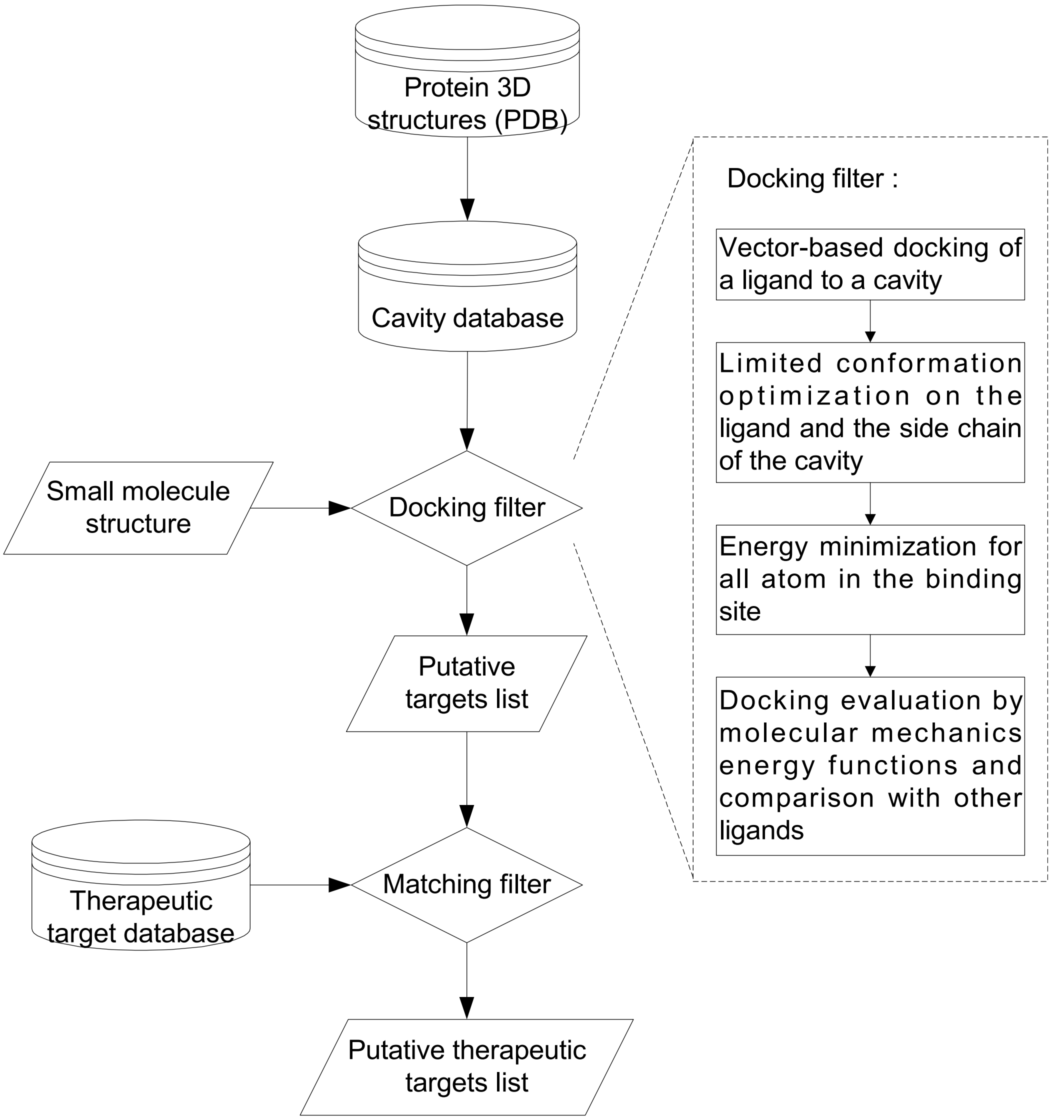

Protein 3D structures (PDB)
Cavity database
Small molecule structure
Docking filter
Putative targets list
Therapeutic target database
Matching filter
Putative therapeutic targets list
Docking filter :
Vector-based docking of a ligand to a cavity
Limited conformation optimization on the ligand and the side chain of the cavity
Energy minimization for all atom in the binding site
Docking evaluation by molecular mechanics energy functions and comparison with other ligands

Figure 4.1: Flowchart of the inverse docking algorithm

To facilitate fast-speed search of potential protein targets of a chemical, a protein cavity database has been developed from the corresponding protein 3D structures in the PDB [111]. Each cavity entry is composed of a sphere cluster filling the cavity. Every cavity in each protein has a corresponding entry in this cavity database.

Docking of a particular chemical to a cavity occurs by the following steps: First, the chemical is aligned within the selected site by matching the position of each atom of the compound with the center of spheres. Because of the relatively low-resolution nature of the conformation sampling of the chemical, certain degree of structural clash is allowed at this stage. A molecular-mechanics conformation optimization is then conducted by a limited torsion space sampling of rotatable bonds in the chemical and those in the side-chain of the receptor amino acid residues at the binding site. Each rotatable bond is sampled in the range of $\pm 15^{o}$. This is followed by 50 iterations of Cartesian coordinate energy minimization on all chemical and protein atoms at the binding site so as to further optimize the compound-protein complex. Energy minimization is by a steepest decent method.

In both torsion optimization and energy minimization, AMBER force fields [120] are used for covalent bond, bond angle, torsion, and non-bonded Van der Waals and electrostatic interactions. The partial charges of drug atoms are assigned following the method described in [446]. Morse potential [321], which is a function of donor-acceptor distance, is used to represent hydrogen bond terms. This potential has been shown to give reasonable description of hydrogen bond energy and dynamics in biomolecules [322,323]. The energy function is:

$$
\begin{aligned}
V = \; & \frac{1}{2}\sum_{Bonds} Kr(R-R_{eq})^2 + \frac{1}{2}\sum_{Angles} K_\theta(\theta-\theta_{eq})^2 + \\
& \frac{1}{2}\sum_{Torsions} V_n[1-\cos(n(\phi-\phi_{eq}))] + \\
& \frac{1}{2}\sum_{H-bonds} [V_0(1-e^{-a(r-r_0)})^2 - V_0] + \\
& \sum_{Non-bonded} [(\frac{A_{ij}}{r_{ij}})^{12} - (\frac{B_{ij}}{r_{ij}})^6 + \frac{q_i q_j}{\varepsilon_r r_{ij}}]
\end{aligned}
$$

In this function, R, θ ..and ϕ denotes bond length, bond angle and torsion angle respectively, R_{eq}, θ_{eq} and ϕ_{eq} are taken as equilibrium bond length, angle and torsion angle respectively and their values are from the original PDB structure and the structure of the drug respectively; Kr and K_θ are covalent and bond angle bending force constant respectively; V_n and n are torsion parameters; r is the hydrogen bond donor-acceptor distance, and V_0, a and r_0 are hydrogen bond potential parameters.

Scoring of docked molecules is based on a ligand-protein interaction energy function ΔE_{LP} composed of the same hydrogen bond and non-bonded terms as those used for structure optimization [111]. Analysis of a large number of PDB ligand-protein complexes shows that the computed ΔE_{LP} is generally below $\Delta E_{Threshold} = -\alpha N$ kcal/mol, where N is the number of ligand atoms and α is a constant ~1.0 [10]. The exact value of α can be determined by fitting the linear equation $\Delta E_{Threshold} = -\alpha N$ to the computed ΔE_{LP} for a large set of PDB structures. This statistically derived energy value can be used empirically as a threshold for screening likely binders. A polynomial form of $\Delta E_{Threshold}$ involving more parameters may also be introduced to derive an energy threshold. ΔE_{LP} can be required to be lower than $\Delta E_{Threshold}$ when selecting successfully docked structures. A discussion on the typical binding energies can be found in [317].

Drug binding is competitive in nature. A drug is less likely to be effective if it binds to its receptor non-competitively against natural ligands and, to some extent, other drugs that bind to the same receptor site. This binding competitiveness may be partially taken into consideration for cavities known as ligand-bound in at least one PDB entry. Ligands in PDB structures are known binders. Therefore PDB ligands bound to the same receptor site as that of a docked molecule may thus be considered as “competitors” of that molecule. In INVDOCK selection of a putative protein target, the computed ΔE_{LP} is not only evaluated against $\Delta E_{Threshold}$ but also compared to the ligand-protein interaction energy of the corresponding PDB ligands that bind in the same cavity in this or other relevant PDB entries. The Ligand-protein interaction energy for the relevant PDB structures is computed by the same energy functions as that for the docked molecule. In addition to the condition that it be lower than $\Delta E_{Threshold}$, ΔE_{LP} of a docked molecule is required to be lower than a “competitor” energy threshold $\Delta E_{Competitor}$ when selecting a putative target. $\Delta E_{Competitor}$ can be taken as highest ligand-protein interaction energy of the corresponding PDB ligands multiplied by a factor β. In order to be able to find weak binders as well as strong binders, a factor $\beta \leq 1$ is introduced to scale the ligand-protein interaction energy of PDB ligands. This is because a weak binder may have slightly higher interaction energy than that of a PDB binder. No experimental data has been found to determine the value of β. Hence β has been tentatively determined by an analysis of the computed energy for a number of compounds. A value of 0.8 has been suggested for β which leads to reasonable results statistically [111].

4.3.3 Validation studies on synthetic chemicals

The effectiveness of INVDOCK prediction of the protein targets of synthetic chemicals can be demonstrated from a recent study on a number of clinical agents [111]. The results of one drug, tamoxifen (Figure 3.1), are quoted here. Tamoxifen is an anticancer drug widely used for treatment of breast cancer [324] and it has been approved as the first cancer preventive drug. Tamoxifen metabolite 4H-tamoxifen is believed to be the major contributor to the anti-oestrogenic effects of tamoxifen inside the human body [324].

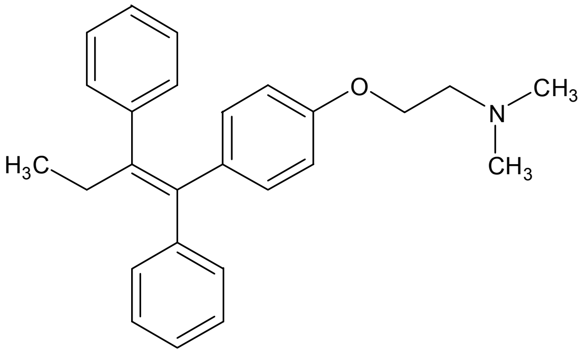


Figure 4.2 Chemical structure of tamoxifen

Potential human and mammalian protein targets of 4H-tamoxifen identified by INVDOCK are given in Table 4.1 along with the respective clinical implications from experiments. The computed binding energies are not listed as energy alone may not be a good indicator on the strength of drug binding *in-vivo* because of the compitions from other natural ligands. A number of known protein targets of tamoxifen are found in the Table. These include estrogen receptor [324], protein kinase C [325], collagenase [326], hydroxysteroid dehydrogenase [327], Alcohol dehydrogenase [328], and prostaglandin synthetase [329]. It has been observed that two other INVDOCK identified proteins glutathione transferase and 3-alpha-hydroxysteroid dehydrogenase exhibited altered activity by tamoxifen [330,331], which may be indicative of direct binding of tamoxifen to these proteins. Experiments showed that the levels of another two identified proteins, dihydrofolate reductase and immunoglobulin, are changed by tamoxifen [332,333]. Ligand binding is known to self-regulate protein levels in certain cases [334]. It remains to be seen whether these two proteins are also the targets of tamoxifen as implicated by INVDOCK search.

Table 4.1: Potential therapeutic targets of 4H-tamoxifen identified from INVDOCK search of human and mammalian proteins.

PDB ID	Target Name	Experimental Findings	Therapeutic Implications
1a52	Estrogen Receptor	Drug target [324]	Breast cancer [324]
1akz	Uracil-DNA Glycosylase		
1ayk	Collagenase	Inhibited activity [327]	Tumor cell invasion and cancer metastasis [327]
1az1	Aldose Reductase		Diabetes

1bnt	Carbonic Anhydrase		
1boz	Dihydrofolate Reductase	Decreased level [332]	Combination therapy for cancer[332]
1d3v	Arginase		
1d6n	Hypoxanthine-guanine phosphoribosyltransferase		
1dda	Alcohol dehydrogenase	Inhibition [328]	Enhanced ethanol's sedative effect [328]
1dht, 1fdt	17-beta-Hydroxysteroid Dehydrogenase	Inhibitor [327]	Promotion of tumor regression [327]
1gsd, 3ljr	Glutathione Transferase A1-1, Glutathione S-Transferase	Suppressed enzyme and activity [330]	Genotoxicity and carcinogenicity [330]
1mch	Immunoglobulin Light Chain	Temporarily enhanced Ig level [333]	Modulation of immune response [333]
1p1g	Macrophage Migration Inhibitory factor		
1ulb	Purine Nucleoside Phosphorylase		
1zqf	DNA Polymerase		Viral infection
2nll	Retinoic Acid Receptor		
1a25	Protein Kinase C	Inhibition [325]	Cancer [325]
1aa8	D-Amino Acid Oxidase		
1afs	3-alpha-Hydroxysteroid Dehydrogenase	Effect on androgen induced activity [331]	Hepatic steroid metabolism [331]
1pth	Prostaglandin H2 Synthase-1	Direct inhibition [329]	Prevention of vasoconstriction [329]

1sep	Sepiapterin Reductase		
2toh	Tyrosine 3-Monooxygenase		

It is noted that a known target of tamoxifen such as calmodulin [325] is not identified by INVDOCK. One possible reason might be that the available PDB structures of calmodulin are not sufficiently close to the tamoxifen-bound conformation. None of these PDB structures is bound by a ligand similar in structure to tamoxifen. The conformation of calmodulin is known to change significantly by binding of ligands [335]. Because of the intrinsic flexibility of this protein, it is highly likely that ligand binding to this protein involves induced-fit. The analysis of two PDB structures of calmodulin bound by a different ligand (PDB id: 1a29 and 2bbm) shows that the conformation of this protein is dependent on its binding ligand. In a recent molecular docking study, a ligand was docked into calmodulin by the consideration of conformation changes that mimic an induced fit [336]. It is also found that 4H-tamoxifen can be placed into calmodulin with slightly less favorable steric interaction than allowed by the INVDOCK scoring function. An appropriate conformation change in calmodulin might allow for the removal of this un-favorable interaction.

The limited number of protein entries available in the cavity database is also expected to result in missed hits. For instance, known tamoxifen metabolizing protein cytochrome P450 [337] is not identified in this work because no corresponding human or mammalian entry is available in the cavity database. A search of bacterial proteins in the database identified this protein (PDB Id: 5cp4 and 1cpt) as a putative target.

As shown in Table 4.1, apart from the ten putative protein targets that have been implicated or confirmed experimentally, INVDOCK identified ten other proteins as putative targets of 4H-tamoxifen. These include aldose reductase, Arginase, carbonic anhydrase, macrophage migration inhibitory factor, purine nucleoside phosphorylase, DNA polymerase, hypoxanthine-guanine phosphoribosyltransferase, retinoic acid oxidase, sepiapterin reductase, and tyrosine 3-monooxygenase. No literature has been found to link tamoxifen to each of these proteins. There is also no report that indicates each of these proteins is not a target of tamoxifen or its analogs. Further investigation is therefore needed to determine whether or not 4H-tamoxifen can bind to these proteins.

In summary, the majority of INVDOCK identified protein targets have been implicated or confirmed by experiments [111,223]. Ligand-protein inverse docking appears to be a useful approach for computer-aided identification of potential protein targets of small synthetic molecules. The ability of identification of therapeutic targets of MP ingredients is evaluated as follows.

4.4 *In silico* prediction of therapeutic targets of MP ingredients

In this work, the therapeutic targets of six MP ingredients have been predicted by the *in silico* method INVDOCK. The results, together with the those of three other MP ingredients published before, have been compared with available experimental findings [338]. These nine ingredients were selected based on a comprehensive

MEDLINE [339] search on MP related publications, and those ingredients with relatively abundant amount of references are selected. This is to ensure that, for each selected ingredient, there is a reasonable amount of experimental findings to make a sensible evaluation of INVDOCK results.

The nine MP ingredients include genistein, ginsenoside Rg1, quercetin, acronycine, baicalin, emodin, allicin, catechin and camptothecin. These compounds have been subjects of various investigations including the probing of their molecular targets. The availability of relevant experimental findings makes it possible to conduct comparative study and to evaluate INVDOCK derived results. The 3D structures of each of these MP ingredients are generated by the following procedure: The 2D structure of each MP ingredient is obtained from the CCD database (http://www.chemnetbase.com/) and it is then converted into 3D structure using CONCORD (http://www.tripos.com/admin/LitCtr/concord.pdf) licensed from SYBYL.

All the proteins with known 3D structure in the Protein Databank PDB are searched by INVDOCK for identification of potential protein targets of the MP ingredients. The therapeutic target database was then searched to provide a list of candidate therapeutic targets from the complete protein targets list.

To evaluate the usefulness of INVDOCK for predicting the therapeutic targets of the selected MP ingredients, literature have been searched for reports about the experimentally determined targets of these MP ingredients. It is noted that a substantial number of experimental binding studies have been conducted at micro-mole level which is significantly higher then that of average *in vivo* drug concentration. Although binding analysis at higher concentrations has routinely been used as an indication about *in vivo* effect at lower concentrations [340,341], certain caution is needed for searching related experimental reports. Two additional requirements have been used in the selection of related experimental reports. One is that a reported experiment be conducted at concentrations no higher than the routinely used ones[340,342]. Another is that the physiological effect of the reported binding has been observed.

Pharmacokinetics is another factor that needs to be considered when searching for related experimental reports. While some reports may indicate the binding of a selected MP ingredient with a protein, it may not be clear whether that particular MP ingredient can reach the target site with sufficient concentration. Therefore, additional experimental evidence such as the reported in vivo effects of the particular MP ingredient on the specific organ have been used as an indication that the MP ingredient likely reach a particular site at certain level of concentration.

4.4.1 Genistein

Genistein (Figure 4.3) is a soy-derived isoflavone of therapeutic interest. Dietary intake of soy is associated with a decreased risk of both hormone-dependent and hormone-independent cancers [343]. At the molecular level, genistein inhibits the activity of ATP binding enzymes such as tyrosine-specific protein kinase, topoisomerase II and enzymes involved in phosphatidylinositol turnover. Moreover, genistein can act via an estrogen receptor-meditated mechanism [344]. At the cellular level, genistein induces apoptosis and differentiation in cancer cells, inhibits cell proliferation, modulates cell cycling, exerts antioxidant effects, inhibits angiogenisis, and suppresses osteoclast and lymphocyte functions [344]. In addition, it acts as an immunosuppressant and shows beneficial effects in the treatment of osteoporosis, cardiovascular disease, and menopause [344].

The pharmacokinetics profile of genistein has been well studied. One experiment on an animal model showed a favorable uptake of genistein [345]. In another study, genistein has been found in various organs of rats, including gut, excretory organs, respiratory organs, peripheral organs, reproductive organs. The pharmacokinetics data were also measured [346]. The good pharmacokinetic profile of genistein also contributes to its wide range of beneficial effects.

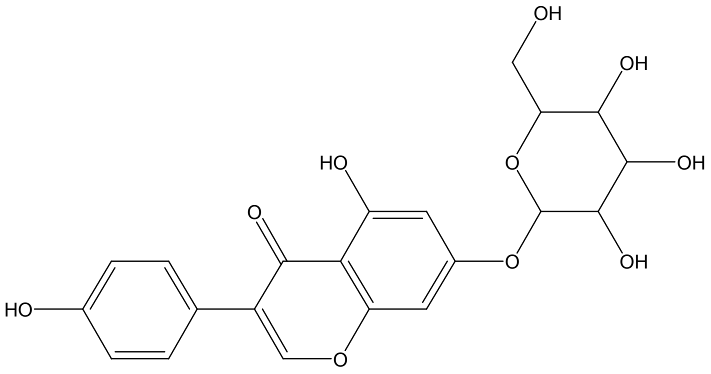


Figure 4.3 Structure of the MP ingredient genistein.

INVDOCK identifies 18 potential therapeutic targets of genistein, which are listed in Table 4.2 together with available experimental findings. Two of these proteins have been reported to be directly inhibited by genistein *in vitro*. These targets are estrogen receptor [347], and FGF receptor 1 [348]. Estrogen receptor beta is reported to bind genistein with an affinity close to that of 17beta-estradiol. However, it remains to be determined whether it is a mediator of genistein's activity in vivo. FGF receptor 1 is a kind of tyrosine kinase. As a general tyrosine kinase inhibitor, genistein is expected to inhibit FGF receptor 1. These two proteins are known anti-cancer targets and there is indeed a report about the effect of genistein on cancer. [349]

Table 4.2: Potential therapeutic targets of Genistein identified from INVDOCK search of human and mammalian proteins.

PDB ID	Target Name	Experimental Findings	Therapeutic Implications
1a35	Topoisomerase I	Genistein has anti-topoisomerase I effect [350].	Cancer [349]
1a7c	Plasminogen activator inhibitor	Genistein shifts urokinase / plasminogen activator inhibitor proteolytic balances [351].	Cancer [349]
1akf	Estrogen receptor	Genistein binds to estrogen receptor beta [347].	Cancer (Breast) [352],d Vascular disease [353]
1d6n	Hypoxanthine-guanine phosphoribosyltransferae	Genistein marginally activates HPRT [354].	Malaria [355]
1di8	Cyclin-dependent kinase 2	Genistein suppresses CDK2 activity [356].	Cardiovascular disease [353]

1fgi	FGF receptor 1	Genistein blocks cytoplasmic receptor domain [348].	Cancer [349] , Angiogenesis [353]
1rts	Thymidylate synthetase		Cancer [349]
1ula	Purine nucleoside phosphorylase		Cancer [349], Malaria [355]
2dhf	Dihydrofolate reductase		Leprosy [357]
1vbt	Cyclophilin A		Cancer [349],
1db4	Phospholipase A2		Inflammation [358]
1diy	Prostaglandin H2 synthase	Genistein decreased the specific activity of prostaglandin H2 synthase prepared from A431 cells. [359]	Inflammation [358]
1d8d	Farnesyltransferase		Cancer [349]
1bpx	DNA polymerase		Viral infection [360]
1b3o	Inosine dehydrogenase		Malaria [355]
1azm	Carbonic anhydrase I		Hypertension, Glaucoma[361], Cancer [349]
1awn	Guanylyl cyclase		Cancer [349]
1a25	Protein kinase C		Cardiovascular disease [353] Cancer [349]

Experiments also showed that the activity or expression level of each of the five INVDOCK identified therapeutic targets given below is affected by genistein. Ligand binding may influence the activity of a protein, and it is also known to self-regulate protein levels in certain cases [334]. Hence there is a possibility that these observed phenomena indicate genistein's binding to each of these proteins as predicted by INVDOCK. The activities of cyclin-dependent kinase 2 [356], topoisomerase I [350] and prostaglandin H2 synthase[359] have been reported to be suppressed by genistein. Cyclin-dependent kinase 2 is a therapeutic target for cardiovascular disease. Genistein has been found to possess an effect preventive of cardiovascular disease [362]. Topoisomerase I is another therapeutic target of cancer. Prostaglandin H2 synthase (COX) is a major therapeutic target for inflammation. Genistein has been reported to have some relationship with inflammation [358]. Also, genistein has been known to induce a shift towards antiproteolysis on urokinase/plasminogen activator inhibitor proteolytic balances [351], which seems to suggest that this protein is a target and it may also account for genistein' efficacy in cancer therapy. Genistein can marginally induce hypoxanthine-guanine phosphoribosyltransferase [354], which seems to suggest a mechanism of genistein effect in treating malaria [355].

Moreover, there are 11 identified therapeutic targets without experimental validation or invalidation. Further experimental investigation is needed to determine whether each of these proteins is a target of genistein as predicted by INVDOCK. These targets have been explored in treatments of eight different diseases. Among them, the effect of genistein on cancer[349], malaria[355], leprosy[357], inflammation[358], herpes viral infection[360], glaucoma[361] and vascular disease[353] has been reported in cell cultivation or animal models. However, whether these effects are due to the predicted interaction of genistein with some of these proteins remain to be studied. Moreover the pharmacokinetic profile of genistein needs also to be studied to assess the clinical effect of genistein on these diseases.

4.4.2 Ginsenoside Rg1

Ginsenoside Rg1 (Figure 4.4) is a major bioactive component of ginseng. Ginseng is a highly valued MP in the Far East and has gained popularity in the West during the last decade [363]. It can be used to combat stress, to enhance both the central and immune systems, and to help maintaining optimal oxidative status against certain chronic disease states and aging[364]. It is also reported to have an effect to prevent cancer[365]. The pharmacokinetic data of ginsenoside Rg1 is not so well studied as genistein. The available data showed that ginsenoside Rg1 had a wide distribution and long half-life in the body [366,367].

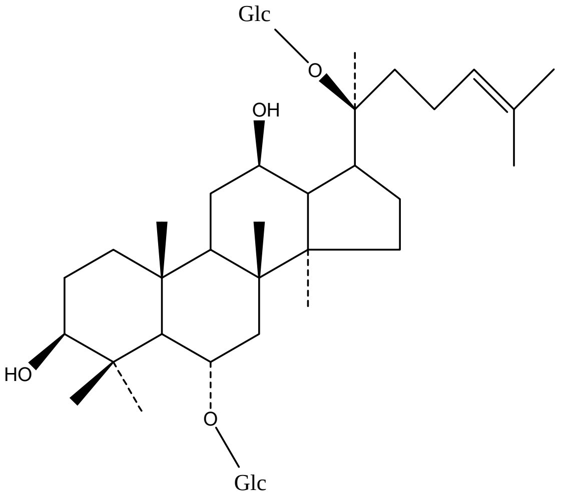


Figure 4.4 Structure of the MP ingredient ginsenoside Rg1.

The predicted therapeutic targets of ginsenoside Rg1 are given in Table 4.3. Three proteins are identified as potential therapeutic targets of this MP ingredient. One is endothelial nitric-oxide synthase, which is known to be inhibited by ginsenoside Rg1[368] and this inhibition may contribute to the observed maintenance of optimal oxidative status against chronic disease states and aging[364]. DNA polymerase beta has not been reported to bind ginsenoside Rg1, however, it has been found that ginsenoside Rg1 can stimulate DNA synthesis[369] and activate DNA polymerase delta[370]. Protein cyclophilin is also identified as a potential target by INVDOCK, while no experimental reports are available to confirm or invalidate it. This protein is related to immunomodulatory activity, which is one of the well-known therapeutic effects of ginsenosides including ginsenoside Rg1[371]. In addition to these therapeutic targets, INVDOCK also predicted an experimentally confirmed non-therapeutic target, 1,4-galactosyltransferase, an in vivo metabolizing enzyme of ginsenoside Rg1 [372].

Table 4.3: Potential therapeutic targets of ginsenoside Rg1 identified from INVDOCK search of human and mammalian proteins.

PDB ID	Target Name	Experimental Findings	Therapeutic Implications

1rpa	DNA polymerase beta	Rg1 stimulates DNA synthesis. Rg1 activates DNA polymerase delta [369,370].	Viral infection (herpes) [373]
1rmh	Cyclophilin A		Immune response [371]
3nos	Endothelial nitric-oxide synthase	Rg1 inhibits NOS dose dependently [368].	Maintaining optimal oxidative status [364]

4.4.3 Quercetin

Quercetin (Figure 4.5) is a flavonoid nearly ubiquitous in plants and it is particularly rich in seeds, citrus fruits, olive oil, tea, and red wine [374]. Certain plants containing flavonoids have been used in traditional medicines and there have been suggestions for exploring these MPs therapeutically [374]. Quercetin is one of the most comprehensively studied flavonoids which are suitable for evaluation of INVDOCK. However, unfortunately, in vivo data on the disposition, absorption, bioavailability, and metabolism of quercetin after intravenous and oral administration in humans are scarce and contradictory [375]. One study on rat showed that quercetin, but not its glycosides, is absorbed from the rat stomach [376]. Another study suggests that quercetin 3-O-beta-glucoside was hydrolysed before or during its intestinal absorption [377]. Also, quercetin and rutin were found in plasma as glucuronides and/or sulfates of quercetin and as unconjugated quercetin aglycone [378].

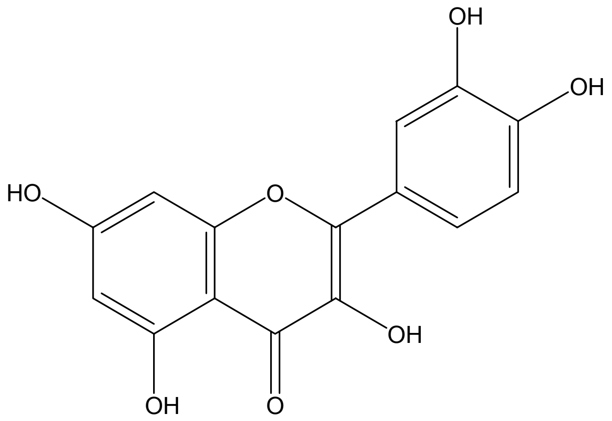


Figure 4.5: Structure of the MP ingredient quercetin.

Eleven therapeutic targets are identified from INVDOCK computation. The results are shown in Table 4.4. Three of them are reported to be inhibited by quercetin directly. They are estrogen receptor, phospholipase A2 and SRC tyrosine kinase[374,379]. SRC tyrosine kinase and estrogen receptor are established therapeutic targets for cancer, which is consistent with the well known anti-cancer effect of quercetin [380]. SRC tyrosine kinase is also a therapeutic target for osteoporosis. Studies on quercetin and its derivatives showed they have beneficial effect on this disease in an animal model [381]. Estrogen receptor has also been explored as a therapeutic target for vascular disease. The efficacy of quercetin in vascular diseases has been well observed [382].

Table 4.4: Potential therapeutic targets of quercetin identified from INVDOCK search of human and mammalian proteins.

PDB ID	Target Name	Experimental Findings	Therapeutic Implications

1a35	DNA topoisomerase I	Quercetin inhibits topoisomerase I - catalyzed DNA religation [383]	Cancer [380]
1akf	Estrogen receptor	Quercetin binds to type II estrogen binding site [379]	Cancer [380] Vascular disease [382]
1azm	Carbonic anhydrase I		Glaucoma, Hypertension [384], Cancer [380]
1bpx	DNA polymerase	Quercetin and myricetin exhibited complex interaction with DNA polymerases [374].	Herpes viral infection [385],[382]
1bvm	Phospholipase A2	Inhibited by quercetin [374].	Inflammation [386] [382]
1csb	Cathepsin B		Autoimmune disease [387], Rheumatoid arthritis
1d3h	Dihydroorotate dehydrogenase		Malaria [388]
1d6n	Hypoxanthine-guanine phosphoribonucleic transferase (HGPRT)	Quercetin induces HGPRT-deficient mutants [389]	Malaria [388]
1klt	Chymase		Asthma [390], Cardiovascular disease [382], Inflammation [382], Dermatitis [391]

1a25	Protein kinase C	Quercetin inhibits the phosphorylating activity [374].	Cancer [380], Vascular disease [382]
1fmk	SRC tyrosine kinase	Inhibited by quercetin [374].	Cancer [380] Osteoporosis [381]

Quercetin has also been reported to inhibit the topoisomerase I catalyzed DNA religation [383], inhibit the phosphorylating activity of protein kinase C [374], as well as exhibit complex interaction with DNA polymerase [374]. These proteins are also identified as potential targets of quercetin by INVDOCK. Therapeutically, they are used in the treatment of cancer, vascular diseases, and viral infection respectively. The uses of quercetin in these diseases are also reported [380,382].

There are 5 other INDOCK predicted therapeutic targets without available experimental finding to either verify or invalidate them. Quercetin is known to induce HGPRT-deficient mutants in rats [389], which might result from the interaction between this protein and quercetin. HGPRT is a therapeutic target for the treatment of malaria. Another therapeutic target for malaria, dihydroorotate dehydrogenase, is also identified as potential target of quercetin. It is noted that experiment has shown that quercetin gave strong antimalarial activity [388]. Other yet-to-be-verified potential targets include chymase, cathepsin B and carbonic anhydrase I. They are therapeutic targets for 9 diseases. Among them the effect of quercetin on hypertension [384], cancer [380], autoimmune diseases [387], asthma [390], cardiovascular diseases [382], inflammation [382] and dermatitis [391] has been reported in animal models or clinical trials.

A number of targets of quercetin unrelated to therapeutics have been reported [374]. The 3D structures of most of these proteins are unavailable and thus beyond the reach of INVDOCK. Nonetheless, the 3D structures of four of these proteins are available. Two of these proteins are identified by INVDOCK as potential targets. They are ribonuclease and nitric oxide synthase. The other two reported targets, aldose reductase and amylase, are missed by INVDOCK. Quercetin has also been reported as a competitor for ATP and GTP in vivo[374]. Therefore it is not surprising that quercetin can bind to some of the ATP-binding or GTP-binding enzymes. INVDOCK identifies two such enzymes as potential targets, which include adenosine kinse and guanylyl cyclase.

4.4.4 Acronycine

Acronycine (Figure 4.6) is an active compound from *Acronychia pedunculata.* It is reported to have anti-cancer effects [392]. Crassum intestine 38 adenocarcinoma cell growth and L1210 leukemia cell growth are reported to be inhibited by this compound through a mechanism of inhibition the synthesis of DNA or RNA [393]. Acronycine is a lipophilic alkaloid. Its etoposide solution was active in multidrug-resistant Chinese hamster ovary cells. The oral bioavailability of an acronycine solution in etoposide diluent was only 50% but both i.p. and p.o. regimens achieved plasma levels greater than 1.0 micrograms/ml. [394].

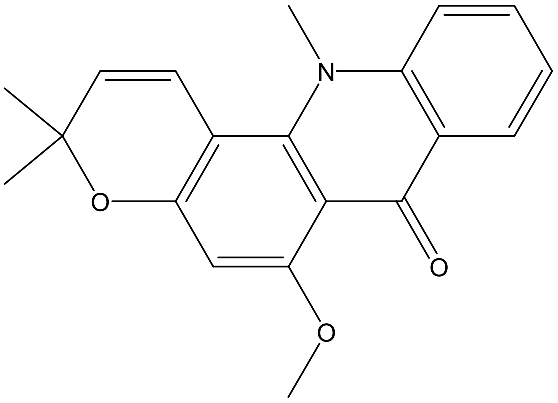


Figure 4.6: Structure of the MP ingredient acronycine.

As shown in Table 4.5, INVDOCK finds three potential therapeutic molecular targets for acronycine. Among them, DNA polymerase is experimentally reported to be a target responsible for its anti-cancer effect [392]. The other two are neither confirmed nor invalidated by experiments. They are beta-catenin, a potential anticancer target, and aldose reductase, a target for diabetes. Further experimental study is needed to clarify whether or not these are targets of acronycine.

Table 4.5: Potential therapeutic targets of acronycine identified from INVDOCK search of human and mammalian proteins.

PDB ID	Target Name	Experimental Findings	Therapeutic Implications
2acq	Aldose reductase		Diabetes
3bct	Beta-catenin		Cancer [392]
1zqo	DNA polymerase beta	Inhibition of DNA synthesis.[393]	Viral infection

4.4.5 Baicalin

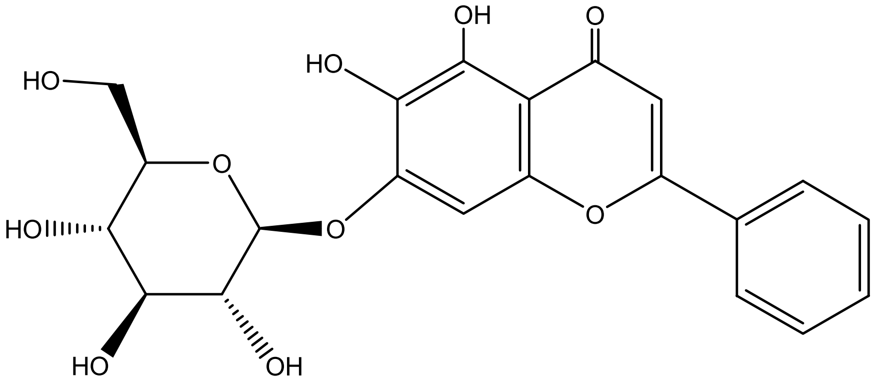


Figure 4.7: Structure of the MP ingredient baicalin.

Baicalin (Figure 4.7) is an active compound from *Scutellaria baicalensis* or *Oroxylum indicum*. It is reported to have anti-cancer [395], anti-inflammation [396], anti-AIDS effects [397], and has been used in the treatment of diabetes [398] and liver problems [399]. Baicalin is absorbed from the rat gastrointestinal tract as the aglycone and restored to its original form [400]. One study showed that its plasma concentration reached a peak of 0.42 microgram/mL 5.3 h after oral administration, 600 mg/kg [401].

As shown in Table 4.6, INVDOCK finds 9 potential therapeutic protein targets for baicalin. Two of these targets were inhibited by baicalin. One is DNA polymerase beta, an anti-virus target, which could be weakly inhibited by baicalin [402]. The other is aldose reductase [399], a target for the treatment of diabetes [398]. It has been reported that baicalin has certain effects on two other therapeutic targets suggested by INVDOCK. Baicalin has been found to down-regulate the expression level of cyclin-dependent kinase 2 [403], which are known anti-cancer targets. This compound has also been reported to have an inhibitory effect on phospholipase A2 [404,405], which is a known anti-inflammatory target. The anti-cancer and anti-inflammatory effects of binding of baicalin to these implicated targets have been observed experimentally [395,396]. The other 5 predicted targets are neither confirmed nor invalidated by experiments. They are all potential anti-cancer targets. Among them, protein kinase C and adenylyl cyclase are also reported to be therapeutic targets in vascular diseases. Baicalin was observed to have an effect on the contractility of rat isolated mesenteric arteries[406]. Further study is needed to determine whether these are targets of baicalin.

Table 4.6: Potential therapeutic targets of baicalin identified from INVDOCK search of human and mammalian proteins.

PDB ID	Target Name	Experimental Findings	Therapeutic Implications
121p	H-Ras p21 protein		Cancer [395]
1ads	Aldose reductase	Baicalin reduced RBC sorbitol levels in diabetic rats as inhibitor of Aldose Reductase.[399]	Diabetes [398]
1agw	FGF receptor 1		Cancer [395]
1a25	Protein kinase C		Vascular disease[406]，Cancer [395]
1awk	Adenylyl cyclase		Vascular disease [406], Heart failure, Erectile dysfunction

1awn	Guanylyl cyclase		Cancer [395]
1irb	Phospholipase A2	Inhibition effect. [404,405]	Inflammation [396]
2bpf	DNA polymerase Beta	Weak inhibition[402]	Viral infection [402]
1jsu	Cyclin-dependent kinase-2	Baicalin decreases expression level of cyclin-dependent kinase[403]	Cancer [395]

4.4.6 Emodin

Emodin (Figure 4.8) is an active compound from *Rheum palmatum*, *Rumex dentatus* and *Cassiatora*. It has been found to have anti-cancer [407], immuno-modulation [408] and laxative effects. Administration of emodin to rabbits by i.v. bolus resulted in a serum profile which could be well described by a two-compartment model. Oral administration of emodin resulted in a very low serum concentration but protein binding assays show that emodin was highly bound to serum protein [409].Liver, kidney and intestinal tract showed higher concentrations than plasma [410].

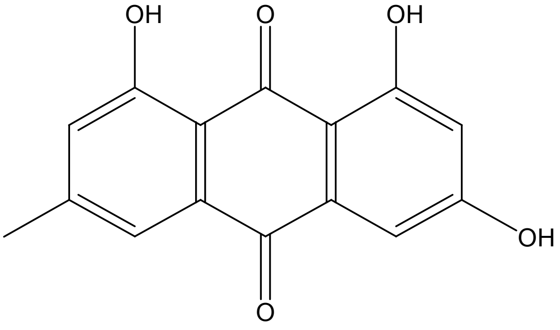


Figure 4.8: Structure of the MP ingredient emodin.

As shown in Table 4.7, INVDOCK identifies five potential therapeutic protein targets for emodin. Two of them have relevant publications, which are protein kinase C [411], and nuclear factor Kappa-B [412]. Emodin was reported to inhibit protein kinase C[411]. Protein kinase C has been explored as a therapeutic target in cancer and vascular diseases. Emodin's beneficial effect in cancer were observed in mice [407]. Emodin was reported to inhibit TNF-induced NF-kappaB activation [412]. NF-kappaB played an important role in a number of diseases. Among them, emodin's beneficial effects in inflammation [413] and atherosclerosis[414] were reported. The other three targets are neither confirmed nor invalidated by experiments. They are therapeutic targets for cancer [407], inflammation [413] and diabetes. It remains to be seen whether or not these are targets of emodin as predicted by INVDOCK.

Table 4.7: Potential therapeutic targets of emodin identified from INVDOCK search of human and mammalian proteins.

PDB ID	Target Name	Experimental Findings	Therapeutic Implications
2acq	Aldose reductase		Diabetic treatment
1pth	Prostaglandin H2 Synthase-1		Inflammation [413]
3bct	Beta-catenin		Cancer [407]

1nfk	Nuclear factor Kappa-B	Emodin inhibits TNF-induced NF-kappaB activation [412]	Inflammation [413], Asthma, Atherosclerosis[414], Neurodegenerative disorders, allergic rhinitis, Migraine
1a25	Protein kinase C	Inhibitor[411]	Cancer [407], Vascular disease

4.4.7 Allicin

Allicin (Figure 4.9) is a bioactive compound from garlic with a number of therapeutic effects. It is known to reduce blood cholesterol, triglycerides levels and systolic blood pressure in hypercholesterolemic rats [415]. This compound has been shown to possess antimicrobial activities [416] especially against H. pylori [417]. It selectively inhibits the GSH-dependent PGH2 to PGE2 isomerase in adenocarcinoma cell line, which has implication in pulmonary vasodilating, anti-inflammatory as well as anti-cancer effects [418]. It has also been found to be an antioxidant agent [419].

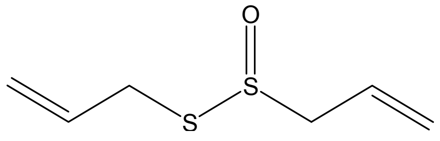


Figure 4.9: Structure of the MP ingredient allicin.

Putative human and mammalian therapeutic protein targets of allicin identified by INVDOCK are given in Table 4.8 along with the respective clinical implications from experiments. Four putative protein targets are identified. One of them, insulin, seems to be implicated by experiment. It has been observed that the level of insulin is increased by allicin [420].

Table 4.8: Potential therapeutic targets of allicin identified from INVDOCK search of human and mammalian proteins.

PDB ID	Target Name	Experimental Findings	Therapeutic Implications
1znj	Insulin	Increased insulin level [420]	Diabetes [420]
1ah3	Aldose Reductase		Diabetes
1cdk	CAMP-Dependent Protein Kinase		Cancer
1rpa	Prostatic Acid Phosphatase		Cancer (prostate cancer)

INVDOCK also finds aldose reductase, CAMP-dependent protein kinase, and prostatic acid phosphatase as putative therapeutic targets. However there is no experimental study to either implicate or invalidate each of these targets. Interaction of allicin with insulin as well as aldose reductase has implications for diabetes, which is consistent with the observed effect of allicin on diabetes [421]. CAMP-dependent protein kinase and prostatic acid phosphatase have implications in anticancer effects, which is consistent with observed anticancer effects of garlic [422]. Further experimental investigation is needed to test whether these three proteins are targets of allicin.

4.4.8 Catechin

Catechin (Figure 4.10), also known as cyanidol, is an active compound from green tea. It has been shown to inhibit the growth of human breast cancer cells [423] and prostate cancer cells [424] partly because of its inhibition of cyclin-dependent kinases [425]). The antitumor activity of this compound may also arise from its inhibition of tyrosine phosphorylation of PDGF beta-receptor [426], induction of apoptosis [427] and inhibition of matrix metalloproteinases [428]. Catechin exhibits anti-inflammatory as well as cancer chemopreventive effects in many animal tumor bioassays, cell culture systems, and epidemiological studies [429]. Some of these effects of catechin are in part from its inhibition of TNF-alpha and NF-kappaB. This compound also has antiplaque and hepatoprotective effects via reduction of membrane fluidity [430]. It has been reported that this compound has antioxidative action mediated by the activation of glutathione peroxidase [431].

INVDOCK search produces seventeen putative therapeutic targets, which are given in Table 4.9. Seven of these targets have been confirmed by experiments, which showed that catechin inhibits each of them. These include cyclin-dependent kinase and FGF receptor [425], neutrophil collagenase [428], protein kinase C [425], CAMP-dependent protein kinase [432] TNF-alpha and NF-kappaB p65 [429]. Inhibition of each of the first five proteins has potential anticancer implications, binding to the sixth protein may produce anti-inflammatory effects, and the interaction with the seventh and eighth proteins may lead to anti-inflammatory as well as cancer chemopreventive effects.

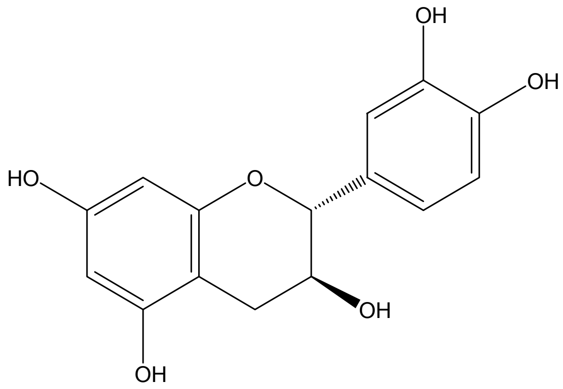


Figure 4.10: Structure of MP ingredient catechin.

Table 4.9: Potential therapeutic targets of catechin identified from INVDOCK search of human and mammalian proteins.

PDB ID	Target Name	Experimental Findings	Therapeutic Implications
1ads	Aldose Reductase		Diabetes
1agw	FGF Receptor 1	Inhibitor[433]	Cancer[433]
1aqc	X11		Anti-clotting
1crp	C-H-Ras p21 Protein	Inhibition of Ras-transformed cells [434]	Cancer [434]
1mcc	Immunoglobulin Lambda Light Chain	Immunoenhancing effect on T and B cell functions [435]	Enhanced immune response [435]
1mnc	Neutrophil Collagenase	Collagenase inhibitor [428]	Cancer [428]
1a25	Protein Kinase C	Inhibitory effect [436]	Cancer [436], Vascular disease

1awn	Guanylyl Cyclase		Cancer
1cdk	CAMP-Dependent Protein Kinase	Inhibitor [432]	Cancer [432]
1p38	MAP kinase p38	Inhibition of activation of p38 mitogen activated protein kinase.	Cancer
1rpa	Prostatic acid phosphatase		Cancer (prostate cancer)
1wav	Insulin	Activator [429]	Diabetes [429]
1ydt	C-AMP-dependent protein kinase		Cancer
1cpj	Cathepsin B	Activator [429]	Cancer [429]
1jsu	Cyclin-dependent kinase-2	Inhibitor [433]	Cancer [433]
1ram	Transcription factor NF-KB p65	Inhibitor [429]	Inflammation [429]
2tnf	Tumor necrosis factor alpha	Inhibitor [429]	Inflammation [429]

Available experimental data also seem to implicate another five of INVDOCK identified therapeutic targets. Catechin has been found to be an activator of cathepsin B and insulin [429]. It is possible that activation of each protein is by direct binding of catechin. Activation of cathepsin may produce anticancer effects, while activation of insulin may help reduce glucose levels and thus have implications in diabetes treatment. Catechin is known to inhibit both the ras-transformed cells and the activation of p38 mitogen-activated protein kinase [434], which may have implication in anticancer properties. One possible reason for these inhibitory effects are due to the binding of catechin to ras p21 protein and MAP kinase p38 respectively as predicted by INVDOCK. Catechin is also known to have immuno-enhancing effect on T and B cell functions [435], which may also result from binding of catechin to immunoglobulin lambda light chain as predicted by INVDOCK. Further investigation is needed to determine whether these proteins are targets of catechin.

Moreover, INVDOCK identified four additional putative therapeutic targets. These are aldose reductase, X11, guanylyl cyclase, and C-AMP-dependent protein kinase. No experimental information has been found to either implicate or invalidate them. Hence, further study is needed to determine whether these proteins are targets of catechin. The potential therapeutic effect of the binding of catechin to each of these proteins is diabetes treatment for the aldose reductase, anti-clotting for X11, and anticancer for guanylyl cyclase and C-AMP-dependent protein kinase respectively.

Some known therapeutic targets of catechin are not found by INVDOCK search. These include matrix metalloproteinase-2, matrix metalloproteinase-9, matrix metalloproteinase-12 and glutathione peroxidase. This occurs because of a lack of relevant structures in the database. The cavity database does not yet have 3D structures of matrix metalloproteinase-2, matrix metalloproteinase-9 and matrix metalloproteinase-12. Although the 3D structures of glutathione peroxidase are available in the database, these are ligand-free structures that may not be a suitable system for accurate analysis of the binding of a compound that affects the function of that protein. Thus these structures are not used in INVDOCK study.

4.4.9 Camptothecin

Camptothecin (Figure 4.11) is a compound from the plant *Canptotheca acuninata*. It has well recognized antitumour activities and has been evaluated in clinical trials [437].

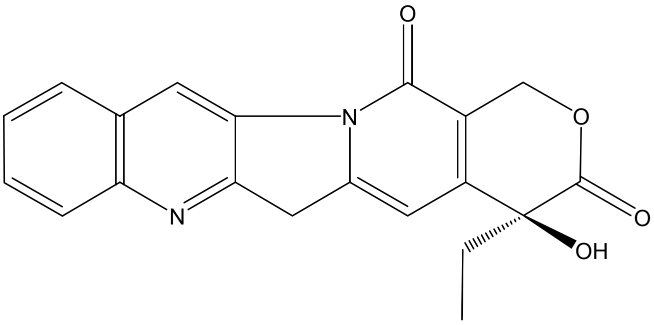


Figure 4.11: Structure of MP ingredient camptothecin.

As shown in Table 4.10, INVDOCK identifies nine putative therapeutic protein targets, eight of which has anticancer implication. Two such putative targets have been confirmed experimentally. These are topoisomerase I [438] and protein kinase C [439,440]. Another identified putative target is implicated by experiment. It has been shown that camptothecin inhibits the activity of calpain [441], which may be indicative of direct binding of camptothecin to this protein. Such a binding is expected to induce apoptosis in leukemic cells.

Table 4.10: Potential therapeutic targets of camptothecin identified from INVDOCK search of human and mammalian proteins.

PDB ID	Target Name	Experimental Findings	Therapeutic Implications
1ads	Aldose Reductase		Diabetes
2gss	Glutathione S-Transferase p1-1	Increased intracellular glutathione [442]	Cancer [442]
7ice	DNA Polymerase Beta		Cancer
1a25	Protein Kinase C	Inhibitor [440]	Cancer [439], Vascular disease
1cdk	CAMP-Dependent Protein Kinase		Cancer
3bct	Beta-Catenin		Cancer
1dvi	Calpain	Inhibition of calpain activities.[441]	Cancer.[441]
1yfo	Receptor Protein Tyrosine Phosphatase	Caused elevation of PTPase in the cytosol and the nucleus which play a critical role in the induction of the differentiation of IW32 erythroleukemia cells.[443]	Cancer [443]
1a35	Topoisomerase I	Inhibitor [438]	Cancer [438]

Two additional putative targets seem to be implicated by experiments as well. Camptothecin has been found to elevate the level of protein tyrosine phosphatase

[443] and to increase intracellular glutathione [442], which might result from its binding to receptor protein tyrosine phosphatase and glutathione S-transferase respectively as predicted by INVDOCK. Binding to these two proteins may have anticancer implication. For instance, it has been found that camptothecin causes elevation of PTPase in the cytosol and the nucleus, which affects the induction of the differentiation of IW32 erythroleukemia cells. Therefore INVDOCK prediction of these two putative targets may partly explain the observed anticancer activities of camptothecin.

Other identified potential therapeutic targets are DNA polymerase beta, CAMP-dependent protein kinase, beta-catenin and aldose reductase. There is no experimental information to either implicate or invalidate these targets. Further study is therefore needed to clarify this. Anticancer activity may potentially be produced by camptothecin binding to DNA polymerase beta, CAMP-dependent protein kinase and beta-catenin respectively. The effect of camptothecin on aldose reductase may have implication in diabetes treatment.

4.5 Limitations and suggested improvement of INVDOCK

Table 4.11 summarizes the comparison between INVDOCK predictions and available experimental findings for the nine MP ingredients presented here. Overall about 51% of INVDOCK identified potential therapeutic targets of these MP ingredients have relevant experimental findings. Moreover, about 70% of the identified therapeutic implications related to these targets have been reported to occur in cultivated cells, animal models or clinical trails. It seems that INVDOCK is capable of providing useful information for experimental researchers in probing the mechanisms of MPs.

Table 4.11: Statistics of therapeutic targets of selected bioactive MP ingredients identified by INVDOCK search. The statistics of experimentally reported or implicated targets is also given for comparison.

MP Ingredient	No. of therapeutic INVDOCK identified targets	No. of therapeutic targets that have relevant literature support	No. of therapeutic effects related to INVDOCK identified targets	No. of therapeutic effects that have relevant literature support
Genistein	18	7	9	8
Ginsenoside Rg1	3	2	3	3
Quercetin	11	6	12	10
Acronycine	3	1	4	2
Baicalin	9	4	7	5
Emodin	5	2	11	4
Allicin	4	1	2	1
Catechin	17	12	5	4
Camptothecine	9	5	3	2
Total	79	40	56	39

Discrepancy between INVDOCK results and available experimental data arises from a number of reasons. It is not expected that exhaustive experiments have been done to determine all protein targets of the studied MP ingredients. The lack of sufficient experimental data is likely an important factor for the discrepancy. Lack of relevant protein structures is likely to be another factor. 3D structure of a large number of known therapeutic targets is not available. Some of the 3D structures may be of little relevance here. These include entries containing incomplete sections or chains, protein mutants that are structurally different from the corresponding proteins investigated in experiments, ligand-bound proteins whose conformations are relevant only to a specific set of compounds, and macromolecular complexes unrelated to a particular biological process studied experimentally. “False hits” may thus be generated if these irrelevant structures are selected by INVDOCK. Anticipated rapid progress in structural genomics[320] is expected to provide a more diverse set of relevant structures. Knowledge from study of protein functions also facilitates the selection of relevant structures in determination of potential protein targets related to a particular cellular or physiological condition.

As in other docking studies, INVDOCK does not take protein profiles, such as gene expression pattern and protein levels, into consideration, which may also be a source of discrepancy between INVDOCK computation and experiments. Some experimental studies of MP ingredients are based on the investigation of cell lines or other assays. Observation of molecular events related to the interaction with a particular protein requires that the protein be expressed at a sufficient level in the system being investigated. If such a level is not reached at a particular setting, the corresponding experiment is not useful in probing the binding of a compound to that protein. Proteins not expressed or at too low levels are unlikely to be a detectable target. Advance in proteomics is providing rapidly growing information about the profiles of proteins inside cells[223]. Incorporation of this information into the INVDOCK procedure may be helpful in improving the prediction accuracy.

Therapeutic action of a chemical requires it to achieve an adequate concentration in the body fluid bathing the target tissue. The concentration of a chemical is determined by its pharmacokinetic profile. Also these chemicals might undergo some extent of metabolism. In some cases, both the original chemical and the metabolited derivative could have the same effect through the same mechanism (e.g. Tamoxifen and 4-H-tamoxifen are both anti-cancer agents). However, in some cases, the metabolited derivative would lose its activity (e.g. quercetin gave strong antimalarial activity, however, its glucosides, showed little significant activity [388]). Therefore the ADME profile, which has been neglected in INVDOCK and other docking studies, needs to be considered. Rapid progress in our understanding of pharmacokinetics and drug metabolism [444] is providing useful information in this regard.

4.6 Summary

A number of *in silico* methods for target identification are being explored or under consideration. Several methods have potential applications in facilitating the identification of therapeutic targets of MP ingredients. These include high-throughput assay based approaches, extended QSAR approaches, statistical learning methods and extended docking methods. One of these methods, INVDOCK, has been specifically used in the identification of protein targets of MP ingredients as well as synthetic chemicals. Testing results suggest the usefulness of INVDOCK as an *in silico* tool in facilitating the identification of potential therapeutic targets of the MP ingredients and thus providing valuable clues to the mechanisms of herbal medicines and their possible secondary therapeutic effects. This may greatly facilitate the mechanistic study of herbal medicines. Performance and applicability of *in silico* methods may be further improved by incorporation of new information from advances in structural genomics, proteomics, protein function, and pharmacokinetics. Efficiency and accuracy of *in silico* methods in analyzing the mechanisms of herbal medicines can also be enhanced from new progress in computational algorithms, parameters, and more accurate models of the interaction between a target and its binding molecules.

Chapter 5

Summary

The course of new drug discovery is still inefficient and costly nowadays. Rational drug design has been introduced to facilitate this process. The essence of rational drug design is the reasoned extrapolation of our knowledge of targeted receptors and "lead" structures to suggest novel chemical structures with defined characteristics as potential drugs.

The selection of therapeutic targets to be worked with is therefore very important. In contemporary new drug discovery processes, it is the first stage of R&D leading to a new drug and directs further investigations. The appropriate selection of effective therapeutic targets and efficient therapeutic intervention strategies is receiving more and more attention. With this regard, this work explored the use of *in silico* approaches in facilitating relevant research.

First, a comprehensive information source of known therapeutic target information devoted to new drug discovery will be undoubtedly helpful to the relevant research communities. However, existing drug target information is still scattered among the huge quantity of biomedical literature. Work needs to be done to collect and sort known drug target information to provide an easy access to relevant communities. As a fundamentally important task, a database of known therapeutic target information, Therapeutic Target Database, was curated. A relational data model was designed specifically for it which aims to maximize the ability to accommodate future extensions and facilitate the integration of information.

Rapid discovery of new therapeutic targets is also very important as it may not only introduce more efficient therapeutic targets for certain diseases, but also increase the flexibility in the design of novel therapeutic intervention strategies by exploiting the synergies between known and newly discovered targets. With known examples of therapeutic targets, statistical learning methods might be able to learn the common features of therapeutic targets and predict the drug-target like proteins in the human genome. This, if feasible, would greatly help the rapid discovery of therapeutic targets based on the human genomic data. A number of statistical learning methods and pre-processing techniques are explored for the application of drug-target like protein prediction. Among them, the support vector machine approach gives the best classification results which are reasonably good to facilitate the *in-sillico* genome scale drug-target screening. Performance of the statistical learning methods may be further improved by incorporation of new information from advances in pharmaceutical sciences and proteomics. The accuracy of statistical learning methods in prediction of drug-target like proteins can also be enhanced from new progress in learning algorithms, descriptors, and pre-processing techniques.

Besides more effective therapeutic targets, delicate therapeutic intervention involving multiple cooperating targets may also help to improve the treatment efficacy. Novel therapeutic mechanisms discovered from studies of herbal medicines have routinely been used in new drug discovery. With known drug target information, *in silico* approaches may also be used in the study of novel medicinal plant mechanisms. While a number of approaches have the potential in this application, our testing results on one of them, the INVDOCK approach, suggests its usefulness in facilitating the identification of potential therapeutic targets of MP ingredients and thus providing valuable clues to the mechanisms of effective herbal medicines. Performance and applicability of *in silico* methods may be further improved with new advances in structural genomics, proteomics, protein function, and pharmacokinetics. Efficiency and accuracy of *in silico* methods in analyzing herbal medicine mechanisms can also be enhanced from new progress in computational algorithms, parameters, and more accurate models of the interaction between a target and its binding molecules.

Currently, the computer aided drug design approaches mainly focus on the structure properties of a drug target and its possible binder to find or design a chemical that could bind the target tightly. Essentially, they are based on the “lock and key” principle proposed by Fisher more than 100 years ago [445]. This principle has such a big influence on medicinal chemists and drug design experts that when a new drug needs to be designed, they always think of the receptors (drug targets) first, and then design a “key” to that receptor to treat that disease.

However, drug designing approaches based on the “lock and key” principle have their innate deficiency. For example, theoretically, drug-receptor interactions are dynamic processes. That is to say, a drug must pass through a “cavity” to reach the active site of the receptor. It means when the drug is passing the “cavity”, the drug-receptor complex is probably in a transition state. The structure of this transition complex can not yet be determined experimentally or be modeled by effective algorithms. Also, the kinetic process of how the drug is administrated, transported, metabolized and excreted is not considered. In other words, the processes prior to and after drug–receptor interaction have not yet been paid enough attention, which are also very important factors that determine the chance of finding a successful drug. Therefore, while the modern drug designing approaches indeed helped much in the research and development of new drugs, the rate of success is still low.

Introducing the consideration of drug mechanisms into rational drug design becomes a popular idea among drug design experts, which is recognized as mechanism based drug design (MBDD). In this regard, the drug target directed *in silico* approaches discussed in this work can be viewed as part of the efforts to embody therapeutic mechanism based drug design. Besides drug targets, the other important factors affecting the success of a drug, such as its ADME profile, toxicity and drug-drug interaction profile, are also critical to the new drug discovery process. Novel approaches incorporating the consideration of these factors into the early stages of the drug discovery process would therefore be expected to further improve the research and development efficiency, which would be interesting topics that follow this work.

References

[1] Drews, J. and Ryser, S., The role of innovation in drug development, *Nat Biotechnol*, 15 (1997) 1318-9.

[2] D, d.P., "Innovate or die" is the first rule of international industial competetion, *Research Technology Management*, 37 (1994) 9-11.

[3] Yevich, J.P., Drug development: from discovery to marketing. In P.L. Krogsgaard-Larsen, T; Madsen,U; (Ed.), *A textbook of drug design and development*, Harwood academic, Australia, 1996, pp. 508.

[4] Ehrlich, P., *Gesammelte Arbeiten zur Immunit鋞sforschung,* A. Hirschwald, Berlin, 1904, 776 pp.

[5] Clore, G.M. and Gronenborn, A.M., Determination of three-dimensional structures of proteins and nucleic acids in solution by nuclear magnetic resonance spectroscopy, C*rit Rev Biochem Mol Biol,* 24 (1989) 479-564.

[6] Aboul-ela, F. and Varani, G., Novel techniques in nuclear magnetic resonance for nucleic acids, C*urr Opin Biotechnol,* 6 (1995) 89-95.

[7] Zhukov, A.V. and Vereschagin, A.G., Current techniques of extraction, purification, and preliminary, fractionation of polar lipids of natural origin, A*dv Lipid Res,* 18 (1981) 247-82.

[8] Hughes, I. and Hunter, D., Techniques for analysis and purification in high-throughput chemistry, C*urr Opin Chem Biol,* 5 (2001) 243-7.

[9] Tadey, T. and Purdy, W.C., Chromatographic techniques for the isolation and purification of lipoproteins, J *Chromatogr B Biomed Appl,* 671 (1995) 237-53.

[10] Friedrich, C. and Seidlein, H.J., [The history of pharmaceutical science. 12. The importance of the discovery of morphine in the development of pharmaceutical science], P*harmazie,* 39 (1984) 340-5.

[11] Sneader, W., D*rug discovery : the evolution of modern medicines,* Wiley, Chichester ; New York, 1985, x, 435 pp.

[12] Diggins, F.W.E., The true history of the discovery of penicillin, with refutation of the misinformation in the literature, B*ritish Journal of Biomedical Science,* 56 (1999) 83-93.

[13] Cohen, S.S., A guide to the history of biochemistry, I*sis,* 91 (2000) 120-124.

[14] Chen, B., Piletsky, S. and Turner, A.P., High molecular recognition: design of "Keys", C*omb Chem High Throughput Screen,* 5 (2002) 409-27.

[15] Summan, M. and Cribb, A.E., Novel non-labile covalent binding of sulfamethoxazole reactive metabolites to cultured human lymphoid cells, C*hem Biol Interact,* 142 (2002) 155-73.

[16] Langer, T. and Hoffmann, R.D., Virtual screening: an effective tool for lead structure discovery?, C*urr Pharm Des,* 7 (2001) 509-27.

[17] Kenny, B.A., Bushfield, M., Parry-Smith, D.J., Fogarty, S. and Treherne, J.M., The application of high-throughput screening to novel lead discovery, P*rog Drug Res,* 51 (1998) 245-69.

[18] Marshall, G.R., Computer-aided drug design, A*nnu Rev Pharmacol Toxicol,* 27 (1987) 193-213.

[19] Loew, G.H., Villar, H.O. and Alkorta, I., Strategies for indirect computer-aided drug design, P*harm Res,* 10 (1993) 475-86.

[20] Jackson, R.C., Update on computer-aided drug design, C*urr Opin Biotechnol,* 6 (1995) 646-51.

[21] Veselovsky, A.V. and Ivanov, A.S., Strategy of computer-aided drug design, C*urr Drug Targets Infect Disord,* 3 (2003) 33-40.

[22] Vedani, A., [Computer-Aided Drug Design: An Alternative to Animal Testing in the Pharmacological Screening], A*ltex,* 8 (1991) 39-60.

[23] Ooms, F., Molecular modeling and computer aided drug design. Examples of their applications in medicinal chemistry, C*urr Med Chem,* 7 (2000) 141-58.

[24] Cohen, N.C. and Tschinke, V., Generation of new-lead structures in computer-aided drug design, P*rog Drug Res,* 45 (1995) 205-43.

[25] Myers, P.L., Will combinatorial chemistry deliver real medicines?, C*urr Opin Biotechnol,* 8 (1997) 701-7.

[26] Gallop, M.A., Barrett, R.W., Dower, W.J., Fodor, S.P. and Gordon, E.M., Applications of

combinatorial technologies to drug discovery. 1. Background and peptide combinatorial libraries, J *Med Chem,* 37 (1994) 1233-51.

[27] Gordon, E.M., Barrett, R.W., Dower, W.J., Fodor, S.P. and Gallop, M.A., Applications of combinatorial technologies to drug discovery. 2. Combinatorial organic synthesis, library screening strategies, and future directions, J *Med Chem,* 37 (1994) 1385-401.

[28] Krogsgaard-Larsen, P., Liljefors, T. and Madsen, U., T*extbook of drug design and discovery,* 3rd edn., Taylor & Francis, London ; New York, 2002, xviii, 572 pp.

[29] Sali, A., 100,000 protein structures for the biologist, N*at Struct Biol,* 5 (1998) 1029-32.

[30] Pieper, U., Eswar, N., Stuart, A.C., Ilyin, V.A. and Sali, A., MODBASE, a database of annotated comparative protein structure models, N*ucleic Acids Res,* 30 (2002) 255-9.

[31] Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. and Ferrin, T.E., A geometric approach to macromolecule-ligand interactions, J *Mol Biol,* 161 (1982) 269-88.

[32] Lybrand, T.P., Ligand-protein docking and rational drug design, C*urr Opin Struct Biol,* 5 (1995) 224-8.

[33] Jones, G. and Willett, P., Docking small-molecule ligands into active sites, C*urr Opin Biotechnol,* 6 (1995) 652-6.

[34] Goodsell, D.S., Morris, G.M. and Olson, A.J., Automated docking of flexible ligands: applications of AutoDock, J *Mol Recognit,* 9 (1996) 1-5.

[35] Nussinov, R. and Wolfson, H.J., Efficient computational algorithms for docking and for generating and matching a library of functional epitopes II. Computer vision-based techniques for the generation and utilization of functional epitopes, C*omb Chem High Throughput Screen,* 2 (1999) 261-9.

[36] Abagyan, R. and Totrov, M., High-throughput docking for lead generation, C*urr Opin Chem Biol,* 5 (2001) 375-82.

[37] Schneider, G. and Bohm, H.J., Virtual screening and fast automated docking methods, D*rug Discov Today,* 7 (2002) 64-70.

[38] Taylor, R.D., Jewsbury, P.J. and Essex, J.W., A review of protein-small molecule docking methods, J *Comput Aided Mol Des,* 16 (2002) 151-66.

[39] Baxter, C.A., Murray, C.W., Clark, D.E., Westhead, D.R. and Eldridge, M.D., Flexible docking using Tabu search and an empirical estimate of binding affinity, P*roteins,* 33 (1998) 367-82.

[40] Lorber, D.M. and Shoichet, B.K., Flexible ligand docking using conformational ensembles, P*rotein Sci,* 7 (1998) 938-50.

[41] Wang, J., Kollman, P.A. and Kuntz, I.D., Flexible ligand docking: a multistep strategy approach, P*roteins,* 36 (1999) 1-19.

[42] Lawrence, M.C. and Davis, P.C., CLIX: a search algorithm for finding novel ligands capable of binding proteins of known three-dimensional structure, P*roteins,* 12 (1992) 31-41.

[43] Bohm, H.J., LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads, J *Comput Aided Mol Des,* 6 (1992) 593-606.

[44] Lauri, G. and Bartlett, P.A., CAVEAT: a program to facilitate the design of organic molecules, J *Comput Aided Mol Des,* 8 (1994) 51-66.

[45] Nishibata, Y. and Itai, A., Confirmation of usefulness of a structure construction program based on three-dimensional receptor structure for rational lead generation, J *Med Chem,* 36 (1993) 2921-8.

[46] Gehlhaar, D.K., Moerder, K.E., Zichi, D., Sherman, C.J., Ogden, R.C., et al., De novo design of enzyme inhibitors by Monte Carlo ligand generation, J *Med Chem,* 38 (1995) 466-72.

[47] Waszkowycz, B., Structure-based approaches to drug design and virtual screening, C*urr Opin Drug Discov Devel,* 5 (2002) 407-13.

[48] van Dongen, M., Weigelt, J., Uppenberg, J., Schultz, J. and Wikstrom, M., Structure-based screening and design in drug discovery, D*rug Discov Today,* 7 (2002) 471-8.

[49] Isiguro, M., [Structure-based drug design], T*anpakushitsu Kakusan Koso,* 45 (2000) 880-6.

[50] Kirkpatrick, D.L., Watson, S. and Ulhaq, S., Structure-based drug design: combinatorial chemistry and molecular modeling, C*omb Chem High Throughput Screen,* 2 (1999) 211-21.

[51] Wlodawer, A. and Vondrasek, J., Inhibitors of HIV-1 protease: a major success of structure-assisted drug design, A*nnu Rev Biophys Biomol Struct,* 27 (1998) 249-84.

[52] Wade, R.C., 'Flu' and structure-based drug design, S*tructure,* 5 (1997) 1139-45.

[53] Marrone, T.J., Briggs, J.M. and McCammon, J.A., Structure-based drug design:

computational advances, A*nnu Rev Pharmacol Toxicol,* 37 (1997) 71-90.

[54] Blundell, T.L., Structure-based drug design, N*ature,* 384 (1996) 23-6.

[55] Lipnick, R.L., Correlative and mechanistic QSAR models in toxicology, S*AR QSAR Environ Res,* 10 (1999) 239-48.

[56] Podlogar, B.L. and Ferguson, D.M., QSAR and CoMFA: a perspective on the practical application to drug discovery, D*rug Des Discov,* 17 (2000) 4-12.

[57] Vedani, A. and Dobler, M., Multi-dimensional QSAR in drug research. Predicting binding affinities, toxicity and pharmacokinetic parameters, P*rog Drug Res,* 55 (2000) 105-35.

[58] Selassie, C.D., Mekapati, S.B. and Verma, R.P., QSAR: then and now, C*urr Top Med Chem,* 2 (2002) 1357-79.

[59] Kellogg, G.E. and Semus, S.F., 3D QSAR in modern drug design, E*xs* (2003) 223-41.

[60] Li, Y. and Harte, W.E., A review of molecular modeling approaches to pharmacophore models and structure-activity relationships of ion channel modulators in CNS, C*urr Pharm Des,* 8 (2002) 99-110.

[61] Froimowitz, M., The pharmacophore for opioid activity, N*IDA Res Monogr,* 134 (1993) 178-94.

[62] Dannhardt, G. and Laufer, S., Structural approaches to explain the selectivity of COX-2 inhibitors: is there a common pharmacophore?, C*urr Med Chem,* 7 (2000) 1101-12.

[63] Kurogi, Y. and Guner, O.F., Pharmacophore modeling and three-dimensional database searching for drug design using catalyst, C*urr Med Chem,* 8 (2001) 1035-55.

[64] Traxler, P., Furet, P., Mett, H., Buchdunger, E., Meyer, T., et al., Design and synthesis of novel tyrosine kinase inhibitors using a pharmacophore model of the ATP-binding site of the EGF-R, J *Pharm Belg,* 52 (1997) 88-96.

[65] Malawska, B. and Scatturin, A., Application of pharmacophore models for the design and synthesis of new anticonvulsant drugs, M*ini Rev Med Chem,* 3 (2003) 341-8.

[66] Abraham, D.J. and Burger, A., B*urger's medicinal chemistry and drug discovery,* 6th edn., Wiley, Hoboken, N.J., 2003.

[67] Hansch, C., Maloney, P.P. and Fujita, T., Correlation of biological activity of phenoxy-acetic acids with Hammett substituent constants and partition coeficients.,

N*ature,* 194 (1962) 178.

[68] Hansch, C. and Fujita, T., P-delta-pi analysis -- correlations of biological activity and chemical structure, J*. Am. Chem. Soc.,* 86 (1964) 1616.

[69] Hansch, C.H. and Leo, A., S*ubstituent constants for correlation analysis in chemistry and biology,* Wiley, New York, 1979, vii, 339 pp.

[70] Chou, J.T. and Jurs, P.C., Computer-assisted computation of partition coefficients from molecular structure using fragment constants, J*. Chem. Inf. Comput. Sci.,* 19 (1979) 172.

[71] Hammett, L.P., P*hysical organic chemistry : reaction rates, equilibria, and mechanisms,* 2nd edn., McGraw-Hill, New York, 1970, 420 pp.

[72] Swain, C.G. and Lupton, E.C.J., Field and resonance compounds of substituent effects, J*. Am. Chem. Soc.,* 90 (1968) 4328.

[73] Newman, M.S., S*teric effects in organic chemistry,* Wiley, New York, 1963, vii, 710 pp.

[74] Ari 鐨 ns, E.J., Dr*ug design,* Academic Press, New York,, 1971, v. pp.

[75] Randic, M., Characterization of molecular branching, J. *Am. Chem. Soc.,* 97 (1975) 6609.

[76] Kier, L.B., Hall, L.H. and Murray, W.J., Molecular connectivity. I. Reaction to non-specific local anesthesia, J. *Pharm. Sci.,* 64 (1975) 1971.

[77] Luco, J.M. and Ferretti, F.H., QSAR based on multiple linear regression and PLS methods for the anti-HIV activity of a large group of HEPT derivatives, J *Chem Inf Comput Sci,* 37 (1997) 392-401.

[78] Hasegawa, K., Matsuoka, S., Arakawa, M. and Funatsu, K., New molecular surface-based 3D-QSAR method using Kohonen neural network and 3-way PLS, Co*mput Chem,* 26 (2002) 583-9.

[79] Mazerska, Z., Mazerski, J. and Ledochowski, A., QSAR of acridines. II. Features of nitracrine analogs for high anti-tumor activity and selectivity on mice, searched by PCA and MRA methods, An*ticancer Drug Des,* 5 (1990) 169-87.

[80] Hemmateenejad, B., Akhond, M., Miri, R. and Shamsipur, M., Genetic algorithm applied to the selection of factors in principal component-artificial neural networks: application to QSAR study of calcium channel antagonist activity of 1,4-dihydropyridines (nifedipine analogous), J *Chem Inf Comput Sci,* 43 (2003) 1328-34.

[81] Gao, H., Application of BCUT metrics and genetic algorithm in binary QSAR analysis, J *Chem Inf Comput Sci,* 41 (2001) 402-7.

[82] Li, Y., Ye, Z. and Lu, J., [Chemical QSAR recognition by using fuzzy min-max neural-network], Sh*eng Wu Yi Xue Gong Cheng Xue Za Zhi,* 19 (2002) 449-51.

[83] Polanski, J., Self-organizing neural network for modeling 3D QSAR of colchicinoids, Ac*ta Biochim Pol,* 47 (2000) 37-45.

[84] Duprat, A.F., Huynh, T. and Dreyfus, G., Toward a principled methodology for neural network design and performance evaluation in QSAR. Application to the prediction of logP, J *Chem Inf Comput Sci,* 38 (1998) 586-94.

[85] Polanski, J., Gieleciak, R. and Bak, A., The comparative molecular surface analysis (COMSA)--a nongrid 3D QSAR method by a coupled neural network and PLS system: predicting pK(a) values of benzoic and alkanoic acids, J *Chem Inf Comput Sci,* 42 (2002) 184-91.

[86] Baskin, II, Ait, A.O., Halberstam, N.M., Palyulin, V.A. and Zefirov, N.S., An approach to the interpretation of backpropagation neural network models in QSAR studies, SA*R QSAR Environ Res,* 13 (2002) 35-41.

[87] Lednicer, D., Ch*ronicles of drug discovery,* Wiley, New York, 1993, v. pp.

[88] Hubbard, R.E., Can drugs be designed?, Cu*rrent Opinion in Biotechnology,* 8 (1997) 696-700.

[89] Drews, J., Drug discovery: a historical perspective, Sc*ience,* 287 (2000) 1960-4.

[90] Chen, X., Ji, Z.L. and Chen, Y.Z., TTD: Therapeutic Target Database, Nu*cleic Acids Res,* 30 (2002) 412-5.

[91] Harrington, J.L., Re*lational database design clearly explained,* AP Professional, San Diego, 1998, xiii, 286 pp.

[92] Gordon, R.S., Oracle 9i: A beginner's guide., Li*brary Journal,* 127 (2002) 125-125.

[93] King, N., Active server pages, In*ternet World,* 8 (1997) 78-78.

[94] Ziener, C., Designing active server pages., Li*brary Journal,* 125 (2000) 176-176.

[95] Ryan, T.E. and Patterson, S.D., Proteomics: drug target discovery on an industrial scale, Tr*ends Biotechnol,* 20 (2002) S45-51.

[96] Dean, P.M., Zanders, E.D. and Bailey, D.S., Industrial-scale, genomics-based drug design and discovery, Tr*ends Biotechnol,* 19 (2001) 288-92.

[97] Wolfsberg, T.G., McEntyre, J. and Schuler, G.D., Guide to the draft human genome, Na*ture,* 409 (2001) 824-6.

[98] Vapnik, V.N., An overview of statistical learning theory, Ie*ee Transactions on Neural Networks,* 10 (1999) 988-999.

[99] Evgeniou, T., Pontil, M. and Poggio, T., Statistical learning theory: A primer, In*ternational Journal of Computer Vision,* 38 (2000) 9-13.

[100] Wu, Y.Q., Ianakiev, K. and Govindaraju, V., Improved k-nearest neighbor classification, Pa*ttern Recognition,* 35 (2002) 2311-2318.

[101] Monson, L., Classifying text with ID3 and C4.5, Dr *Dobbs Journal,* 22 (1997) 117-&.

[102] Nagendra, S.M.S. and Khare, M., Principal component analysis of urban traffic characteristics and meteorological data, Tr*ansportation Research Part D-Transport and Environment,* 8 (2003) 285-297.

[103] Cao, J., Murata, N., Amari, S., Cichocki, A. and Takeda, T., A robust approach to independent component analysis of signals with high-level noise measurements, Ie*ee Transactions on Neural Networks,* 14 (2003) 631-645.

[104] Vapnik, V.N., St*atistical learning theory,* Wiley, New York, 1998, xxiv, 736 pp.

[105] Evans, F.J., Natural products as probes for new drug target identification, J *Ethnopharmacol,* 32 (1991) 91-101.

[106] Clark, A.M., Natural products as a resource for new drugs, Ph*arm Res,* 13 (1996) 1133-44.

[107] Borris, R.P., Natural products research: perspectives from a major pharmaceutical company, J *Ethnopharmacol,* 51 (1996) 29-38.

[108] Strohl, W.R., The role of natural products in a modern drug discovery program, Dr*ug Discov Today,* 5 (2000) 39-41.

[109] Butte, A., The use and analysis of microarray data, Na*t Rev Drug Discov,* 1 (2002) 951-60.

[110] Shoemaker, R.H., Scudiero, D.A., Melillo, G., Currens, M.J., Monks, A.P., et al., Application of high-throughput, molecular-targeted screening to anticancer drug

discovery, Cu*rr Top Med Chem,* 2 (2002) 229-46.

[111] Chen, Y.Z. and Zhi, D.G., Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule, Pr*oteins,* 43 (2001) 217-26.

[112] Chen, X., Ung, C.Y. and Chen, Y.Z., Can an In-Silico Drug-Target Search Method be Used to Probe Potential Mechanisms of Medicinal Plant Ingredients?, *Nat. Prod. Rep.,* 20 (2003) 432-444.

[113] Ohlstein, E.H., Ruffolo, R.R., Jr. and Elliott, J.D., Drug discovery in the next millennium, An*nu Rev Pharmacol Toxicol,* 40 (2000) 177-91.

[114] Peltonen, L. and McKusick, V.A., Genomics and medicine. Dissecting human disease in the postgenomic era, Sc*ience,* 291 (2001) 1224-9.

[115] Zanders, E., Impact of genomics on medicine, Ph*armacogenomics,* 3 (2002) 443-6.

[116] Koonin, E.V., Tatusov, R.L. and Galperin, M.Y., Beyond complete genomes: from sequence to structure and function, Cu*rr Opin Struct Biol,* 8 (1998) 355-63.

[117] Wallace, K.B. and Starkov, A.A., Mitochondrial targets of drug toxicity, An*nu Rev Pharmacol Toxicol,* 40 (2000) 353-88.

[118]Vesell, E.S., Advances in pharmacogenetics and pharmacogenomics, J *Clin Pharmacol,* 40 (2000) 930-8.

[119] Fagan, R. and Swindells, M., Bioinformatics, target discovery and the pharmaceutical/biotechnology industry, Cu*rr Opin Mol Ther,* 2 (2000) 655-61.

[120] Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., et al., A second generation force field for the simulation of proteins, nucleic acids, and organic molecules (vol 117, pg 5179, 1995), Jo*urnal of the American Chemical Society,* 118 (1996) 2309-2309.

[121] Lim, H.A. and Venkatesh, T.V., Bioinformatics in the pre- and post-genomic eras, Tr*ends Biotechnol,* 18 (2000) 133-5.

[122] Miller, C.J. and Attwood, T.K., Bioinformatics goes back to the future, Na*t Rev Mol Cell Biol,* 4 (2003) 157-62.

[123] Wu, T.D., Bioinformatics in the post-genomic era, Tr*ends Biotechnol,* 19 (2001) 479-80.

[124] Levinthal, C., Molecular model-building by computer, Sc*i Am,* 214 (1966) 42-52.

[125] Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Jr., Brice, M.D., et al., The Protein Data Bank. A computer-based archival file for macromolecular structures, Eu*r J Biochem,* 80 (1977) 319-24.

[126] Maxam, A.M. and Gilbert, W., A new method for sequencing DNA, Pr*oc Natl Acad Sci U S A,* 74 (1977) 560-4.

[127] Sanger, F., Nicklen, S. and Coulson, A.R., DNA sequencing with chain-terminating inhibitors, Pr*oc Natl Acad Sci U S A,* 74 (1977) 5463-7.

[128] Kraft, W., The technology of new fluorescence illumination systems, Mi*kroskopie,* 31 (1975) 129-46.

[129] Luckey, J.A., Drossman, H., Kostichka, A.J., Mead, D.A., D'Cunha, J., et al., High speed DNA sequencing by capillary electrophoresis, Nu*cleic Acids Res,* 18 (1990) 4417-21.

[130] Swerdlow, H. and Gesteland, R., Capillary gel electrophoresis for rapid, high resolution DNA sequencing, Nu*cleic Acids Res,* 18 (1990) 1415-9.

[131] Cohen, A.S., Najarian, D.R. and Karger, B.L., Separation and analysis of DNA sequence reaction products by capillary gel electrophoresis, J *Chromatogr,* 516 (1990) 49-60.

[132] Karsch-Mizrachi, I. and Ouellette, B.F., The GenBank sequence database, Me*thods Biochem Anal,* 43 (2001) 45-63.

[133] Harger, C., Skupski, M., Bingham, J., Farmer, A., Hoisie, S., et al., The Genome Sequence DataBase (GSDB): improving data quality and data access, Nu*cleic Acids Res,* 26 (1998) 21-6.

[134] Burks, C., The GenBank database and the flow of sequence data for the human genome, Ba*sic Life Sci,* 46 (1988) 51-6.

[135] Mungall, C.J., Misra, S., Berman, B.P., Carlson, J., Frise, E., et al., An integrated computational pipeline and database to support whole-genome sequence annotation, Ge*nome Biol,* 3 (2002) RESEARCH0081.

[136] de Haen, C., Swanson, E. and Teller, D.C., The evolutionary origin of proinsulin. Amino acid sequence homology with the trypsin-related serine proteases detected and evaluated by new statistical methods, J *Mol Biol,* 106 (1976) 639-61.

[137] Stephens, J.C., Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion, Mo*l Biol Evol,* 2 (1985) 539-56.

[138] Gojobori, T., Moriyama, E.N. and Kimura, M., Statistical methods for estimating sequence divergence, Me*thods Enzymol,* 183 (1990) 531-50.

[139] Chow, S.C. and Shao, J., Statistical methods for two-sequence three-period cross-over designs with incomplete data, St*at Med,* 16 (1997) 1031-9.

[140] Engel, L.W., The Human Genome Project. History, goals, and progress to date, Ar*ch Pathol Lab Med,* 117 (1993) 459-65.

[141] Roberts, L., Davenport, R.J., Pennisi, E. and Marshall, E., A history of the Human Genome Project, Sc*ience,* 291 (2001) 1195.

[142] Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., et al., The sequence of the human genome, Sc*ience,* 291 (2001) 1304-51.

[143] McPherson, J.D., Marra, M., Hillier, L., Waterston, R.H., Chinwalla, A., et al., A physical map of the human genome, Na*ture,* 409 (2001) 934-41.

[144] Crowe, G.D., A History of Computer-Technology - from the Simplest Counting Devices to Complex Relay Systems - Apokin,Ia, Maistrov,Le, Is*is,* 83 (1992) 306-307.

[145] Biggerstaff, T.J., Moore's law: Change or die, Ie*ee Software,* 13 (1996) 4-6.

[146] Glowniak, J., History, structure, and function of the Internet, Se*minars in Nuclear Medicine,* 28 (1998) 135-144.

[147] Leiner, B.M., Cerf, V.G., Clark, D.D., Kahn, R.E., Kleinrock, L., et al., The past and future history of the Internet, Co*mmunications of the Acm,* 40 (1997) 102-108.

[148] Ratzek, W., ARPA KADABRA - The history of the Internet, Nf*d Information-Wissenschaft Und Praxis,* 50 (1999) 307-307.

[149] Wiggins, M., Unix and the Internet: A brief history, Ie*ee Internet Computing,* 2 (1998) 52-52.

[150] Rundle, D., Internet history (A survey of the latest multimedia innovations, are they useful to historians), Hi*story Today,* 48 (1998) 14-15.

[151] Greene, R., Web work, a history of internet art, Ar*tforum,* 38 (2000) 162-+.

[152] Hawley, T., Special edition using HTML and XHTML, Te*chnical Communication,* 50 (2003) 288-290.

[153] Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D., et al., ExPASy: the

proteomics server for in-depth protein knowledge and analysis, Nu*cleic Acids Res,* 31 (2003) 3784-8.

[154] Ostell, J.M., Wheelan, S.J. and Kans, J.A., The NCBI data model, Me*thods Biochem Anal,* 43 (2001) 19-43.

[155] Baxevanis, A.D., The molecular biology database collection: an online compilation of relevant database resources, Nu*cleic Acids Res,* 28 (2000) 1-7.

[156] Baxevanis, A.D., The Molecular Biology Database Collection: 2002 update, Nu*cleic Acids Res,* 30 (2002) 1-12.

[157] Baxevanis, A.D., The Molecular Biology Database Collection: 2003 update, Nu*cleic Acids Res,* 31 (2003) 1-12.

[158] Burbidge, R., Trotter, M., Buxton, B. and Holden, S., Drug design by machine learning: support vector machines for pharmaceutical data analysis, Co*mput Chem,* 26 (2001) 5-14.

[159] Colville-Nash, P.R. and Gilroy, D.W., Cyclooxygenase enzymes as targets for therapeutic intervention in inflammation, Dr*ug News Perspect,* 13 (2000) 587-97.

[160] Kanehisa, M., The KEGG database, No*vartis Found Symp,* 247 (2002) 91-101; discussion 101-3, 119-28, 244-52.

[161] Martin, C., Berridge, G., Higgins, C.F., Mistry, P., Charlton, P., et al., Communication between multiple drug binding sites on P-glycoprotein, Mo*l Pharmacol,* 58 (2000) 624-32.

[162] He, F., Seryshev, A.B., Cowan, C.W. and Wensel, T.G., Multiple zinc binding sites in retinal rod cGMP phosphodiesterase, PDE6alpha beta, J *Biol Chem,* 275 (2000) 20572-7.

[163] Aleshin, A.E., Kirby, C., Liu, X., Bourenkov, G.P., Bartunik, H.D., et al., Crystal structures of mutant monomeric hexokinase I reveal multiple ADP binding sites and conformational changes relevant to allosteric regulation, J *Mol Biol,* 296 (2000) 1001-15.

[164] Xu, Y., Gurusiddappa, S., Rich, R.L., Owens, R.T., Keene, D.R., et al., Multiple binding sites in collagen type I for the integrins alpha1beta1 and alpha2beta1, J *Biol Chem,* 275 (2000) 38981-9.

[165] Coassolo, P., Briand, C., Bourdeaux, M. and Sari, J.C., Microcalorimetric method to determine competitive binding. Action of a psychotropic drug (dipotassium chlorazepate)

on L-tryptophan . human serum albumin complex, Bi*ochim Biophys Acta,* 538 (1978) 512-20.

[166] Menke, G., Worner, W., Kratzer, W. and Rietbrock, N., Kinetics of drug binding to human serum albumin: allosteric and competitive inhibition at the benzodiazepine binding site by free fatty acids of various chain lengths, Na*unyn Schmiedebergs Arch Pharmacol,* 339 (1989) 42-7.

[167] Steinmann, L. and Thormann, W., Characterization of competitive binding, fluorescent drug immunoassays based on micellar electrokinetic capillary chromatography, El*ectrophoresis,* 17 (1996) 1348-56.

[168] Erim, F.B. and Kraak, J.C., Vacancy affinity capillary electrophoresis to study competitive protein-drug binding, J *Chromatogr B Biomed Sci Appl,* 710 (1998) 205-10.

[169] Angelakou, A., Valsami, G., Macheras, P. and Koupparis, M., A displacement approach for competitive drug-protein binding studies using the potentiometric 1-anilino-8-naphthalene-sulfonate probe technique, Eu*r J Pharm Sci,* 9 (1999) 123-30.

[170] Dura, E., Natural language in information retrieval, Co*mputational Linguistics and Intelligent Text Processing, Proceedings,* 2588 (2003) 537-540.

[171] Voorhees, E.M., Natural language processing and information retrieval, In*formation Extraction: Towards Scalable, Adaptable Systems,* 1714 (1999) 32-48.

[172] Kreymer, O., An evaluation of help mechanisms in natural language information retrieval systems, On*line Information Review,* 26 (2002) 30-39.

[173] Franzen, K., Eriksson, G., Olsson, F., Asker, L., Liden, P., et al., Protein names and how to find them, In*ternational Journal of Medical Informatics,* 67 (2002) 49-61.

[174] Ono, T., Hishigaki, H., Tanigami, A. and Takagi, T., Automated extraction of information on protein-protein interactions from the biological literature, Bi*oinformatics,* 17 (2001) 155-161.

[175] Pereira, F.C.N. and Grosz, B.J., Na*tural language processing,* 1st MIT Press edn., MIT Press, Cambridge, Mass., 1994, vi, 531 pp.

[176] Ceusters, W., Spyns, P. and De Moor, G., From syntactic-semantic tagging to knowledge discovery in medical texts, In*ternational Journal of Medical Informatics,* 52 (1998) 149-157.

[177] Volot, F., Joubert, M. and Fieschi, M., Review of biomedical knowledge and data representation with conceptual graphs, Me*thods of Information in Medicine,* 37 (1998)

86-96.

[178] Ramu, C., SIR: a simple indexing and retrieval system for biological flat file databases, Bi*oinformatics,* 17 (2001) 756-8.

[179] Etzold, T. and Argos, P., SRS--an indexing and retrieval tool for flat file data libraries, Co*mput Appl Biosci,* 9 (1993) 49-57.

[180] Chyka, P.A., Holimon, T.D., Tepedino, J.T. and Petersen, H., Relational database for drug-use review of Tennessee Medicaid claims, Am *J Health Syst Pharm,* 53 (1996) 164-6.

[181] Macdonald, A.M. and Hamer, S.A., Development of computerized storage facilities for twin data: a relational database system for a twin register, Be*hav Genet,* 27 (1997) 1-13.

[182] Carazo, J.M. and Stelzer, E.H., The BioImage Database Project: organizing multidimensional biological images in an object-relational database, J *Struct Biol,* 125 (1999) 97-102.

[183] Altmann, U., Wachter, W., Tafazzoli, A.G., Katz, F.R., Schweiger, R., et al., A model for integration and continuous development of standards for tumour documentation using relational database techniques and extensible markup language, St*ud Health Technol Inform,* 68 (1999) 895-8.

[184] Hahn, J. and Cole-Williams, A., Developing and implementing a relational database for heart failure outcomes in an integrated healthcare system, Ou*tcomes Manag,* 7 (2003) 61-7.

[185] Baran, M.C., Moseley, H.N., Sahota, G. and Montelione, G.T., SPINS: standardized protein NMR storage. A data dictionary and object-oriented relational database for archiving protein NMR spectra, J *Biomol NMR,* 24 (2002) 113-21.

[186] Zdobnov, E.M., Lopez, R., Apweiler, R. and Etzold, T., The EBI SRS server--recent developments, Bi*oinformatics,* 18 (2002) 368-73.

[187] Schuler, G.D., Epstein, J.A., Ohkawa, H. and Kans, J.A., Entrez: molecular biology database and retrieval system, Me*thods Enzymol,* 266 (1996) 141-62.

[188] Libkin, L., Expressive power of SQL, Th*eoretical Computer Science,* 296 (2003) 379-404.

[189] Gordon, R.S., SQL: Visual QuickStart guide., Li*brary Journal,* 127 (2002) 171-171.

[190] Gordon, R.S., Oracle 9i DBA 101: Learn the essentials of oracle database administration., Li*brary Journal,* 127 (2002) 125-125.

[191] Brathwaite, K.S., In*formation engineering,* CRC Press, Boca Raton, Fla., 1992, 3 v. pp.

[192] Hares, J.S., In*formation engineering for the advanced practitioner,* Wiley, Chichester, 1992, xiii, 410 pp.

[193] Thalheim, B., En*tity-relationship modeling : foundations of database technology,* Springer, Berlin ; New York, 2000, xii, 627 pp.

[194] Spaccapietra, S. and Association fran鏰ise pour la cybern閳ique 閳onomique et technique., Entit*y-relationship approach : ten years of experience in information modeling, Nor*th-Holland, Amsterdam, 1987, xiv, 557 pp.

[195] L閳nard, M., Databa*se design theory, Macm*illan, Basingstoke, 1992, 259 pp.

[196] Stephens, R.K. and Plew, R.R., Databa*se design, Sams* Pub., Indianapolis, Ind., 2001, xi, 508 pp.

[197] Wong, E. and Katz, R.H., Logical Design and Schema Conversion for Relational and DBTG databases, Procee*dings of the international conference on entity-relationship approach to system analysis and design. Los Angeles. (1980*) 311-321.

[198] Elmasri, R. and Navathe, S.B., Fundam*entals of database systems, 3rd* edn., Addison-Wesley, Menlo Park, Calif. ; Harlow, 1999, xxvii, 960 pp.

[199] Gabaude, J.M., Towards a history of ontology, Revue *Philosophique De La France Et De L Etranger, 127* (2002) 108-109.

[200] Sechser, O., Principles of Semantic Networks - Explorations in the Representation of Knowledge - Sowa,Jf, Knowle*dge Organization, 20 (*1993) 60-61.

[201] Lehmann, F., Semantic Networks, Comput*ers & Mathematics with Applications, 23 (*1992) 1-50.

[202] Horty, J.F., Thomason, R.H. and Touretzky, D.S., A Skeptical Theory of Inheritance in Nonmonotonic Semantic Networks, Artific*ial Intelligence, 42 (*1990) 311-348.

[203] Brachman, R.J., On the *epistemological status of semantic networks, Acad*emic Press, New York, 1979.

[204] Date, C.J., An int*roduction to database systems, 3d e*dn., Addison-Wesley Pub. Co., Reading, Mass., 1981, 2 v. pp.

[205] Hanna, P., JSP 2*.0 : the complete reference, McGr*aw-Hill/Osborne, Berkeley, Calif., 2003, xix, 841 pp.

[206] Argerich, L., Profes*sional PHP4, Wrox* Press, Birmingham, 2002, xv, 974 pp.

[207] Anderson, R., ASP 3*.0 programmer's reference, Wrox* Press, Birmingham, 2000, xxvii, 1297 pp.

[208] Descartes, A., Bunce, T. and NetLibrary Inc., Progra*mming the Perl DBI, O'Re*illy, Cambridge, MA, 2000, xvi, 346 pp.

[209] Sussman, D., Profes*sional ADO 2.5, Wrox* Press, Birmingham, 2000, xxv, 973 pp.

[210] Microsoft Corporation., Progra*mmer's reference : Microsoft Open Database Connectivity software development kit, version 2.0 : for Microsoft Windows and Windows NT operating systems, Micr*osoft Press, Redmond, Wash., 1994, xvii, 713, 95 pp.

[211] Wood, C., OLE DB *and ODBC developer's guide, M&T* Books, Foster City, Calif., 1999, xxiii, 668 pp.

[212] Weinman, W., The CG*I book, New* Riders Pub., Indianapolis, Ind., 1996, xii, 304 pp.

[213] Drews, J. and Ryser, S., Classic Drug Targets: Special Pullout., Nat Bi*otechnol, 15 (*1997).

[214] Ji, Z.L., Han, L.Y., Yap, C.W., Sun, L.Z., Chen, X., et al., Drug Adverse Reaction Target Database (DART) : Proteins Related to Adverse Drug Reactions, Drug S*af, 26 (*2003) 685-90.

[215] Sun, L.Z., Ji, Z.L., Chen, X., Wang, J.F. and Chen, Y.Z., ADME-AP: a database of ADME associated proteins, Bioinf*ormatics, 18 (*2002) 1699-700.

[216] Pumford, N.R. and Halmes, N.C., Protein targets of xenobiotic reactive intermediates, Annu Re*v Pharmacol Toxicol, 37 (*1997) 91-117.

[217] Park, B.K., Kitteringham, N.R., Powell, H. and Pirmohamed, M., Advances in molecular toxicology-towards understanding idiosyncratic drug toxicity, Toxico*logy, 153* (2000) 39-60.

[218] Rang, H.P.D., M.M.; Rotter, J.M., Pharma*cology, 4th* edn., Churchill Livingstone, 1999.

[219] Casarett, L.J., Doull, J. and Klaassen, C.D., Casare*tt and Doull's toxicology : the basic science of poisons, 6th* edn., McGraw-Hill Medical Pub. Division, New York, 2001, xix,

1236 pp.

[220] Gerhold, D., Rushmore, T. and Caskey, C.T., DNA chips: promising toys have become powerful tools, Trends *Biochem Sci, 24 (*1999) 168-73.

[221] Nuwaysir, E.F., Bittner, M., Trent, J., Barrett, J.C. and Afshari, C.A., Microarrays and toxicology: the advent of toxicogenomics, Mol Ca*rcinog, 24 (*1999) 153-9.

[222] Barratt, M.D., Integrating computer prediction systems with in vitro methods towards a better understanding of toxicology, Toxico*l Lett, 102-*103 (1998) 617-21.

[223] Chen, Y.Z.U., C.Y.;, Prediction of potential toxicity and side effect protein targets of a small molecule by a ligand-protein inverse docking approach., J.Mol.*Graph.Mod., 20 (*2001) 199-218.

[224] Zhang, L., Brett, C.M. and Giacomini, K.M., Role of organic cation transporters in drug absorption and elimination, Annu R*ev Pharmacol Toxicol, 38 (*1998) 431-60.

[225] Ito, K., Iwatsubo, T., Kanamitsu, S., Nakajima, Y. and Sugiyama, Y., Quantitative prediction of in vivo drug clearance and drug interactions from in vitro data on metabolism, together with binding and transport, Annu R*ev Pharmacol Toxicol, 38 (*1998) 461-99.

[226] Lin, J.H. and Lu, A.Y., Role of pharmacokinetics and metabolism in drug discovery and development, Pharma*col Rev, 49 (*1997) 403-49.

[227] Tamai, I. and Tsuji, A., Transporter-mediated permeation of drugs across the blood-brain barrier, J Phar*m Sci, 89 (*2000) 1371-88.

[228] de Wolf, F.A. and Brett, G.M., Ligand-binding proteins: their potential for application in systems for controlled delivery and uptake of ligands, Pharma*col Rev, 52 (*2000) 207-36.

[229] Vizi, E.S., Role of high-affinity receptors and membrane transporters in nonsynaptic communication and drug action in the central nervous system, Pharma*col Rev, 52 (*2000) 63-89.

[230] Caldwell, J., Gardner, I. and Swales, N., An introduction to drug disposition: the basic principles of absorption, distribution, metabolism, and excretion, Toxico*l Pathol, 23 (*1995) 102-14.

[231] Eddershaw, P.J., Beresford, A.P. and Bayliss, M.K., ADME/PK as part of a rational approach to drug discovery, Drug D*iscov Today, 5 (*2000) 409-414.

[232] Clark, D.E. and Pickett, S.D., Computational methods for the prediction of 'drug-likeness', Drug D*iscov Today, 5 (*2000) 49-58.

[233] Norris, D.A., Leesman, G.D., Sinko, P.J. and Grass, G.M., Development of predictive pharmacokinetic simulation models for drug discovery, J Cont*rol Release, 65 (*2000) 55-62.

[234] Ekins, S., Waller, C.L., Swaan, P.W., Cruciani, G., Wrighton, S.A., et al., Progress in predicting human ADME parameters in silico, J Phar*macol Toxicol Methods, 44 (*2000) 251-72.

[235] Li, A.P., Screening for human ADME/Tox drug properties in drug discovery, Drug D*iscov Today, 6 (*2001) 357-366.

[236] Wang, J.H. and Hewick, R.M., Proteomics in drug discovery, Drug D*iscov Today, 4 (1*999) 129-133.

[237] Scharpe, S. and De Meester, I., Peptide truncation by dipeptidyl peptidase IV: a new pathway for drug discovery?, Verh K *Acad Geneeskd Belg, 63 (*2001) 5-32; discussion 32-3.

[238] Williams, M., Genome-based drug discovery: prioritizing disease-susceptibility/disease-associated genes as novel drug targets for schizophrenia, Curr O*pin Investig Drugs, 4 (2*003) 31-6.

[239] Conkright, M.D., Guzman, E., Flechner, L., Su, A.I., Hogenesch, J.B., et al., Genome-wide analysis of CREB target genes reveals a core promoter requirement for cAMP responsiveness, Mol Ce*ll, 11 (*2003) 1101-8.

[240] Klein, P., Kanehisa, M. and DeLisi, C., Prediction of protein function from sequence properties. Discriminant analysis of a data base, Biochi*m Biophys Acta, 787* (1984) 221-6.

[241] Nakai, K., Kidera, A. and Kanehisa, M., Cluster analysis of amino acid indices for prediction of protein structure and function, Protei*n Eng, 2 (1*988) 93-100.

[242] Fetrow, J.S. and Skolnick, J., Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases, J Mol *Biol, 281* (1998) 949-68.

[243] Edwards, Y.J. and Cottage, A., Prediction of protein structure and function by using bioinformatics, Method*s Mol Biol, 175* (2001) 341-75.

[244] Baxter, S.M. and Fetrow, J.S., Sequence- and structure-based protein function

prediction from genomic information, Curr O*pin Drug Discov Devel, 4 (*2001) 291-5.

[245] Vapnik, V.N., The na*ture of statistical learning theory, 2nd* edn., Springer, New York, 2000, xix, 314 pp.

[246] Breiman, L., Classi*fication and regression trees, Wad*sworth, Belmont, Calif., 1984, x, 358 pp.

[247] Quinlan, J.R., Induction of decision trees., Machine *Learning, 1 (1*986) 81-106.

[248] Quinlan, J.R., C4.5 : *programs for machine learning, Morg*an Kaufmann Publishers, San Mateo, Calif., 1993, x, 302 pp.

[249] Cameron-Jones, R.M. and Quinlan, J.R., Efficient top-down induction of logical programs, SIGART *Bulletin, 5(1)* (1994) 33-42.

[250] Cohen, W.W., Fast effective rule induction., In Pro*c. of The 12th International Conference on Machine Learning, Morg*an Kauffman, 1995, pp. 115-123.

[251] McCallum, A.K., BOW: A toolkit for statistical language modeling, text retrieval, classification and clustering., http:/*/www.cs.cmu.edu/~mccallum/bow (1996*).

[252] Michie, D., Spiegelhalter, D.J. and Taylor, C.C., Machine *learning, neural and statistical classification, Pren*tice Hall, Englewood Cliffs, N.J., 1994, xiv, 289 pp.

[253] Fausett, L.V., Fundam*entals of neural networks : architectures, algorithms, and applications, Pren*tice-Hall, Englewood Cliffs, N.J., 1994, xvi, 461 pp.

[254] Goldberg, D.E., Genetic *algorithms in search, optimization, and machine learning, Addi*son-Wesley Pub. Co., Reading, Mass., 1989, xiii, 412 pp.

[255] Dasarathy, B.V., Neares*t neighbor(NN) norms : NN pattern classification techniques, IEEE* Computer Society Press, Los Alamitos, Calif., 1991, xii, 447 pp.

[256] Wettschereck, D., Aha, D.W. and Mohri, T., A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms, Artific*ial Intelligence Review, 11 (*1997) 273-314.

[257] Baldwin, J.F., Lawry, J. and Martin, T.P., A mass assignment based ID3 algorithm for decision tree induction, Intern*ational Journal of Intelligent Systems, 12 (*1997) 523-552.

[258] Lubbe, J.C.A.v.d., Inform*ation theory, Camb*ridge University Press, Cambridge, 1997, xii, 350 pp.

[259] Fayyad, U.M. and Irani, K.B., On the Handling of Continuous-Valued Attributes in Decision Tree Generation, Machin*e Learning, 8 (*1992) 87-102.

[260] Kim, H. and Koehler, G.J., An Investigation on the Conditions of Pruning an Induced Decision Tree, Europe*an Journal of Operational Research, 77 (*1994) 82-95.

[261] Elomaa, T., The biases of decision tree pruning strategies, Advanc*es in Intelligent Data Analysis, Proceedings, 1642* (1999) 63-74.

[262] Satoh, Y., Matsumoto, G., Mori, H. and Ito, K., Nearest neighbor analysis of the SecYEG complex. 1. Identification of a SecY-SecG interface, Bioche*mistry, 42 (*2003) 7434-7441.

[263] Bao, Y.G. and Ishii, N., Combining multiple k-nearest neighbor classifiers for text classification by reducts, Discov*ery Science, Proceedings, 2534* (2002) 340-347.

[264] Malinen, J., Maltamo, M. and Verkasalo, E., Predicting the internal quality and value of Norway spruce trees by using two non-parametric nearest neighbor methods, Forest *Products Journal, 53 (*2003) 85-94.

[265] Kim, K.I., Jung, K., Park, S.H. and Kim, H.J., Support vector machine-based text detection in digital video, Patter*n Recognition, 34 (*2001) 527-529.

[266] de Vel, O., Anderson, A., Corney, M. and Mohay, G., Mining e-mail content for author identification forensics, Sigmod *Record, 30 (*2001) 55-64.

[267] Ben-Yacoub, S., Abdeljaoued, Y. and Mayoraz, E., Fusion of face and speech data for person identity verification, Ieee T*ransactions on Neural Networks, 10 (*1999) 1065-1074.

[268] Karlsen, R.E., Gorsich, D.J. and Gerhart, G.R., Target classification via support vector machines, Optica*l Engineering, 39 (*2000) 704-711.

[269] Liong, S.Y. and Sivapragasam, C., Flood stage forecasting with support vector machines, Journa*l of the American Water Resources Association, 38 (*2002) 173-186.

[270] Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., et al., Knowledge-based analysis of microarray gene expression data by using support vector machines, Procee*dings of the National Academy of Sciences of the United States of America, 97 (*2000) 262-267.

[271] Yuan, Z., Burrage, K. and Mattick, J.S., Prediction of protein solvent accessibility using support vector machines, Protei*ns, 48 (*2002) 566-70.

[272] Ding, C.H. and Dubchak, I., Multi-class protein fold recognition using support vector machines and neural networks, Bioinf*ormatics, 17 (*2001) 349-58.

[273] Hua, S. and Sun, Z., A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach, J Mol *Biol, 308* (2001) 397-407.

[274] Bock, J.R. and Gough, D.A., Predicting protein--protein interactions from primary structure, Bioinf*ormatics, 17 (*2001) 455-60.

[275] Cai, C.Z., Han, L.Y., Ji, Z.L., Chen, X. and Chen, Y.Z., SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence, Nuclei*c Acids Res, 31 (*2003) 3692-7.

[276] Alpay, D., Reprod*ucing kernel spaces and applications, Birk*hauser Verlag, Boston, MA, 2003.

[277] Alpay, D., The Sc*hur algorithm, reproducing kernel spaces, and system theory, Amer*ican Mathematical Society, Providence, R.I., 2001, viii, 150 p. pp.

[278] Gunn, S.R., Suppor*t Vector Machines for Classification and Regression: Technical Report., UNIV*ERSITY OF SOUTHAMPTON, 1998.

[279] Chang, C.C. and Lin, C.J., LIBSVM: a Library for Support Vector Machines., Computer Science and Information Engineering, National Taiwan University., 2003.

[280] Bellman, R.E., Adapti*ve control processes: a guided tour, Prin*ceton University Press, Princeton, N.J.,, 1961, 255 pp.

[281] Krishnaiah, P.R. and Kanal, L.N., Classi*fication, pattern recognition, and reduction of dimensionality, Nort*h-Holland Pub. Co. ;

Sole distributors for the U.S.A. and Canada Elsevier Science Pub. Co., Amsterdam ; New York

New York, N.Y., 1982, xxii, 903 pp.

[282] Kartasasmita, M., Dimensionality reduction by linear transformation for pattern classification with applications to Thematic Mapper data. Davis, Calif., 1986, pp. 183 leaves,.

[283] Lee, E.S., Reduct*ion in dimensionality, dynamic programming and quasilinearization, Kans*as State University, Manhattan,, 1967, 32 [2] leaves pp.

[284] Stone, C.J., The di*mensionality reduction principle for generalized additive models, Dept*. of Statistics University of California, Berkeley, Calif., 1985, 28 leaves. pp.

[285] Jolliffe, I.T., Princi*pal component analysis, 2nd* edn., Springer-Verlag, New York, 2002, xxix, 487 pp.

[286] Vidal, R.E., Ma, Y. and Sastry, S., Genera*lized principal component analysis (GPCA), Elec*tronics Research Laboratory College of Engineering University of California, Berkeley, 2002, 24 pp.

[287] Comon, P., Independent Component Analysis, a New Concept, Signal *Processing, 36 (*1994) 287-314.

[288] Jutten, C. and Herault, J., Blind Separation of Sources .1. An Adaptive Algorithm Based on Neuromimetic Architecture, Signal *Processing, 24 (*1991) 1-10.

[289] Papoulis, A. and Pillai, S.U., Probab*ility, random variables, and stochastic processes, 4th* edn., McGraw-Hill, Boston, 2002, x, 852 pp.

[290] Hyvarinen, A., New approximations of differential entropy for independent component analysis and projection pursuit., In Adv*ances in neural information processing systems., MIT* Press, 1998, pp. v.

[291] Zhao, Y. and Atkeson, C.G., Implementing projection pursuit learning, Ieee T*ransactions on Neural Networks, 7 (1*996) 362-373.

[292] Trizna, D.B., Bachmann, C., Sletten, M., Allan, N., Toporkov, J., et al., Projection pursuit classification of multiband polarimetric SAR land images, Ieee T*ransactions on Geoscience and Remote Sensing, 39 (*2001) 2380-2386.

[293] Huber, P.J., Projection pursuit., The An*nals of Statistics, 13(2*) (1985) 435-475.

[294] Jones, M.C. and Sibson, R., What is projection pursuit?, J. of *the Royal Statistical Society ser. A, 150* (1987) 1-36.

[295] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., et al., The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, Nuclei*c Acids Res, 31 (*2003) 365-70.

[296] Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., et al., The Pfam protein families database, Nucleic *Acids Res, 30 (*2002) 276-80.

[297] Karchin, R., Karplus, K. and Haussler, D., Classifying G-protein coupled receptors with support vector machines, Bioinf*ormatics, 18 (*2002) 147-59.

[298] Dubchak, I., Muchnik, I., Holbrook, S.R. and Kim, S.H., Prediction of protein folding class using global description of amino acid sequence, Proc N*atl Acad Sci U S A, 92 (*1995) 8700-4.

[299] Hanselman, D.C. and Littlefield, B., Master*ing MATLAB 6 : a comprehensive tutorial and reference, Pren*tice Hall, Upper Saddle River, N.J., 2001, xviii, 814 p. pp.

[300] Hyvarinen, A., Fast and robust fixed-point algorithms for independent component analysis, Ieee T*ransactions on Neural Networks, 10 (*1999) 626-634.

[301] Cai, C.Z., Wang, W.L. and Chen, Y.Z., Support Vector Machine Calssification of Physical and Biological Datasets., Inter. *J. Mod. Phys. C, Acc*epted (2003).

[302] Heinrich, M., Ethnobotany and its role in drug development, Phytot*her Res, 14 (*2000) 479-88.

[303] Cheng, J.T., Review: drug therapy in Chinese traditional medicine, J Clin *Pharmacol, 40 (*2000) 445-50.

[304] Yuan, R. and Lin, Y., Traditional Chinese medicine: an approach to scientific proof and clinical validation, Pharma*col Ther, 86 (*2000) 191-8.

[305] Sutter, M.C. and Wang, Y.X., Recent cardiovascular drugs from Chinese medicinal plants, Cardio*vasc Res, 27 (*1993) 1891-901.

[306] Zhu, D.Y.B., D.L.; Tang, X.C., Recent studies on traditional Chinese medicinal plants., Drug D*ev. Res., 39 (*1996) 147-57.

[307] Li, F., Sun, S., Wang, J. and Wang, D., Chromatography of medicinal plants and Chinese traditional medicines, Biomed *Chromatogr, 12 (*1998) 78-85.

[308] Gong, X. and Sucher, N.J., Stroke therapy in traditional Chinese medicine (TCM): prospects for drug discovery and development, Trends *Pharmacol Sci, 20 (*1999) 191-6.

[309] Lee, K.H., Novel antitumor agents from higher plants, Med Re*s Rev, 19 (*1999) 569-96.

[310] Walker, M.G., Pharmaceutical target identification by gene expression analysis, Mini R*ev Med Chem, 1 (2*001) 197-205.

[311] Hatfield, G.W., Hung, S.P. and Baldi, P., Differential analysis of DNA microarray gene expression data, Mol Mi*crobiol, 47 (*2003) 871-7.

[312] Valafar, F., Pattern recognition techniques in microarray data analysis: a survey, Ann N

Y Acad Sci, 980 (2002) 41-64.

[313] Nishiu, M., Yanagawa, R., Nakatsuka, S., Yao, M., Tsunoda, T., et al., Microarray Analysis of Gene-expression Profiles in Diffuse Large B-cell Lymphoma: Identification of Genes Related to Disease Progression, Jpn J *Cancer Res, 93 (*2002) 894-901.

[314] Zhu, H. and Snyder, M., Protein chip technology, Curr O*pin Chem Biol, 7 (2*003) 55-63.

[315] Green, S.M. and Marshall, G.R., 3D-QSAR: a current perspective, Trends *Pharmacol Sci, 16 (*1995) 285-91.

[316] Bock, J.R. and Gough, D.A., A new method to estimate ligand-receptor energetics, Mol Ce*ll Proteomics, 1 (2*002) 904-10.

[317] Chen, X., Ji, Z.L., Zhi, D.G. and Chen, Y.Z., CLiBE: a database of computed ligand binding energy for ligand-receptor complexes, Comput*ers & Chemistry, 26 (*2002) 661-666.

[318] Chen, Y.Z., Li, Z.R. and Ung, C.Y., Computational Method for Drug Target Search and Application in Drug Discovery., J. The*or. Comp. Chem., 1 (2*002) 213-224.

[319] Berman, H.M., Bhat, T.N., Bourne, P.E., Feng, Z., Gilliland, G., et al., The Protein Data Bank and the challenge of structural genomics, Nat St*ruct Biol, 7 Su*ppl (2000) 957-9.

[320] Rost, B. and Sander, C., Bridging the protein sequence-structure gap by structure predictions, Annu R*ev Biophys Biomol Struct, 25 (*1996) 113-36.

[321] Baird, N., Simulation of hydrogen bonding in biological systems: Ab initio calculations for NH3-NH3 and NH3-NH4+. Int. J*. Quantum Chem. Symp., 1 (1*974) 49-53.

[322] Chen, Y.Z. and Prohofsky, E.W., The role of a minor groove spine of hydration in stabilizing poly(dA).poly(dT) against fluctuational interbase H-bond disruption in the premelting temperature regime, Nuclei*c Acids Res, 20 (*1992) 415-9.

[323] Chen, Y.Z. and Prohofsky, E.W., Premelting base pair opening probability and drug binding constant of a daunomycin-poly d(GCAT).poly d(ATGC) complex, Biophy*s J, 66 (*1994) 820-6.

[324] Favoni, R.E. and Cupis, A.D., Sterioidal and nonsteroidal oestrogen antagonists in breast cancer: basic and clinical appraisal., Trends *Pharmacol Sci., 19 (*1998) 406-415.

[325] Rowlands, M.G., Budworth, J., Jarman, M., Hardcastle, I.R., McCague, R., et al., Comparison between inhibition of protein kinase C and antagonism of calmodulin by tamoxifen analogues., Bioche*m. Pharmacol., 50 (*1995) 723-726.

[326] Abbas Abidi, S.M., Howard, E.W., Dmytryk, J.J. and Pento, J.T., Differential influence of antiestrogens on the in vitro release of gelatinases (type IV collagenases) by invasive and non-invasive breast cancer cells, Clin E*xp Metastasis, 15 (*1997) 432-9.

[327] Santner, S.J. and Santen, R.J., Inhibition of estrone sulfatase and 17 beta-hydroxysteroid dehydrogenase by antiestrogens, J Ster*oid Biochem Mol Biol, 45 (*1993) 383-90.

[328] Messiha, F.S., Leu-enkephalin, tamoxifen and ethanol interactions: effects on motility and hepatic ethanol metabolizing enzymes, Gen Ph*armacol, 21 (*1990) 45-8.

[329] Ritchie, G.A., The direct inhibition of prostaglandin synthetase of human breast cancer tumor tissue by tamoxifen, Recent *Results Cancer Res, 71 (*1980) 96-101.

[330] Nuwaysir, E.F., Daggett, D.A., Jordan, V.C. and Pitot, H.C., Phase II enzyme expression in rat liver in response to the antiestrogen tamoxifen, Cancer *Res, 56 (*1996) 3704-10.

[331] Lax, E.R., Rumstadt, F., Plasczyk, H., Peetz, A. and Schriefers, H., Antagonistic action of estrogens, flutamide, and human growth hormone on androgen-induced changes in the activities of some enzymes of hepatic steroid metabolism in the rat, Endocr*inology, 113* (1983) 1043-55.

[332] Levine, R.M., Rubalcaba, E., Lippman, M.E. and Cowan, K.H., Effects of estrogen and tamoxifen on the regulation of dihydrofolate reductase gene expression in a human breast cancer cell line, Cancer *Res, 45 (*1985) 1644-50.

[333] Paavonen, T., Aronen, H., Pyrhonen, S., Hajba, A. and Andersson, L.C., The effect of toremifene therapy on serum immunoglobulin levels in breast cancer, Apmis, *99 (*1991) 849-53.

[334] Schmidt, T.J. and Meyer, A.S., Autoregulation of corticosteroid receptors. How, when, where, and why?, Recept*or, 4 (1*994) 229-57.

[335] Meador, W.E., Means, A.R. and Quiocho, F.A., Modulation of calmodulin plasticity in molecular recognition on the basis of x-ray structures, Science*, 262* (1993) 1718-21.

[336] Sandak, B., Wolfson, H.J. and Nussinov, R., Flexible docking allowing induced fit in proteins: insights from an open to closed conformational isomers, Protei*ns, 32 (*1998) 159-74.

[337] Crommentuyn, K.M., Schellens, J.H., van den Berg, J.D. and Beijnen, J.H., In-vitro metabolism of anti-cancer drugs, methods and applications: paclitaxel, docetaxel, tamoxifen and ifosfamide, Cancer *Treat Rev, 24 (*1998) 345-66.

[338] Chen, Y.Z. and Ung, C.Y., Computer automated prediction of potential therapeutic and toxicity protein targets of bioactive compounds from Chinese medicinal plants, Am J C*hin Med, 30 (*2002) 139-54.

[339] McEntyre, J. and Lipman, D., PubMed: bridging the information gap, Cmaj, *164* (2001) 1317-9.

[340] Molokanova, E. and Kramer, R.H., Mechanism of inhibition of cyclic nucleotide-gated channel by protein tyrosine kinase probed with genistein, J Gen *Physiol, 117* (2001) 219-34.

[341] Cammalleri, C. and Germinario, R.J., The effects of protease inhibitors on basal and insulin-stimulated lipid metabolism, insulin binding, and signaling, J Lipi*d Res, 44 (*2003) 103-8.

[342] Lashley, M.R., Niedzinski, E.J., Rogers, J.M., Denison, M.S. and Nantz, M.H., Synthesis and estrogen receptor affinity of a 4-hydroxytamoxifen-Labeled ligand for diagnostic imaging, Bioorg *Med Chem, 10 (*2002) 4075-82.

[343] Barnes, S. and Peterson, T.G., Biochemical targets of the isoflavone genistein in tumor cell lines, Proc S*oc Exp Biol Med, 208* (1995) 103-8.

[344] Polkowski, K. and Mazurek, A.P., Biological properties of genistein. A review of in vitro and in vivo data, Acta P*ol Pharm, 57 (*2000) 135-55.

[345] Andlauer, W., Kolb, J., Stehle, P. and Furst, P., Absorption and metabolism of genistein in isolated rat small intestine, J Nutr*, 130* (2000) 843-6.

[346] Coldham, N.G. and Sauer, M.J., Pharmacokinetics of [(14)C]Genistein in the rat: gender-related differences, potential mechanisms of biological action, and implications for human health, Toxico*l Appl Pharmacol, 164* (2000) 206-15.

[347] Barnes, S., Boersma, B., Patel, R., Kirk, M., Darley-Usmar, V.M., et al., Isoflavonoids and chronic disease: mechanisms of action, Biofa*ctors, 12 (*2000) 209-15.

[348] Munoz, R., Klingenberg, O., Wiedlocha, A., Rapak, A., Falnes, P.O., et al., Effect of mutation of cytoplasmic receptor domain and of genistein on transport of acidic fibroblast growth factor into cells, Oncog*ene, 15 (*1997) 525-36.

[349] Theodorescu, D., Laderoute, K.R., Calaoagan, J.M. and Guilding, K.M., Inhibition of human bladder cancer cell motility by genistein is dependent on epidermal growth factor receptor but not p21ras gene expression, Int J *Cancer, 78 (*1998) 775-82.

[350] Kikuchi, H. and Hossain, A., Signal transduction-mediated CYP1A1 induction by

omeprazole in human HepG2 cells, Exp To*xicol Pathol, 51 (*1999) 342-6.

[351]Fajardo, I., Quesada, A.R., Nunez de Castro, I., Sanchez-Jimenez, F. and Medina, M.A., A comparative study of the effects of genistein and 2-methoxyestradiol on the proteolytic balance and tumour cell proliferation, Br J C*ancer, 80 (*1999) 17-24.

[352] Morito, K., Hirose, T., Kinjo, J., Hirakawa, T., Okawa, M., et al., Interaction of phytoestrogens with estrogen receptors alpha and beta, Biol P*harm Bull, 24 (*2001) 351-6.

[353] Kapiotis, S., Hermann, M., Held, I., Seelos, C., Ehringer, H., et al., Genistein, the dietary-derived angiogenesis inhibitor, prevents LDL oxidation and protects endothelial cells from damage by atherogenic LDL, Arteri*oscler Thromb Vasc Biol, 17 (*1997) 2868-74.

[354] Kulling, S.E. and Metzler, M., Induction of micronuclei, DNA strand breaks and HPRT mutations in cultured Chinese hamster V79 cells by the phytoestrogen coumoestrol, Food C*hem Toxicol, 35 (*1997) 605-13.

[355] Dluzewski, A.R. and Garcia, C.R., Inhibition of invasion and intraerythrocytic development of Plasmodium falciparum by kinase inhibitors, Experi*entia, 52 (*1996) 621-3.

[356] Kuzumaki, T., Kobayashi, T. and Ishikawa, K., Genistein induces p21(Cip1/WAF1) expression and blocks the G1 to S phase transition in mouse fibroblast and melanoma cells, Bioche*m Biophys Res Commun, 251* (1998) 291-5.

[357] Prabhakaran, K., Harris, E.B. and Randhawa, B., Regulation by protein kinase of phagocytosis of Mycobacterium leprae by macrophages, J Med *Microbiol, 49 (*2000) 339-42.

[358] Sadowska-Krowicka, H., Mannick, E.E., Oliver, P.D., Sandoval, M., Zhang, X.J., et al., Genistein and gut inflammation: role of nitric oxide, Proc S*oc Exp Biol Med, 217* (1998) 351-7.

[359] Kniss, D.A., Zimmerman, P.D., Su, H.C. and Fertel, R.H., Genistein suppresses EGF-induced prostaglandin biosynthesis by a mechanism independent of EGF receptor tyrosine kinase inhibition, Prosta*glandins, 51 (*1996) 87-105.

[360] Akula, S.M., Hurley, D.J., Wixon, R.L., Wang, C. and Chase, C.C., Effect of genistein on replication of bovine herpesvirus type 1, Am J V*et Res, 63 (*2002) 1124-8.

[361] Yousufzai, S.Y. and Abdel-Latif, A.A., Tyrosine kinase inhibitors suppress prostaglandin F2alpha-induced phosphoinositide hydrolysis, Ca2+ elevation and contraction in iris

sphincter smooth muscle, Eur J *Pharmacol, 360* (1998) 185-93.

[362] Pan, W., Ikeda, K., Takebe, M. and Yamori, Y., Genistein, daidzein and glycitein inhibit growth and DNA synthesis of aortic smooth muscle cells from stroke-prone spontaneously hypertensive rats, J Nutr*, 131* (2001) 1154-8.

[363] Attele, A.S., Wu, J.A. and Yuan, C.S., Ginseng pharmacology: multiple constituents and multiple actions, Bioche*m Pharmacol, 58 (*1999) 1685-93.

[364] Kitts, D. and Hu, C., Efficacy and safety of ginseng, Public *Health Nutr, 3 (*2000) 473-85.

[365] Shin, H.R., Kim, J.Y., Yun, T.K., Morgan, G. and Vainio, H., The cancer-preventive potential of Panax ginseng: a review of human and experimental evidence, Cancer *Causes Control, 11 (*2000) 565-76.

[366] Takino, Y., [Studies on the pharmacodynamics of ginsenoside-Rg1, -Rb1 and -Rb2 in rats], Yakuga*ku Zasshi, 114* (1994) 550-64.

[367] Huo, Y.S., Zhang, S.C., Zhou, D., Yao, D.L., You, G.Y., et al., [Pharmacokinetics and tissue distribution of [3H]ginsenoside Rg1], Zhongg*uo Yao Li Xue Bao, 7 (1*986) 519-21.

[368] Li, J.Q., Li, Z.K., Duan, H. and Zhang, J.T., [Effect of age and ginsenoside Rg1 on nitric oxide content and nitric oxide synthase activity of cerebral cortex in rats], Yao Xu*e Xue Bao, 32 (*1997) 251-4.

[369] Lee, K.Y. and Lee, S.K., Ginsenoside-Rg1 positively regulates cyclin E-dependent kinase activity in human hepatoma SK-HEP-1 cells, Bioche*m Mol Biol Int, 39 (*1996) 539-46.

[370] Cho, S.W., Cho, E.H. and Choi, S.Y., Ginsenosides activate DNA polymerase delta from bovine placenta, Life S*ci, 57 (*1995) 1359-65.

[371] Kenarova, B., Neychev, H., Hadjiivanova, C. and Petkov, V.D., Immunomodulating activity of ginsenoside Rg1 from Panax ginseng, Jpn J *Pharmacol, 54 (*1990) 447-54.

[372] Danieli, B., Falcone, L., Monti, D., Riva, S., Gebhardt, S., et al., Regioselective enzymatic glycosylation of natural polyhydroxylated compounds: galactosylation and glucosylation of protopanaxatriol ginsenosides, J Org *Chem, 66 (*2001) 262-9.

[373] Study on chemoprevention of hepatocellular carcinoma by ginseng: an introduction to the protocol, J Kore*an Med Sci, 16 S*uppl (2001) S70-4.

[374] Middleton, E., Jr., Kandaswami, C. and Theoharides, T.C., The effects of plant flavonoids on mammalian cells: implications for inflammation, heart disease, and cancer, Pharma*col Rev, 52 (*2000) 673-751.

[375] Graefe, E.U., Derendorf, H. and Veit, M., Pharmacokinetics and bioavailability of the flavonol quercetin in humans, Int J *Clin Pharmacol Ther, 37 (*1999) 219-33.

[376] Crespy, V., Morand, C., Besson, C., Manach, C., Demigne, C., et al., Quercetin, but not its glycosides, is absorbed from the rat stomach, J Agric *Food Chem, 50 (*2002) 618-21.

[377] Morand, C., Manach, C., Crespy, V. and Remesy, C., Quercetin 3-O-beta-glucoside is better absorbed than other quercetin forms and is not present in rat plasma, Free R*adic Res, 33 (*2000) 667-76.

[378] Erlund, I., Kosonen, T., Alfthan, G., Maenpaa, J., Perttunen, K., et al., Pharmacokinetics of quercetin from quercetin aglycone and rutin in healthy volunteers, Eur J *Clin Pharmacol, 56 (*2000) 545-53.

[379] Caltagirone, S., Ranelletti, F.O., Rinelli, A., Maggiano, N., Colasante, A., et al., Interaction with type II estrogen binding sites and antiproliferative activity of tamoxifen and quercetin in human non-small-cell lung cancer, Am J R*espir Cell Mol Biol, 17 (*1997) 51-9.

[380] Lamson, D.W. and Brignall, M.S., Antioxidants and cancer, part 3: quercetin, Altern *Med Rev, 5 (2*000) 196-208.

[381] Horcajada-Molteni, M.N., Crespy, V., Coxam, V., Davicco, M.J., Remesy, C., et al., Rutin inhibits ovariectomy-induced osteopenia in rats, J Bone *Miner Res, 15 (*2000) 2251-8.

[382] Formica, J.V. and Regelson, W., Review of the biology of Quercetin and related bioflavonoids, Food C*hem Toxicol, 33 (*1995) 1061-80.

[383] Boege, F., Straub, T., Kehr, A., Boesenberg, C., Christiansen, K., et al., Selected novel flavones inhibit the DNA binding or the DNA religation step of eukaryotic topoisomerase I, J Biol *Chem, 271* (1996) 2262-70.

[384] Duarte, J., Perez-Palencia, R., Vargas, F., Ocete, M.A., Perez-Vizcaino, F., et al., Antihypertensive effects of the flavonoid quercetin in spontaneously hypertensive rats, Br J P*harmacol, 133* (2001) 117-24.

[385] Ohnishi, E. and Bannai, H., Quercetin potentiates TNF-induced antiviral activity, Antivi*ral Res, 22 (*1993) 327-31.

[386] Taguchi, K., Hagiwara, Y., Kajiyama, K. and Suzuki, Y., [Pharmacological studies of Houttuyniae herba: the anti-inflammatory effect of quercitrin], Yakuga*ku Zasshi, 113* (1993) 327-33.

[387] Shoskes, D.A., Effect of bioflavonoids quercetin and curcumin on ischemic renal injury: a new class of renoprotective agents, Transp*lantation, 66 (*1998) 147-52.

[388] Castro, O., Barrios, M., Chinchilla, M. and Guerrero, O., [Chemical and biological evaluation of the effect of plant extracts against Plasmodium berghei], Rev Bi*ol Trop, 44 (*1996) 361-7.

[389] van der Hoeven, J.C., Bruggeman, I.M. and Debets, F.M., Genotoxicity of quercetin in cultured mammalian cells, Mutat *Res, 136* (1984) 9-21.

[390] Knekt, P., Kumpulainen, J., Jarvinen, R., Rissanen, H., Heliovaara, M., et al., Flavonoid intake and risk of chronic diseases, Am J C*lin Nutr, 76 (*2002) 560-8.

[391] Katsarou, A., Davoy, E., Xenos, K., Armenaka, M. and Theoharides, T.C., Effect of an antioxidant (quercetin) on sodium-lauryl-sulfate-induced skin irritation, Contac*t Dermatitis, 42 (*2000) 85-9.

[392] Guilbaud, N., Kraus-Berthier, L., Meyer-Losic, F., Malivet, V., Chacun, C., et al., Marked antitumor activity of a new potent acronycine derivative in orthotopic models of human solid tumors, Clin C*ancer Res, 7 (*2001) 2573-80.

[393] Shieh, H.L., Pezzuto, J.M. and Cordell, G.A., Evaluation of the cytotoxic mechanisms mediated by the broad-spectrum antitumor alkaloid acronycine and selected semisynthetic derivatives, Chem B*iol Interact, 81 (*1992) 35-55.

[394] Dorr, R.T., Liddil, J.D., Von Hoff, D.D., Soble, M. and Osborne, C.K., Antitumor activity and murine pharmacokinetics of parenteral acronycine, Cancer *Res, 49 (*1989) 340-4.

[395] Ikemoto, S., Sugimura, K., Yoshida, N., Yasumoto, R., Wada, S., et al., Antitumor effects of Scutellariae radix and its components baicalein, baicalin, and wogonin on bladder cancer cell lines, Urolog*y, 55 (*2000) 951-5.

[396] Lin, C.C. and Shieh, D.E., The anti-inflammatory activity of Scutellaria rivularis extracts and its active components, baicalin, baicalein and wogonin, Am J C*hin Med, 24 (*1996) 31-6.

[397] De Clercq, E., Current lead natural products for the chemotherapy of human immunodeficiency virus (HIV) infection, Med Re*s Rev, 20 (*2000) 323-49.

[398] Zhou, Y.P. and Zhang, J.Q., Oral baicalin and liquid extract of licorice reduce sorbitol

levels in red blood cell of diabetic rats, Chin M*ed J (Engl), 102* (1989) 203-6.

[399] Nagai, T., Yamada, H. and Otsuka, Y., Inhibition of mouse liver sialidase by the root of Scutellaria baicalensis, Planta *Med, 55 (*1989) 27-9.

[400] Akao, T., Kawabata, K., Yanagisawa, E., Ishihara, K., Mizuhara, Y., et al., Baicalin, the predominant flavone glucuronide of scutellariae radix, is absorbed from the rat gastrointestinal tract as the aglycone and restored to its original form, J Phar*m Pharmacol, 52 (*2000) 1563-8.

[401] Wu, J., Chen, D. and Zhang, R., Study on the bioavailability of baicalin-phospholipid complex by using HPLC, Biomed *Chromatogr, 13 (*1999) 493-5.

[402] Kitamura, K., Honda, M., Yoshizaki, H., Yamamoto, S., Nakane, H., et al., Baicalin, an inhibitor of HIV-1 production in vitro, Antivir*al Res, 37 (*1998) 131-40.

[403] Liu, W., Kato, M., Akhand, A.A., Hayakawa, A., Takemura, M., et al., The herbal medicine sho-saiko-to inhibits the growth of malignant melanoma cells by upregulating Fas-mediated apoptosis and arresting cell cycle through downregulation of cyclin dependent kinases, Int J *Oncol, 12 (*1998) 1321-6.

[404] Nakahata, N., Kutsuwa, M., Kyo, R., Kubo, M., Hayashi, K., et al., Analysis of inhibitory effects of scutellariae radix and baicalein on prostaglandin E2 production in rat C6 glioma cells, Am J C*hin Med, 26 (*1998) 311-23.

[405] Kyo, R., Nakahata, N., Sakakibara, I., Kubo, M. and Ohizumi, Y., Baicalin and baicalein, constituents of an important medicinal plant, inhibit intracellular Ca2+ elevation by reducing phospholipase C activity in C6 rat glioma cells, J Phar*m Pharmacol, 50 (*1998) 1179-82.

[406] Huang, Y., Tsang, S.Y., Yao, X., Lau, C.W., Su, Y.L., et al., Baicalin-induced vascular response in rat mesenteric artery: role of endothelial nitric oxide, Clin E*xp Pharmacol Physiol, 29 (*2002) 721-4.

[407] Zhang, L., Lau, Y.K., Xia, W., Hortobagyi, G.N. and Hung, M.C., Tyrosine kinase inhibitor emodin suppresses growth of HER-2/neu-overexpressing breast cancer cells in athymic mice and sensitizes these cells to the inhibitory effect of paclitaxel, Clin C*ancer Res, 5 (1*999) 343-53.

[408] Huang, H.C., Chang, J.H., Tung, S.F., Wu, R.T., Foegh, M.L., et al., Immunosuppressive effect of emodin, a free radical generator, Eur J *Pharmacol, 211* (1992) 359-64.

[409] Liang, J.W., Hsiu, S.L., Wu, P.P. and Chao, P.D., Emodin pharmacokinetics in rabbits,

Planta *Med, 61 (*1995) 406-8.

[410] Lang, W., Pharmacokinetic-metabolic studies with 14C-aloe emodin after oral administration to male and female rats, Pharma*cology, 47 S*uppl 1 (1993) 110-9.

[411] Jinsart, W., Ternai, B. and Polya, G.M., Inhibition of myosin light chain kinase, cAMP-dependent protein kinase, protein kinase C and of plant Ca(2+)-dependent protein kinase by anthraquinones, Biol C*hem Hoppe Seyler, 373* (1992) 903-10.

[412] Kumar, A., Dhawan, S. and Aggarwal, B.B., Emodin (3-methyl-1,6,8-trihydroxyanthraquinone) inhibits TNF-induced NF-kappaB activation, IkappaB degradation, and expression of cell surface adhesion proteins in human vascular endothelial cells, Oncog*ene, 17 (*1998) 913-8.

[413] Goel, R.K., Das Gupta, G., Ram, S.N. and Pandey, V.B., Antiulcerogenic and anti-inflammatory effects of emodin, isolated from Rhamnus triquerta wall, Indian *J Exp Biol, 29 (*1991) 230-2.

[414] Guo, D., Xu, C. and Chen, Y., [A study on the effect of emodin on smooth muscle cell proliferation], Zhongh*ua Nei Ke Za Zhi, 35 (*1996) 157-9.

[415] Ali, M., Al-Qattan, K.K., Al-Enezi, F., Khanafer, R.M. and Mustafa, T., Effect of allicin from garlic powder on serum lipids and blood pressure in rats fed with a high cholesterol diet, Prosta*glandins Leukot Essent Fatty Acids, 62 (*2000) 253-9.

[416] Ankri, S. and Mirelman, D., Antimicrobial properties of allicin from garlic, Microb*es Infect, 1 (1*999) 125-9.

[417] Jonkers, D., van den Broek, E., van Dooren, I., Thijs, C., Dorant, E., et al., Antibacterial effect of garlic and omeprazole on Helicobacter pylori, J Anti*microb Chemother, 43 (*1999) 837-9.

[418] Shalinsky, D.R., McNamara, D.B. and Agrawal, K.C., Inhibition of GSH-dependent PGH2 isomerase in mammary adenocarcinoma cells by allicin, Prosta*glandins, 37 (*1989) 135-48.

[419] Prasad, K., Laxdal, V.A., Yu, M. and Raney, B.L., Antioxidant activity of allicin, an active principle in garlic, Mol Ce*ll Biochem, 148* (1995) 183-9.

[420] Mathew, P.T. and Augusti, K.T., Studies on the effect of allicin (diallyl disulphide-oxide) on alloxan diabetes. I. Hypoglycaemic action and enhancement of serum insulin effect and glycogen synthesis, Indian *J Biochem Biophys, 10 (*1973) 209-12.

[421] Augusti, K.T., Studies on the effect of allicin (diallyl disulphide-oxide) on alloxan

diabetes, Experi*entia, 31 (*1975) 1263-5.

[422] Agarwal, K.C., Therapeutic actions of garlic constituents, Med Re*s Rev, 16 (*1996) 111-24.

[423] Damianaki, A., Bakogeorgou, E., Kampa, M., Notas, G., Hatzoglou, A., et al., Potent inhibitory action of red wine polyphenols on human breast cancer cells, J Cell *Biochem, 78 (*2000) 429-41.

[424] Sakamoto, K., Synergistic effects of thearubigin and genistein on human prostate tumor cell (PC-3) growth via cell cycle arrest, Cancer *Lett, 151* (2000) 103-9.

[425] Liang, Y.C., Lin-Shiau, S.Y., Chen, C.F. and Lin, J.K., Inhibition of cyclin-dependent kinases 2 and 4 activities as well as induction of Cdk inhibitors p21 and p27 during growth arrest of human breast carcinoma cells by (-)-epigallocatechin-3-gallate, J Cell *Biochem, 75 (*1999) 1-12.

[426] Sachinidis, A., Seul, C., Seewald, S., Ahn, H., Ko, Y., et al., Green tea compounds inhibit tyrosine phosphorylation of PDGF beta-receptor and transformation of A172 human glioblastoma, FEBS L*ett, 471* (2000) 51-5.

[427] Gupta, S., Ahmad, N., Nieminen, A.L. and Mukhtar, H., Growth inhibition, cell-cycle dysregulation, and induction of apoptosis by green tea constituent (-)-epigallocatechin-3-gallate in androgen-sensitive and androgen-insensitive human prostate carcinoma cells, Toxico*l Appl Pharmacol, 164* (2000) 82-90.

[428] Demeule, M., Brossard, M., Page, M., Gingras, D. and Beliveau, R., Matrix metalloproteinase inhibition by green tea catechins, Biochi*m Biophys Acta, 1478* (2000) 51-60.

[429] Ahmad, N., Gupta, S. and Mukhtar, H., Green tea polyphenol epigallocatechin-3-gallate differentially modulates nuclear factor kappaB in cancer cells versus normal cells, Arch B*iochem Biophys, 376* (2000) 338-46.

[430] Tsuchiya, H., Effects of green tea catechins on membrane fluidity, Pharma*cology, 59 (*1999) 34-44.

[431] Nagata, H., Takekoshi, S., Takagi, T., Honma, T. and Watanabe, K., Antioxidative action of flavonoids, quercetin and catechin, mediated by the activation of glutathione peroxidase, Tokai *J Exp Clin Med, 24 (*1999) 1-11.

[432] Polya, G.M. and Foo, L.Y., Inhibition of eukaryote signal-regulated protein kinases by plant-derived catechin-related compounds, Phyto*chemistry, 35 (*1994) 1399-405.

[433] Liang, Y.C., Lin-shiau, S.Y., Chen, C.F. and Lin, J.K., Suppression of extracellular signals and cell proliferation through EGF receptor binding by (-)-epigallocatechin gallate in human A431 epidermoid carcinoma cells., J Cell *Biochem., 67 (*1997) 55-65.

[434] Chung, J.Y., Huang, C., Meng, X., Dong, Z. and Yang, C.S., Inhibition of activator protein 1 activity and cell growth by purified green tea and black tea polyphenols in H-ras-transformed cells: structure-activity relationship and mechanisms involved, Cancer *Res, 59 (*1999) 4610-7.

[435] Brattig, N.W., Diao, G.J. and Berg, P.A., Immunoenhancing effect of flavonoid compounds on lymphocyte proliferation and immunoglobulin synthesis, Int J *Immunopharmacol, 6 (1*984) 205-15.

[436] Komori, A., Yatsunami, J., Okabe, S., Abe, S., Hara, K., et al., Anticarcinogenic activity of green tea polyphenols, Jpn J *Clin Oncol, 23 (*1993) 186-90.

[437] Mantle, D., Lennard, T.W. and Pickering, A.T., Therapeutic applications of medicinal plants in the treatment of breast cancer: a review of their pharmacology, efficacy and tolerability, Advers*e Drug React Toxicol Rev, 19 (*2000) 223-40.

[438] Sinha, B.K., Topoisomerase inhibitors. A review of their therapeutic potential in cancer., Drugs, *49 (*1995) 11-19.

[439] Martelli, A.M., Bortul, R., Bareggi, R., Tabellini, G., Grill, V., et al., The pro-apoptotic drug camptothecin stimulates phospholipase D activity and diacylglycerol production in the nucleus of HL-60 human promyelocytic leukemia cells, Cancer *Res, 59 (*1999) 3961-7.

[440] Nieves-Neira, W. and Pommier, Y., Apoptotic response to camptothecin and 7-hydroxystaurosporine (UCN-01) in the 8 human breast cancer cell lines of the NCI Anticancer Drug Screen: multifactorial relationships with topoisomerase I, protein kinase C, Bcl-2, p53, MDM-2 and caspase pathways, Int J *Cancer, 82 (*1999) 396-404.

[441] Eymin, B., Dubrez, L., Allouche, M. and Solary, E., Increased gadd153 messenger RNA level is associated with apoptosis in human leukemic cells treated with etoposide, Cancer *Res., 57 (*1997) 686-695.

[442] Matsumoto, Y., Fujiwara, T. and Nagao, S., Determinants of drug response in camptothecin-11-resistant glioma cell lines, J Neur*ooncol, 23 (*1995) 1-8.

[443] Wang, M.C., Liu, J.H. and Wang, F.F., Protein tyrosine phosphatase-dependent activation of beta-globin and delta-aminolevulinic acid synthase genes in the camptothecin-induced IW32 erythroleukemia cell differentiation, Mol Ph*armacol, 51 (*1997) 558-66.

[444] Persidis, A., Proteomics, Nat Bi*otechnol, 16 (*1998) 393-4.

[445] Behr, J.-P., The lo*ck-and-key principle : the state of the art--100 years on, Wile*y, Chichester England ; New York, 1994, ix, 325 pp.

[446] Singh, U.C. and Peter, A.K., An approach to computing electrostatic charges for molecules, *J. Comput. Chem*.*,* 5(2) (1984) 129-145.