DATABASE DEVELOPMENT AND MACHINE LEARNING CLASSIFICATION OF MEDICINAL CHEMICALS AND BIOMOLECULES

PANKAJ KUMAR
(M.Pharm, BITS-Pilani; B.Pharm, IT-BHU)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF PHARMACY

NATIONAL UNIVERSITY OF SINGAPORE

2009

Acknowledgements

I would like to present my sincere thanks to my supervisor, Professor Chen Yu Zong, for his invaluable guidance and being a wonderful mentor. I have benefited tremendously from his profound knowledge, expertise in research, as well as his enormous support. My appreciation for his mentorship goes beyond my words.

Special thanks go to our present and previous BIDD Group members. In particulars, I would like to thank Dr. Yap Chun Wei, Dr. Li Hu, Dr. Ung CY, Ms Xiaohua Ma, Ms Jiajia, Mr Zhu Feng, Ms Shi Zhe, Ms Liu Xin, Mr. Xiang hui, Mr. Han Bucong, and our research staffs. A special appreciation goes to my wife, my parents, and my friends for love and support.

Table of Contents

Summary

The drug discovery is a long and time-consuming process that also requires huge sums of financial investment. Advances in bioinformatics areas such as database development and machine learning methods have played a great role in reducing the time and money invested, rationalizing the entire approach, and increasing efficiency for drug discovery processes. Focus of my work has been to aid the drug discovery processes applying various computational methods. A particular focus has been given to improvise the storing, managing and providing the customized data by developing web accessible databases of medicinal chemicals and biomolecules; i.e. (i) Updating of Kinetic Database of Biomolecular Interactions(KDBI), and (ii) Indian Herbs and their Chemical Database(IHCD) . Also, focus has been given on the use of machine learning classification by predicting the medicinal chemicals for (i) genotoxicity, and (ii) p38 inhibitors.

Database development for biological and chemical data is explored from the beginning of data collection to deploying of web application. Biological and chemical data which can be helpful in drug discovery process are used for this purpose. The complexities involved such as biological data collection, filtering, cross-linking to other database, providing web accessibility, facilitating data download, and modeling of databases are explained in detail. The two databases, IHCD and KDBI, developed have different kind of data content and cover a broad area of biological and chemical databases space. IHCD contain information on a total of 2326 herbs from 430 therapeutic classes and 3978 chemical ingredients. IHCD also contain information about chemical ingredient through cross-linking to chemical, pathway, and molecular binding databases PUBCHEM, NCBI bioassay, KEGG pathways, BIND, and bindingDB databases respectively. IHCD also provides 3D structure, computed molecular descriptors for all ingredients, and computer predicted potential protein targets and binding structures for select ingredients. The other database, KDBI, contain information on 19263 experimental kinetic data, which include 2635 protein-protein, 1711 protein-nucleic acid, 11873 protein-small molecule, and 1995 nucleic acid-small molecule interactions. KDBI also has 63 literature reported pathway simulation model kinetic parameter data set and provides facility to download each pathway kinetic dataset in SBML file format.

Machine Learning Classification methods are employed in areas that are directly linked to early stage of drug discovery such as predicting genotoxic compounds and p38 MAPK inhibitor by collecting more than 4000 genotoxic compounds and about 1100 p38 MAPK inhibitors. Different types of machine learning methods such as SVM, kNN, PNN and decision trees are applied for these studies, although the special focus is on SVM. Also, machine learning based virtual screening is done on PUBCHEM and MDDR database. A total of 522 molecular descriptors were calculated for each compound to represent compounds and either entire 522 or selected 100 descriptors were used for machine learning classification.

List of Tables

List of Figures

List of Abbreviations

API: Application Programming Interface

DT: Decision Tree

FDA: Food and Drug Administration

FP: False Positive

FN: False Negative

GT: Genotoxicity

IHCD: Indian Herbs and Chemical Database

KDBI: Kinetic Database of Biomolecular Interactions

k-NN: k Nearest Neighbor

MAPK: Mitogen Activated Protein Kinase

MLC: Machine Learning Classification

MLM: Machine Learning Methods

MCC: Matthews's correlation coefficient

PNN: Probabilistic Neural Network

SBML: System Biology Markup Language

SVM: Support Vector Machine

SEN: Sensitivity

SP: Specificity

TN: True Negative

TP: True Positive

WEKA: Waikato Environment for Knowledge Analysis

XML: Extensible Mark-up Language

List of Publications

1. Update of KDBI: Kinetic Data of Bio-molecular Interaction Database. **Pankaj Kumar**, Z.L. Ji, B.C. Han, Z. Shi, J. Jia, Y.P, Wang, Y.T. Zhang, L. Liang, and Y. Z. Chen. ***Nucleic Acids Res***. 2009 37: D636-D641; (PUBMED ID: 18971255).

2. Automation in Understanding the Molecular Mechanisms of Herbal Ingredients and Herbal Plants: Novel approach. **Pankaj Kumar**, Y. Z. Chen. 19th Singapore Pharmacy Congress 2007.

3. Update of TTD: Therapeutic Target Database. F. Zhu, B.C. Han, P. Kumar, X.H. Liu, X.H. Ma, X.N. Wei, L. Huang, Y.F. Guo, L.Y. Han, C.J. Zheng, Y.Z. Chen. *Nucleic Acids Res.* 38(Database issue):D787-91(2010). Pubmed

4. Effect of Training Data Size and Noise Level on Support Vector Machines Virtual Screening of Genotoxic Agents from Large Compound Libraries. Kumar, Pankaj; Ma, Xiaohua; Liu, XiangHui; jia, Jia; Bucong, Han; Ying, Xue; Li, Ze-Rong; Yang, Shengyong; Yap, Chun Wei; Chen, Yu Zong (Submitted to *Chemical Research in Toxicology*)

Chapter 1 Introduction

Drug discovery is a long and time-consuming process that requires huge sums of monetary/financial investment. Many studies have been done to find the strategies for reducing the time, for reducing the cost and for increasing the efficiency to cover a number of drugs in the drug discovery process. This work on "Database development and machine learning classification of medicinal chemicals and biomolecules" is one of such kind of strategy which is introduced in this chapter along with the background of Drug Discovery and Bioinformatics. This chapter consists five parts: (1) Drug Discovery (Section 1.1) (2) Bioinformatics in Drug Discovery (Section 1.2) (3) Database development of medicinal chemicals and biomolecules and their roles in drug discovery (Section 1.3) (4) Machine learning classification of medicinal chemicals as a tool in drug discovery (Section 1.4). (5) Objectives of my PhD projects (Section 1.5)

1.1 Drug discovery

A typical drug discovery process involves the identification of candidates, synthesis, characterization, screening, and assays for therapeutic efficacy. Once a compound has shown its value in these initial assays, it will go for the process of drug development prior to clinical trials. The whole process takes about 10-17 years, $800 million (as per conservative estimates), and has less than 10% overall probability of success. There is a significant productivity gap in drug discovery and is of major concern for biopharmaceutical industry. The global pharmaceutical market is worth US$ 712 billion (Malik 2008). Compared to the huge R&D investment in implementing new technologies for drug discovery, return is insignificant (Ashburn and Thor 2004). Search of novel undiscovered compounds has motivated many pharmaceutical companies and scientists for the last few decades, but difficulties in getting new molecules out with respect to time and money has slowed the momentum of drug discovery in recent times and this slowdown trend is expected to continue (Malik 2008). **Figure 1** shows the investment done in drug discovery and corresponding number of new chemical entities (NCEs) approved by Food and Drug Administration (FDA) every year starting from 1992.

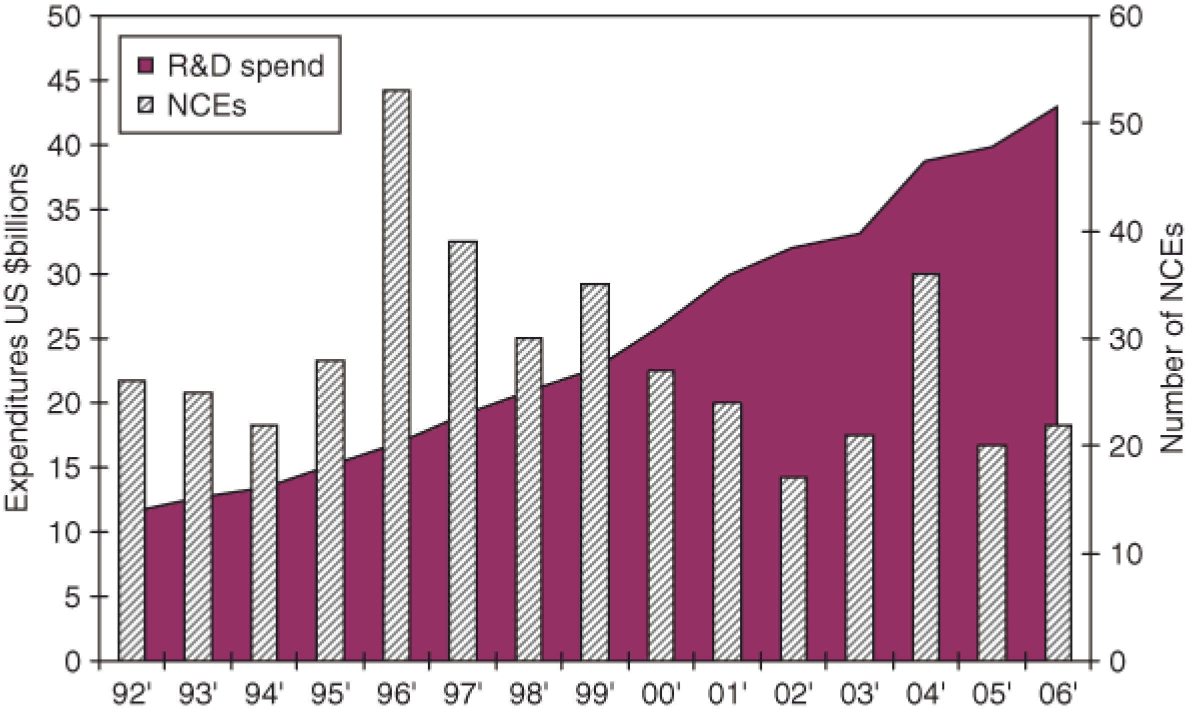


Figure 1: Number of new chemical entities (NCEs) in relation to research and development (R&D) spending (1992–2006). Source: Pharmaceutical Research and Manufacturers of America and the US Food and Drug Administration (Sollano, Kirsch et al. 2008).

Drugs, in the past, have been discovered either by finding the active ingredient from traditional medicines or by serendipitous discovery (Kaul 1998). Long before the advent of pharmaceutical industry, the usage of these drugs discovered by trial and error were passed down by verbal and written records (Ratti and Trist 2001). Lack of data management about these discovery and traditional medicines have been a reason of underutilization of these findings by pharmaceutical industries. In mid 20th century, this drug discovery process by trial and error started having little rationalization by screening the known drug like compounds by randomly testing for activity. In this progression, lead molecules found by chance or from screening the diverse chemical libraries were followed by lead optimization. Slowly, when the understanding of diseases and mechanism of action for drugs started becoming clearer, the rational approach was sought for drug discovery.

In this rational approach, *in vitro* assays on animal tissues became the standard way and well-liked for the process of getting valuable information on structure–activity relationships and pharmacophore construction. By this approach, even if the lead molecule fails there is adequate information about the cause of failure in terms of structure or physiochemical descriptors which should be modified in the molecules. In similar way, many such strategies got developed in time to rationalize the drug discovery process.

Recently, the strategy of finding a therapeutic role of an existing compound has become popular (**Figure 2**). Moreover, finding new therapeutic role for an existing drug has also become desired area of research. The number of drug like candidates is increasing very rapidly (around 170,000) (MDL Information System Inc 2004; 2004) in comparison to limited number of potential therapeutic target (around 1500) (Hopkins and Groom 2002). Some researchers speculate that existing drugs and candidates may have covered a significant number of potential drug targets (Ji, Kong et al. 2007; McArdle and Quinn 2007; Park and Kim 2008) and single drug can bind to multiple receptors(Paolini, Shapland et al. 2006; Yildirim, Goh et al. 2007) for producing the effects. The present chemical space of drugs like candidates constitutes highly diversified compounds and mining of this space may produce good drugs (Kong, Li et al. 2009).

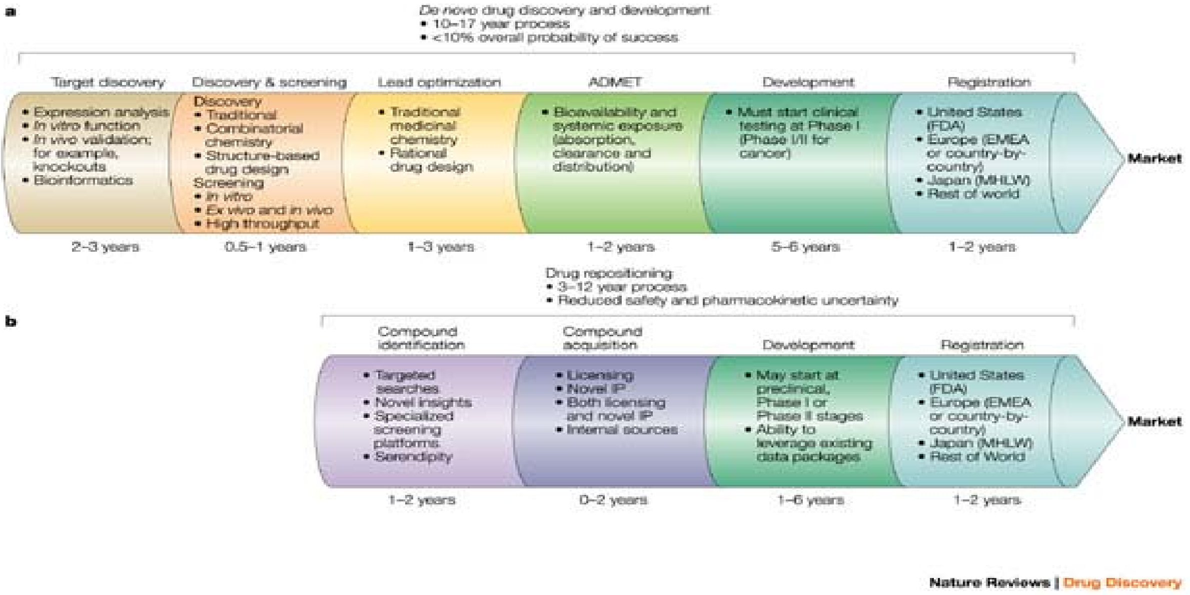

a
De novo drug discovery and development
• 10–17 year process
• <10% overall probability of success
Target discovery
• Expression analysis
• In vitro function
• In vivo validation; for example, knockouts
• Bioinformatics
2–3 years
Discovery & screening
Discovery
• Traditional
• Combinatorial chemistry
• Structure-based drug design
Screening
• In vitro
• Ex vivo and in vivo
• High throughput
0.5–1 years
Lead optimization
• Traditional medicinal chemistry
• Rational drug design
1–3 years
ADMET
• Bioavailability and systemic exposure (absorption, clearance and distribution)
1–2 years
Development
• Must start clinical testing at Phase I (Phase I/II for cancer)
5–6 years
Registration
• United States (FDA)
• Europe (EMEA or country-by-country)
• Japan (MHLW)
• Rest of world
1–2 years
Market
b
Drug repositioning
• 3–12 year process
• Reduced safety and pharmacokinetic uncertainty
Compound identification
• Targeted searches
• Novel insights
• Specialized screening platforms
• Serendipity
1–2 years
Compound acquisition
• Licensing
• Novel IP
• Both licensing and novel IP
• Internal sources
0–2 years
Development
• May start at preclinical, Phase I or Phase II stages
• Ability to leverage existing data packages
1–6 years
Registration
• United States (FDA)
• Europe (EMEA or country-by-country)
• Japan (MHLW)
• Rest of World
1–2 years
Market
Nature Reviews | Drug Discovery

Figure 2 : A comparison of traditional (a) *de novo* drug discovery and development versus (b) drug repositioning. (Ashburn and Thor 2004)

In 1990s, areas like molecular biology, cellular biology and genomics grew rapidly which helped in understanding disease pathways and processes into their molecular and genetic components to recognize the cause of malfunction precisely, and problematic point seeking therapeutic intervention. This progress helped in finding many new molecular targets and number of molecular targets increased significantly (from approximately 500 to more than 10,000 targets) which could be utilized for the discovery of novel methods for the prevention, diagnosis, and treatment of human diseases (Newman 2008). This was accompanied by development of ultra high throughput screening (ultra-HTS) for screening extensive chemical libraries upon a small number of biological targets such as enzyme or a cell-surface receptor. The method usually follows combinatorial chemistry which produces chemical compounds of interest with extremely high speed, and these compounds may respond positively in assay upon the desired target. While there has been some success with this approach, the number of innovative discoveries has been confined (Koehn and Carter 2005).

To further improvise the drug discovery processes, systems biology has a comprehensive approach by analyzing biological operation, cellular processes and disease-mediated processes at a systems-level to understand the difficult to determine underlying causes, and research options for treatment (Davidov, Holland et al. 2003). This is facilitated by combining feedback from genomics (global gene expression analysis and whole genome functional analysis), proteomics (protein structure and function), and metabolomics (measurement of metabolite concentrations and fluxes and secretions in cells and tissues that have a direct connection to genetic, protein, and metabolic activity) to incorporate data such as structurally defined chemical libraries with specific biological pathway information (Nicholson and Wilson 2003). Systems biology integrates massive quantities of complex data generated by genomic, proteomic and metabolic analyses to understand phenotypic variation and build comprehensive models of cellular organization and function. The objective of studying complex relationships is to use research findings to better define targets with the intent of developing more effective therapies (Harrill and Rusyn 2008). Furthermore, systems biology is newly forming as an access to drug discovery that will assist pharmaceutical companies to produce more effective drugs with small side effects in addition to lower the development time and costs. Systems biology uses a combining approach to know the performance of biological systems as they answer to perturbations in their surrounding condition such as the administration of drugs. System biology has caused encouragement in the drug discovery society; though drug companies for the most part are not following this approach. While the study is commonly accepted to be yielding, the time it will take for the research to turn applicable to drug companies is not perceived. There can be increase in number of companies based on systems biology which can help in early stage of drug discovery (Cho, Labow et al. 2006; Schrattenholz and Soskic 2008).

An important archetype in drug discovery is the design of selective agents to act on individual drug targets. In contrast, some drugs have effect on multiple targets, such as Gleevec (Petrelli and Giordano 2008; Zhang, Crespo et al. 2008). Advances in systems biology are revealing phenotypic robustness and network structures that strongly suggest that elegantly selective compounds, compared with multi-target drugs, may produce lower than desired clinical efficacy. This new appreciation of the role of pharmacology has significant implications for handling the two prime sources of attritions in drug development - efficacy and toxicity. A promising way to develop more effective and less toxic candidates for druggable targets is the integration of system biology and pharmacology based on the explosively growing biomedical data (Jenwitheesuk, Horst et al. 2008; Schadt, Friend et al. 2009). Even if a compound shows high selectivity and specificity to a disease-causing protein in pre-clinical studies, there is no guarantee that the compound can succeed as a drug in clinical phase. This is due to several important aspects in pharmacology: pharmacokinetics, pharmacodynamics and toxicity. Toxicity is the side effects that can be caused by the multiple targets of the drug candidates through interfering cells normal functions. Phase I clinical trials for a compound involves years of painstaking preclinical testing and yet has only an 8% chance of reaching the market. Toxicity results in the further reduction by 20% of such molecules during late development stages. Therefore, the implementation of toxicity testing as early as possible in the drug development process is of primary significance (Custer and Sweder 2008).

Huge amounts of compounds necessary for *in vivo* studies, dearth of reliable high-throughput assays, and the inability of *in vitro* and animal models to correctly predict toxicities in human are the main reasons that prevent pharmaceutical companies from conducting earlier screening for toxicity. These problems can be addressed through the development of computational or *in silico* toxicity prediction tools, either structure-based or ligand-based approaches which involve the application of modeling techniques on human data. These serve as main approaches to extract potentially toxic effects in humans even before the physical availability of compounds.

By looking at challenges involved in drug discovery processes, there should be innovative ways in drug discovery which cut down the time and financial investment. One of the great ways of achieving this is using bioinformatics in drug discovery.

1.2 Bioinformatics in Drug discovery

Computational methods and bioinformatics tools like predictions of biological activity and virtual screening can help in reducing the cost and time taken in drug discovery process. This can help in pursuing only the most promising experiments and can eliminate many unnecessary experiments beforehand. According to the BCC research report, the worldwide value of bioinformatics is expected to increase from $1.02 billion in 2002 to $3.0 billion in 2010, at an average annual growth rate (AAGR) of 15.8% (**Figure 3**). The use of bioinformatics in drug discovery is likely to reduce the annual cost by 33%, and the time by 30% for developing a new drug.

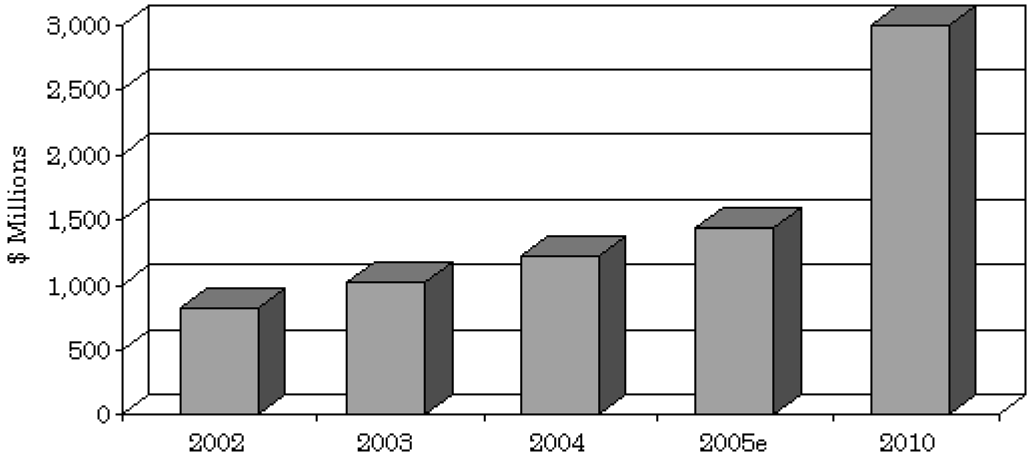


Figure 3: Worldwide value of bioinformatics Source (BCC Research[1])

The increasing pressure to discover or invent more drugs in less time has resulted in noteworthy significance of bioinformatics. By applying bioinformatics tools, it is now possible to start with the compound which explicitly targets a desired protein or group of protein (multi-targeting). Thus the whole process is no longer on a trial and error based like the traditional approach of drug discovery in which a compound with probable pharmacological activity is

[1] http://www.bccresearch.com/report/BIO051A.html

isolated and then tested on animals and subsequently in human during clinical trials. Bioinformatics has helped in making a rational approach for the drug discovery process. Bioinformatics tools are getting developed which are capable to congregate all the required information regarding potential targets like nucleotide and protein sequencing, homologue mapping (Muller, MacCallum et al. 1999; Friedberg, Kaplan et al. 2000), function prediction(Li, Lin et al. 2006; Chen, Chen et al. 2008), pathway information (Cerami, Bader et al. 2006), structural information (Cases, Pisano et al. 2007) and disease associations (Nakazato, Takinaka et al. 2008). The availability of the information about potential targets into databases can help pharmaceutical companies in saving time and money exerting efforts on targets that will fail later.

Rapid development in bioinformatics have accumulated huge amount of biological data. It becomes necessary to organize these data which is also an area of great interest in bioinformatics. With the growth of biological databases and data mining approaches, to extract or filter valuable targets or compounds by combining biological thoughts with computational tools or methods has changed the way drug discovery is conducted. Here, in this thesis, the work has been done to aid the drug discovery processes in general by applying various computational methods. A particular focus has been given to improvising the storing, managing and providing the customized data by developing web accessible databases of medicinal chemicals and biomolecules. The second focus has been given on the use machine learning classification as helper in drug development processes by classifying medicinal chemicals.

1.3 Database development of medicinal chemicals and biomolecules and their role in drug discovery

Role of database development is vital in drug discovery for managing and analyzing the expanding magnitudes of diverse chemical and biological data. Databases of medicinal chemicals and biomolecules are very important to accelerate the medicinal research. It helps in fast search of medicinal chemicals and biomolecules for their categories, mechanism, sources like information. Many public and commercial databases have been developed for these purposes (Southan, Varkonyi et al. 2007). Some of these databases provide comprehensive information for broad category of medicinal chemicals, biomolecules or literature. One of the most widely used literature based public database is Pubmed database which has more than 18 million citations from more than 20,400 life science journals. Over 9.8 million of these citations have abstracts, and 8.7 million of these abstracts have links to their full text articles (Sayers, Barrett et al. 2009). Other very popular databases like, Pubchem and CAS database are most general chemical information databases. Pubchem is a public database by NIH which contain information about chemical, structural and biological properties of small molecules, in particular their roles as diagnostic and therapeutic agents. Pubchem itself has three categorized databases: PCSubstance for substance information, PCCompound for compound structures and PCBioAssay for bioactivity data. Pubchem databases hold records for nearly 41 million substances containing over 19 million unique structures. More than 750 000 of these substances have bioactivity data in at least one of the nearly 1200 Pubchem Bioassays (Sayers, Barrett et al. 2009). Another leading chemical database is CAS which is short form for Chemical Abstract Service by American Chemical Society. CAS is the largest databases of chemistry-related information, and provides searchable interface through SciFinder (a commercial search and retrieval software) and STN (Scientific & Technical Information Network) which provides links to the original literature and patents.

Most of these big databases provide extensive cross-linking and cross-referencing. The search output is generally full of hyperlinks which can link to other databases for detailed information. Pubmed has controlled vocabulary indexing of articles in the form of Medicine Medical Subject Headings (MeSH), which link compound names to journal articles. Similarly, the Protein Data Bank (PDB) (Berman, Westbrook et al. 2000) which stores protein structure data is linked to Uniprot for protein sequences (Bairoch, Apweiler et al. 2005; 2009).

Some database just covers specific areas with in-depth information. For example, NCI and SuperNatural (Dunkel, Fullbeck et al. 2006) are specific databases about chemical information of cancer related and natural compounds resources respectively. Uniprot and KEGG are very popular databases which contain information about biomolecules like proteins and enzyme respectively. Databases of biomolecules are very important for understanding the biological systems and pathways or pharmacological and pharmacokinetic aspect of drugs. Databases addressing specific biological and medicinal problems require innovative databases perspectives.

The vast amount of biological information and their widespread usage by scientists for research purpose is creating new challenges for the database development. Several gene, protein, and small-molecule dealings databases have been justified for these pursuits. The data are generally collected from different sources like public databanks, proprietary data providers, biological, pharmacological, synthetic or simulation experiments. These data can be of various types, including very organized data type like relational database tables and XML files, disorganized web pages or flat files, and small or large objects like three-dimensional (3D) biochemical structures. Most of these data often lack common data formats or the common record identifiers that are required for interoperability. Also, there is a high rate of development of system biology, which demands and produces computer readable data format and thus further increases the complexity of data management. To combine information regarding disjointed biological case, databases are required to fill in information gaps to the growing application of systems-level research. Databases based on machine input/output data assist researchers in using data directly into the software without further processing e.g. database on Systems Biology Markup Language (SBML) helps in creating machine-executable simulation models rather than simple human-readable file format.

Majority of these high quality biological or chemical database which are very useful to scientific community are being published by leading journals like Nucleic Acids Research, Bioinformatics and Journal of Chemical Informatics and Modeling for biological, bioinformatics and chemical databases respectively. Nucleic Acids Research, which is one of the leading journal for biological community, started its annual database issue in 1993 with 24 database has now 179 database published in 2009 making the total sum of 1170 databases (Galperin and Cochrane 2009). Research community is well aware of the importance of database and its availability to user instantly. For this purpose, Nucleic Acid research has made database papers as open access and also generally publishes web accessible databases (Galperin and Cochrane 2009).

Recent trend is that the databases should be accessible through web browser. This web accessible feature has outstanding advantages over the local databases. Web accessible databases become instantly available to user though internal browsers. Current web interfaces of biological data sources generally provide many user-specified criteria as part of queries. With such capability, the accessibility of customized records from the query results becomes a very easy process even for naive users. Researchers who want to use data from web databases for their research generally take advantage of advanced features like data retrieval in other than plain format, programs to collect the data because the manual collection of large number of records is not convenient.

Some specific databases may provide data to be readily used in many computational methods or studies directly or with little preprocessing which otherwise would require manual data collection from literature. In pace with database development, computational methods like machine learning classification is flourishing which generally require large amount of categorized data to make prediction models. Development in machine learning classification method is serving a great need in drug discovery processes. The detailed introduction of machine learning classification is provided in next section.

1.4 Machine learning classification of medicinal chemicals and biomolecules as tools in drug discovery

Machine learning has been defined in number of ways. Some of these definitions are , ‘The ability of a program to learn from experience — that is, to modify its execution on the basis of newly acquired information[2] ’, ‘The ability of a machine to improve its performance based on previous results[3] ’ , ‘The process by which computer systems can be directed to improve their performance over time[4] ’ , and ‘Machine learning is a branch of computer science covering software that uses data to improve its accuracy at some given task[5] ’. Machine learning has been applied in many fields e.g. robotics (Miglino, Lund et al. 1995; Vidovszky, Smith et al. 2006; Zeng, Teo et al. 2008), stock market analysis , machine perception, detecting credit card fraud, brain-machine interfaces (Zhao, Rattanatamrong et al. 2008), natural language processing (Pestian, Matykiewicz et al. 2008; Jiao and Wild 2009; Xu, Wang et al. 2009; Yang, Spasic et al. 2009), search engines, medical diagnosis (Kononenko 2001; Kloppel, Stonnington et al. 2008), syntactic pattern recognition (Badr and Oommen 2006), bioinformatics (Bhaskar, Hoyle et al. 2006; Larranaga, Calvo et al. 2006; Hamelryck 2009; Valentini, Tagliaferri et al. 2009), object recognition in computer vision, game playing, software engineering and speech and handwriting recognition. The widespread use of machine learning is due to its high accuracy, capability of handling complex data, low cost in applying, and fast performance.

Machine Learning Classification (MLC) methods are increasingly used in early drug discovery stage for target and lead discovery. Some of these successful application includes

[2] http://www.nature.com/nrg/journal/v5/n4/glossary/nrg1315_glossary.html

[3] http://dli.grainger.uiuc.edu/glossary.htm

[4] amsglossary.allenpress.com/glossary/browse

[5] http://www.broadinstitute.org/annotation/conrad/glossary.html

classification of cytochrome P450 1A2 inhibitors and non-inhibitors (Vasanthanathan, Taboureau et al. 2009), protein expression profiling (Bradley, Kalampanayil et al. 2009), virtual screening of GPCRs (Shacham, Marantz et al. 2004; Evers, Hessler et al. 2005; Jacob, Hoffmann et al. 2008), prediction of interactions with ABC-transporters (Ecker, Stockner et al. 2008), early detection of drug-induced idiosyncratic liver toxicity (Cruz-Monteagudo, Cordeiro et al. 2008), prediction of toxicological properties and adverse drug reactions of pharmaceutical agents (Ma, Wang et al. 2008), target discovery (Chen, Fang et al. 2007; Ekins, Mestres et al. 2007; Han, Zheng et al. 2007; Chen and Chen 2008; Yousef, Showe et al. 2009), prediction of P-glycoprotein substrates (Xue, Yap et al. 2004; Huang, Ma et al. 2007), prediction of drug-likeness (Matter, Baringhaus et al. 2001; Walters and Murcko 2002; Zernov, Balakin et al. 2003). The motivation for the adoption of machine learning classification methods in drug discovery is due to its capability to model complex relationships in biological data.

Machine learning classification methods require known information to train the machine and make a prediction model; based on which the model will be able to predict the class of unknown data. The robustness of prediction model comes through the quality of data used to train the machine. The most common machine learning methods are Support Vector Machines (SVM), Artificial Neural Network (ANN), Probabilistic Neural Network (PNN), k nearest neighbor (k-NN), C4.5 decision tree (C4.5DT) which have shown good performance in various fields.

Machine learning classification methods have become increasingly important in the drug discovery and development process by predicting the class of chemicals or biomolecules. In target discoveries, machine learning classification methods have been applied for analyzing microarray data, non-invasive images, and mass spectral data to find biomarkers. In lead identification, machine learning classification methods are used to assess potential lead suspects, and for performing ligand based virtual screening to find possible hits. In addition machine learning classification methods are used to eliminate toxic compounds at very early stage of drug discovery. Even if a compound shows high selectivity and specificity to a disease-causing protein, there is significant probability of it failing in clinical phase. With the advent of combinatorial chemistry huge number of research compounds is being synthesized. These compounds should ideally be assessed for the activity or toxicity before it goes to expensive wet lab assay and clinical trials. Many studies has suggested the use of computational pre-assessment of compound e.g. the need of genetic toxicity prediction method (Van Gompel, Woestenborghs et al. 2005). This way, machine learning methods by its robust prediction capability can help as in selecting useful compounds and eliminating unwanted compounds.

1.5 Objectives of my PhD projects

The main objectives of this study are to contribute to efficient drug discovery processes by

(i) To contribute to efficient drug discovery processes by assessing the role of database development and machine learning methods

   a. To develop a database which would create a bridge between traditional medicine and modern medicine

   b. To develop a database which would trigger new pathway discovery process

(ii) To contribute to efficient drug discovery processes by providing some useful databases and machine learning classification studies.

   a. To develop a machine learning approach to solve an important toxicity related issues in early drug discovery process

   b. To develop a machine learning approach for lead identification for an important therapeutic target

With these objectives, databases were developed e.g. Indian Herbs and their Chemical Database (IHCD) and Kinetic Database of Biomolecular Interaction database was updated; and machine learning classification methods were applied for genotoxicity and p38 MAPKs inhibitor predictions. In addition, some secondary objectives are as follows:

1. To employ wide spectrum of biological or chemical data space for database development.
2. To evaluate the different data collection procedures in terms of speed, accuracy and loss of information in the process.
3. To observe the difference of web technologies employed in developing databases in terms of handling biological and chemical data complexity.
4. To observe the effect of diversity of dataset in machine learning classification methods.

5. To observe the effect of number of molecular descriptors used in machine learning methods.
6. To compare different machine learning methods performance
7. To evaluate different machine learning performance in virtual screening of large databases.

Chapter 2 Methods

2.1 Database development

2.1.1 Data collection

Data collection for making databases can be done by various ways e.g. manual data collection from literature, experiments or software output, part of the data taken from other databases, customized data collected programmatically from other databases either locally or over the web, and text mining by programs. Manual data collection from literature or manual curation of collected data is considered of the best quality. However, manual annotations is time consuming and expensive (Seringhaus and Gerstein 2007). A number of solutions for this problem are in practice. Data curation and annotation can be done in collaboration with other groups or providing online facility to edit or submission of data (Baumgartner, Cohen et al. 2007). In this work, most of the data is collected manually to ensure good quality. However, biological data is generally very large in number and it is not always possible to collect data manually. One such solution is the use of web services which is used extensively used in this work for collecting data from National Library of Medicine (NLM).

Web Services: It is a way to automatically access or facilitate data through the web. The term web service was originally created as a specific W3C standard (Stockinger, Attwood et al. 2008). Lately it has been used as a method of programmatic access over web technologies. In recent times, new web technologies such as Web 2.0, Service Oriented Architectures (SOA) and other web-related technologies have been introduced. Since many bioinformatics tools and biological databases are deployed as web accessible and depend on the internet, these new technologies seem to be of considerable importance for users as well as for developers of databases.

In other instances, data was also collected from some static web pages by writing html parser. Some commercial software are also available for this purpose e.g. Kapow Robo Suite, but in this work programs were written in Perl or Java to collect and parse html pages. Writing an html parser is a challenge because html file generally have unstructured data format. An efficient use of regular expression is necessary to retrieve structured data out of html.

2.1.2 Data Integration

Data integration is necessary where data from different sources need to be standardized before using it in making databases. Biological and chemical data comes from varied sources and its integration to a single database sometimes become big challenge. Improper integration can lead to loss of some part of data or even can introduce mistakes. The correct way of data integration for biological databases can generally be divided into two parts: (i) Syntactic integration in which data from different sources and of different file formats are standardized to have single file format. (ii) Semantic integration in which data from different databases are formalized to have a relational schema which holds relational tables and integrity constraints.

For syntactic integration, the standardized file format to which other data should be converted is generally XML. XML is short form of Extensible Markup Language. The structure of XML is such that it can hold data of various types of data such as simple plain table data, tree like data, relational tables and web pages. This easy conversion capability of XML makes it extremely useful format for exchange of data over web e.g. web pages file with aspx or jspx extension to html pages, for communication between different database software e.g. MySQL and Oracle, and for communicating between software which takes input XML file and produces result in XML format. In this work, the powerful feature of XML has been utilized for various purposes e.g. collection of Pubmed extracts for the Indian medicinal plants and their chemical ingredient name as keywords using NCBI entrez utilities, presenting pathway models in KDBI database in System Biology Markup Language (SBML) which is an extension of XML and customized to keep system biology data.

Semantic data integration on the other hand gives leverage to keep data in semi structured way. Sometime it is not possible to standardize a part of data to the convention of unified single file format. In these cases semantic data integration gives the flexibility to mix complex biological data. Well known databases like Uniprot and GO are good example of utilizing this kind of semantic integration.

In addition to the abovementioned ways of data integration, data can be integrated manually as well. It is very time consuming and tedious to do that but sometimes it becomes indispensible. Moreover, it has the advantage of including high quality data which otherwise would be missed. Manual data integration is generally achieved through scripting languages like Perl or Python. These scripting languages are handy to use yet very powerful. Perl has modules like DBI, DBD: MYSQL, DBD: ORACLE by which it can connect to databases such as MySQL and Oracle. One can easily write script to manipulate database tables by integrating plain unformatted text taken from literature or html we page. The power of programming languages like Perl and Java has led major public database provided by NCBI and EMBL to provide database access though user written program. For example entrez programming utilities by NCBI provide many example scripts to get customized data by constructing pipeline over its database. **Figure 4** shows the database model of NCBI databases and their interconnectivity, this snapshot taken shows linkage of pubmed database to other databases of NCBI. The detail about the NCBI databases can be found at http://www.ncbi.nlm.nih.gov/Database/ . A pipeline can be created by connecting several databases together for a string or IDs. This way of data integration can also be a part of data mining method which is explained in detail in next section.

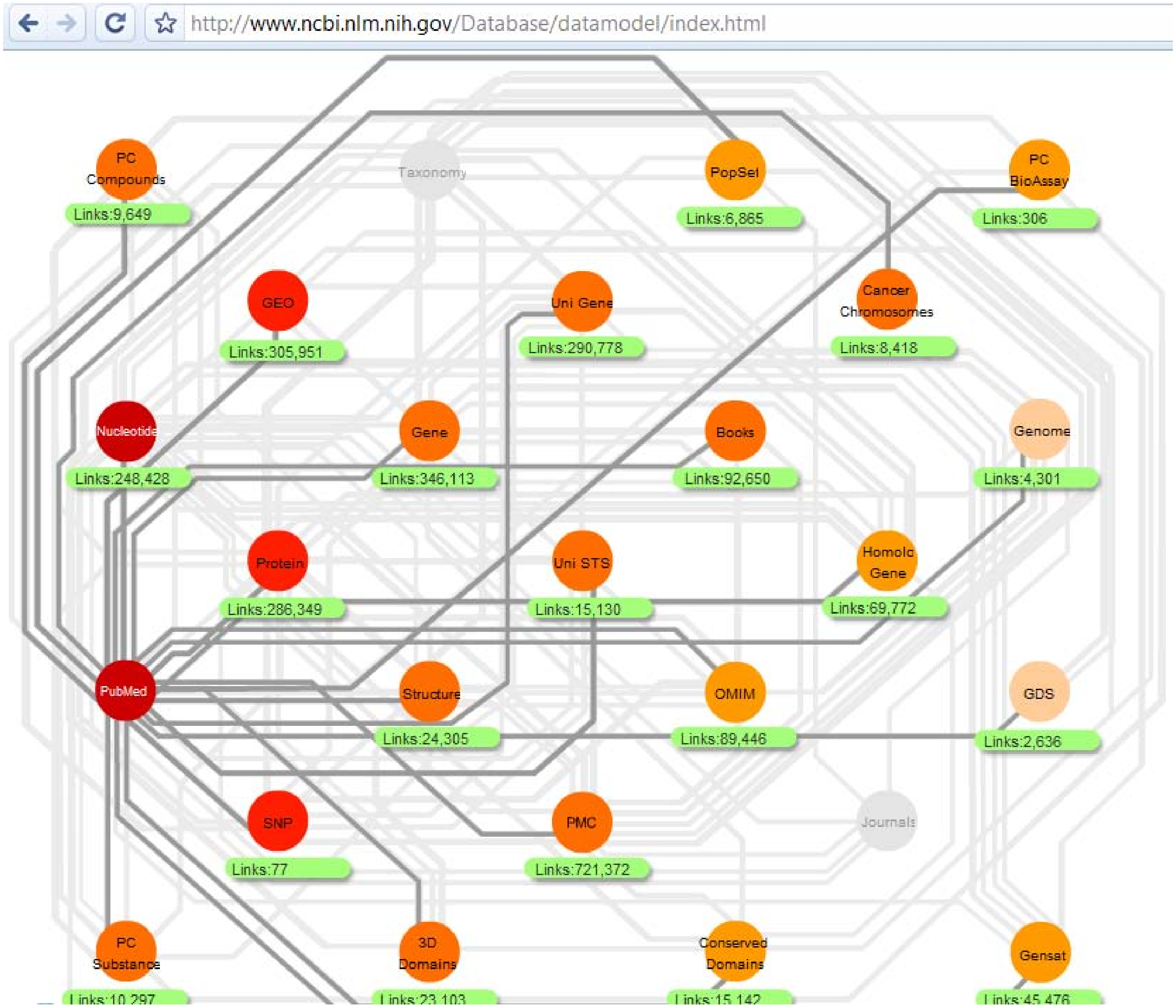


Figure 4: Database model of NCBI databases for entrez search. This screenshot is taken at web address displayed in the figure by placing mouse on the Pubmed when then displays cross-linking of Pubmed to other databases. The linked objects are different NCBI databases.

2.1.3 Data mining

Simple understanding of data mining can be perceived as the method to extract the data from any source which cannot be retrieved using straightforward manner. Data mining also include finding the relationship or pattern in data by association, clustering, classification, forecasting and so on. Some of the biological and chemical data mining technique includes sequence similarity search using BLAST, chemical structure similarity using fingerprint and text similarity search using regular expression.

Sequence similarity of Proteins

The BLAST program is used to do sequence-similarity searches against protein and nucleotide databases, which align the input sequence with database on the server with great speed. It is one of the most widely used programs for data mining in genomics and proteomics. The different versions and modifications in the BLAST program have made various variants of BLAST. Different server can store different databases for their BLAST program e.g. BLAST for nucleotide search human genome and transcript sequences, BLAST for protein searches GenBank, Swiss-Prot, PDB, PRF and PIR proteins. The result of BLAST is normally pair wise alignment, multiple sequence alignment formats, hit table and a report explaining hits by taxonomy. The BLAST hit is based on bit score and expectation value which is the measure of probability of alignment by chance. Short input sequence will generally have high expectation value because of its high probability of being present in any sequence. The NCBI BLAST programs are also available freely to download; it can be installed locally and can be used as standalone command line programs. One can download a sequence database on which the BLAST program will align an input sequence, or sequence database can be custom created for a set of protein and nucleotide of interest. One such application of local standalone BLAST has been introduced in this work is PIK-BLAST (a web server to find kinetic parameters from a pool of protein interacting pairs) which keeps custom sequence database of protein interacting pair.

Similarity of small molecules

Chemical similarity search using fingerprint represents chemical compound in a binary format of differing length depending on the program e.g. Pubchem structural fingerprint is of 1536 bits which is combination of 1024 bit fingerprint based on Molecular Design Limited

(MDL) and a 512 bit fingerprint representing 317 structural features as Smiles Arbitrary Target Specification (SMARTS) pattern[6]. In chemical similarity search, fingerprint or bit-string is generated for the input structure and is compared to fingerprints stored of other compounds in database using the Tanimoto coefficient which is a similarity index and can be defined as:

$$T = \frac{N_{xy}}{N_x + N_y - N_{xy}}$$

where N_x and N_y describe the number of bits, set to 1 in the fingerprint, of compound x and y, respectively. N_{xy} is the number of bit positions set to 1 in both fingerprints. When a structural feature is present or absent in the molecule, the fingerprint or bit-string of that molecule will have 1 (present) or 0 (absent) at the specific position (each structural feature will correspond to one position in bit-string).

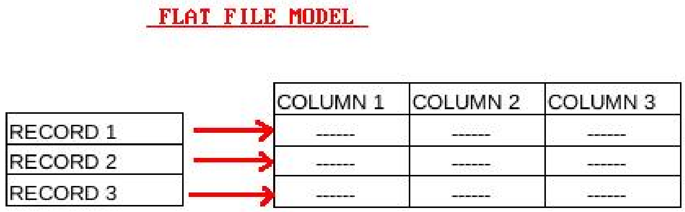
Text matching is necessary at many places for file or table editing. It is generally achieved by using regular expression which can be defined as sequence of characters that depict a pattern in text. Almost all programming languages has regular expression based search capability but some of them like Perl has become very popular because of its easiness, speed and flexibility to perform same thing in many ways. In regular expression, metacharacters (like ^, &, (,), * etc.) are utilized to construct efficient search which is very useful in complex, hard to edit, time consuming text searching (Stephens, Chen et al. 2005).

2.1.4 Data model

The data model in the database development is the incorporated concepts to describe relationship and constraints involved in the data. There are many different types of data model possible for making databases such as flat file model, network model, hierarchical model, and relational model. In this work, we have applied relational data model.

[6] http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html

Flat-file model: It is the simplest type of data model and uses just plain table to describe or to keep the data (**Figure 5**). One single row in the table represents one record. Each record can have set of features which are called attributes or fields which are kept in separate column. If the record does not have a particular feature then this field will be null. This flat data model is very convenient when the data is not very complex. Moreover, depending on the number of features involved there can be huge increase in number of records because records may be different by just one different feature. This way table usually becomes very big and speed of database decreases and subsequently becomes critical issue. Biological data is generally very complex and in this work we have not employed flat file model.



	COLUMN 1	COLUMN 2	COLUMN 3
RECORD 1	------	------	------
RECORD 2	------	------	------
RECORD 3	------	------	------

Figure 5: Flat file model

Hierarchical model: Hierarchical data model is very much like tree structure (**Figure 6**). This data model incorporates data very well and keeps the data in ‘one to many’ relationship. This data model is very much capable in mapping real world data complexities. Because of this nesting capability it has now become the standard of XML file. In this hierarchical model, one always needs to know the full path for accessing a record which put some limitation this type of model.

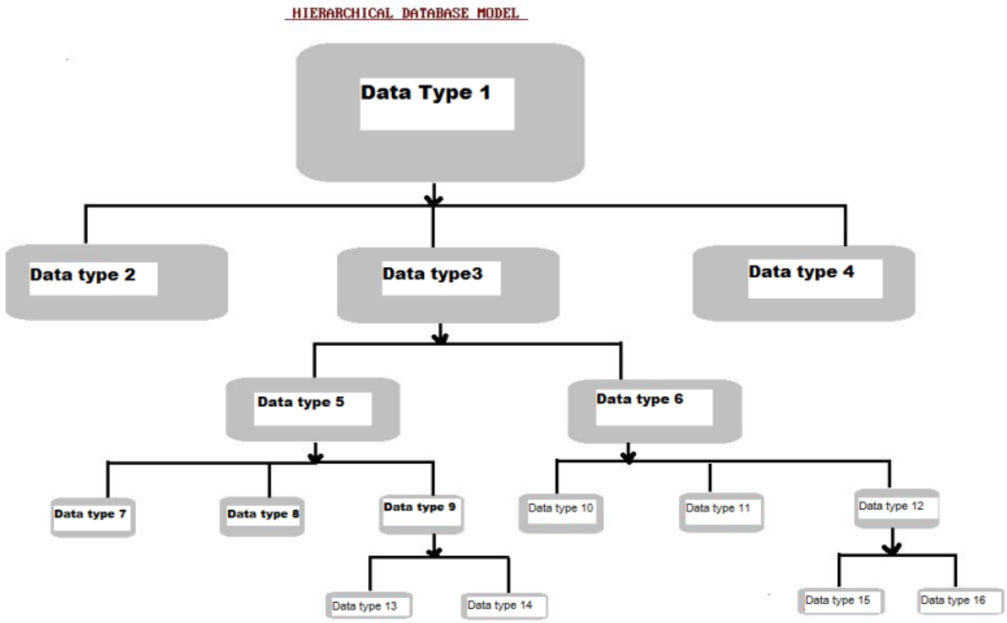


Figure 6: Hierarchical data model

Network model: The network model looks like hierarchical model but it differs significantly in that branches of the tree can be linked to multiple nodes in upward link. **Figure 7** shows the network data model in which 'Data type 9' is linked to two upper level 'Data type 5' and 'Data type 7'. The network data model can represent redundant data more efficiently than hierarchical data model. The network model operates in navigational style i.e. a program upholds a current position on one record and moves to another record according to the relationships present.

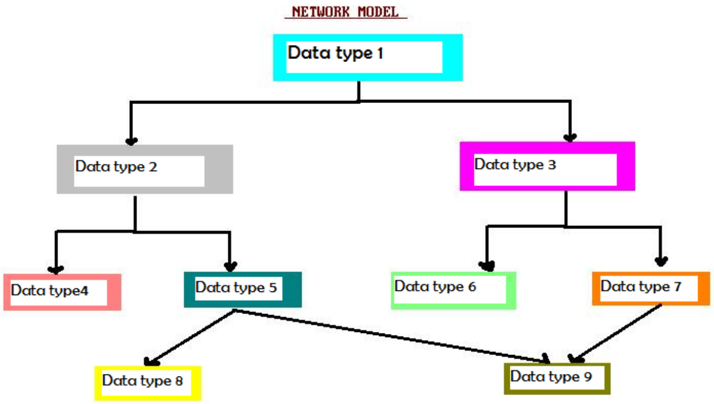


Figure 7: Network data model

Relational model: The relational data model is a powerful approximation of mathematical model to make database tables well connected by some rules in order to be unaffected by kind of web application employed or built upon it. The databases used by making use of relational data model are often called as relational database. There are three important terms in relational data models i.e. relations, attributes, and domains. A relation is a table of rows containing records and columns whose name are called attributes. The attributes can take certain range of values which are called as domains. A relational data model generally consist many tables with some relationship to each other. There is some basic rules to construct relational data model e.g. each table should not contain duplicate records, there should be primary keys in each table which must be unique, primary key of table may be present in another table and which will be the basis of linkage (**Figure 8**). The keys of each table play a crucial role in relational data model by creating connections as well as fast retrieval of data upon request. The primary keys are automatically indexed which is a feature of providing fast access to record of table by jumping directly to index number rather than crawling at each record and searching. The other attributes can additionally be indexed as well but is only necessary if the search is being done on that attribute. Overall, the relational data model is very robust data structure and that is why it has been applied in this work to construct databases.

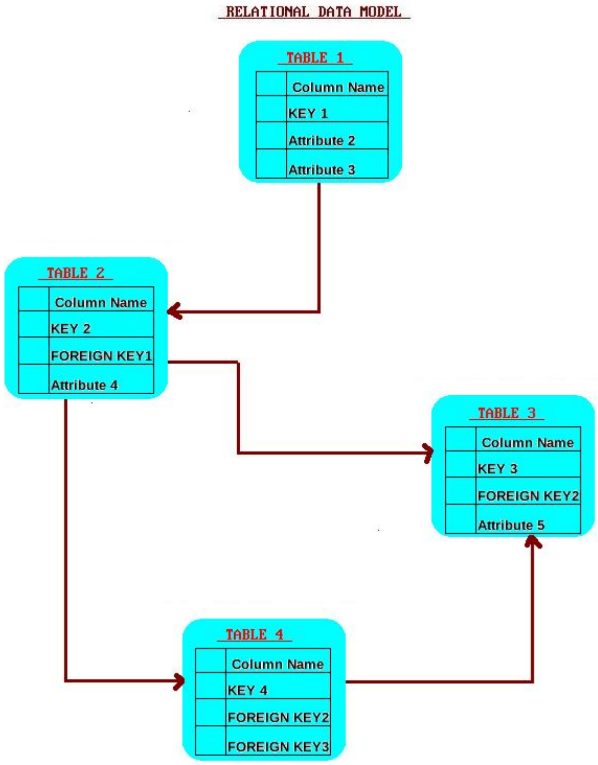


Figure 8: Relational data model

2.1.4 Database interface

Database interface or web interface (because this work represents web accessible databases) is what user sees and interacts with the database. The database web interface should be very convenient to understand and user should have certain level of flexibility of getting customized data. User interaction capability can put web pages in two categories: static pages and dynamic pages. Static pages are the type of web page which will be same to all users i.e. user cannot get custom or advanced feature. Dynamic pages are the type of web pages which presents different web page content to different user according to the form submitted by them which may differ in keywords or selection of features. Here in this work, web accessible databases are mostly presented as dynamic web pages. These dynamic web pages have been built upon using both server side as well as client side technologies. Server side dynamic web page creation can be achieved by various technologies like Active Server Pages (ASP), Java Server Pages (JSP), PHP and CGI (Common Gateway Interfaces) while client side dynamic web page creation is generally achieved through JavaScript. In this work ASP and JSP technologies are used for server side dynamic web page creation and JavaScript is used for client side dynamic web page creation. Server side dynamic web page creation over database involves submission of user supplied query to web server which further interacts with database software such as MySQL and Oracle. In contrast, client side dynamic web page creation does not include interaction with web server. The client side technology uses user internet browsers e.g. Internet Explorer, Mozilla Firefox and Google Chrome to run its code and display the data. The client side dynamic web page is thus very simple and generally used to present data in a beautiful manner and provides helps about the content such as change in color or short string giving help when mouse is place on some part of the content. In contrast, server side dynamic web page creation requires efficient programming like java code for JSP and vbscript for ASP technology. Server side dynamic web page creation also require good tuning or configuration of web servers which handles user request to provide correct data.

2.2 Machine learning classification methods

Machine learning classification methods employ computational and statistical methods to construct mathematical models from training samples which is used to classify independent test sample. The training samples are represented by vectors which can be binary, categorical or continuous. Machine learning can be of two types: Supervised and Unsupervised. Supervised machine learning, as the name indicates, generally needs feeding which is availability of already labeled or classified data for training. Example of supervised machine learning includes Support Vector Machine, Artificial Neural Network, Decision tree learning, Inductive logic programming, Boosting, Gaussian process regression etc. Unsupervised machine learning, as the name indicates, gets unlabeled training data and the learning task involve to find the organization of data. Examples of unsupervised machine learning include Clustering, Adaptive Resonance Theory, and Self Organized Map (SOM). Some of machine learning methods employed in this work are Support Vector Machine (SVM), Probabilistic Neural Network (PNN), k nearest neighbor (KNN), Decision trees and Hierarchical clustering. These are explained below in subsequent sub sections.

2.2.1 Support vector machine

Support Vector Machine is a very specific class of supervised learning algorithms which separates labeled input data by a hyperplane. The input data can be of any number of dimensions, SVM by its robust algorithm can still find a hyper plane by the use of different kernel functions. On either side of this separating hyperplane, a hyperplane is constructed to push the corresponding labeled data so that the maximum margin (distance) is achieved between either sides of hyperplane (**Figure 9** and **Figure 10**). The labeled vector data points on these two hyperplanes are called as support vectors.(Cristianini 2000).

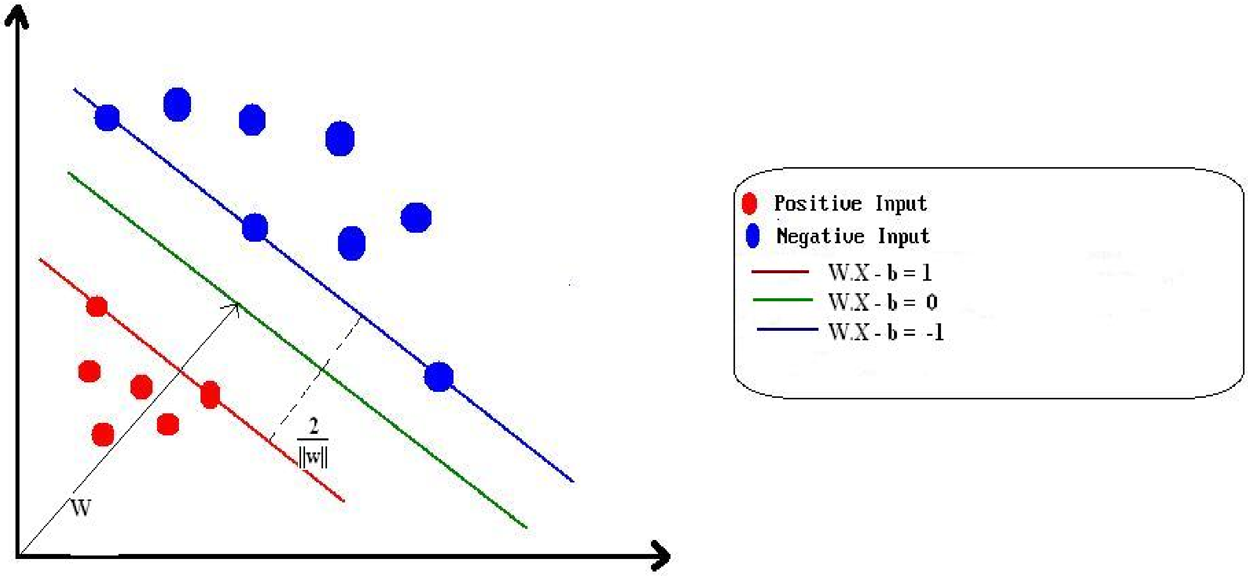


Figure 9: SVM hyperplanes separating positive and negative. The green line shows the separating hyperplane. On either side of this hyperplane, two hyperplanes are shown with red and blue line.

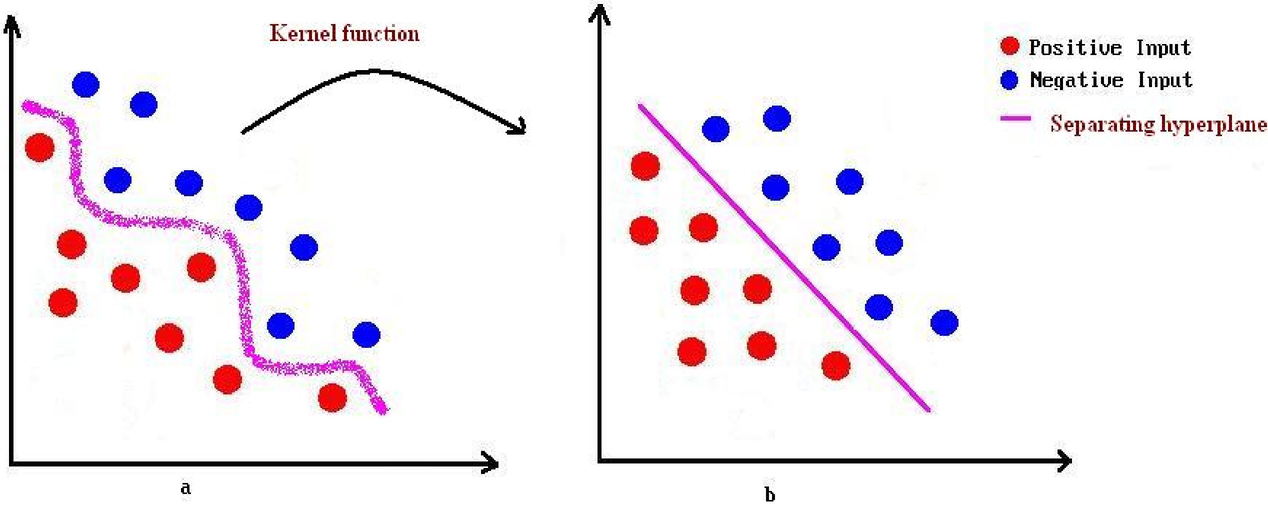


Figure 10 : Use of kernel functions in SVM in high dimensional space to convert non-linear hyperplane to linear hyperplane

For some training data, data points D can be presented in the form

$$Data\ points, D = \{(\mathbf{X}_{\boldsymbol{i}}\ , a_i\ )|\mathbf{X}_{\boldsymbol{i}}\ \in\ \mathbf{R}^{\mathrm{m}}\ ,\ a_i\ \in \{\ \mathbf{-1}, \mathbf{1}\}\}_{\boldsymbol{i=1}}^{\boldsymbol{n}} \qquad \text{Equation 1}$$

where for each point $\mathbf{X}_i$ is a multidimensional vector, the value of a_i is either 1 or−1, indicative of the class to which it belongs. In order to construct the maximum-margin hyperplane which separates the points having a_i = -1 from those having a_i = 1 (Figure 9), an equation of separating hyperplane can be written as the set of points $\mathbf{X}$ satisfying.

$$\mathbf{W}.\mathbf{X} - \mathrm{b} = 0 \qquad \text{Equation 2}$$

where $\mathbf{W}$ is a normal vector which is perpendicular to the hyperplane. The parameter $\frac{b}{||\mathbf{W}||}$ determines the offset of the hyperplane from the origin in the direction of normal vector $\mathbf{W}$. The value of $\mathbf{W}$ and b should be chosen to maximize the margin, or distance as much as possible between the parallel hyperplanes on either side of this separating hyperplane and separating the data simultaneously. These two parallel hyperplanes on either side of separating hyperplane can be written as

$$\mathbf{W}.\mathbf{X} - \mathrm{b} = 1 \qquad \text{Equation 3}$$

$$\&\quad \mathbf{W}.\mathbf{X} - \mathrm{b} = -1 \qquad \text{Equation 4}$$

If the training data can be separated linearly then the margin of the two hyperplanes can be selected in such a way that there are no data points between them and maximum distance is achieved between them. The distance between these two hyperplanes can be calculated as $\frac{2}{||\mathbf{W}||}$, so $||\mathbf{w}||$ should be minimized. To prevent data points falling into the margin, we can add the following constraint:

$$\mathbf{W}.\mathbf{X}_{\boldsymbol{i}} - \mathrm{b}\ \geq 1 \qquad \text{Equation 5}$$

$$\boldsymbol{\&}\quad \mathbf{W}.\mathbf{X}_{\boldsymbol{i}} - \mathrm{b}\ \leq\ -1 \qquad \text{Equation 6}$$

this can be rewritten as:

$a_i(\mathbf{W}.\mathbf{X}_i - \text{b}) \geq 1$, for all $1 \leq i \leq n$ Equation 7

While minimizing ||**w**|| (in w, b), Equation 7 becomes the optimization problem. This optimization problem is hard to solve because it depends on ||**w**||, the norm of **w**, which involves a square root. However, it is can be solved by little change in the equation by replacing ||**w**|| with $\frac{1}{2}||w||^2$ without changing the solution which then becomes quadratic programming optimization problem. If the equation is written in its unconstrained dual form then it can be seen that the classification depends just on the support vectors. This unconstrained dual form can be seen to have the following optimization problem:

Maximize (in α*ᵢ*)

$\sum_{i=1}^{n} \propto_i - \frac{1}{2\sum_{i,j}\propto_i\propto_j a_i a_j \mathbf{X}_i^T \mathbf{X}_j}$, for all $1 \leq i \leq n, \propto_i \geq 0, and \sum_{i=1}^{n} \propto_i a_i = 0$ Equation 8

The α terms represent a dual form for the weight vector:

$\mathbf{w} = \sum_{\mathrm{i}} \propto_{\mathrm{i}} \mathrm{a}_{\mathrm{i}} \mathbf{x}_{\mathrm{i}}$ Equation 9

The positive or negative value of Equation 8 specify that the vector $\mathbf{X}_i$ go to positive or negative class (either side of separating hyperplane) respectively.

2.2.2 Decision Trees

Decision tree is a type of supervised machine learning method which is very good in solving the problems in which instances are characterized by attribute-value pairs and are explained by a unchanging set of attributes (e.g., chemical descriptors like solubility) and their values (e.g., water soluble). Decision trees can be very easily applied in situations when every attribute has small number of disjoint values (e.g., water soluble, oil soluble). Nevertheless, it can handle other type of attributes like log p, chi value efficiently as well. Decision tree models are also good in tackling missing values, and little error in either independent or dependent variable or in both. Moreover, the clarity and visualization of decision making process is easily comprehendible which makes it more adoptable in comparison to algorithm like artificial neural network (ANN) which is very complex to understand. (Frank 2005)

Construction of decision tree Model

As a first step, the whole data is split into two or more disjoint sub-samples. The whole data set is termed as root node and the sub-samples are known as a node. This division of whole data is done on the basis of one of the independent variables which are called the splitting attribute. Based on different values of this splitting attribute separate branches are made. Then every data point or instance in root node is placed into one of the directly attached node based on the value of splitting attribute. The selection of splitting attribute is made to achieve best homogeneous sub-samples after partitioning of root node.

In second step, the partitioning done in first step is repeated for every node by taking into consideration only the instances present in that particular node alone. This process continues till there is no violation of any stopping-rule imposed by the algorithm. When there is such violation, the further partitioning on that particular node is stopped and that node is termed as leaf node. This whole process of decision tree is finished when only leaf nodes are present i.e. no node is left for further partitioning.(Frank 2005). An example decision tree is shown in **Figure 11** which depicts the decision making process for a compound’s positive or negative class.

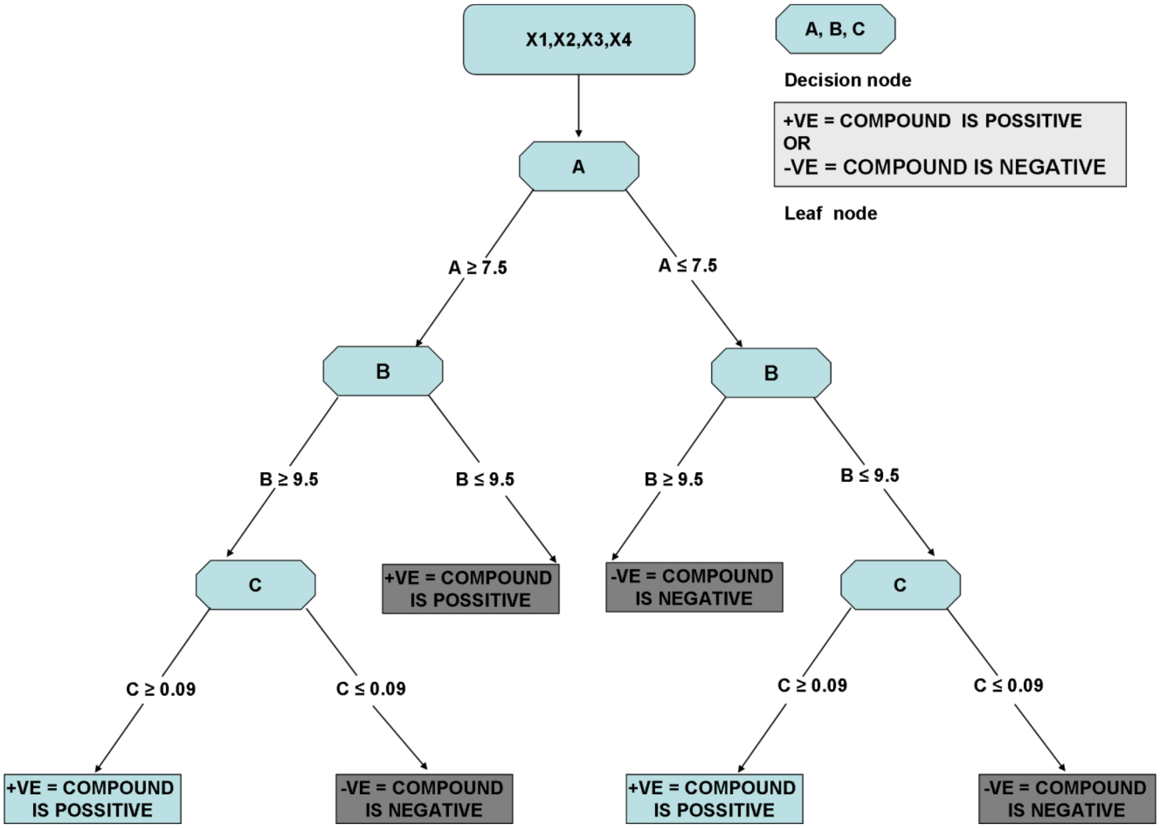

X1,X2,X3,X4
A, B, C
Decision node
+VE = COMPOUND IS POSSITIVE
OR
-VE = COMPOUND IS NEGATIVE
Leaf node
A
A ≥ 7.5
A ≤ 7.5
B
B
B ≥ 9.5
B ≤ 9.5
B ≥ 9.5
B ≤ 9.5
C
+VE = COMPOUND IS POSSITIVE
-VE = COMPOUND IS NEGATIVE
C
C ≥ 0.09
C ≤ 0.09
C ≥ 0.09
C ≤ 0.09
+VE = COMPOUND IS POSSITIVE
-VE = COMPOUND IS NEGATIVE
+VE = COMPOUND IS POSSITIVE
-VE = COMPOUND IS NEGATIVE

Figure 11: Decision tree

The mathematical algorithm of the decision tree can be comprehended briefly in terms of entropy or any other tree splitting parameter choosing function. During tree building in decision tree, data is portioned repeatedly until the dataset present in each partition belong to single class or the partition node has very small dataset. The decision for splitting of a node can be based on entropy as splitting parameter choosing function or splitting index.

$$Entropy(T) = -\sum f_j x \log_2(f_j) \qquad \text{Equation 10}$$

where f_j is the relative frequency for class j. The best split is one which has maximum information gain:

$$I(S) - \sum_{i=1}^{n} \frac{S_i}{S} I(S_i) \qquad \text{Equation 11}$$

where the split partitions S can have n different class Si (i = 1 to n) with I() as splitting index.

After that partitioning, tree pruning is done to remove statistical noise which may only be particular to the training set. Tree pruning is helpful in finding a sub tree which has least estimated error rate.

There are various decision tree algorithms available to do the abovementioned steps but they differ significantly in criteria of selecting splitting attribute or splitting index, imposing a stopping rule and how nodes are depicted of a particular class. Some of the popular algorithm includes C4.5 developed by Quinlan (Quinlan 1993) , Random Forest, Naive Bayes trees, and logistic model trees. The decision trees applied in this work are from Waikato Environment for Knowledge Analysis(WEKA) (Frank 2005) implementation of these decision tree algorithms in the form of classes like J48 (C4.5)(Quinlan 1993), Random Forest(Breiman 2001), ID3(Quinlan 1986), NBTree(Kohavi 1996), Random Tree, LMT, RepTree, ADTree(Freund 1999), BFTree(Tibshirani 2000) and M5P(Quinlan 1992).

2.2.3 k-nearest neighbor (k-NN)

KNN is a supervised machine learning method which classifies data by grouping close neighbors together. Based on the label of input training data points, the new test data is classified by the count of labeled of k nearest neighbored training data (**Figure 12**). Ideally, the value of k should be decided on the number of labeled training data and is optimized during training. The algorithm implementing kNN can vary in number of ways e.g. on the basis distance calculation methods like Euclidian or Manhattan. Different K-nearest neighbor algorithm have been used for the classification of biological and chemical data (Chin, Wang et al. 2006; Chou and Shen 2006; Karakoc, Cherkasov et al. 2006). In this work, k-NN is used by WEKA class IBk(Kibler 1991).

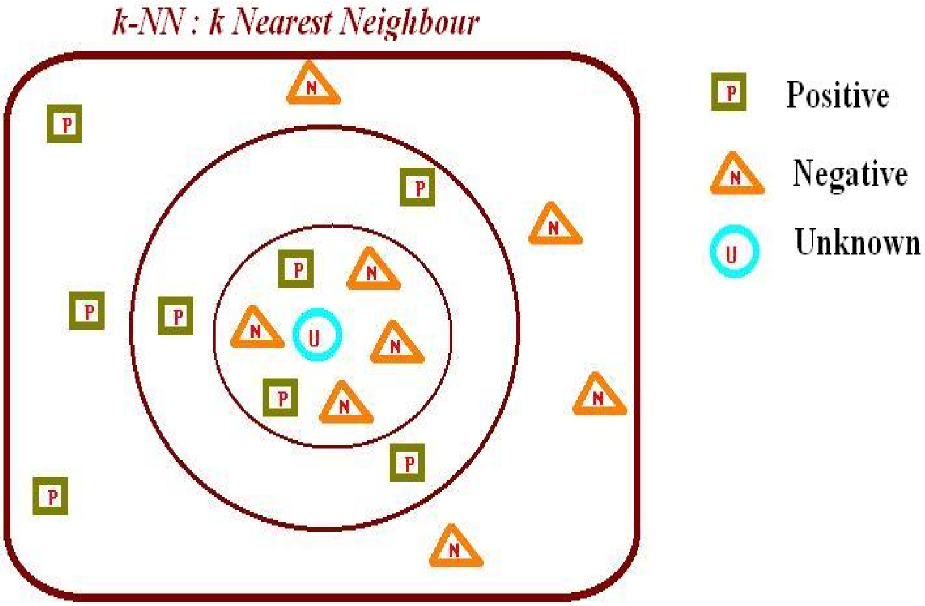


Figure 12: k-Nearest Neighbor

2.2.4 Feed forward Neural Networks

Neural networks are a type of supervised machine learning method and Feed forward Neural Network is one of its subtypes. PNN has been applied in this work by WEKA class Multi Layer Perceptron (Frank 2005). A multilayer perceptron maps sets of input data onto a set of appropriate output. Multilayer perceptron is a modified standard linear perceptron in that it uses three or more layers of neurons (nodes) with nonlinear activation functions, and is more powerful than the perceptron in that it can distinguish data that is not linearly separable.

Multilayer perceptron design has three or more layers which are input, output, and one or more hidden layers (**Figure 13**). Nodes of one layer connects o every node in the following layer with certain weight. Learning occurs in the perceptron by changing connection weights after each piece of data is processed, based on the amount of error in the output compared to the expected result. Generalization of the least mean squares algorithm in the linear perceptron results in back propagation.

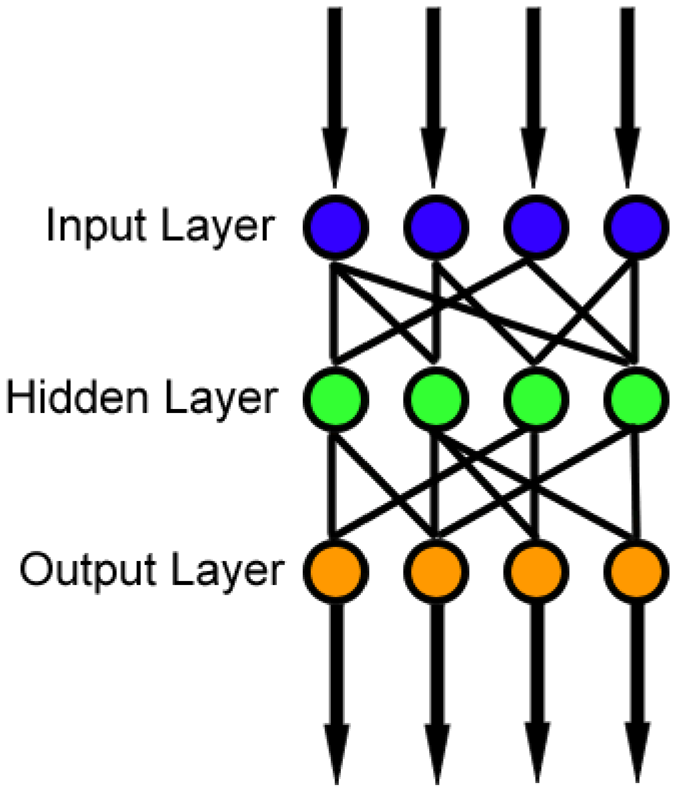


Figure 13: Feed forward neural network

2.2.5 Hierarchical Clustering

Hierarchical clustering is a kind of clustering method which builds a hierarchy of clusters for a given dataset. This hierarchical structure can also be seen as tree kind of structure called dendogram. In this tree kind of structure root node contains all the data points which break down via different branches. Hierarchical clustering is generally of two types: agglomerative and divisive (**Figure 14**). In agglomerative type, clustering starts from leaf which keep on adding together till it reaches to root. The divisive type is reverse of agglomerative type that is starting from root and going towards leaf. In this work, hierarchical clustering is applied via WEKA class COBWEB (Fisher 1990) which does clustering in divisive way.

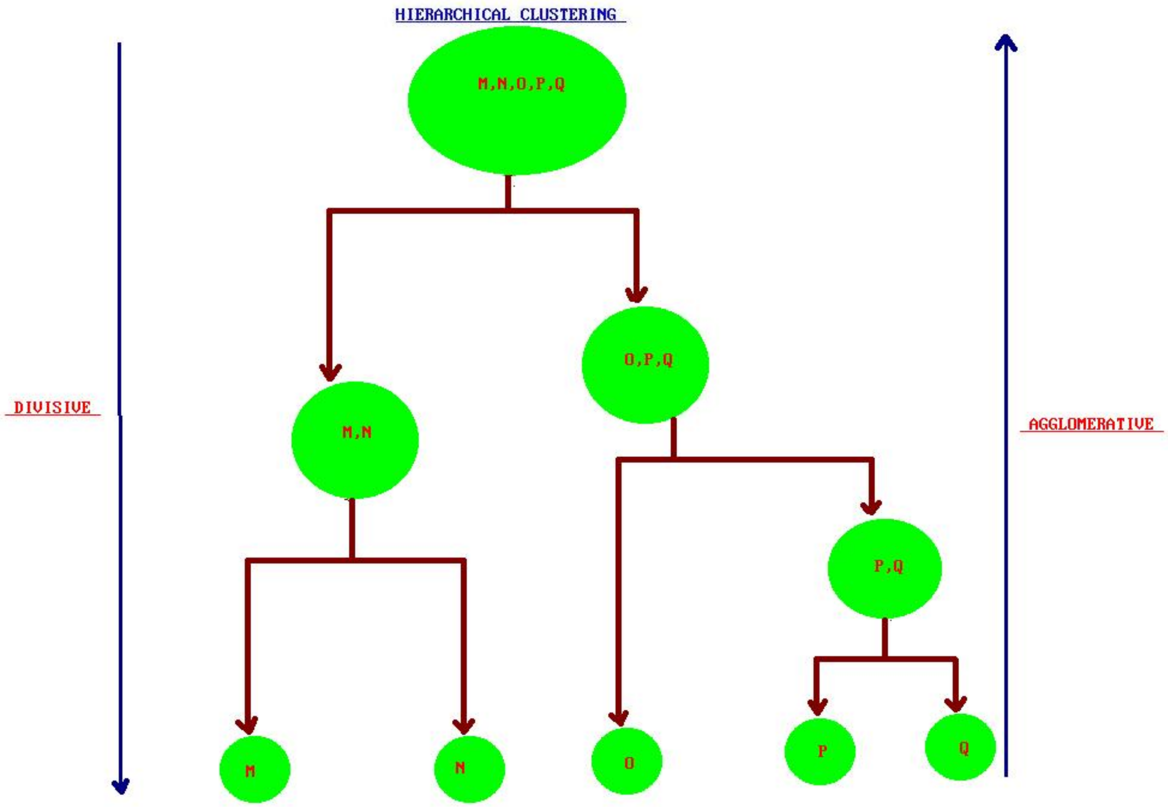


Figure 14: Hierarchical Clustering: Agglomerative and Divisive

2.2.6 Data collection for machine learning

Dataset used for machine learning classification is of utmost importance. Various factors like quality, size and relevance of the dataset can affect machine learning process greatly. Dataset quality is generally assessed at the time of data collection. Data collected from less reliable source will give rise to faulty models which will lose its predictive power when assessed for true independent set. Here, in this work care has been taken to include data from very reliable sources like good journals and manually annotated databases. For example while collecting data for p38 inhibitors from journals like Bioorganic and medicinal chemistry and Journal of Medicinal chemistry, chemical compounds were drawn manually. Data collected manually are generally considered of very high quality, but in compounds data collections from papers may need additional care to ensure high quality. Chemical compounds in these synthesis related journal usually have series of compounds with the same ring and only little variation in side chain. Manual drawing of compounds in these scenario are prone to errors. By keeping this fact in mind, sketched compound were checked thrice to ensure correct compound structure. Such practice is very necessary to propagate good quality to the built model. Quality of dataset can also be ensured by correct labeling. In certain cases, data points falls into grey area i.e. neither positive nor negative. These hard to label data should be labeled carefully with some cut off like IC_{50} (concentration at which 50 % of the enzyme is inhibited) value or can be excluded permanently from dataset.

2.2.7 Data representation: Molecular descriptors

Molecular descriptors are frequently used to describe various physicochemical or structural properties of molecules for many computational studies small molecules. There are many types of chemical descriptors such as composition based descriptors, electronic descriptors, and geometrical descriptors. Broadly these chemical descriptors can be classified into three categories: one dimensional, two dimensional or three dimensional. Chemical composition like number of carbon atom, number of oxygen atom etc are one dimensional chemical descriptors; geometric descriptors topological descriptors like molecular connectivity chi indices, molecular shape Kappa indices, electrotopological state indices, and atom type electrotopological state indices are two dimensional chemical descriptors; and molecular volume, dipole moment, polar surface are three dimensional descriptors. A number of programs e.g. OpenBabel, MODEL (Li, Han et al. 2007), Chemistry Development Kit(CDK) (Steinbeck, Han et al. 2003; Steinbeck, Hoppe et al. 2006) etc are available to calculate chemical descriptors. In this work, varying number of chemical descriptors are used which was calculated from MODEL.

Molecular descriptors have been extensively used in deriving structure-activity relationships (Fang, Tong et al. 2001; Tong, Xie et al. 2004), quantitative structure activity relationships (Hu and Aizawa 2003; Jacobs 2004), and machine learning prediction models for pharmaceutical agents (Doniger, Hofmann et al. 2002; Byvatov, Fechner et al. 2003; He, Jurs et al. 2003; Zernov, Balakin et al. 2003; Snyder, Pearl et al. 2004; Xue, Li et al. 2004; Yap, Cai et al. 2004; Yap and Chen 2005). A total of 522 chemical descriptors was derived by using program developed by BIDD group (Xue, Yap et al. 2004) , of which either entire 522 or selected 100 descriptors were used in this work (See **Table 14** and **Table A1** in Appendix for the detail of descriptors used).

2.2.8 Data processing:

2.2.8.1 Redundancy and similarity within datasets

Compounds are checked for redundancy by comparing exact match of chemical descriptors. In this work, scripts are written to find exact match of chemical descriptors to remove redundancy from dataset.

Similarity of the compounds can be checked by Tanimoto-based similarity searching method (Willett, Barnard et al. 1998) :

$$Tanimoto\ Similarity(i,j) = (\ \sum_{d=1}^{n} x_{di}x_{dj})/(\ \sum_{d=1}^{n}(x_{di})^2 + \sum_{d=1}^{n}\left(x_{dj}\right)^2 - \sum_{d=1}^{n} x_{di}x_{dj})$$

where n is the number of molecular descriptors and x is representing molecular descriptor. The compound *i* is evaluated as similar to the compound *j* if the tanimoto similarity calculated is greater than the decided cut-off value. In this work, the tanimoto similarity search was conducted for MDDR compounds with genotoxic compounds. The different cut-off values 0.7, 0.8 and 0.9 were tried for searching similarity of compounds (Bostrom, Hogner et al. 2006; Huang, Shoichet et al. 2006).

2.2.8.2 Scaling

Chemical descriptors are normally scaled before they can be employed for machine learning. Scaling of chemical descriptors ensures that each of descriptor have unbiased contribution in creating the prediction models(Dutta, Guha et al. 2006). Scaling can be done by number of ways e.g auto-scaling, range scaling, Pareto scaling, and feature weighting (van den Berg, Hoefsloot et al. 2006; Parsons, Ludwig et al. 2007). In this work, range scaling is used to scale the chemical descriptor data. Range scaling is done by dividing the difference between descriptor value and the minimum value of that descriptor with the range of that descriptor:

$$d_{ij}^{scaled} = \frac{d_{ij} - d_{j,min}}{d_{j,max} - d_{j,min}}$$

Where d_{ij}^{scaled}, d_{ij} , $d_{j,\max}$ and $d_{j,\min}$ are the scale descriptor value of compound i, absolute descriptor value of compound i , maximum and minimum values of descriptor j respectively. The scaled descriptor value falls in the range of 0 and 1.

2.2.9 Model validation

One of the usual ways to assess or to find the optimum parameters for a model built by machine learning is to see its performance either by independent validation set or cross validation. In this work, models were validated by using both independent validation set (manually segregated a part of data based on some criteria like recently published), and by cross validation. There are various types of cross validation commonly used in many statistical studies such as repeated random sub-sampling cross validation, k-fold cross validation, and leave one out cross validation. In this work, we have applied k-fold cross validation with k value is equal to 5, thus making it 5-fold cross-validation (**Figure 15**). For 5-fold cross-validation, these compounds are randomly divided into five subsets of equal size. Each of these folds contains equal number of positive and negative data, thereby rendering it a stratified cross-validation. Four subsets are selected as the training set and the fifth as the validation set. This process is repeated five times such that every subset is selected as a validation set once. The SVM models were saved in each case and prediction were done for validation data.

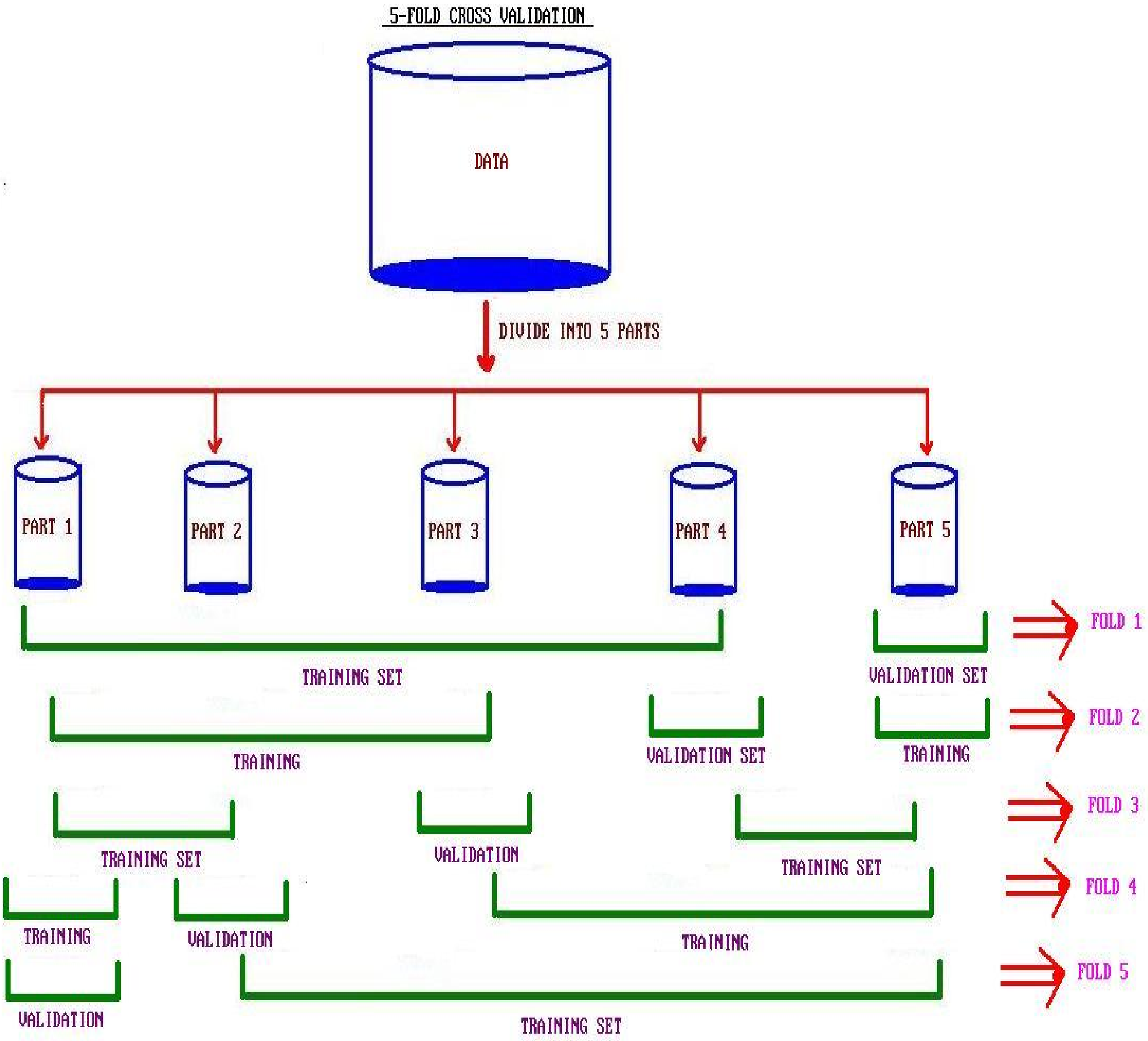


Figure 15: 5-Fold cross validation

2.2.10 Performance evaluation methods

In this work, performance of machine learning models is evaluated by using following formulas:

SE (positive accuracy) = $TP/ (TP + FN)$

SP (negative accuracy) = $TN/ (TN + FP)$

Q (overall accuracy) = $(TP + TN)/ (TP+TN+FP+FN)$

$MCC = (TP \times TN - FP \times FN)/ (TP + FN) (TP + FP) (TN + FP) (TN + FN)$

where TP (true positive), TN (true negative), FP (false positive), and FN (false negative), and Matthews correlation coefficient (MCC) correspond to correctly predicted positive, correctly predicted negative, negative samples incorrectly predicted as positive, and positive samples incorrectly predicted as negative, and randomness of prediction respectively. MCC value has range of -1 to 1. Positive values of MCC signify the agreement between measurement and prediction, negative values signify the disagreement between measurement and prediction, and zero value signify the prediction is same as guess.

2.2.11 Overfitting problems and strategies for detecting and avoiding them

Overfitting (**Figure 16)** is major concern in machine learning classification method. In the course of model building using cross validation, many times machine over fits the model with very high accuracy in cross validation results but show poor accuracy while tested with independent dataset. That is why; sometime it is good practice to choose the model which performs better with independent data. The reason for overfitting is usually linked with the model having high number of degrees of freedom compared to the number of records. Other possible reason for overfitting could be the conformability of the model in accordance to data shape, and the extent of model error matched up to the expected level of data error.

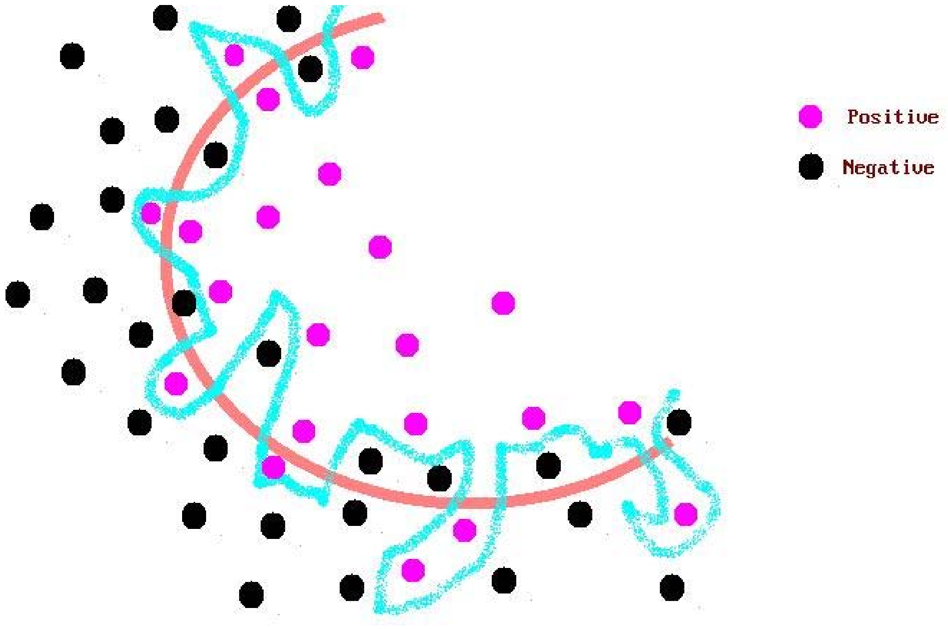


Figure 16: Overfitting of machine learning classification methods. Red line: Normal separating line, Blue Line: Overfitted separating line

2.2.12 Machine learning classification-based virtual Screening platform

Virtual screening is basically of two types: Structure based and Ligand based. In structure based, small molecule database is docked on a protein structure. Based on the scoring functions of docked complex compounds are selected as hits. In ligand based virtual screening there is no need of protein structure. Based on the existing experimental hits, a model or an equation is generated and this is used for screening small molecule database. So, the ligand based virtual screening is kind of similarity or pattern searching. The virtual screening by machine learning methods falls into the category of ligand-based virtual screening. The models are developed by using SVM for the best parameter range found by 5-fold cross-validation which is used for the Virtual Screening of MDDR and Pubchem database. The models developed for virtual screening are different from models that have been developed using the 5-fold cross-validation. The models developed for virtual screening use all the data accumulated for training purposes, while 5-fold cross-validation study keeps four folds for training and one for validation. The common compounds found in MDDR and genotoxicity positive data are both removed and used in the development of the SVM model. This has been done to make it true independent database used in virtual screening.

The performance of virtual screening is sometime presented with additional parameters than simple number of hits e.g. hit rate, yield and enrichment factor. These can be defined as follows:

Hit rate = Ratio of predicted known hits to all the predicted hits.

Yield = Percentage of known hits predicted

Enrichment factor = Magnitude of hit rate improvement over random selection

Chapter 3 Database development of medicinal chemicals: Indian medicinal herbs and their chemical ingredients

3.1. Introduction of Indian medicinal herbs

Traditional medicines have been extensively used in various countries and are gaining popularity in industrial countries. The global market for traditional medicine has reached US $60 billion with 5-10% annual growth rate (Kartal 2007). One of the popular traditional medicines is Ayurvedic medicine which is widely used in India (Mishra, Singh et al. 2001). Like other traditional medicines, Ayurvedic medicines mostly explore single medicinal plant or mixture of medicinal plant extracts for achieving the claimed therapeutic actions. However, rigorous investigations are needed for investigating the therapeutic effectiveness of Ayurvedic medicines and the mechanism of actions, which requires the knowledge about the bioactive ingredients and their mechanism of actions and have thus attracted strong interests in the relevant research (Smit, Woerdenbag et al. 1995; Arora, Kaur et al. 2003) particularly in studying the collective effects of multiple herbs and ingredients based on the currently limited knowledge about the ingredients and their targeted biomolecules and biological networks (Ichikawa, Nakamura et al. 2007). Therefore, easily accessible resources that provide comprehensive and integrated information about the herbs and ingredients of Ayurvedic medicines and their targeted biomolecules and biological networks are highly useful for facilitating the relevant research that have been hindered by the insufficiency of the relevant information (Koehn and Carter 2005).

Most of the available Ayurvedic medicine databases tend to emphasize more on formulations and less on ingredients and their mechanism of actions (Jayaraman 2006). The latter is important for investigating the claims of Ayurvedic remedies and discovering new drug leads (McGuffin 2008). Moreover, there is a lack of resources for facilitating the search of the biomolecules and biological pathways targeted by the herbs and ingredients of Ayurvedic remedies. As part of the efforts to fill-in these gaps to complement the available databases, we developed a new database IHCD (Indian Herbs and Chemical Database) freely accessible at (http://bidd.cz3.nus.edu.sg/ihcd) for facilitating the access of comprehensive and integrated information about herbs, ingredients, therapeutic actions, chemical descriptors and the possible biomolecules and biological targets of the relevant herbs and ingredients.

3.2 Data collection and database construction methods

The relevant herbs and ingredients were collected from reputed books such as *Indian Herbal Pharmacopoeia (1999)* , *Indian Medicinal Plants: An illustrated dictionary (Khare 2007)* and journals such as *Journal of Ethnopharmacology*, *Journal of Alternative and Complementary Medicine* and through comprehensive search of Medline. The information of a total of 2326 herbs from 430 therapeutic classes and 3978 ingredients were collected. Further information about each ingredient was provided via cross-link to chemical, pathway, and molecular binding databases PUBCHEM, NCBI bioassay, KEGG pathways, BIND, and bindingDB databases. IHCD also provides 3D structure, computed molecular descriptors for all ingredients, and computer predicted potential protein targets and binding structures for selected ingredients. The crosslink was established by the following procedure: The chemical name and synonyms of each ingredient is mapped to those in the Pubchem substance database. The matched ones were subsequently mapped to other databases like MESH and Pubchem bioassay databases. MESH mapping from Pubchem substance ids were done by NCBI e-utilities. IHCD also contain information of pubmed abstracts related to herb. The pubmed abstracts were collected for herbs botanical name. The abstracts were downloaded in xml format by NCBI e-utilities, parsed and imported to database (Oliver, Bhalotia et al. 2004).

The predicted potential targets for each ingredient were derived from virtual screening of PDB database by using INVDOCK software (Y.Z. Chen 2001). INVDOCK has high accuracy of about 83% in predicting the protein targets of a small molecule when the scope of search is confined to all available 3D structures of protein (Chen and Ung 2001).

3.3 Database Access and Construction

IHCD website is at http://bidd.cz3.nus.edu.sg/ihcd. The web interface was developed by using Java server pages at front-end and MySQL database on backend (**Figure 17**).

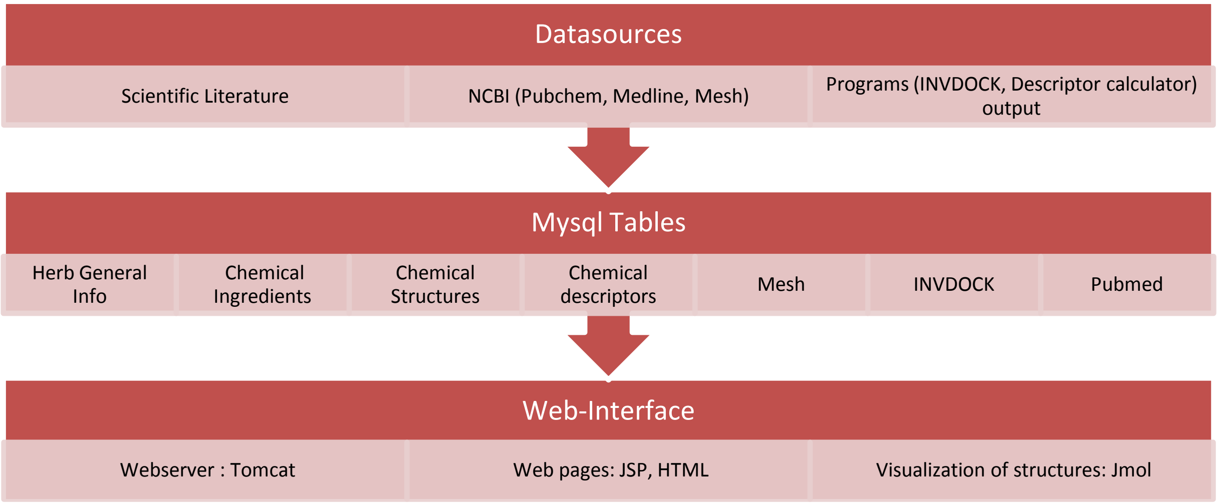
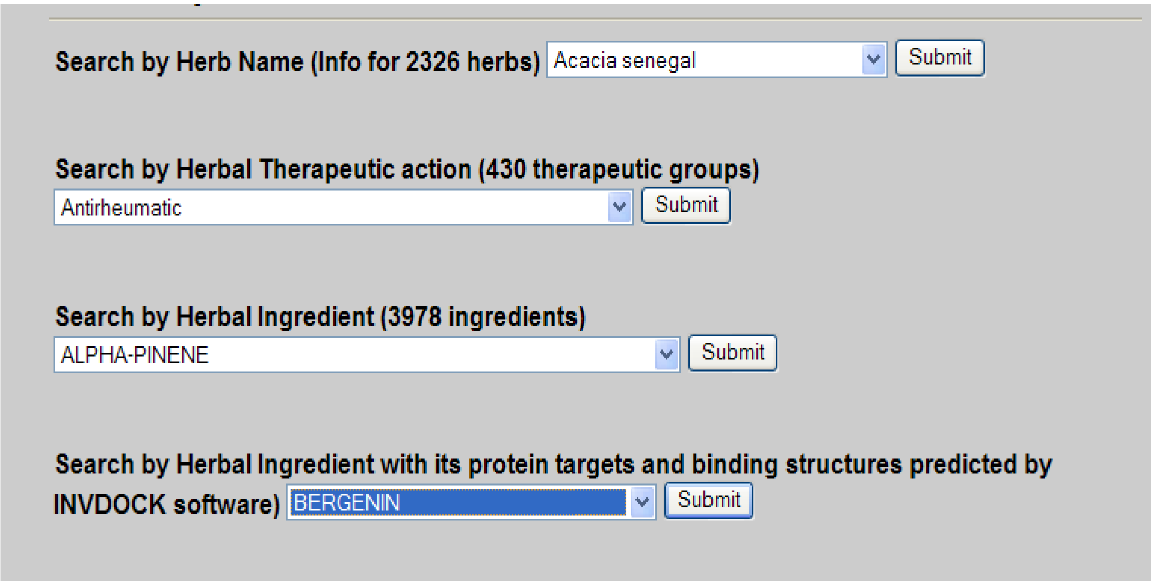


Figure 17: Overview of IHCD database model

Search fields were provided for searching the information in four different categories: herb name, therapeutic class, active ingredient, and ingredient with information about computer predicted targets (**Figure 18)**. When entered by choosing herb botanical name, the herb general information like botanical name, family name, Indian name, therapeutic activity is provided and further prompted to choose chemical ingredients for displaying chemical structure, descriptors, Pubchem substance mapping and other cross-linking information.



Search by Herb Name (Info for 2326 herbs) Acacia senegal

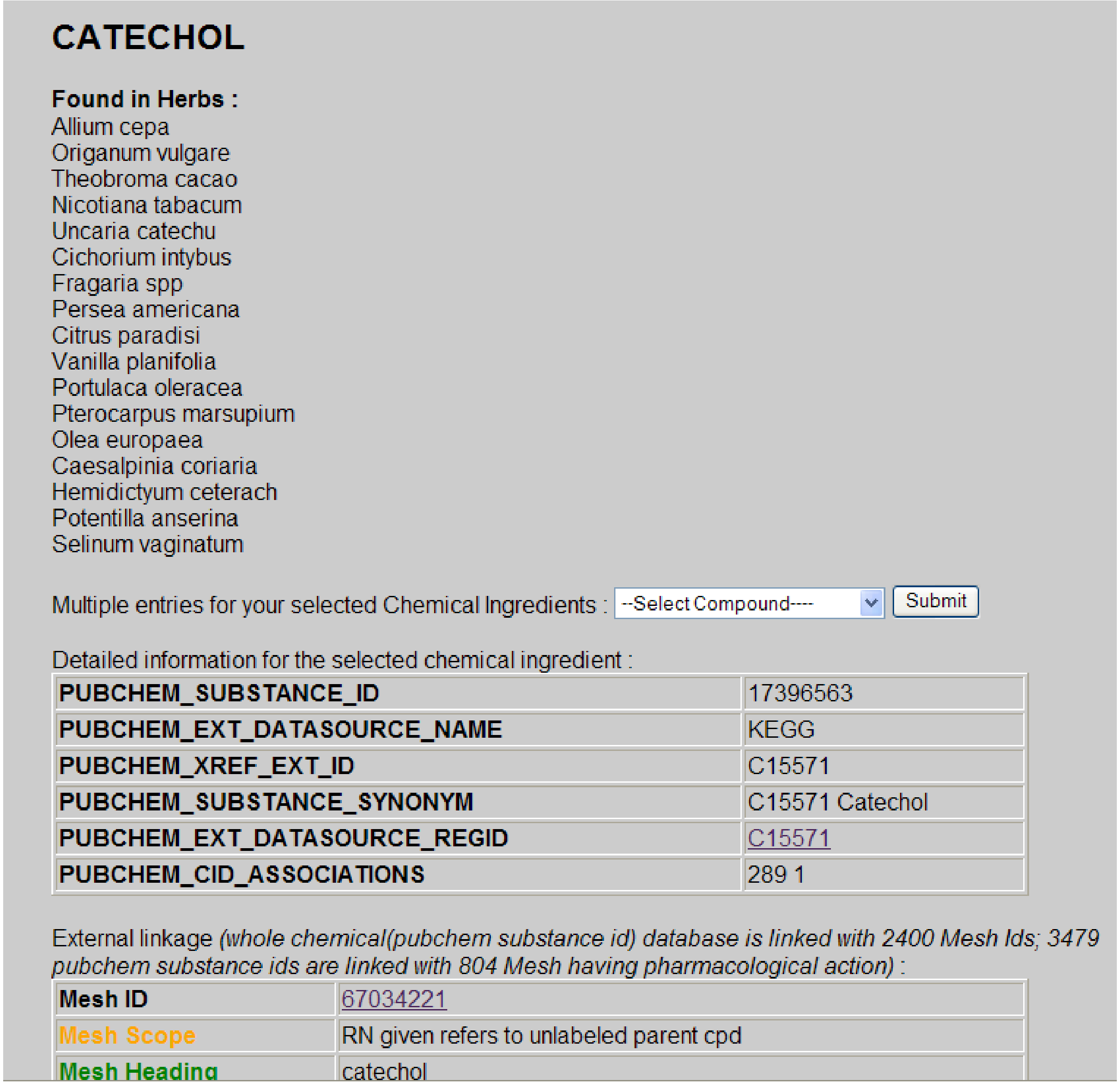
Search by Herbal Therapeutic action (430 therapeutic groups) Antirheumatic

Search by Herbal Ingredient (3978 ingredients) ALPHA-PINENE

Search by Herbal Ingredient with its protein targets and binding structures predicted by INVDOCK software) BERGENIN

Figure 18: The screenshot of IHCD main page

Similarly, when entered by selecting chemical ingredient, it will display the herbs in which the selected chemical is present, chemical information like structure, descriptors, and cross-linked Pubchem data and subsequently to other databases (**Figure 19**).



CATECHOL

Found in Herbs :
Allium cepa
Origanum vulgare
Theobroma cacao
Nicotiana tabacum
Uncaria catechu
Cichorium intybus
Fragaria spp
Persea americana
Citrus paradisi
Vanilla planifolia
Portulaca oleracea
Pterocarpus marsupium
Olea europaea
Caesalpinia coriaria
Hemidictyum ceterach
Potentilla anserina
Selinum vaginatum

Multiple entries for your selected Chemical Ingredients :

Detailed information for the selected chemical ingredient :

PUBCHEM_SUBSTANCE_ID	17396563
PUBCHEM_EXT_DATASOURCE_NAME	KEGG
PUBCHEM_XREF_EXT_ID	C15571
PUBCHEM_SUBSTANCE_SYNONYM	C15571 Catechol
PUBCHEM_EXT_DATASOURCE_REGID	C15571
PUBCHEM_CID_ASSOCIATIONS	289 1

External linkage *(whole chemical(pubchem substance id) database is linked with 2400 Mesh Ids; 3479 pubchem substance ids are linked with 804 Mesh having pharmacological action)* :

Mesh ID	67034221
Mesh Scope	RN given refers to unlabeled parent cpd
Mesh Heading	catechol

Figure 19: Screenshot of search result for a chemical ingredient

The main purpose of cross-linking our database with Pubchem Substance database is to further crosslink with database like Pubchem Bioassay, Mesh and Pubmed (Southan, Varkonyi et al. 2007; Zhou, Zhou et al. 2007). Although the user can use Pubchem substance id to get related other important interlinked information through Pubchem web site, some of the important feature of Pubchem are facilitated in our database to make it convenient. For example, IHCD is mapped to Mesh (Medical Subject Heading) database through Pubchem substance id. These 11590 substance ids are mapped to 2400 different Mesh Ids. Out of 11590 substance ids 3479 are linked to 804 Mesh terms having Pharmacological actions (**Figure 20**). We have just provided Mesh heading, subheading and scope wherever applicable and have created the hyperlink of Mesh id to NCBI mesh database for detailed information. For bioactivity analysis, these 11590 substances when searched on Pubchem bioassay server, it returned 990 tested molecules, of which 576 have, detailed information. The chemical ingredients page of our database http://bidd.cz3.nus.edu.sg/ihcd/mechdup.jsp has two hyperlinks showing this batch analysis:

1. http://bidd.cz3.nus.edu.sg/ihcd/bioactivity/Analysis.htm
2. http://bidd.cz3.nus.edu.sg/ihcd/bioactivity/Structure-Activity.htm .

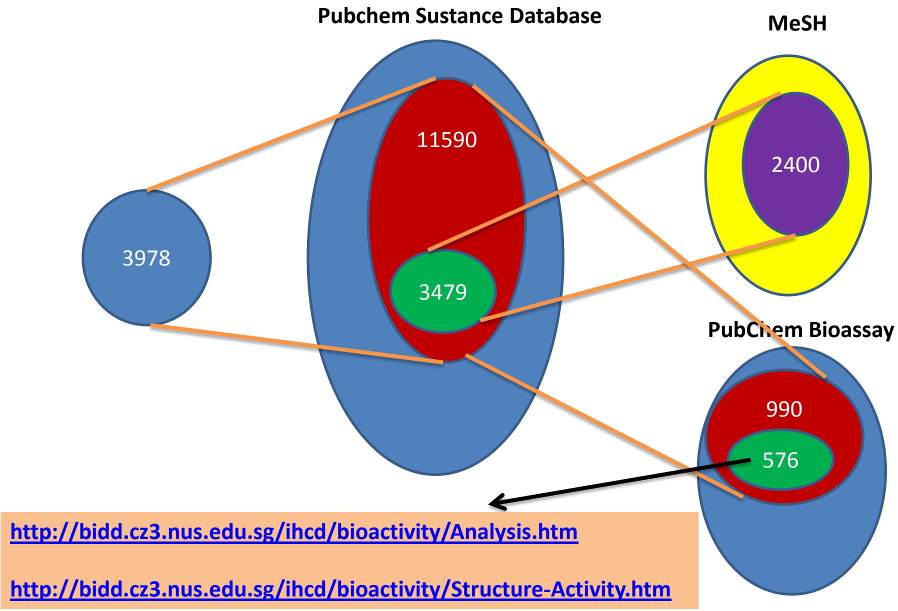
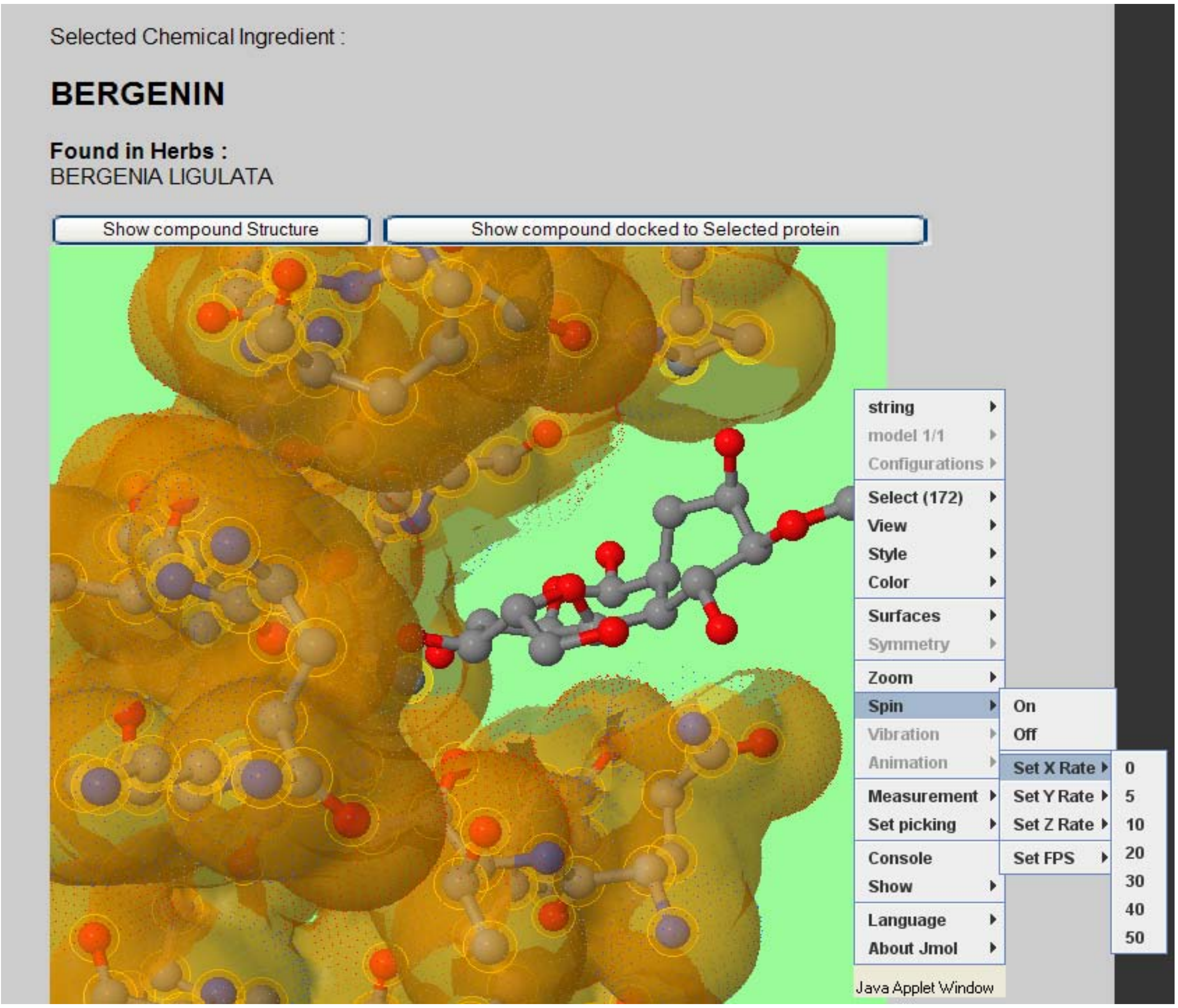

Pubchem Sustance Database
MeSH
11590
2400
3978
3479
PubChem Bioassay
990
576
http://bidd.cz3.nus.edu.sg/ihcd/bioactivity/Analysis.htm
http://bidd.cz3.nus.edu.sg/ihcd/bioactivity/Structure-Activity.htm

Figure 20: Chemical ingredients mapped to Pubchem Substance Database and which is linked to Medical Subject Heading (MeSH) database and Pubchem Bioassay.

We also provided the field 'Pubchem_ext_datasource' as well as 'Pubchem_ext_datasource_regid'. So, wherever 'Pubchem_ext_datasource' is DTP/NCI, user can take ‘Pubchem_ext_datasource_regid’ and can search for individual bioassay or can click the hyperlink already made. Other than this wherever external data source is bindingDB and KEGG, the ‘Pubchem_ext_datasource_regid’ is hyperlinked to their respective database. The detailed distribution of source can be seen on http://bidd.cz3.nus.edu.sg/ihcd/help.jsp .

When selected the chemical ingredients for which virtual screening has been done by INVDOCK software, additional feature will appear for selecting pdb id. All the chemical structure and docked ligand-protein complex in IHCD are visualized through jmol(2007). Once selected, user can view either the compound structure alone or compound docked into the protein cavity. User can view the compound-protein complex in various ways by right clicking on jmol applet window and interacting with jmol defined options. For example, in order to view compound and protein separately in docked complex, right clicking and selecting hetero ligand from the complex and then inverting the selection will turn into rest of cavity to be selected whose surface can be rendered as van der walls surface (**Figure 21**).



Selected Chemical Ingredient :

BERGENIN

Found in Herbs :
BERGENIA LIGULATA

Figure 21: Screenshot of visualization of a potential target of the bergenin found by INVDOCK software

The virtual screening hits by INVDOCK are primarily based on shape and energy cut off (Y.Z. Chen 2001; Chen, Ung et al. 2003).

Inverse Docking Procedure: INVDOCK is well established method in identifying multiple protein targets of a compound. A cavity database is being utilized by INVDOCK which has been created by Protein Data Bank (PDB). In this inverse docking procedure compounds shape is matched against the cavity and energy is minimized in situ for both compound and amino acid residues at that particular cavity of the protein(Chen and Zhi 2001). The energy function of INVDOCK is determined by following equation:

$$V = \frac{1}{2}\sum\nolimits_{bonds} K_r (R - R_{eq})^2 + \frac{1}{2}\sum\nolimits_{angles} K_\theta (\theta - \theta_{eq})^2 + \frac{1}{2}\sum\nolimits_{torsions} V_n [1 - \cos\Big(n(\varphi -$$

$$\varphi eq]+H\ bonds[V0(1-e-a(r-r0))2-V0]+non\ bonded[Aijrij12-Bijrij6+qiqj\varepsilon rrij]$$

Where R = bond length , θ = angle, and φ = torsion angle, R_{eq} = equilibrium bond length, θ_{eq} = equilibrium angle, φ_{eq} = equilibrium torsion angle , K_r = covalent bond angle , K_θ = bond angle bending force constant, r is hydrogen bond donor–acceptor distance, V_n and n are torsion parameters, and V_0, a and r_0 are hydrogen bond potential parameters. The values of R, θ, and φ are from the original PDB structure while the values of R_{eq}, θ_{eq} , and φ_{eq} are from structure of the drug.

In INVDOCK, there is an option to select single conformer or multiple conformer of a compound. We applied multiple conformer option for each of the studied compound. Currently, a virtual hit of compounds and literature relevance are not cross-linked in our database and is one of future work in further development of our database. As of preliminary work to illustrate example use of mapping the virtual hits of INVDOCK to literature for understanding the mechanism of chemical ingredients, the methodology is provided below with chemical ingredient Bergenin taken as example. General procedure of doing mechanistic analysis is as follows:

Filtering the INVDOCK result:

The protein targets found by INVDOCK for each compound were imported in Oracle database tables. Also, the Therapeutic Target Database (TTD) was imported in Oracle table. For each compound the protein targets were filtered which were present in TTD through SQL query. Further, the protein targets were filtered where the organism sources were human.

Importing the INVDOCK results into Pathway Studio: To analyze INVDOCK results we imported protein list to Pathway Studio software. The ‘Pathway Studio’ by Ariadne Genomics utilizes Medscan technology, a natural language processing method, to find biological interactions like protein-protein and protein-small molecule from literature. Medscan has accuracy of 90% (Medscan 1.8) and gives only 10% false positive interactions. Moreover, if it retrieves interaction from repeated sentence then it has 100% accuracy to predict the molecular interaction (Yuryev, Mulyukov et al. 2006). The INVDOCK shows the result in PDB id format but Pathway Studio does not have the functionality to import PDB ids. The Pathway Studio recognizes Locuslink (Entrez gene) id, Hugo id, Genbank id, Microarray ID, Name or alias, and Swissprot accession. Thus, it is necessary to convert the PDB ids into any of the format which Pathway Studio recognizes. There are some online id mapping services which have the option to convert PDB id to other formats. The PDB ids were mapped to SWISSPROT accession number by the online id mapping service http://pir.georgetown.edu/pirwww/search/idmapping.shtml. Then these SWISSPROT accession numbers were used to import proteins in Pathway Studio. However, only 70% of these SWISSPROT accession number were recognized by Pathway Studio. Therefore, we tried to convert PDB ids to Entrez gene id which is comprehensively recognizable by Pathway Studio. The id mapping to convert ‘SWISSPROT accession number’ to ‘Entrez gene id was done by, http://www.pir.uniprot.org/search/idmapping.shtml and http://www.ariadnegenomics.com/services/idmap.html. In our case, Pathway Studio imports about 95% protein targets when it is in Entrez Gene id format. Each compound protein targets were imported in Pathway Studio separately.

Studying mechanism of Bergenin with INVDOK and Pathway Studio: Bergenin (**Figure 22)** is an important constituent of *Bergenia Ligulata*. The INVDOCK protein targets of Bergenin were imported in pathway studio and grouped together which is shown in **Table 1**. There were 53 abstracts in Pubmed when searched for the word ‘Bergenin’. Based on Pubmed abstracts information Bergenin- protein interaction graph had been created by Pathway studio (**Figure 23**). The detailed references for every interaction of Figure 2 are shown in Table 2.

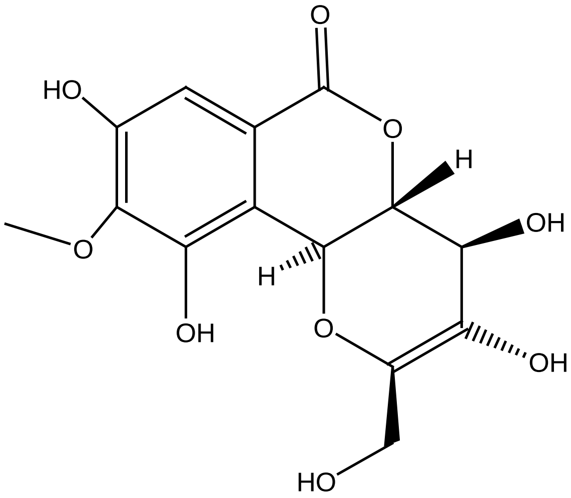


Figure 22: Chemical structure of Bergenin

Table 1: Bergenin INVDOCK targets (mammalian)

#	Name	Description	LocusLink ID
1	ACAT2	acetyl-Coenzyme A acetyltransferase 2 (acetoacetyl Coenzyme A thiolase)	106825, 21456, 110510, 39, 110460, 11414, 224530, 11415
2	SAT	spermidine/spermine N1-acetyltransferase	25188, 6303, 106503, 302642, 20229
3	TP53	tumor protein p53 (Li-Fraumeni syndrome)	24842, 7157, 22059, 289761, 224883, 301300
4	ESR1	estrogen receptor 1	13982, 24890, 2099, 103092
5	MPG	N-methylpurine-DNA glycosylase	17477, 103693, 268395, 24561, 4350
6	CASP3	caspase 3, apoptosis-related cysteine protease	836, 12367, 25402
7	CASP7	caspase 7, apoptosis-related cysteine protease	64026, 840, 12369, 107145
8	CTNNA1	catenin (cadherin-associated protein), alpha 1, 102kDa	12385, 1495, 307505, 106962, 106853
9	CTNNB1	catenin beta	252926, 360543, 112387, 84353, 209012, 1499, 12387
10	TGFA	transforming growth factor, alpha	7039, 21802, 24827
11	AXL	AXL receptor tyrosine kinase	101531, 308444, 22231, 83625, 26362,

			558
12	FGFR1	fibroblast growth factor receptor 1 (fms-related tyrosine kinase 2, Pfeiffer syndrome)	2260, 14182, 360286, 51033, 84151, 497708, 102305, 79114
13	BAIAP2	BAI1-associated protein 2	108100, 94087, 97767, 117542, 10458
14	MAN1B1	mannosidase, alpha, class 1B, member 1	227619, 26016, 51697, 11253
15	ADAM17	a disintegrin and metalloproteinase domain 17 (tumor necrosis factor, alpha, converting enzyme)	6868, 111491, 57027, 11491
16	MMP2	matrix metalloproteinase 2 (gelatinase A, 72kDa gelatinase, 72kDa type IV collagenase)	4313, 381686, 81686, 17390
17	CHIT1	chitinase 1 (chitotriosidase)	7831, 1118, 289032, 71884
18	Bche	butyrylcholinesterase	590, 65036, 12038
19	CMA1	chymase 1, mast cell	29267, 1215, 25627
20	APOA1	apolipoprotein A-I	335, 11806, 25081
21	SERPINA1	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1	5265, 116807, 64311, 24648
22	CTSG	cathepsin G	13035, 1511, 290257
23	SERPINC1	serpin peptidase inhibitor, clade C (antithrombin), member 1	304917, 462, 98260, 11905
24	CP	ceruloplasmin (ferroxidase)	51906, 294942, 12870, 24268, 1356
25	PKND	cathepsin K (pycnodysostosis)	13038, 1513, 99590, 29175, 94319
26	CTSC	cathepsin C	13032, 5065, 1075, 101486, 25423, 50958
27	CTSS	cathepsin S	50654, 13040, 1520, 50653
28	GZMB	granzyme B (granzyme 2, cytotoxic T-lymphocyte-associated serine esterase 1)	171528, 14939, 3002, 105531
29	CTSF	cathepsin F	8722, 107211, 56464
30	ABO	ABO blood group (transferase A, alpha 1-3-N-acetylgalactosaminyltransferase; transferase B, alpha 1-3-galactosyltransferase)	296504, 311792, 28, 65270, 80908
31	B3GAT3	beta-1,3-glucuronyltransferase 3 (glucuronosyltransferase I)	26229, 293722, 72727
32	B3GAT1	beta-1,3-glucuronyltransferase 1 (glucuronosyltransferase P)	102604, 27087, 117108, 76898, 964
33	NPR3	natriuretic peptide receptor C/guanylate cyclase C (atrionatriuretic peptide receptor C)	289058, 4883, 155012, 192290, 498240, 16861, 18162, 20902, 360263, 25339,
34	FSHR	follicle stimulating hormone receptor	25449, 14309, 4959, 2492

35	Braf	v-raf murine sarcoma viral oncogene homolog B1; belongs to the Serine/Threonine family of protein kinases.	52385, 232705, 12187, 319686, 109880, 58892, 330290, 673, 114486, 97330
36	SRC	Rous sarcoma oncogene	83805, 99351, 20779, 320779, 6714
37	CDC42	cell division cycle 42 (GTP binding protein, 25kDa)	100285, 100196, 12540, 998, 332881, 212710, 64465
38	Rac1	ras-related C3 botulinum toxin substrate 1 (rho family, small GTP binding protein Rac1)	363875, 5879, 19353, 171377, 100781, 52352, 319353
39	ARL2	ADP-ribosylation factor-like 2	107390, 402, 65142, 69901, 80563, 107120, 56327
40	POR1	ADP-ribosylation factor interacting protein 2 (arfaptin 2)	76932, 23647, 293344
41	PSCD2	pleckstrin homology, Sec7 and coiled-coil domains 2 (cytohesin-2)	9266, 116692, 19158
42	PLCE1	phospholipase C, epsilon 1	51196, 56231, 74055, 114633
43	EGFR	epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian)	360274, 13649, 103781, 1956, 170565, 24329
44	CDK6	cyclin-dependent kinase 6	330039, 1021, 100686, 12571, 114483
45	PTK2	PTK2 protein tyrosine kinase 2	114083, 414083, 14083, 25614, 5747
46	MAP2K1	mitogen-activated protein kinase kinase 1	5604, 326395, 19101, 26395, 170851
47	PDPK1	3-phosphoinositide dependent protein kinase-1	5170, 81745, 18607, 28993
48	Abl1	v-abl Abelson murine leukemia oncogene 1	98922, 11350, 311860, 111350, 368055, 24155, 25
49	KIT	v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog	64030, 330256, 16590, 72135, 3815
50	PRKR	protein kinase, interferon-inducible double stranded RNA dependent	106646, 76759, 21850, 5610, 54287, 106605, 19106
51	HCK	hemopoietic cell kinase	99093, 3055, 25734, 15162
52	GBA	glucosidase, beta; acid (includes glucosylceramidase)	2629, 14466
53	STK6	serine/threonine kinase 6	99385, 261730, 99193, 6790
54	CSK	C-Src tyrosine Kinase. A ubiquitously expressed intracellular protein involved in tyrosine phosphorylation; contains a Src homology 2 (SH2) and SH3 domain at its C-terminus.	1445, 315707, 12988, 102764
55	EPHA2	EPH receptor A2	13836, 100429, 1969
56	ACK1	tyrosine kinase, non-receptor, 2	51789, 53909,

			303882, 106433, 10188, 224114
57	CSNK2B	casein kinase 2, beta polypeptide	81650, 1460, 257555, 257616, 13001
58	EPOR	The Erythropoietin receptor, a member of the cytokine receptor family, plays an important role in erythroid cell survival. Upon erythropoietin binding, the erythropoietin receptor activates Jak2 tyrosine kinase which activates different intracellular p...	13857, 113857, 24336, 2057
59	BCR	breakpoint cluster region	110279, 12058, 103260, 103308, 613, 309696, 71258
60	CLK1	CDC-like kinase 1	98487, 301434, 1195, 12747
61	HSPCA	heat shock 90kDa protein 1, alpha	104921, 299331, 15524, 104922, 104409, 3320, 15519
62	YWHAQ	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, theta polypeptide	25577, 76805, 10971, 22630, 104726, 104947, 97839
63	YWHAH	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, eta polypeptide	25576, 7533, 22629, 194104
64	CSNK1G2	casein kinase 1, gamma 2	1455, 65278, 72764, 103236
65	F3	coagulation factor III (thromboplastin, tissue factor)	14066, 25584, 2152, 99486
66	F7	coagulation factor VII (serum prothrombin conversion accelerator)	14068, 2155, 260320, 101998
67	NPPC	natriuretic peptide precursor C	4880, 114593, 18159
68	TNFSF13B	tumor necrosis factor (ligand) superfamily, member 13b	24099, 52115, 10673, 89794
69	GPI	glucose phosphate isomerase	2821, 292804, 24403, 110643, 110644, 14754, 14753, 14751, 110600, 14752
70	CGA	glycoprotein hormones, alpha polypeptide	1081, 116700, 12640
71	FSHB	follicle stimulating hormone, beta polypeptide	25447, 14308, 2488
72	NMNAT3	nicotinamide nucleotide adenylyltransferase 3	74080, 349565
73	RNASE2	ribonuclease, RNase A family, 2 (liver, eosinophil-derived neurotoxin)	53877, 13587, 6036
74	PDE5A	phosphodiesterase 5A, cGMP-specific	171115, 8654, 242202
75	PLA2G10	phospholipase A2, group X	8399, 26969, 29359, 26565
76	DCK	deoxycytidine kinase	1633, 79127, 13178
77	CYP2C9	cytochrome P450, family 2, subfamily C, polypeptide 9	29277, 1560, 1559, 13096, 29298, 29296, 171521, 29297
78	ALAD	aminolevulinate, delta-, dehydratase	17025, 25374, 210
79	FHIT	fragile histidine triad gene	14198, 2272, 60398, 105644, 2385

80	ACADM	acyl-Coenzyme A dehydrogenase, C-4 to C-12 straight chain	51779, 34, 99793, 11364, 24158
81	CYP2C8	cytochrome P450, family 2, subfamily C, polypeptide 8	1558
82	SULT2A1	sulfotransferase family, cytosolic, 2A, dehydroepiandrosterone (DHEA)-preferring, member 1	107959, 20864, 20859, 6822
83	HK1	hexokinase 1	15275, 3098, 25058
84	CYP3A4	cytochrome P450, family 3, subfamily A, polypeptide 4	171352, 1576, 13113, 25642, 1575, 229675
85	ALDOA	aldolase A, fructose-bisphosphate	24189, 226, 11674
86	SFN	stratifin	313017, 2810, 55948
87	PAH	phenylalanine hydroxylase	18478, 5053, 103418, 24616
88	BCAT2	branched chain aminotransferase 2, mitochondrial	12036, 587, 64203
89	FKBP4	FK506 binding protein 4, 59kDa	260321, 101346, 2288, 14228, 107270
90	NUDT3	nudix (nucleoside diphosphate linked moiety X)-type motif 3	11165, 294292, 56409, 10909, 106513, 68495
91	HSD17B1	hydroxysteroid (17-beta) dehydrogenase 1	15485, 3292, 25322
92	PGDS	prostaglandin D2 synthase, hematopoietic	27306, 54486, 58962
93	GSS	glutathione synthetase	2937, 14854, 25458, 98903
94	HADHSC	L-3-hydroxyacyl-Coenzyme A dehydrogenase, short chain	99932, 113965, 99798, 99484, 360353, 3033, 15107
95	NT5M	5',3'-nucleotidase, mitochondrial	56953, 287368, 69877, 103850
96	DECR1	2,4-dienoyl CoA reductase 1, mitochondrial	1666, 117543, 67460
97	GMPR2	guanosine monophosphate reductase 2	108706, 69081, 105446, 192357, 70653, 10784, 319199, 51292
98	PDE6D	phosphodiesterase 6D, cGMP-specific, rod, delta	18582, 5147, 98438
99	ACY1	aminoacylase 1	109652, 95, 24164, 300981, 11483, 66130
100	COMTD1	catechol-O-methyltransferase domain containing 1	305685, 69156, 118881
101	AMPH	amphiphysin (Stiff-Man syndrome with breast cancer 128kDa autoantigen)	218038, 109629, 11718, 60668, 273
102	CAPN1	calpain 1, (mu/I) large subunit	12333, 29153, 823
103	BACE	beta-site APP-cleaving enzyme 1	29392, 23621, 97509, 23821
104	PAPSS1	3'-phosphoadenosine 5'-phosphosulfate synthase 1	99599, 295443, 23971, 9061
105	BHMT	betaine-homocysteine methyltransferase	12116, 81508, 328308, 268685, 635, 218451
106	Dut	dUTP pyrophosphatase	93804, 71267, 52842, 80993, 67757, 1854,

			23864, 94200, 110074
107	CFTR	cystic fibrosis transmembrane conductance regulator, ATP-binding cassette (sub-family C, member 7)	1080, 368064, 24255, 101370, 547216, 12638
108	AKR1B1	aldo-keto reductase family 1, member B1 (aldose reductase)	11677, 231
109	NR3C1	nuclear receptor subfamily 3, group C, member 1 (glucocorticoid receptor)	2908, 14815, 389335, 24413
110	ARHGEF1	Rho guanine nucleotide exchange factor (GEF) 1	16801, 60323, 9138
111	EPHX2	epoxide hydrolase 2, cytoplasmic	13850, 65030, 105655, 2053
112	MASA	E-1 enzyme	58478, 305177, 101037, 97253, 67870
113	ACE	angiotensin I converting enzyme (peptidyl-dipeptidase A) 1	11421, 192774, 217246, 116576, 1636, 24310, 104604
114	TPSAB1	tryptase alpha/beta 1	17230, 54271, 7177, 7176
115	FOLH1	folate hydrolase (prostate-specific membrane antigen) 1	85309, 53320, 2346
116	GZMA	granzyme A (granzyme 1, cytotoxic T-lymphocyte-associated serine esterase 3)	266708, 3001, 105363, 14938
117	CASP2	caspase 2, apoptosis-related cysteine protease (neural precursor cell expressed, developmentally down-regulated 2)	835, 12366, 64314
118	ACE2	angiotensin I converting enzyme (peptidyl-dipeptidase A) 2	302668, 59272, 26125, 70008
119	TPSB2	tryptase beta 2	64499
120	DF	D component of complement (adipsin)	1675
121	CREG1	cellular repressor of E1A-stimulated genes 1	433375, 8804, 289185
122	F9	coagulation factor IX (plasma thromboplastic component, Christmas disease, hemophilia B)	2158, 24946, 14071, 103022
123	CD4	CD4 antigen	12504, 24932, 920
124	GP1BA	glycoprotein Ib (platelet), alpha polypeptide	287460, 14723, 2811
125	BST1	bone marrow stromal cell antigen 1	12182, 683, 269647, 81506
126	CD209L	C-type lectin domain family 4, member M	10332
127	HIF1AN	hypoxia-inducible factor 1, alpha subunit inhibitor	309434, 368022, 77039, 84175, 55662, 319594
128	Chd1	chromodomain helicase DNA binding protein 1	106815, 12648, 75119, 1105, 106666, 308215
129	ETFB	electron-transfer-flavoprotein, beta polypeptide	13988, 2109, 68360, 110826, 72756, 292845
130	CLIC1	chloride intracellular channel 1	1192, 114584, 406864
131	GGA3	golgi associated, gamma adaptin ear containing, ARF binding protein 3	260302, 23163

132	NCBP1	nuclear cap binding protein subunit 1, 80kDa	298075, 110519, 60346, 4686
133	Etfa	electron-transfer-flavoprotein, alpha polypeptide (glutaric aciduria II)	300726, 2108, 235393, 259204, 13987, 110842, 52321
134	BIRC7	baculoviral IAP repeat-containing 7 (livin)	79444, 64126, 329581
135	APC	adenomatosis polyposis coli	107030, 106874, 324, 106987, 24205, 11789
136	S100A8	S100 calcium binding protein A8 (calgranulin A)	20201, 99591, 116547, 6279, 104427
137	GPX1	glutathione peroxidase 1	2876, 24404, 14775, 102648, 102449
138	HSPA1A	heat shock 70kDa protein 1A	193740, 15514, 24472, 24964, 3303
139	ANXA5	annexin A5	25673, 308, 11747, 97115
140	BAG1	BCL2-associated athanogene	12017, 297994, 573
141	PROCR	protein C receptor, endothelial (EPCR)	19124, 98921, 10544
142	CENPE	centromere protein E, 312kDa	109951, 1062, 12619, 16550, 229841
143	H3FA		8350
144	GAS6	growth arrest-specific 6	2621, 58935, 14456
145	RAP1GA1	RAP1, GTPase activating protein 1	78775, 9676, 100110, 76280, 5909, 110351, 298570, 19393
146	NCBP2	nuclear cap binding protein subunit 2, 20kDa	68092, 98015, 288040, 106266, 106124, 22916
147	TH	tyrosine hydroxylase	7054, 25085, 21823
148	AP2B1		163
149	H3FD	histone 1, H3e	8353, 319151
150	HSPA1B	heat shock 70kDa protein 1B	3304, 294254, 15511
151	HRSP12	heat-responsive protein 12	15473, 65151, 10247
152	P5326		83638
153	H3FL		8358
154	LOC285362		285362

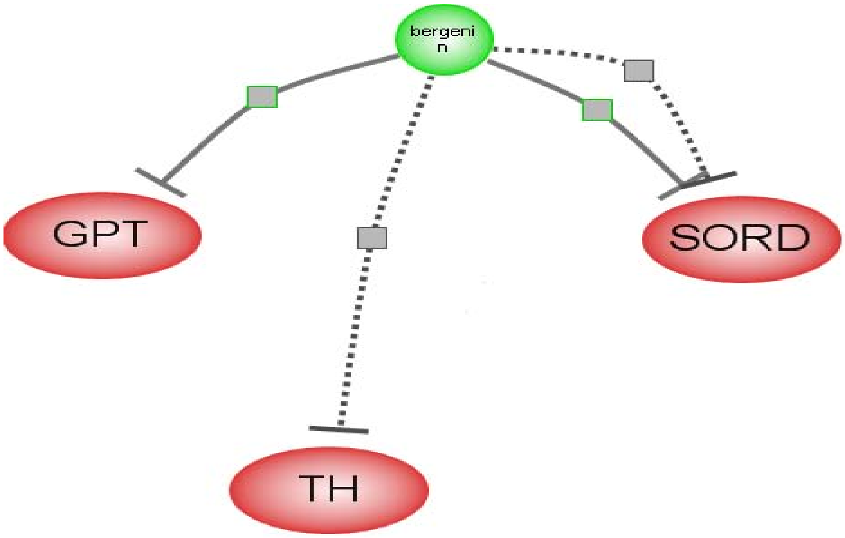


Figure 23: Graph generated by Pathway Studio for the Pubmed search word 'bergenin'. Green color circle- *small molecule*. Red color circle- *protein*. Grey dotted line – *Regulation.* Solid grey line- *MolTransport*. Negative regulation is shown as "---|". Negative MolTransport is shown as "-|". SORD: Sorbitol dehydrogenase, TH: Tyrosine hydroxylase, GPT: Glutamic pyruvic transaminase.

Table 2: Corresponding reference of Figure 22

Entities	Type	MedLine Reference	Sentence
bergenin ---| TH	Regulation	13680837:2	Bergenin and norbergenin inhibited the TH activity by 29.0% and 53.4% at a concentration of 20 microg/mL, respectively, and exhibited noncompetitive inhibition of TH activity with the substrate l-tyrosine.
SORD |--- bergenin	MolTransport	10720791:1	Bergenin (100 microM) decreased the release of glutamic pyruvic transaminase and sorbitol dehydrogenase by 62 and 50%, respectively, into hepatocyte medium incubated for 14 h with 1.5 mM galactosamine.
bergenin ---| GPT	MolTransport	10720791:1	Bergenin (100 microM) decreased the release of glutamic pyruvic transaminase and sorbitol dehydrogenase by 62 and 50%, respectively, into hepatocyte medium incubated for 14 h with 1.5 mM galactosamine.
SORD |--- bergenin	Regulation	10720791:1	Bergenin (100 microM) decreased the release of glutamic pyruvic transaminase and sorbitol dehydrogenase by 62 and 50%, respectively, into hepatocyte medium incubated for 14 h with 1.5 mM galactosamine.
SORD |--- bergenin	Regulation	10661887:1	Bergenin significantly reduced the activities of glutamic pyruvic transaminase and sorbitol dehydrogenase released from the CCl4-intoxicated hepatocytes.

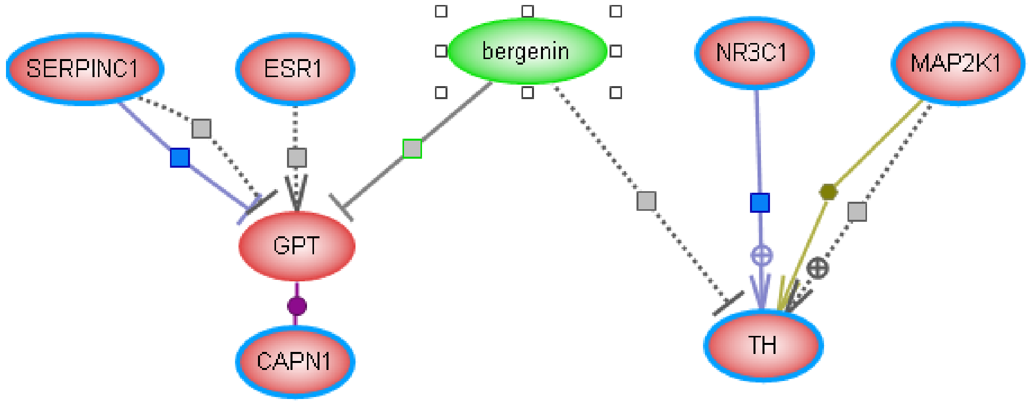


Figure 24: Mapping of Bergenin INVDOCK targets to literature. INVDOCK targets of bergenin are highlighted in blue (TH, CAPN1, SERPINC1, ESR1, NR3C1, MAP2K1). Green color circle- *small molecule*. Red color circle- *protein*. Grey dotted line – *Regulation*.. Solid grey line- *MolTransport*. . Blue arrow – *Expression* relation. Brown arrow – *MolSynthesis.*Arrow with "+" indicate positive relation and negative relation is shown as "-|"

By examining **Figure 23** and **Table 1**, we get limited information about molecular mechanism of Bergenin. Bergenin non-competitively inhibits Tyrosine Hydroxylase (Zhang, Fang et al. 2003). Other two proteins Glutathione disulfide reductase (GSH) and sorbitol dehydrogenase (SORD) are indicators of hepatotoxicity. In case of hepatotoxicity (liver cell damage), glutamic pyruvic transaminase (GPT) and sorbitol dehydrogenase (SDH) are released from hepatocytes to extracellular spaces. Therefore, these proteins are important for determining the hepatotoxicity levels of the toxicant and liver protective effects of test compound. Next, GSH is important to prevent lipid peroxidation. In case of hepatotoxicity, GSH decreases and bergenin has been found to preserve the activity of GSH. However, these experimental finding are not sufficient to clarify about the protein targets of Bergenin for its effectiveness in liver disorders. By INVDOCK we found the protein targets of bergenin and is been imported to Pathway Studio (Table 1)

One of the ways to understand the molecular mechanism of bergenin would be to map experimental findings in **Figure 23** with INVDOCK results in **Table 1**. To do this, a pathway was built by finding all entities connected to bergenin. The filter was set to find only proteins connected to bergenin and maximum number of steps was 2. Another pathway was built by finding shortest pathway between bergenin and its imported INVDOCK protein targets. These two pathways were intersected which is shown in **Figure 24**. Bergenin's INVDOCK protein targets are highlighted in blue. Tyrosine Hydroxylase (TH) has been found as the target of Bergenin by literature and also by INVDOCK. According to literature Bergenin non-competitively inhibits Tyrosine Hydroxylase, corresponding INVDOCK results is shown in **Table 3.**

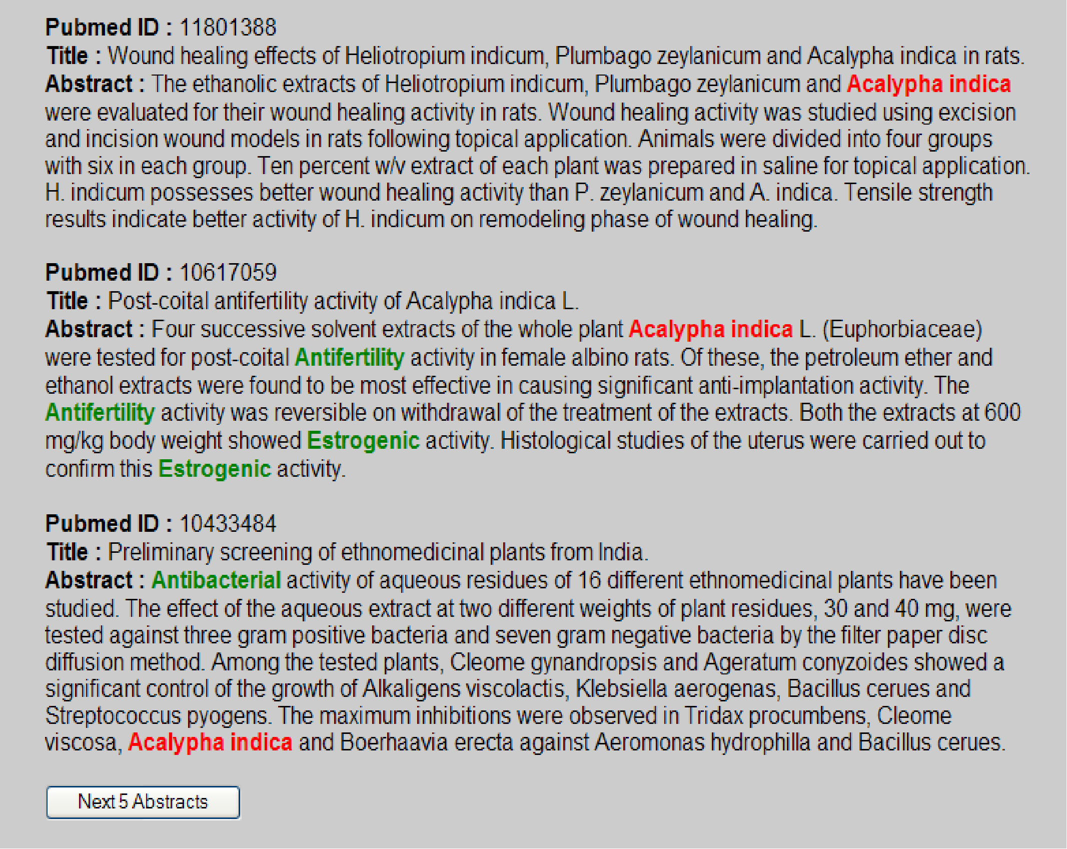
Table 3: Bergenin inhibits tyrosine hydroxylase, corresponding PDB entries are shown

PDB	Classification	Name	Species	Energy
6pah	MONOOXYGENASE	PHENYLALANINE 4-MONOOXYGENASE	HUMAN	-51.4
1dmw	OXIDOREDUCTASE	PHENYLALANINE HYDROXYLASE	HUMAN	-51
1ltz	OXIDOREDUCTASE	PHENYLALANINE-4-HYDROXYLASE	BACTERIA	-46.2
2toh	HYDROXYLASE	TYROSINE 3-MONOOXYGENASE	RAT	-50

By this method the INVDOCK targets are mapped which have literature implications in context of bergenin. So, the possible reason of decrease in levels of GPT in intoxicated liver cells may be due to modulations through CAPN1 (calpain 1, (mu/I) large subunit), SERPINC1 (serpin peptidase inhibitor, clade C (antithrombin), member 1) and ESR1 (estrogen receptor 1).

Mapping IHCD to Pubmed:

As of preliminary literature correlation, the text mining of pubmed abstract for herb name and herb with disease term are done. The Pubmed abstracts were retrieved programmatically with NCBI Entrez facility for the entire herbs name and their combination with disease and chemical ingredient terms. User can search the pubmed abstracts by selecting the herb name. The herb name and disease term in corresponding pubmed abstract are highlighted (**Figure 25**).



Pubmed ID : 11801388
Title : Wound healing effects of Heliotropium indicum, Plumbago zeylanicum and Acalypha indica in rats.
Abstract : The ethanolic extracts of Heliotropium indicum, Plumbago zeylanicum and **Acalypha indica** were evaluated for their wound healing activity in rats. Wound healing activity was studied using excision and incision wound models in rats following topical application. Animals were divided into four groups with six in each group. Ten percent w/v extract of each plant was prepared in saline for topical application. H. indicum possesses better wound healing activity than P. zeylanicum and A. indica. Tensile strength results indicate better activity of H. indicum on remodeling phase of wound healing.

Pubmed ID : 10617059
Title : Post-coital antifertility activity of Acalypha indica L.
Abstract : Four successive solvent extracts of the whole plant **Acalypha indica** L. (Euphorbiaceae) were tested for post-coital **Antifertility** activity in female albino rats. Of these, the petroleum ether and ethanol extracts were found to be most effective in causing significant anti-implantation activity. The **Antifertility** activity was reversible on withdrawal of the treatment of the extracts. Both the extracts at 600 mg/kg body weight showed **Estrogenic** activity. Histological studies of the uterus were carried out to confirm this **Estrogenic** activity.

Pubmed ID : 10433484
Title : Preliminary screening of ethnomedicinal plants from India.
Abstract : **Antibacterial** activity of aqueous residues of 16 different ethnomedicinal plants have been studied. The effect of the aqueous extract at two different weights of plant residues, 30 and 40 mg, were tested against three gram positive bacteria and seven gram negative bacteria by the filter paper disc diffusion method. Among the tested plants, Cleome gynandropsis and Ageratum conyzoides showed a significant control of the growth of Alkaligens viscolactis, Klebsiella aerogenas, Bacillus cerues and Streptococcus pyogens. The maximum inhibitions were observed in Tridax procumbens, Cleome viscosa, **Acalypha indica** and Boerhaavia erecta against Aeromonas hydrophilla and Bacillus cerues.

Figure 25: Screenshot of pubmed abstracts display page on IHCD. Herb name is highlighted in red and disease terms are highlighted in green

In context of speed of IHCD, most of the queries performed are very fast which has been achieved by proper indexing of every field involved in query process(Rao 2004). The only exception to speed will be the first time loading of Jmol applet which can take 5-15 second but subsequent search will be very fast as the applet resides inside local java virtual machine.

3.4 Discussion and Conclusion

The usefulness of the IHCD in facilitating general information about herbs and their chemical ingredients is evident through IHCD website. In addition, IHCD attempt to provide automation and rationalization in understanding the mechanisms of herbs and herbal ingredient. In case of herbs we are generally aware of their therapeutic activity as well as negligible toxicity profile in respect of being traditional medicine. These herbs are generally understood as having multiple targets in human body system. The overall therapeutic activity of the herb may come from these multiple targets. In our example analysis for bergenin, it was attempted to combine all the targets by individual principal ingredients (e.g. bergenin) of the herb (e.g. *Bergenia Liguta*) to represent mechanism for whole herb. Therapeutic indications of bergenin available in literature were successfully covered by INVDOCK protein targets. If there is information about the targets or the pathway through which herbal ingredients exhibit their therapeutic activity, the appropriate targets can be selected to perform experimental studies like binding assays. The information at IHCD can be utilized by researchers working in the area of plant based drug discovery. User (any researcher) will get the information about herbs, herbal ingredients, their therapeutic targets as well as interactions based on Pubmed abstracts and from INVDOCK software through website. Furthermore, the practical significance of the results lies in its ability of predicting the novel targets and unexplored therapeutic indication of the particular herb and herbal ingredients.

Generally, the logic behind the usage of traditional medicine is being confirmed by functional assays. These functional assays consolidate the confidence about their usage and their ingredients. However, functional assays are unable to solve mystery of their mechanisms which is very important in drug discovery. Our result is presumably predicting the direct binding to the protein based on threshold binding energy by INVDOCK. In our present study we have binding energy information for drug-protein complex but we are not using as selection criteria for a protein being a better target. Due to present INVDOCK algorithm, all the binding energy above threshold is treated equal. In future it can be tuned to make a section of a target over other based on the binding energy. In addition, another way of making a selection of one target over other in the mechanistic pathway can be based on number of references supporting the particular protein-protein interaction. Also, the process of id mapping is indirect i.e. PDB id is first converted to Swissprot id and then to Entrez gene id which is time consuming and introduces little discrepancy. This could be solved by mapping of Protein Data Bank by Pathway Studio and by adding the way to import protein list based on PDB id.

Compounds from medicinal plants are important and resources about their information are needed. The IHCD database provides information about Indian herbs and their chemical ingredients. It connects chemical ingredients of the commonly used herbs in Ayurveda to therapeutic classes, biological pathway and activity related databases like Pubchem bioassay, KEGG, BIND, bindingDB. The database is addressing both the general information as well as mechanistic approach of herbs used in Ayurveda. It is expected that the building of such an integrated database, which can be constantly updated, could provide an understanding of herbs and their ingredient’s therapeutic action. An important aspect of IHCD database design is the ability to expand seamlessly either by manual addition of data or by cross-linking to other databases.

Chapter 4 Database development of medicinal biomolecules: Kinetic database of biomolecular interactions

4.1. Introduction to biomolecular interactions and their kinetics

Biomolecular interactions, via individual and network actions, play fundamental roles in biological, disease, and therapeutic processes (Lengeler 2000; Downward 2001; Legrain, Wojcik et al. 2001; Kitano 2007). Extensive experimental and computational studies have significantly advanced our understanding of the characteristics, organization, evolution and complexity of biomolecular interaction networks in biological systems (Drees, Sundin et al. 2001; Gavin, Bosche et al. 2002; Qian, Lin et al. 2003; Beyer, Bandyopadhyay et al. 2007), and enabled the generation of genome-scale protein-protein interactions and the development prediction tools (Dandekar, Snel et al. 1998; Pellegrini, Marcotte et al. 1999; Drees, Sundin et al. 2001; Gavin, Bosche et al. 2002; Phizicky, Bastiaens et al. 2003; Lo, Cai et al. 2005) .

Many databases have been developed for providing information about biomolecular interactions (e.g. MIPS(Mewes, Frishman et al. 2002), DIP (Salwinski, Miller et al. 2004), BIND (Alfarano, Andrade et al. 2005) , Biocyc (Karp, Ouzounis et al. 2005), MINT (Zanzoni, Montecchi-Palazzi et al. 2002), Biomodels (Le Novere, Bornstein et al. 2006), STRING (von Mering, Jensen et al. 2007), and IntAct (Kerrien, Alam-Faruque et al. 2007)), and biological networks and pathways (KEGG (Okuda, Yamada et al. 2008), BioGRID (Breitkreutz, Stark et al. 2008), NetworKIN (Linding, Jensen et al. 2008), STITCH (Kuhn, von Mering et al. 2008), DOMINE (Raghavachari, Tasneem et al. 2008), CellCircuits (Mak, Daly et al. 2007), Reactome (Joshi-Tope, Gillespie et al. 2005) and enzyme reactions (Goto, Okuno et al. 2002)).

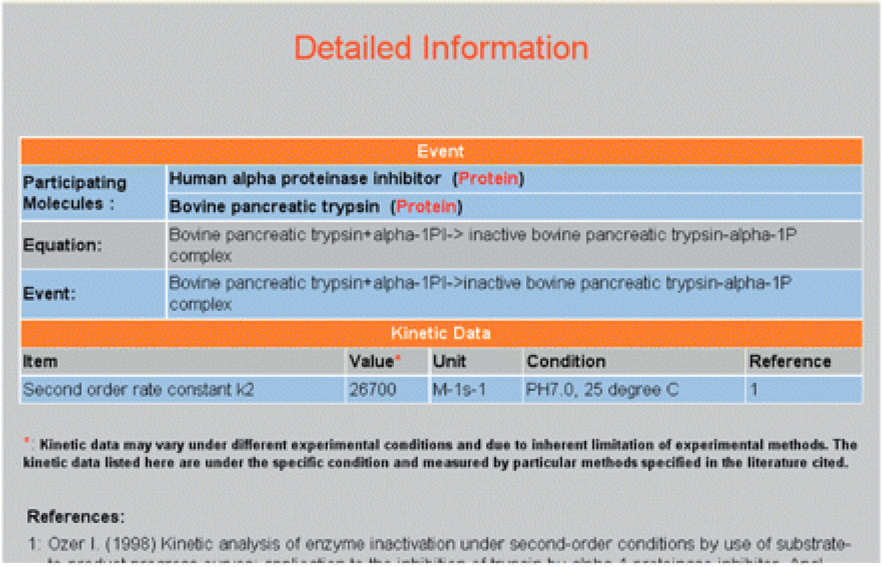
In view that quantitative as well as mechanistic understanding of biomolecular interactions is important for exploration and engineering of biological networks and for the development of novel therapeutics to combat diseases (Fabrizi, Bunnapradist et al. 2003; Zhou, Chan et al. 2004), kinetic data of biomolecular interactions have been provided in some databases. For instance, BRENDA (Schomburg, Chang et al. 2002) and SABIO-RK (Rojas, Golebiewski et al. 2007) provide kinetic constants of enzymatic activities, DOQCS contains kinetic parameters of simulation models of cellular signaling derived from experimental and other sources (Sivakumaran, Hariharaputran et al. 2003). To complement these databases for providing the kinetic data not yet covered by other databases, Kinetic Data of Bio-molecular Interactions database (KDBI) (Ji, Chen et al. 2003) have been developed to provide experimentally measured kinetic data for protein-protein, protein-nucleic acid, and protein-small molecule interactions aimed at facilitating mechanistic investigation, quantitative study and simulation of cellular processes and events (Fussenegger, Bailey et al. 2000; Haugh, Wells et al. 2000; Sahm, Eggeling et al. 2000; Schoeberl, Eichler-Jonsson et al. 2002; Schomburg, Chang et al. 2002; Sivakumaran, Hariharaputran et al. 2003; van den Broek, Noom et al. 2005; Rojas, Golebiewski et al. 2007). Kinetic data in KDBI have been manually collected from literatures, a substantial percentage of which are not yet available in other databases (e.g. some protein-protein interactions in thrombin, translation initiation, DNA repair, and ion transport pathways, and individual protein-nucleic acid interactions).

In the updated KDBI(Kumar, Han et al. 2009), apart from 2.3 fold increase of experimental kinetic data, four new features are added. The first is the access of KDBI entries via the list of nucleic acid and pathway names. The second is the inclusion of literature-reported kinetic parameter sets of 63 pathway simulation models (Fussenegger, Bailey et al. 2000; Haugh, Wells et al. 2000; Sahm, Eggeling et al. 2000; Schoeberl, Eichler-Jonsson et al. 2002; Altan-Bonnet and Germain 2005; Sasagawa, Ozaki et al. 2005; van den Broek, Noom et al. 2005; Birtwistle, Hatakeyama et al. 2007; Suresh, Babar et al. 2008; Ung, Li et al. 2008) for facilitating the applications, assessments, and further development of these pathway models. The third is the facility for collectively accessing the available kinetic data of multi-step processes (e.g. metabolism, pathway segments) collected in KDBI. The fourth is the availability of SBML (Bornstein, Keating et al. 2008) files for all records of the kinetic parameter sets of pathway simulation models for facilitating the use of the relevant data in such software tools as Celldesigner (Funahashi, Matsuoka et al. 2008), Copasi (Hoops, Sahle et al. 2006), cPath (Cerami, Bader et al. 2006), PaVESy (Ludemann, Weicht et al. 2004), and SBMLeditor (Nicolas, Donizelli et al. 2007)

1.2 Database content and access

4.2.1 Experimental kinetic data and access

Additional sets of the experimentally determined kinetic data of biomolecular interactions were collected from published literatures. Compared to the last version of KDBI, the number of entries in the updated KDBI is increased by 2.3 fold to 19263, which include 2635 protein-protein, 1711 protein-nucleic acid, 11873 protein-small molecule, and 1995 nucleic acid-small molecule interactions. Each entry provides detailed description about binding or reaction event, participating molecules, binding or reaction equation, kinetic data, and related references. As shown in **Figure 26-28**, kinetic data for protein-protein, small molecule-nucleic acid and protein-small molecule interactions is provided in terms of one or a combination of kinetic quantities as given in the literature of a particular event. These quantities include association/dissociation rate constant, on/off rate constant, first/second/third/… order rate constant, catalytic rate constant, equilibrium association/dissociation constant, inhibition constant, and binding affinity constant, IC50, etc. and experimental conditions (ph value and temperature).



Detailed Information

Event	
Participating Molecules :	Human alpha proteinase inhibitor (Protein)
	Bovine pancreatic trypsin (Protein)
Equation:	Bovine pancreatic trypsin+alpha-1PI-> inactive bovine pancreatic trypsin-alpha-1P complex
Event:	Bovine pancreatic trypsin+alpha-1PI->inactive bovine pancreatic trypsin-alpha-1P complex

Kinetic Data

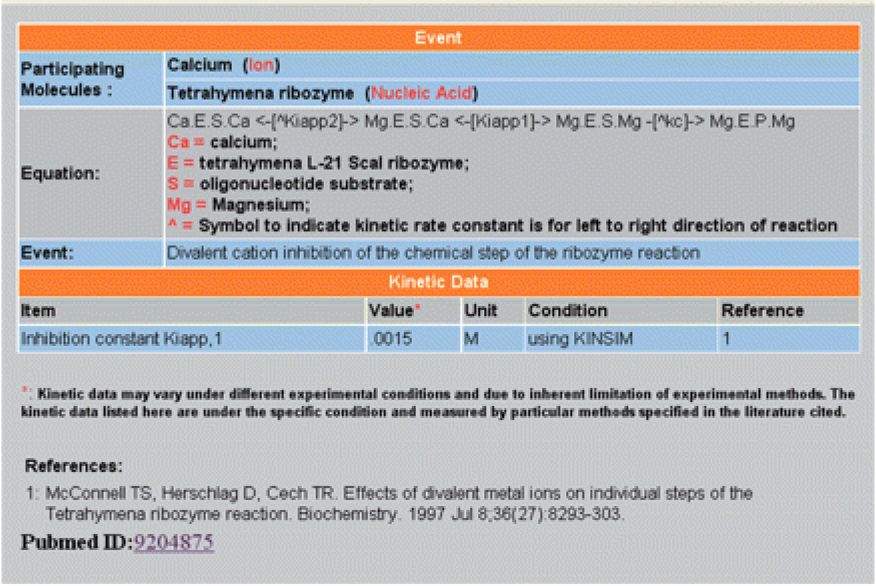
Item	Value*	Unit	Condition	Reference
Second order rate constant k2	26700	M-1s-1	PH7.0, 25 degree C	1

*: Kinetic data may vary under different experimental conditions and due to inherent limitation of experimental methods. The kinetic data listed here are under the specific condition and measured by particular methods specified in the literature cited.

References:

1: Ozer I. (1998) Kinetic analysis of enzyme inactivation under second-order conditions by use of substrate-

Figure 26: Experimental kinetic data page showing protein–protein interaction. This page provides kinetic data and reaction equation (while available) as well as the name of participating molecules and description of event.



Event	
Participating Molecules :	Calcium (Ion)
	Tetrahymena ribozyme (Nucleic Acid)

Equation:	Ca.E.S.Ca <-[^Kiapp2]-> Mg.E.S.Ca <-[Kiapp1]-> Mg.E.S.Mg -[^kc]-> Mg.E.P.Mg
	Ca = calcium;
	E = tetrahymena L-21 ScaI ribozyme;
	S = oligonucleotide substrate;
	Mg = Magnesium;
	^ = Symbol to indicate kinetic rate constant is for left to right direction of reaction

Event:	Divalent cation inhibition of the chemical step of the ribozyme reaction

Kinetic Data

Item	Value*	Unit	Condition	Reference
Inhibition constant Kiapp,1	.0015	M	using KINSIM	1

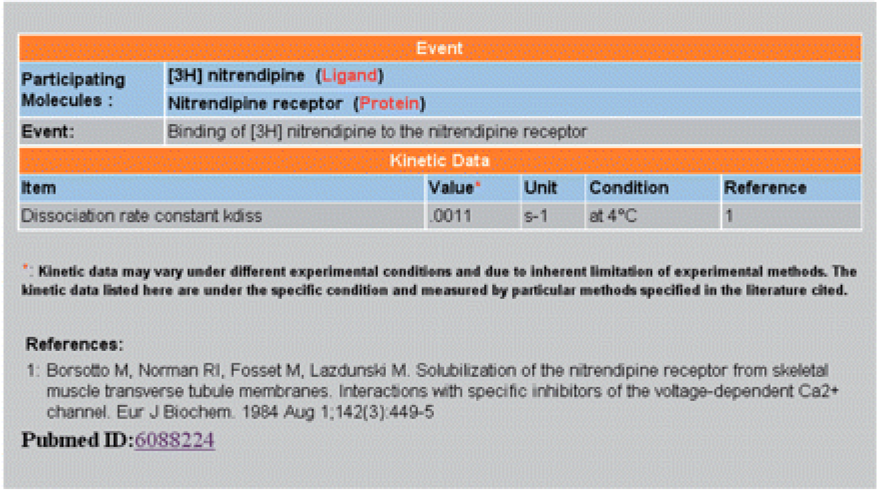
*: Kinetic data may vary under different experimental conditions and due to inherent limitation of experimental methods. The kinetic data listed here are under the specific condition and measured by particular methods specified in the literature cited.

References:

1: McConnell TS, Herschlag D, Cech TR. Effects of divalent metal ions on individual steps of the Tetrahymena ribozyme reaction. Biochemistry. 1997 Jul 8;36(27):8293-303.

Pubmed ID:9204875

Figure 27: Experimental kinetic data page showing small molecule–nucleic acid interaction. This page provides kinetic data and reaction equation (while available) as well as the name of participating molecules and description of event.



Event	
Participating Molecules :	[3H] nitrendipine (Ligand) Nitrendipine receptor (Protein)
Event:	Binding of [3H] nitrendipine to the nitrendipine receptor

Kinetic Data

Item	Value*	Unit	Condition	Reference
Dissociation rate constant kdiss	.0011	s-1	at 4°C	1

*: Kinetic data may vary under different experimental conditions and due to inherent limitation of experimental methods. The kinetic data listed here are under the specific condition and measured by particular methods specified in the literature cited.

References:

1: Borsotto M, Norman RI, Fosset M, Lazdunski M. Solubilization of the nitrendipine receptor from skeletal muscle transverse tubule membranes. Interactions with specific inhibitors of the voltage-dependent Ca2+ channel. Eur J Biochem. 1984 Aug 1;142(3):449-5

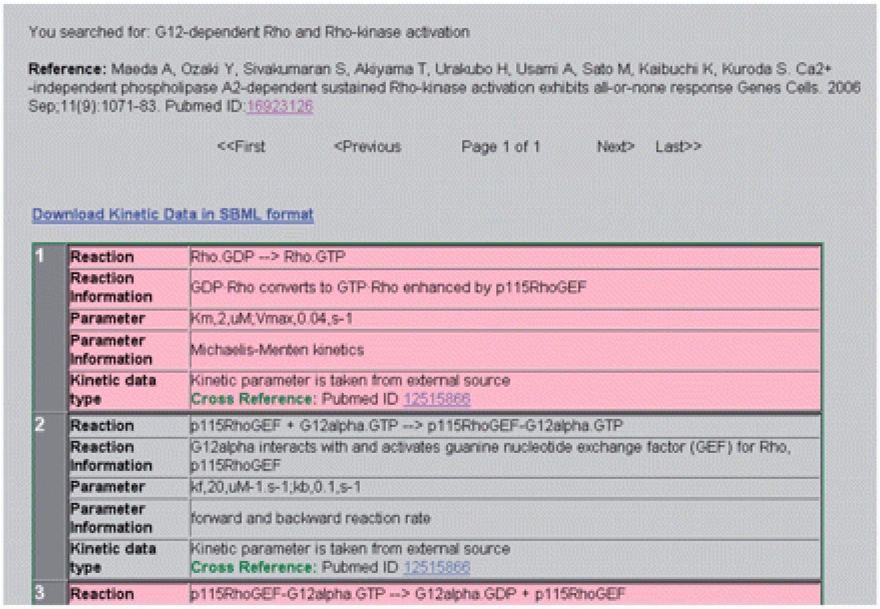
Pubmed ID:6088224

Figure 28: Experimental kinetic data page showing protein–small molecule interaction. This page provides kinetic data and reaction equation (while available) as well as the name of participating molecules and description of event.

These data can be accessed via input of names of molecules and bio-events (association, dissociation, complex formation, electron transfer, inhibition etc), and via selection of pathway and protein name from the pathway list and protein list fields in KDBI webpage. The kinetic data of an event is searchable by several methods. One method is via the name of participating molecules (protein, nucleic acid, small peptide, ligand or ion) or pathway involved in an event. In some events described in the literature, a participating entity is an unidentified molecule located in the membrane of a cell or on the surface of a virus. In these entries, only the name of the cell or virus is given. An entry can also be searched through a Swiss-Prot AC number for a protein or the CAS number for a small molecule ligand. Moreover, keyword-based text search is also supported. To facilitate convenient access of relevant data, partial lists of proteins and nucleic acid are provided. Searches involving combination of these methods or selection fields are also supported.

4.2.2 Parameter sets of pathway simulation models

As part of the efforts for facilitating the understanding and quantitative analysis of complex biological processes and network responses, mathematical simulation models of various pathways have been developed and extensively used for studying and quantitative understanding of signaling dynamics (Fussenegger, Bailey et al. 2000; Haugh, Wells et al. 2000; Sahm, Eggeling et al. 2000; Schoeberl, Eichler-Jonsson et al. 2002; van den Broek, Noom et al. 2005), signal specific sensing (Sasagawa, Ozaki et al. 2005) and discrimination (Altan-Bonnet and Germain 2005), feedback regulations and crosstalks (Suresh, Babar et al. 2008; Ung, Li et al. 2008), and receptor cross-activation (Birtwistle, Hatakeyama et al. 2007) and internalization (Ung, Li et al. 2008). These mathematical models typically use ordinary differential equations (ODEs) to describe the temporal dynamic behavior of molecular species in the pathway. The kinetic rate constants of protein–protein, protein-small molecule, protein-nucleic acid, and other interactions (e.g. binding association rate K_a, binding dissociation rate K_d, reaction rate K, reaction turnover rate K_{cat}, Michaelis–Menten constant K_m) are needed to establish these ODEs, which have been primarily generated by combinations of experimental data, computed theoretical values, and empirically fitted values computational (Schoeberl, Eichler-Jonsson et al. 2002; Altan-Bonnet and Germain 2005; Sasagawa, Ozaki et al. 2005; Birtwistle, Hatakeyama et al. 2007; Suresh, Babar et al. 2008; Ung, Li et al. 2008) . To facilitate further applications, developments, and assessments of the published pathway models, we collected and included in KDBI the parameter sets of 63 published ODE-based models, which can be accessed from the pathway list in the “Pathway Simulation Parameters” field in KDBI webpage. Moreover, we added kinetic data type to every entry to clearly distinguish its original source (experimental or simulation model). In particular, for the kinetic data of a simulation model that have been obtained from other publications, cross reference to the original source is provided. A typical search result is shown in **Figure 29**



You searched for: G12-dependent Rho and Rho-kinase activation

Reference: Maeda A, Ozaki Y, Sivakumaran S, Akiyama T, Urakubo H, Usami A, Sato M, Kaibuchi K, Kuroda S. Ca2+ -independent phospholipase A2-dependent sustained Rho-kinase activation exhibits all-or-none response Genes Cells. 2006 Sep;11(9):1071-83. Pubmed ID:16923126

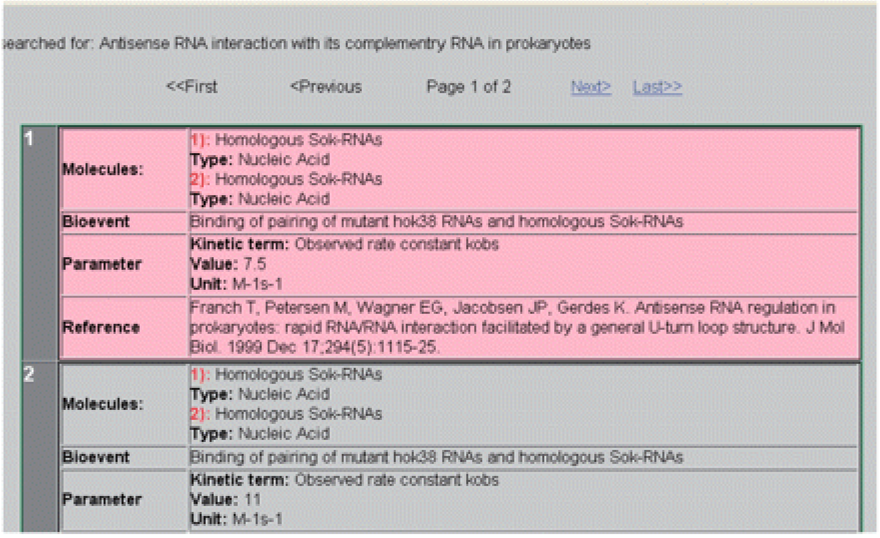
Download Kinetic Data in SBML format

1	**Reaction**	Rho.GDP --> Rho.GTP
	Reaction information	GDP·Rho converts to GTP·Rho enhanced by p115RhoGEF
	Parameter	Km,2,uM;Vmax,0.04,s-1
	Parameter information	Michaelis-Menten kinetics
	Kinetic data type	Kinetic parameter is taken from external source **Cross Reference:** Pubmed ID 12515866
2	**Reaction**	p115RhoGEF + G12alpha.GTP --> p115RhoGEF-G12alpha.GTP
	Reaction information	G12alpha interacts with and activates guanine nucleotide exchange factor (GEF) for Rho, p115RhoGEF
	Parameter	kf,20,uM-1.s-1;kb,0.1,s-1
	Parameter information	forward and backward reaction rate
	Kinetic data type	Kinetic parameter is taken from external source **Cross Reference:** Pubmed ID 12515866
3	**Reaction**	p115RhoGEF-G12alpha.GTP --> G12alpha.GDP + p115RhoGEF

Figure 29: Pathway parameter set page. This page provides kinetic data and reaction equation (while available) as well as the name of participating molecules and description of event.

4.2.3 Kinetic data for multi-step processes

Some published studies provide information about the experimental kinetic data for multiple components of multi-step processes (Hoshino, Kawata et al. 1996; Franch, Petersen et al. 1999; Korneeva, Lamphear et al. 2001). Examples of these processes include RNA binding activity to translation initiation factors eIF4G, 70-kDa Heat Shock Protein polymerization, and control of platelet function by cyclic AMP, GroEL interaction with conformational states of horse cytochrome c, intermolecular catalysis by hairpin ribozymes, antisense RNA interaction with its complimentary RNA, nucleotide binding to actin. To facilitate the development of pathway simulation models based on these building blocks, we provided direct access to the collection of the kinetic data for each of these processes, which can be accessed via a separate search field “Multi-step processes” in KDBI webpage. A typical search result is shown in **Figure 30**.



earched for: Antisense RNA interaction with its complementry RNA in prokaryotes

1	**Molecules:**	1): Homologous Sok-RNAs **Type:** Nucleic Acid 2): Homologous Sok-RNAs **Type:** Nucleic Acid
	Bioevent	Binding of pairing of mutant hok38 RNAs and homologous Sok-RNAs
	Parameter	**Kinetic term:** Observed rate constant kobs **Value:** 7.5 **Unit:** M-1s-1
	Reference	Franch T, Petersen M, Wagner EG, Jacobsen JP, Gerdes K. Antisense RNA regulation in prokaryotes: rapid RNA/RNA interaction facilitated by a general U-turn loop structure. J Mol Biol. 1999 Dec 17;294(5):1115-25.
2	**Molecules:**	1): Homologous Sok-RNAs **Type:** Nucleic Acid 2): Homologous Sok-RNAs **Type:** Nucleic Acid
	Bioevent	Binding of pairing of mutant hok38 RNAs and homologous Sok-RNAs
	Parameter	**Kinetic term:** Observed rate constant kobs **Value:** 11 **Unit:** M-1s-1

Figure 30: Multi-process kinetic data page. This page provides kinetic data and reaction equation (while available) as well as the name of participating molecules and description of event.

4.3 Kinetic data files in SBML format

Systems Biology Markup Language (SBML) has been developed as a free, open, XML-based format for representing biochemical reaction networks, and it is a software-independent language for describing models common to computational biology research, including cell signaling pathways, metabolic pathways, gene regulation, and others (Hucka, Finney et al. 2003). Many pathway simulation and analysis software tools have built-in SBML compatibility features to allow the input, manipulation, simulation and analysis of different pathway models and parameters (Hucka, Finney et al. 2003; Alves, Antunes et al. 2006; Deckard, Bergmann et al. 2006; Zi and Klipp 2006; Schmidt, Drews et al. 2007; Bornstein, Keating et al. 2008). To facilitate the input of the pathway parameter sets into these software tools, we created the SBML file for the parameter sets of all 63 pathway simulation models included in KDBI, which can be downloaded via the link provided on the top of the page that displays the relevant kinetic data.

The SBML files follow the norm of SBML API version 2.3.3. The code was written in JAVA programming language with the help of Java library of SBML API to generate SBML file from flat files. Although, it is created for advanced user, a viewer (SBMLBIDDviewer) is also written to visualize SBML file. This viewer is freely available to download from KDBI website.

4.4 Remarks

The updated version of KDBI is intended to be a more useful resource for convenient access of available biomolecular kinetic data to complement other biomolecular interaction and pathway databases in facilitating quantitative studies of biomolecular interactions and networks. New technologies have been developed in employing surface plasmon resonance technology for deriving real-time dynamics and kinetic data (Huber and Mueller 2006) , and in using protein microarrays (Yu, Xu et al. 2006) and solution NMR spectroscopy (Pellecchia 2005) for monitoring and characterizing biomolecular interactions. Moreover, new experimental designs of the well established technologies such as isothermal titration calorimetry allow the measurement and estimate of previously inaccessible kinetic parameters (Buurma and Haq 2007).   Resources for collecting and accessing the increasing amount of kinetic data can better serve the need for mechanistic investigation, quantitative study and simulation of biological processes and events.

Chapter 5 Machine Learning Classification: Prediction of genotoxicity

5.1 Introduction of genotoxicity and drug discovery

Drug discovery and approval processes involve the evaluation of adverse drug reactions (ADRs), one of which is genotoxicity. The molecular mechanisms that are a part of genotoxicity include DNA intercalation that takes place due to an aromatic ring of a drug, DNA methylation, DNA adduct formation and strand breakage as well as an unscheduled DNA synthesis (Bolzan and Bianchi 2002).

The significance of genotoxicity testing lies in the identification of potentially hazardous drug candidates. The results generated from genetic toxicology tests, in combination with other toxicity data are used as the basis for approval of clinical trials of drug candidates (Custer and Sweder 2008). The importance of the optimization of molecules during early drug development for efficacy and with regard to their pharmacokinetic and toxicological properties has gained wide recognition. A balance of target potency, selectivity, favorable ADME (absorption distribution metabolism excretion) and (pre)clinical safety properties that will ultimately result in the selection and clinical development of a potential new drug has been suggested. Phase I clinical trials for a compound involves years of rigorous preclinical testing and yet has only an 8% chance of reaching the market. Toxicity results in the dropping of 20% of such molecules during late development stages. Therefore, the implementation of toxicity testing as early as possible in the drug development process is of prime significance (Custer and Sweder 2008). Huge amounts of compounds necessary for *in vivo* studies, dearth of reliable high-throughput *in vitro* assays, and the inability of *in vitro* and animal models to correctly predict some human toxicity are the main reasons that prevent pharmaceutical companies from conducting earlier screening for toxicity.

Among different toxicity tests, genotoxicity test has been of prime importance. According to ICH guideline Genotoxicity tests can be defined as *in vitro* and *in vivo* tests intended to detect compounds which make genetic damage directly or indirectly by different mechanisms. These tests should be able to detect damage to DNA and its fixation. The processes like fixation of DNA damage by gene mutations, recombination, extensive chromosomal damage, and numerical chromosome changes are generally measured to be important in the multi-step process of malignancy and for heritable effects. There are different genotoxicity test types described in **Table 4**:

Table 4: Genotoxicity testing types

In vitro	The Salmonella/E. coli Mutagenicity Test or Ames Test
	Mouse Lymphoma
	Chinese Hamster Ovary Cell cytogentics (1)Chromosomal Aberration (CA) test (2) sister chromatid exchanges (SCE) test
	in vitro micronucleus (MN)
In vivo	Drosophila melanogaster (1)sex-linked recessive lethal (SLRL) mutations (2) chromosomal reciprocal translocations (RT)
	Micronuclues

Also ICH defines Standard battery of genotoxicity tests: (i) a test for gene mutation in bacteria (ii) an *in vitro* test with cytogenetic evaluation of chromosomal damage with mammalian cells or an in vitro mouse lymphoma tk assay (iii) an *in vivo* test for chromosomal damage using rodent hematopoietic cells. The compounds which give negative results in all of this 3-test battery will typically be safe and will not have genotoxic activity. Compounds which give positive results in the standard test battery may, depending on their therapeutic use, require extensive tests.

The genotoxicity tests can be utilized to decide about compounds potential to be human carcinogens and/or mutagens. There is evidence that human being exposed to compounds which is found positive in genotoxicity test, also had cancer and the vast majority of these are detected by both the *Salmonella* assay and rodent micronucleus tests. These evidences suggested strong correlation between genotoxicity and carcinogenesis, but an analogous connection has not been established for heritable diseases. Thus, genotoxicity tests have been used primarily for the prediction of carcinogenicity(Kirkland, Aardema et al. 2005).

In recent times, genotoxicity testing methods has been argued upon very low specificity of all mammalian cell tests. In contrast, the specificity of the Ames test has been found reasonable. The extremely low specificity reveals deficiencies in the current prediction from and understanding of such *in vitro* results for the *in vivo* situation. (Kirkland, Aardema et al. 2005)

In vivo genotoxicity tests play a pivotal role in genotoxicity testing batteries. They are used both to determine if potential genotoxicity observed *in vitro* is realized *in vivo* and to detect any genotoxic carcinogens that are poorly detected *in vitro*. It is recognized that individual in vivo genotoxicity tests have limited sensitivity but good specificity. Thus, a positive result from the established in vivo assays is taken as strong evidence for genotoxic carcinogenicity of the compound tested.(Tweats, Blakey et al. 2007)

One of the main objectives of short term in vitro studies is to replace long term (2 year) animal assays thus reducing the animal sacrifice and also the time. The failure of in vitro tests in achieving this objective must be addressed while developing the prediction model. However, it would be useful to observe the developed model performance with and without addressing this major issue of extremely low specificity. The prediction model based on just in vitro positive in training dataset will produce many false positive upon scanning chemical databases which would block many potential compounds to be developed as drugs. This model will produce lots of genotoxicity hits upon virtual screening.

These problems can be addressed through the development of computational or *in silico* toxicity prediction tools, either structure-based or which involve the application of modeling techniques on human data. These serve as main approaches to extract potentially toxic effects in humans even before the physical availability of compounds. *In silico* techniques like knowledge-based expert systems (quantitative) structure activity relationship tools and modeling approaches help to significantly reduce drug development costs in predicting adverse drug reactions in preclinical studies. (Muster, Breidenbach et al. 2008). Over the years, computational toxicology prediction systems have tremendously increased their predictive power but have not yet achieved a major breakthrough due to lack of sufficiently large datasets. The development of such systems take coordinated efforts since they are dependent on the gold standard, low throughput data but once set up, could reduce investment as well as the use of animals.

Primarily, the significance of computational tools arises from their applicability during the early stages of development. At the stage when chemical series are initially screened concerning undesired activities, information on possible adverse properties should be obtained through the use of globally valid computational tools. An excellent correlation with ‘wet-lab’ data that is, high sensitivity, as well as high specificity, an easy to use and easy to interpret *in silico* model are key requirements for its usefulness. As a non-expert tool, the need for it to be available to the medicinal chemist via computer networks has been acknowledged.

A variety of computational tools for a quick and efficient prediction of drug genotoxic potential have been developed (Cash 2001; He, Jurs et al. 2003; Mattioni, Kauffman et al. 2003; Li, Ung et al. 2005).

In search for a new approach for the prediction of genotoxicity, machine learning methods have been developed, without compromising on the structures or types of molecules. Methods such as these classify molecules into GT+ and non-genotoxic (GT–) agents, based on their general structural and physicochemical properties, without considering their structural and chemical types. As such, these methods are expected to be applicable to a diverse set of molecules. Nevertheless, the quality of the molecular descriptors influences the performance of such methods, in addition to training and testing data, and the efficiency of machine learning algorithms.

Thus far, machine learning methods, the likes of linear discriminate analysis (LDA), *k*-nearest neighbor classification (*k*-NN), support vector machines (SVM), and probabilistic neural networks (PNNs), have been in use and have achieved a prediction accuracy of up to 73.8% for GT+ and 92.8% for GŦ agents, respectively (Li, Ung et al. 2005). However, these methods have been developed and tested by using no more than 860 known GT+ and GT– agents.

A more diverse set of molecules would significantly enhance the levels of accuracy. A training set comprising of an even more diverse set of GT+ agents would further heighten accuracy levels and its prediction capability for true independent dataset. Support Vector Machines (SVM) and k-NN are among those machine learning methods that have shown great potential in these types of studies. The importance of SVM is evident in studies that have been carried out for the prediction of antibiotic resistance proteins (Zhang, Lin et al. 2008), mitochondrial toxicity (Zhang, Chen et al. 2008), blood-brain barrier permeability (Kortagere, Chekmarev et al. 2008), torsade-causing potential of drugs (Yap, Cai et al. 2004), P-glycoprotein substrates (Xue, Yap et al. 2004).

Our work has involved the evaluation and use of several Machine Learning Methods (MLMs). These include SVM, PNN, *k*-NN, and Decision Tree (DT). Support Vector Machines (SVM) has been used to a great extent due to its applicability to a variety of classification problems. A huge improvement on earlier studies conducted has been in the number of compounds used, consideration of issue of extremely low specificity of genotoxicity tests, comparison of different machine learning methods by using the exactly same data set and descriptors. In addition to the development of genotoxicity prediction models by MLMs, these models have been further utilized for the Virtual Screening (VS) of chemical libraries.

Virtual Screening techniques might be categorized into two broad types: ligand-based and structure-based. In the ligand-based technique, a model of a receptor can be built, based on a set of structurally diverse ligands that bind to the receptor. The structure-based technique, on the other hand, involves the docking of candidate ligands into a protein target. Support Vector Machines (SVM) has been utilized as ligand-based VS (LBVS) tools to complement or to be used in combination with structure-based VS (SBVS) and other LBVS tools. In genotoxicity, however, one might rule out the use of SBVS since the targets are not well-defined. In the current study, SVM has been implemented as the ligand-based VS (LBVS).

By classifying active compounds based on the differentiating physicochemical profiles between active and inactive compounds rather than structural similarity to active compounds per se, SVM acquires specific importance. The knowledge of target structure and activity-related molecular descriptors and the computation of binding affinity and solvation effects are not required. SVM's fast speed results in an efficient search of vast chemical space. Some of these advantages have been realized through good VS performance in screening large compound libraries. The performance of SVM is significantly influenced by the levels of the training active and inactive compounds in representing the physicochemical profiles of the remaining compounds in the chemical space.

5.2 Genotoxicity data set

Collection of genotoxicity compounds

Genotoxicity data were collected from different sources such as National Toxicology Program, Bursi Mutagenicity dataset (Kazius, McGuire et al. 2005), NLM leased data, EAFUS, Helma CPDB Mutagenicity Subset(Helma, Cramer et al. 2004), GRAS and from a number of publications. **Table 5** and **Table 6** show different sources for genotoxicity positive and genotoxicity negative data collection respectively.

Table 5: Genotoxicity Positive Data Set

Source	Type	Number of compounds	Compounds considered (3d structures and unique)
Mutation Research 584 (2005) 1–256)	rodent carcinogenic and positive in at least one ICH standard battery of tests	433	426
	Rodent carcinogenic and positive in all in vitro tests	20	20
NLM leased data	genotoxicity positive	1989	1989
	in vivo positive	442	442
	positive (in any test) by more than one references	786	786
	positive (in any test) by more than one references and negative (in any test) by 1 or 0 references	611	611
Mutagenesis vol. 22 no. 6 pp. 409–416, 2007	Green screen assay and Ames positive	42	42
CPDBAS (Carcinogenic potency database)	Ames positive	394	394
Kazius, J.; McGuire, R.; Bursi, R. Derivation and Validation of Toxicophores for Mutagenicity Prediction. J. Med. Chem. 2005, 48(1), 312-320 -	Ames positive	2401	2401
Mutation Research 653 (2008) 99–108 (Recommended lists	Independent data set	19	19

of genotoxic and non-genotoxic chemicals for assessment of the performance of new or improved genotoxicity tests: A follow-up to an ECVAM workshop)			
Other recent journals of 2008 and 2009 (newly synthesized and found genotoxic by tests)	Independent data set	19	19

Table 6: Genotoxicity negative data set

Source	Type	Number of compounds	Compounds (3d structures and unique)
Everything added to food (EAFUS)		2328	2328
Drugbank fda approved drugs		1293	1293
GRAS clean		369	177
Kazius, J.; McGuire, R.; Bursi, R. Derivation and Validation of Toxicophores for Mutagenicity Prediction. J. Med. Chem. 2005, 48(1), 312-320 -	Ames negative	1926	1926
Mutation Research 584 (2005) 1–256	rodent non-carcinogenic	177	177
An update on the genotoxicity and carcinogenicity of marketed pharmaceuticals with reference to in silico predictivity.*Environmental and molecular mutagenesis* (2009)	Approved PDR drugs	545	540
Mutation Research 653 (2008) 99–108 (Recommended lists of genotoxic and non-genotoxic chemicals for assessment of the performance of new or improved genotoxicity tests: A follow-up to an ECVAM workshop)	Independent data set	23	23
Clinical Trials (Phase 1 , 2, 3)	Independent data set	2387	2039

5.3 Methods

The methods to do genotoxicity studies were designed to handle issues like diversity and extremely low specificities. The genotoxicity study was performed in three instances.

1. Study with 100 descriptor and smaller dataset :

Positive (Total 2776): Positive in any genotoxicity test

Negative (Total 4116): Approved drugs + gras

Independent: Part of positive and negative (no true independent dataset)

Compounds representation: 100 descriptors

Objective: To assess various machine learning method performances.

2. High diversity high noise (HDHN)(positive in any assay) model :

Positive (Total 4763): Positive in any genotoxicity test

Negative (Total 8232): Approved drugs + EAFUS + non-mutagenic + gras

Independent positive (Total 38): from recent journals

Independent negative (Total 2008): Clinical trial drugs

Compounds representation: 522 descriptors

Objective: To obtain broadly applicable SVM model with little compromise on specificity

3. Low diversity low noise (LDLN) (positive in Ames or in vivo) model:

Positive (Total 3321): Ames + in vivo

Negative (Total 8232): Approved drugs + EAFUS + non-mutagenic + gras

Independent positive (Total 38): from recent journals

Independent negative (Total 2008): Clinical trial drugs

Compounds representation: 522 descriptors

Objective: To address the low specificity issues of different in vitro genotoxicity tests

The detail method of machine learning algorithms employed and cross-validation is provided in Chapter 2.

5.4 Results and discussion

This result and discussion is divided into three subsection based on dataset and descriptors used in different run. The first part is smaller dataset of all the collected dataset (about 50%) in which 100 descriptors were used. The second part is for HDHN in which entire 522 molecular descriptors (see appendix) were used. The third part is for LDLN in which 522 molecular descriptors were used.

5.4.1 Results of the study with 100 descriptors and smaller dataset

5.4.1.1 Comparative study of SVM with other machine learning methods

SVM has been used by the application of LibSVM, in addition to other machine learning methods like kNN, decision trees, feedforward backpropagation neural network by using Weka software. A 5-fold cross-validation was performed for each of the MLMs used in this study while, the dataset remained the same for the purpose of efficient comparison. The prediction accuracy of the 5-fold cross-validation by SVM is shown in Table 2. The SVM parameter sigma was scanned from 0.1 to 5 with an increase of 0.1 at each step. In Table 2, results are presented in two ways: 1. the best prediction accuracies while scanning SVM parameter sigma for each of the folds. 2. The average of prediction accuracy for all the models with different sigma values in each case. These results have been presented to show the average, maximum, minimum and standard deviation for all the folds. High prediction accuracies were observed for sigma values 1.3 to 1.8. The best prediction accuracy (positive accuracy corresponding to 85.77, negative accuracy equal to 91.62 and overall prediction accuracy corresponding to 89.26) was achieved with a sigma value 1.8.

In **Table 7** and **Table 8**, the results of the application of various other MLMs (using 5-fold cross-validation) for determining genotoxicity prediction have been shown. As is indicative from the figures generated, apart from the high efficiency levels as displayed by the use of KNN and Multilayer perceptron methods, Random Forest, one among many Decision Tree (DT) methods has shown great potential

Table 7: SVM Five-fold cross validation on genotoxicity by using 100 descriptors

SVM 5-fold cross validation						
	Accuracy of models (Best models of each fold)			Average accuracy		
	Positive Accuracy	Negative Accuracy	Overall Accuracy	Positive Accuracy	Negative Accuracy	Overall Accuracy
Fold1	84.14	89.43	87.3	81.72	87.88	85.39
Fold2	84.14	90.28	87.81	81.72	87.89	85.40
Fold3	83.06	91.13	87.88	81.69	87.89	85.39
Fold4	85.05	89.91	87.95	81.63	87.89	85.37
Fold5	85.77	91.62	89.26	81.58	87.86	85.33
Average	84.43	90.47	88.04	81.67	87.88	85.38
Max	85.77	91.62	89.26	81.72	87.89	85.40
Min	83.06	89.43	87.3	81.58	87.86	85.33
STDEV	1.03	0.89	0.73	0.06	0.01	0.03

Table 8: Other MLM 5-fold cross validation by using 100 descriptors

5-Fold cross validation accuracy (Average accuracy)			
	Positive Accuracy	Negative Accuracy	Overall Accuracy
IBk (KNN)	79.82	85.57	83.25
MultilayerPerceptron	79.82	85.57	83.25
RandomForest	74.20	92.05	84.86
ADTree	56.76	91.23	77.34
BFTree	70.45	88.17	81.03
DecisionStump	72.21	88.52	81.95
FT	70.69	89.11	81.69
J48	68.86	89.81	81.37
J48graft	67.79	89.37	80.68
LMT	70.00	88.99	81.34
NBTree	69.91	89.16	81.41
REPTree	69.45	89.29	81.30

5.4.1.2 Virtual Screening of MDDR and PUBCHEM database

The models that have been developed using SVM for the sigma values of 1.3 to 1.8 (the best parameter range found by 5-fold cross-validation) were used for the Virtual Screening of MDDR and Pubchem database. **Table 9** depicts the results of Virtual Screening of MDDR database. Virtual Screening evaluation/performance is shown through ‘Yield’, “Hit Rate’ and

‘Enrichment Factor’. There are 79 common compounds that have been found in the MDDR database and our genotoxicity positive data collected. The results in **Table 9** indicate those that were arrived at after removal of the 79 common compounds as against those that were arrived at after retaining the compounds. There is clear difference for the prediction of 79 actual genotoxic positive compounds by the SVM models when these actual genotoxic positive 79 compounds are included and excluded in training data set. When these actual genotoxic compounds are included in SVM models training dataset, it can predict about 80% of these compounds as geneotoxic positive. In contrast, when these actual genotoxic compounds are excluded in SVM models training dataset, it can predict only about 59% of these compounds as geneotoxic positive.

Table 10 shows the Virtual Screening of the MDDR database by the use of different Tanimoto similarity coefficients as the threshold for Tanimoto similarity searching, using fingerprints of chemical compounds.

Table 9: Virtual Screening of MDDR database

	Sigma	MDDR Total	MDDR Hits	MDDR Intersection GT+	Actual GT+ in MDDR which got Predicted as GT+ by model	Yield	Hit Rate	Enrichment Factor
VS by SVM models while **79** common compounds of MDDR and GT+ are **present** in Training dataset	1.1	168016	37450	79	63	79.75	0.00168	3.577
	1.2	168016	34387	79	62	78.48	0.00180	3.834
	1.3	168016	34076	79	59	74.68	0.00173	3.682
	1.4	168016	35019	79	57	72.15	0.00163	3.461
	1.5	168016	30978	79	57	72.15	0.00184	3.913
	1.6	168016	30309	79	59	74.68	0.00195	4.140
	1.7	168016	28722	79	58	73.42	0.00202	4.294
	1.8	168016	31206	79	54	68.35	0.00173	3.680
	1.9	168016	30681	79	56	70.89	0.00183	3.881
VS by SVM models while **79** common compounds of MDDR and GT+ are **removed** from Training dataset	1	168016	31635	79	40	50.63	0.00126	2.689
	1.1	168016	32902	79	41	51.90	0.00125	2.650
	1.2	168016	30356	79	41	51.90	0.00135	2.872
	1.3	168016	29077	79	42	53.16	0.00144	3.072
	1.4	168016	31539	79	44	55.70	0.00140	2.967
	1.5	168016	27632	79	44	55.70	0.00159	3.3866
	1.6	168016	26632	79	41	51.90	0.00154	3.274188
	1.7	168016	26013	79	47	59.49	0.00181	3.842651
	1.8	168016	28108	79	43	54.43	0.00153	3.253584
	1.9	168016	27346	79	41	51.90	0.00150	3.1887

Table 10: Tanimoto similarity with MDDR database based on fingerprint

Tanimot Similarity Coefficient as cut-off	MDDR Total	MDDR Hits	MDDR Hits unique	MDDR Intersection GT+		Yield	Hit Rate	Enrichment factor
0.7	168016	161934	38463	79	58	73.42	0.0015	3.207
0.8	168016	47251	13769	79	45	56.96	0.0033	6.951
0.9	168016	13984	4195	79	38	48.10	0.0091	19.265

5.4.1.3 Performance evaluation

An examination of the accuracy levels of machine learning methods for genotoxicity prediction of a diverse set of molecules is required to gauge whether the accuracy achieved by these methods is at a similar level as those derived by the use of a significantly smaller set of molecules. It is noted that a direct comparison with results from previous studies is inappropriate because of the differences in the data set and molecular descriptors used. However, the current study has been undertaken by using the same molecular descriptors for all the MLMs, including SVM along with the same number and same distribution of data sets in each fold of stratified 5-fold cross-validation. The positive accuracy levels for all the MLMs, especially SVM, have increased to levels (range ~83-85) unsurpassed by any previous study (range ~72-75). The negative accuracy remains unchanged at 90-92% (in comparison with previous studies).

Genotoxicity assessment of a broad ranges of molecules through the implementation of machine learning methods, particularly SVM, *k*-NN, PNN, and DT such as Random Forest and Decision Stump has thus been established through our study. The prediction accuracy of these methods is at a similar, if not superior level, as those of earlier studies that were tested by using a much smaller number of molecules. An added advantage of these methods is that they do not require knowledge about the molecular mechanism or SAR of a particular drug property. The classification speed of SVM is fast compared to other MLMs that use Weka. It has been noticed that the speed for MLMs such as *k*-NN has been exceptionally slower than the rest, whereas that of J48 and Random Forest has been fast although all of these use Weka. Virtual Screening has been done for some types of DT, the data for which can be viewed as additional supplementary data online.

5.4.2 High diversity high noise (HDHN) (positive in any assay) model prediction performance

The result in this section is presented for 5-fold cross validation accuracy, testing on independent dataset, and virtual screening on Pubchem and MDDR.

5.4.2.1 Five fold

The 5-fold cross validation accuracy is shown in **Table 11**. The negative, positive, overall , and average accuracy over different sigma values are shown in **Figure 31**, **Figure 32**, **Figure 33**, and **Figure 34** respectively.

Table 11: 5-fold cross validation for genotoxicity prediction models on more diverse dataset (positive in any assay)

SVM 5-fold cross validation						
	Accuracy of models (Best models of each fold)			Average accuracy		
	Positive Accuracy	Negative Accuracy	Overall Accuracy	Positive Accuracy	Negative Accuracy	Overall Accuracy
Fold1	77.63	87.96	84.15	73.86	86.06	81.56
Fold2	78.57	86.79	83.76	75.33	85.18	81.54
Fold3	77.63	88.08	84.22	74.62	86.13	81.88
Fold4	78.15	86.73	83.57	74.44	84.85	81.01
Fold5	77.63	86.06	82.95	74.98	85.12	81.38
Average	77.92	87.12	83.73	74.65	85.47	81.47
Max	78.57	88.08	84.22	75.33	86.13	81.88
Min	77.63	86.06	82.95	73.86	84.85	81.01
STDEV	0.43	0.87	0.51	0.56	0.59	0.32

Negative Accuracy

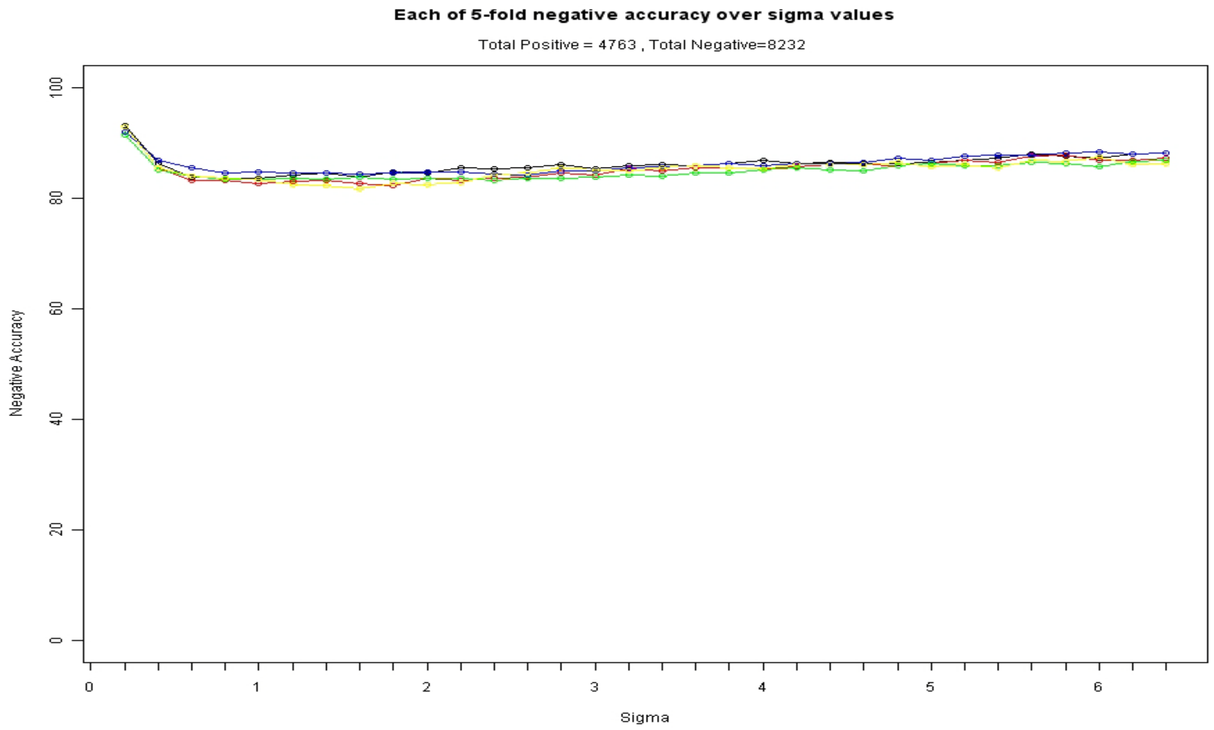


Figure 31: Fivefold negative accuracy (Genotoxicity, SVM, More diverse (positive in any assay) way). Negative accuracy (red color), positive accuracy (blue color) and overall accuracy.

Positive Accuracy

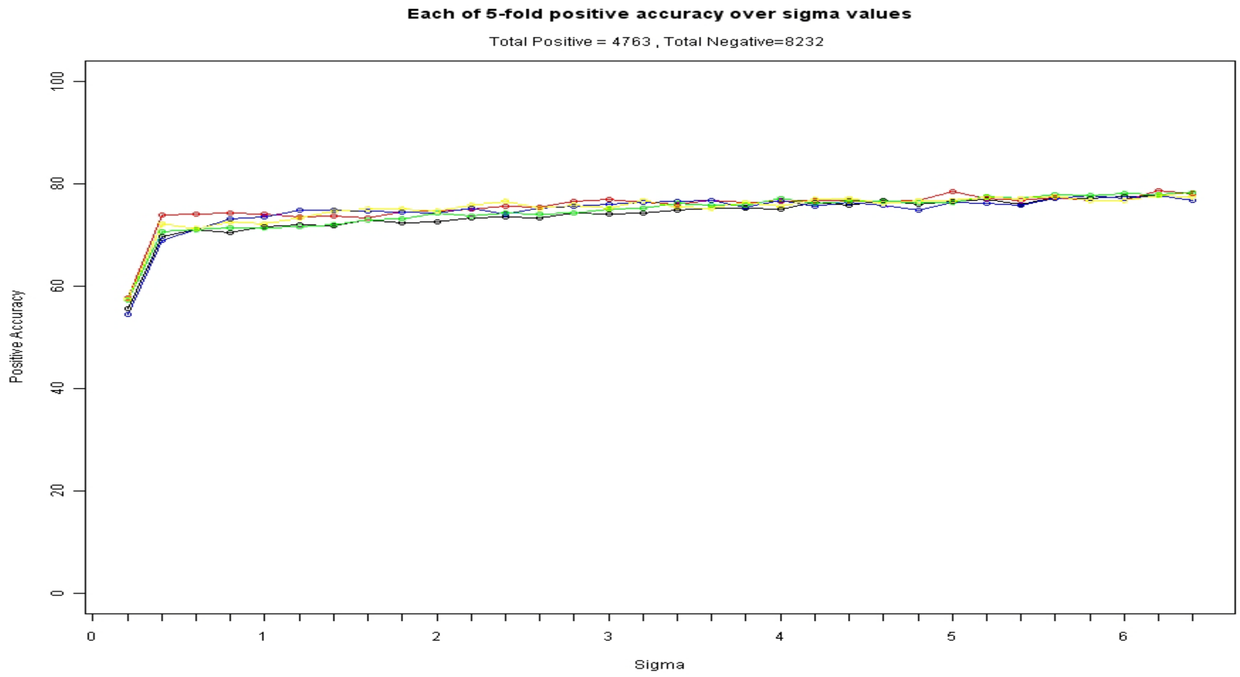


Figure 32: Fivefold positive accuracy (Genotoxicity, SVM, High diversity high noise (HDHN) (positive in any assay) model). Negative accuracy (red color), positive accuracy (blue color) and overall accuracy.

Overall Accuracy

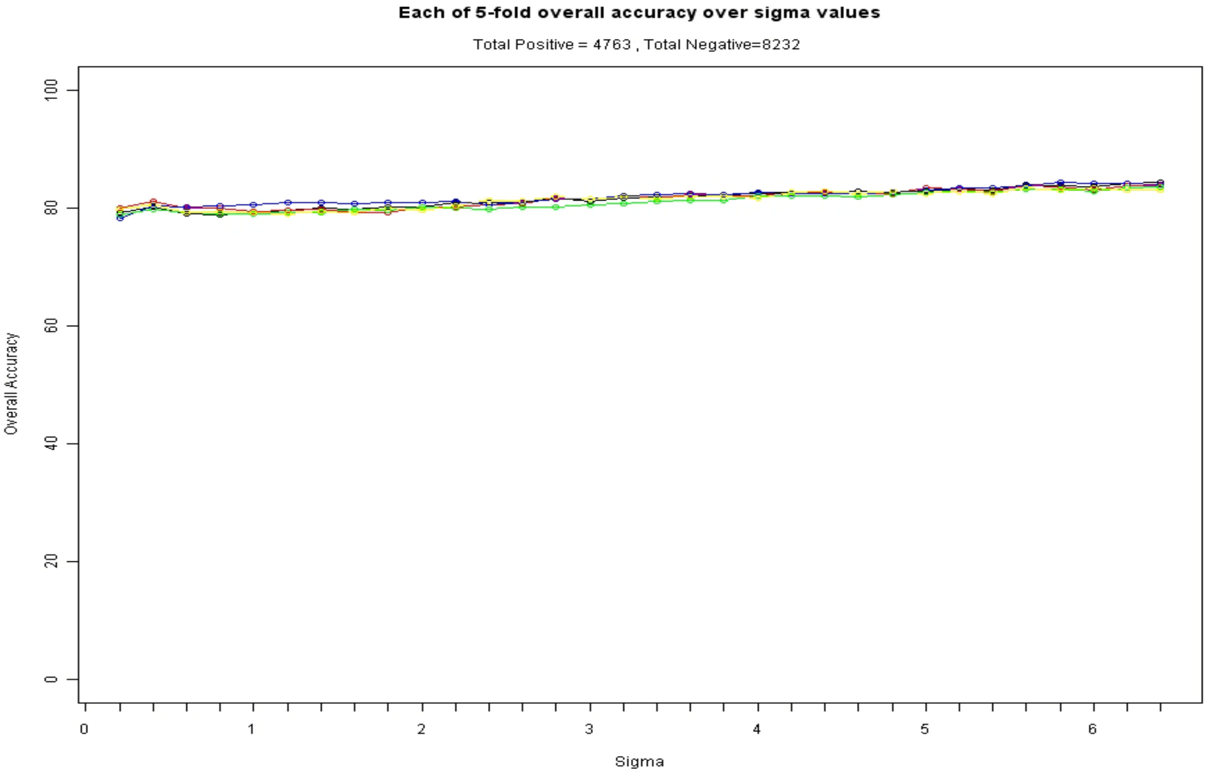


Figure 33: Fivefold overall accuracy (Genotoxicity, SVM, High diversity high noise (HDHN) (positive in any assay) model). Negative accuracy (red color), positive accuracy (blue color) and overall accuracy.

Average Accuracy

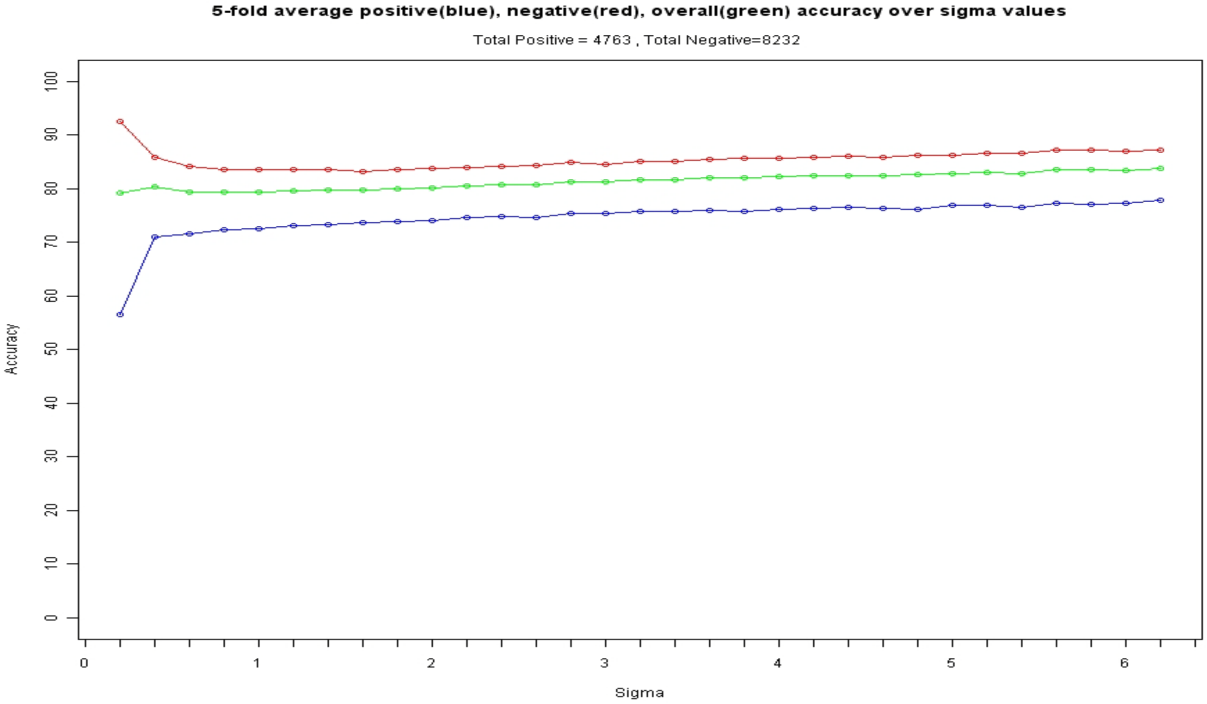


Figure 34: Fivefold average accuracy (Genotoxicity, SVM, High diversity high noise (HDHN) (positive in any assay) model). Negative accuracy (red color), positive accuracy (blue color) and overall accuracy.

In **Figure 31**, **Figure 32**, **Figure 33**, and **Figure 34** negative, positive, overall and average accuracy over different sigma values are relatively stable for the five fold cross validation.

5.4.2.2 Testing on Independent data

After checking the performance of 5-fold, SVM model was built using all the training data (4763 GT positive and 8232 GT negative compounds) for testing on independent dataset (38 GT positive and 2008 clinical trial negative compounds) for different sigma values (**Figure 35**).

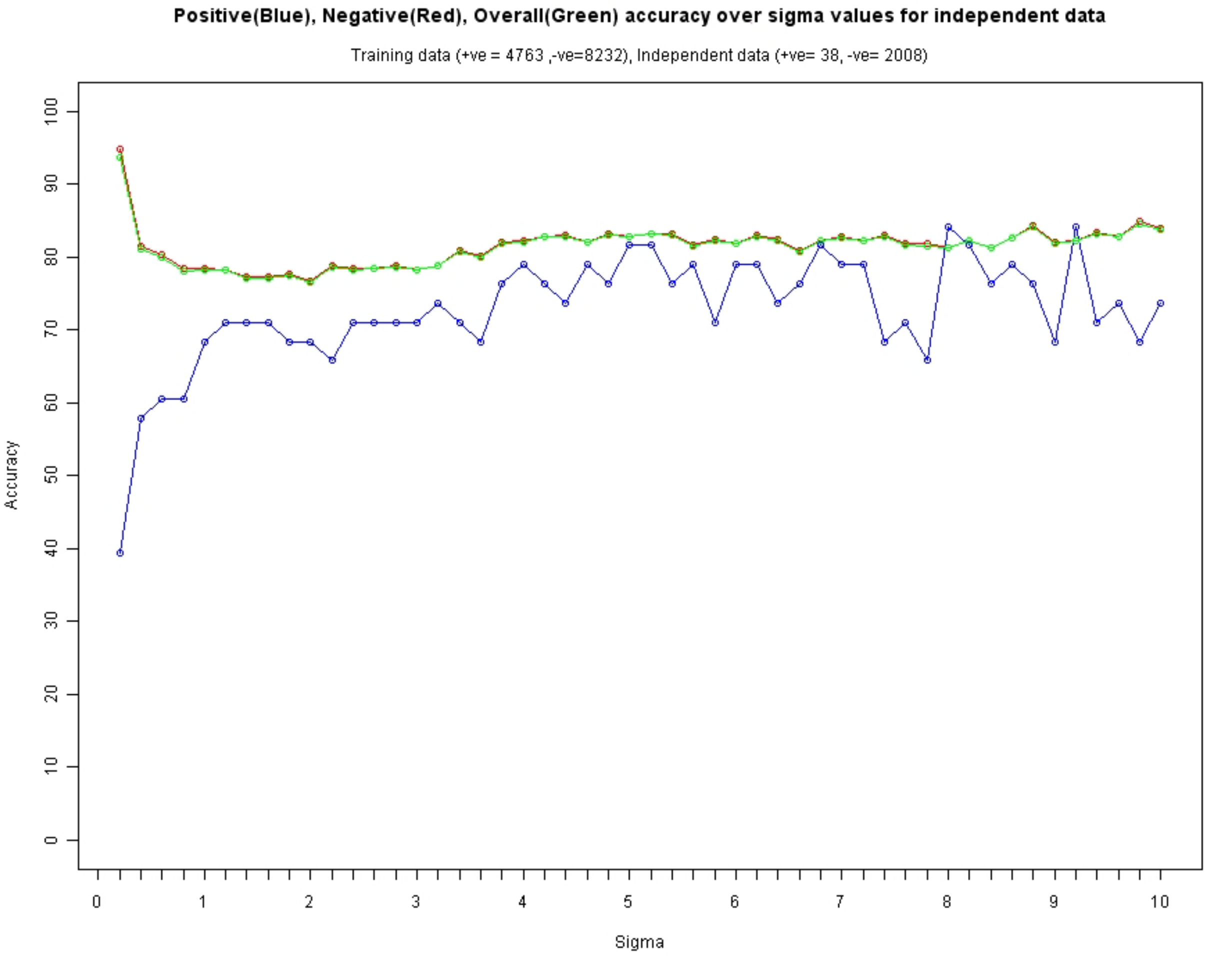


Figure 35: Testing on Independent data set (Genotoxicity, SVM, High diversity high noise (HDHN) (positive in any assay) model)

5.4.2.3 Virtual Screening on Pubchem and MDDR

The Pubchem and MDDR database were scanned by models created for different sigma values (**Figure 36** and **Figure 37**). For scanning this database, models were created by including independent dataset in first instance and later by just including positive independent dataset (leaving the negative clinical trial dataset).

The scanning with the model created by including clinical trial data in negative dataset will introduce a bias towards finding a compound able to reach till clinical trial. This percentage is also useful for pharmaceutical industry or regulatory bodies because of the fact of very less rate of compounds reaching to clinical trial. The scanning with the model without clinical trial data is supposedly more accurate way of scanning, because the fate clinical trial compound is not sure i.e. whether it will be genotoxic or non-genotoxic.

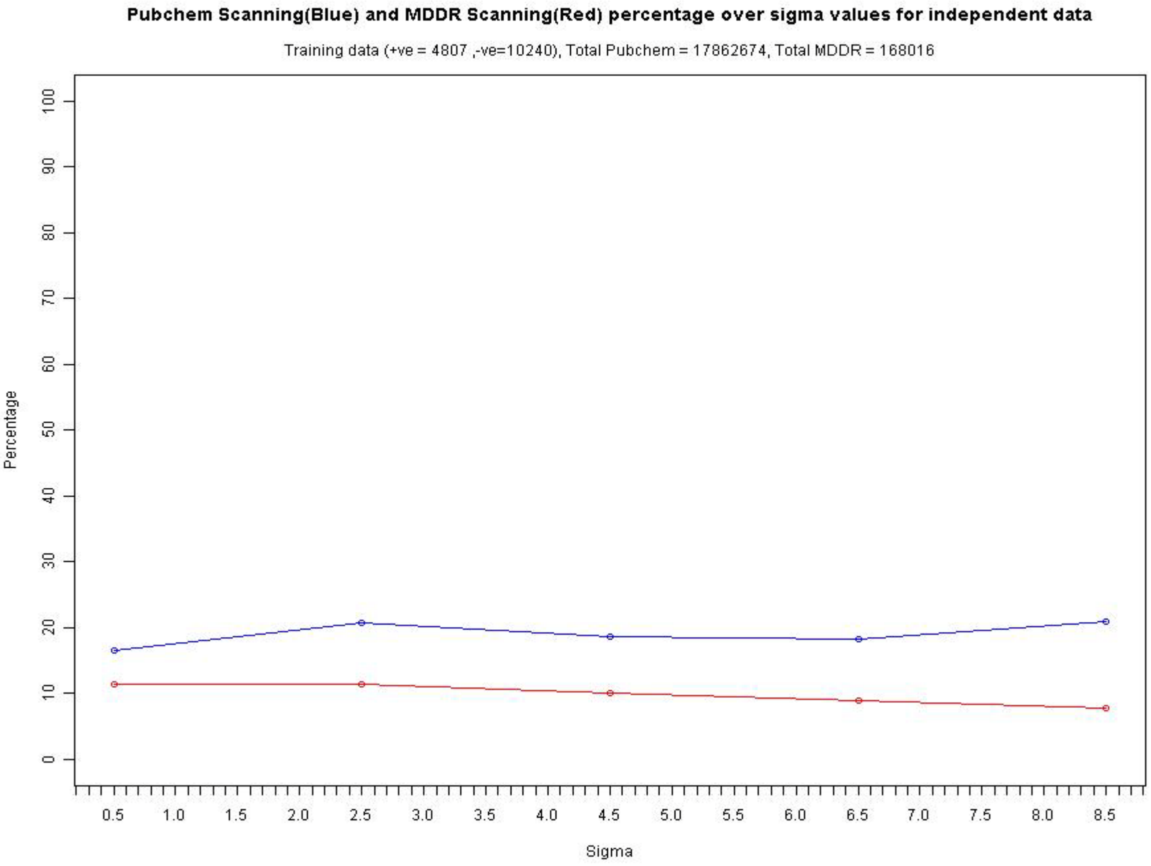


Figure 36: Scanning Pubchem and MDDR (Genotoxicity, SVM, High diversity high noise (HDHN)(positive in any assay) model). The graph shows the percentage of total number of compounds in database found as genotoxic positive over different sigma values. Blue dots and line represent percentage of Pubchem

compounds predicted as genotoxic positive. Red dots and percentage represent percentage of MDDR compounds predicted as genotoxic positive.

Scanning without clinical trial

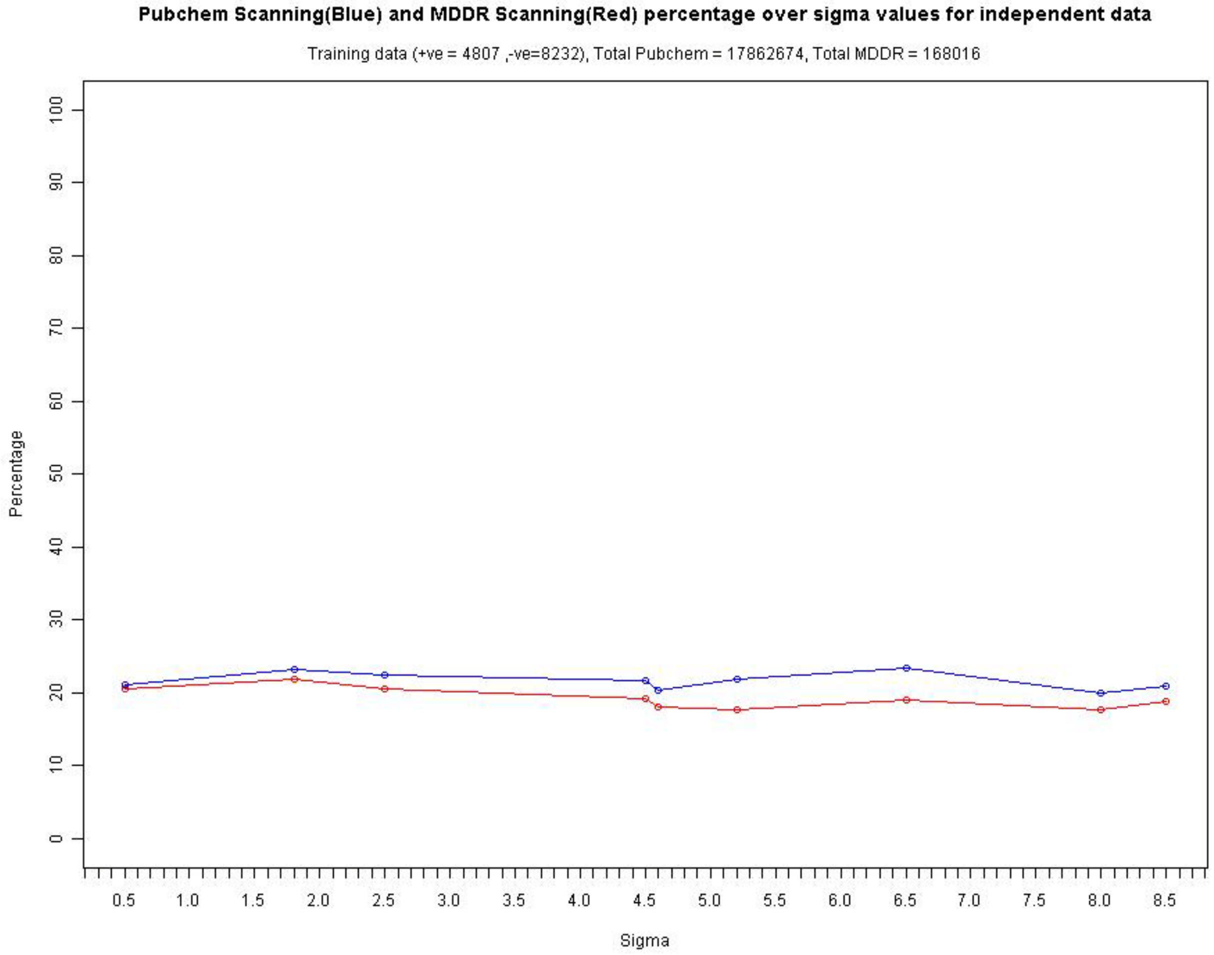


Figure 37: Scanning Pubchem and MDDR (Clinical trial data set excluded while constructing models) (Genotoxicity, SVM, High diversity high noise (HDHN)(positive in any assay) model)

5.4.3 Low diversity low noise (LDLN) (positive in Ames or in vivo) model prediction performance

The result in this section is presented for 5-fold cross validation accuracy, testing on independent dataset, and virtual screening on Pubchem and MDDR.

5.4.3.1 Five fold

Table 12: 5-fold cross validation for genotoxicity prediction models on less diverse dataset (positive in Ames or *in vivo*)

SVM 5-fold cross validation						
	Accuracy of models (Best models of each fold)			Average accuracy		
	Positive Accuracy	Negative Accuracy	Overall Accuracy	Positive Accuracy	Negative Accuracy	Overall Accuracy
Fold1	80.57	91.13	88.10	77.16	89.68	86.08
Fold2	80.12	91.86	88.48	75.35	90.70	86.29
Fold3	81.93	90.89	88.31	78.30	90.18	86.77
Fold4	79.07	91.07	87.62	75.85	89.96	85.90
Fold5	79.22	92.59	88.74	75.90	91.36	86.92
Average	80.18	91.51	88.25	76.51	90.38	86.39
Max	81.93	92.59	88.74	78.30	91.36	86.92
Min	79.07	90.89	87.62	75.35	89.68	85.90
STDEV	1.16	0.71	0.42	1.20	0.67	0.44

Negative Accuracy

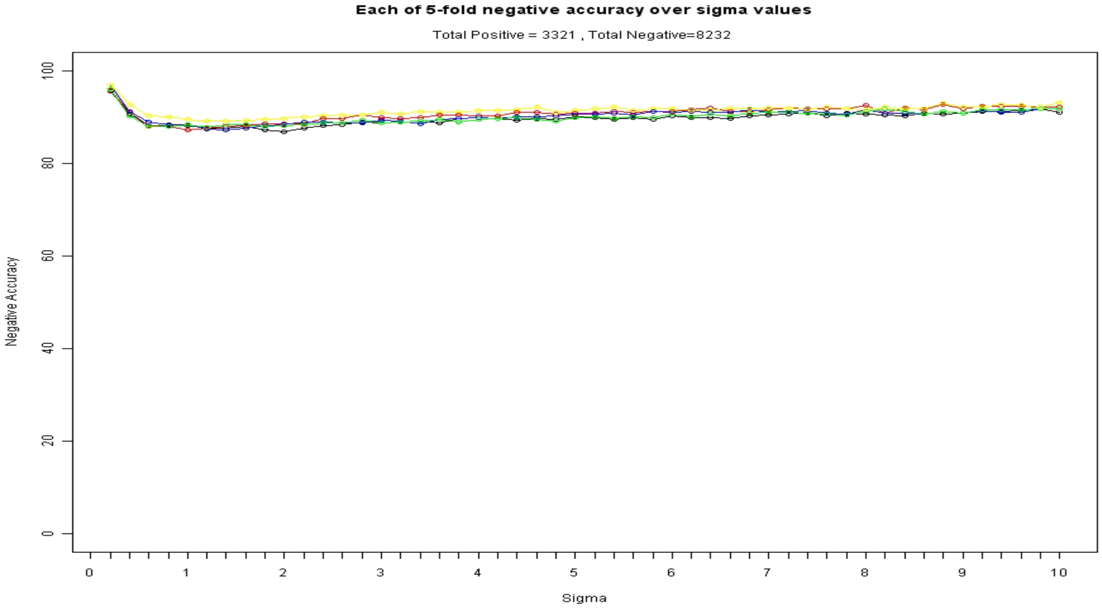


Figure 38: Fivefold negative accuracy (Genotoxicity, SVM, Low diversity low noise (LDLN) (positive in Ames or in vivo) model). Negative accuracy (red color), positive accuracy (blue color) and overall accuracy.

Positive Accuracy

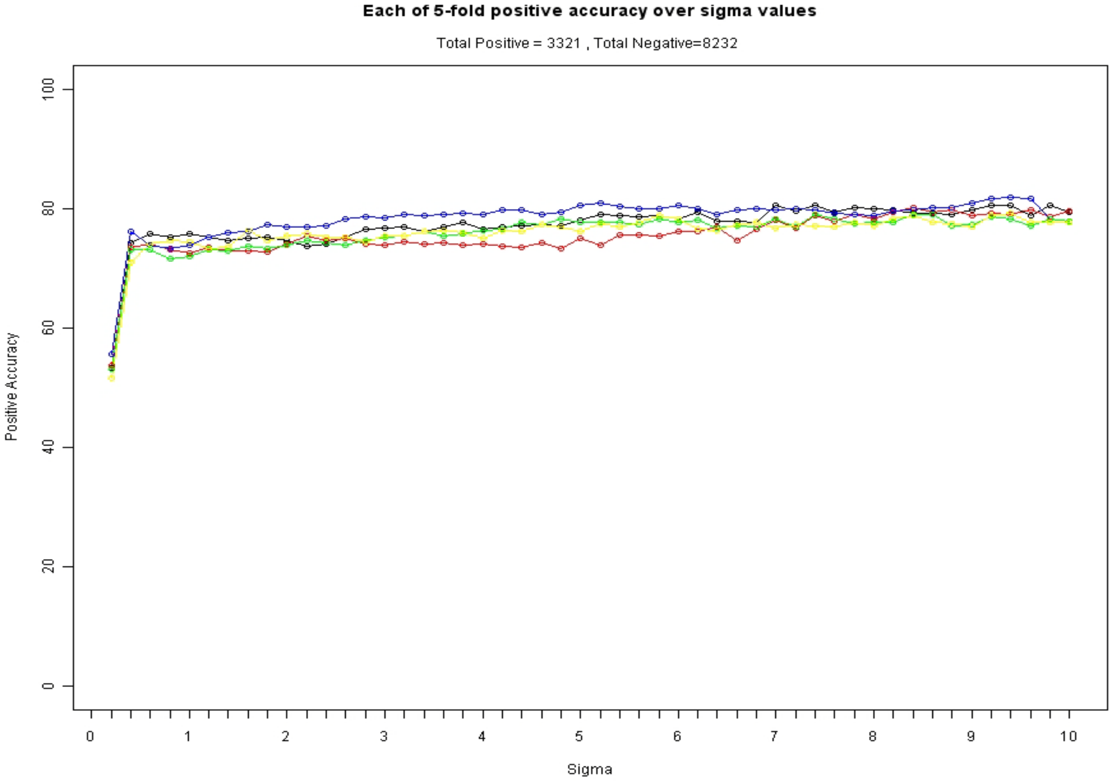


Figure 39: Fivefold positive accuracy (Genotoxicity, SVM, Low diversity low noise (LDLN) (positive in Ames or in vivo) model). Negative accuracy (red color), positive accuracy (blue color) and overall accuracy.

Overall Accuracy

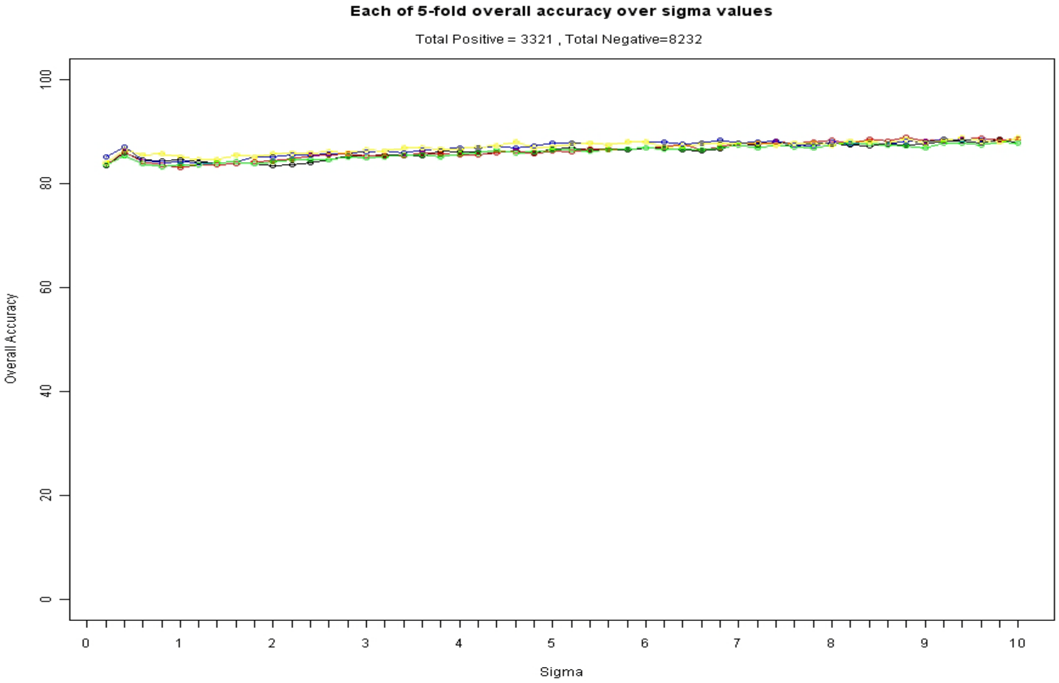


Figure 40: Fivefold overall accuracy (Genotoxicity, SVM, Low diversity low noise (LDLN) (positive in Ames or in vivo) model). Negative accuracy (red color), positive accuracy (blue color) and overall accuracy.

Average Accuracy

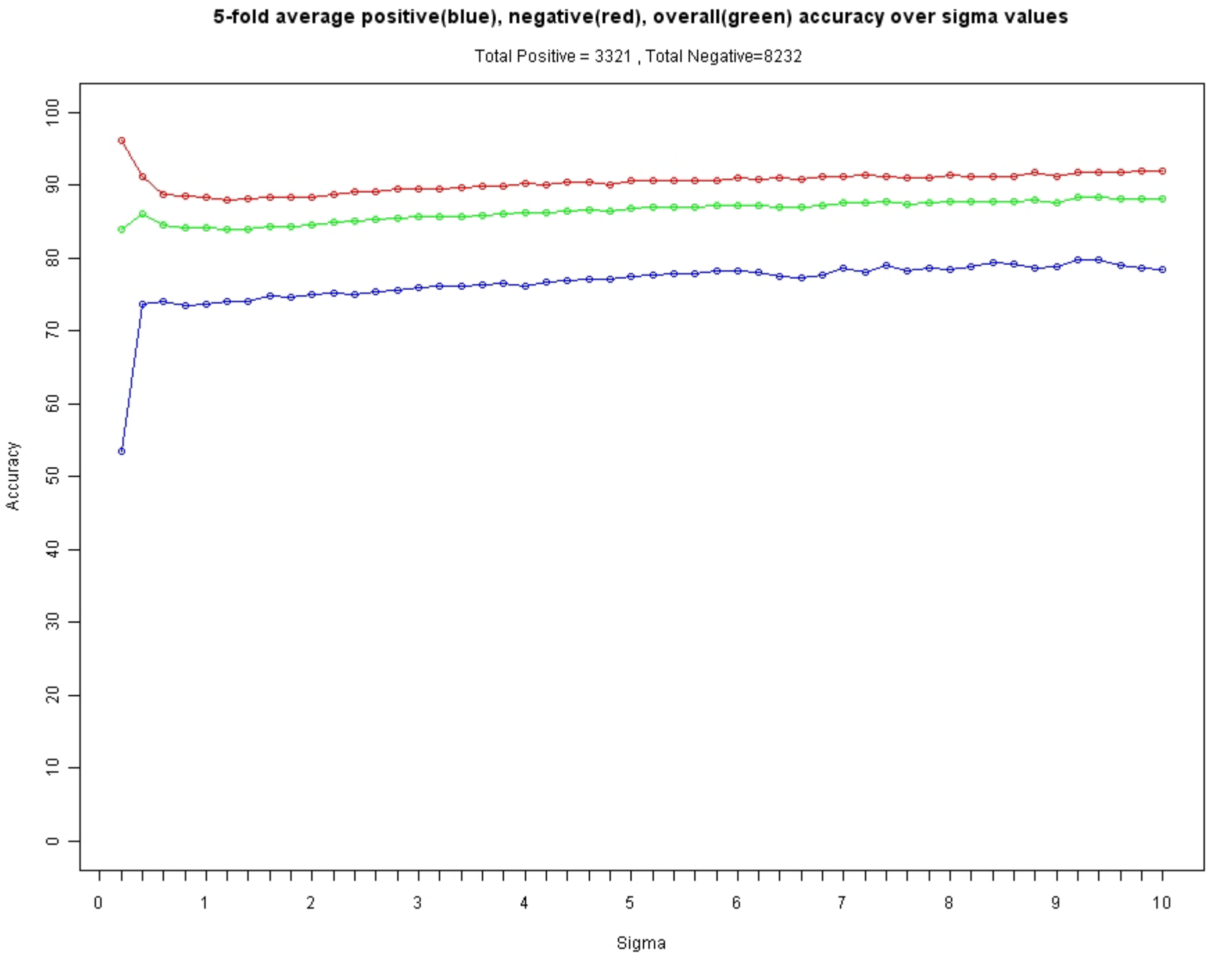


Figure 41: Fivefold average accuracy (Genotoxicity, SVM, Low diversity low noise (LDLN) (positive in Ames or in vivo) model). Negative accuracy (red color), positive accuracy (blue color) and overall accuracy.

5.4.3.2 Testing on Independent data

After checking the performance of 5-fold, SVM model was built using all the training data (3221 GT positive and 8232 GT negative compounds) for testing on independent dataset (38 GT positive and 2008 clinical trial negative compounds) for different sigma values

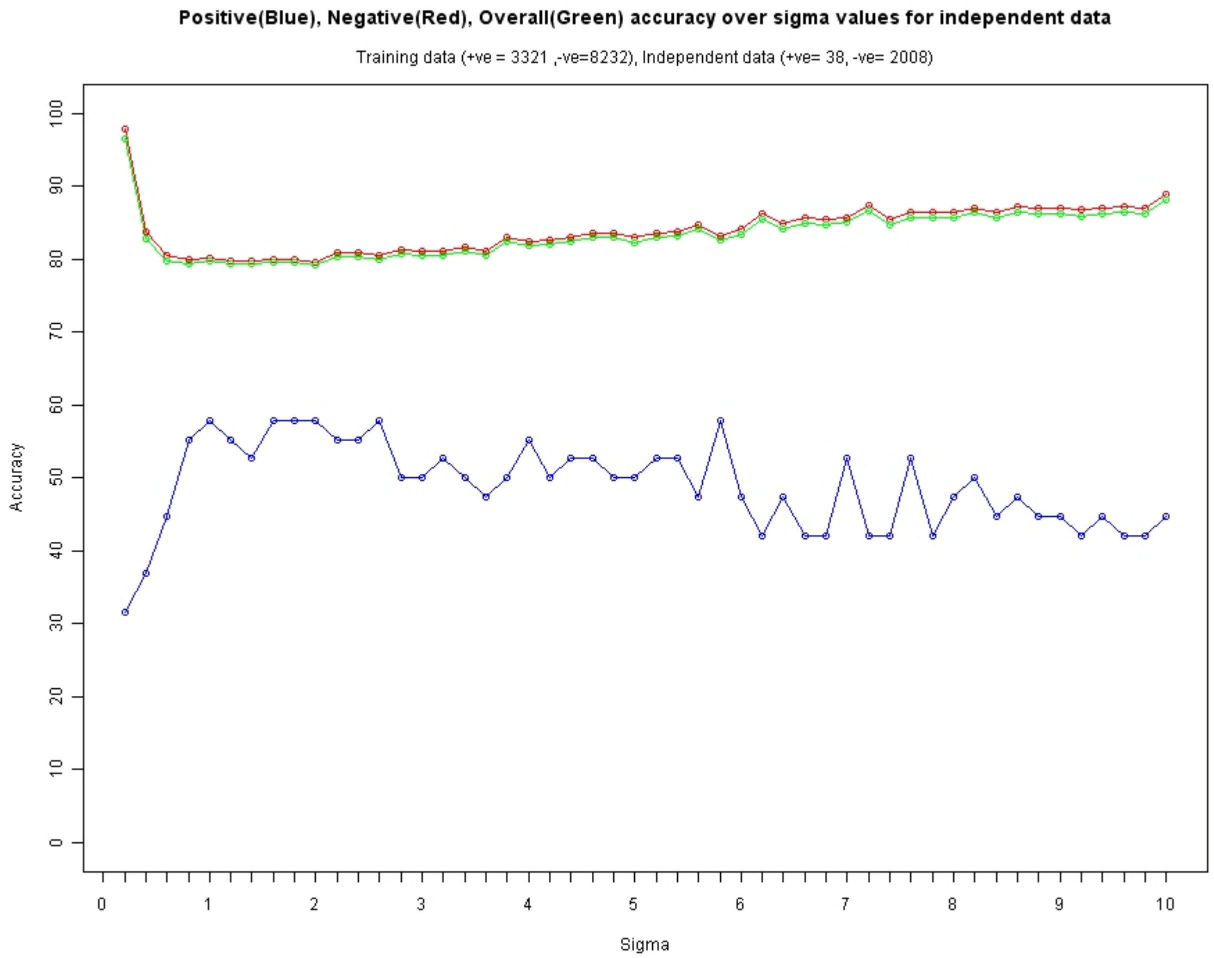


Figure 42: Testing on independent data set (Genotoxicity, SVM, Low diversity low noise (LDLN) (positive in Ames or in vivo) model)

5.4.3.3 Virtual Screening on Pubchem and MDDR

The Pubchem and MDDR database were scanned by models created for different sigma values (**Figure 43** and **Figure 44**). For scanning this database, models were created by including independent dataset in first instance and later by just including positive independent dataset (leaving the negative clinical trial dataset). The reason for scanning done by these two ways is same as in virtual screening of more diverse dataset (Section 5.4.2.3).

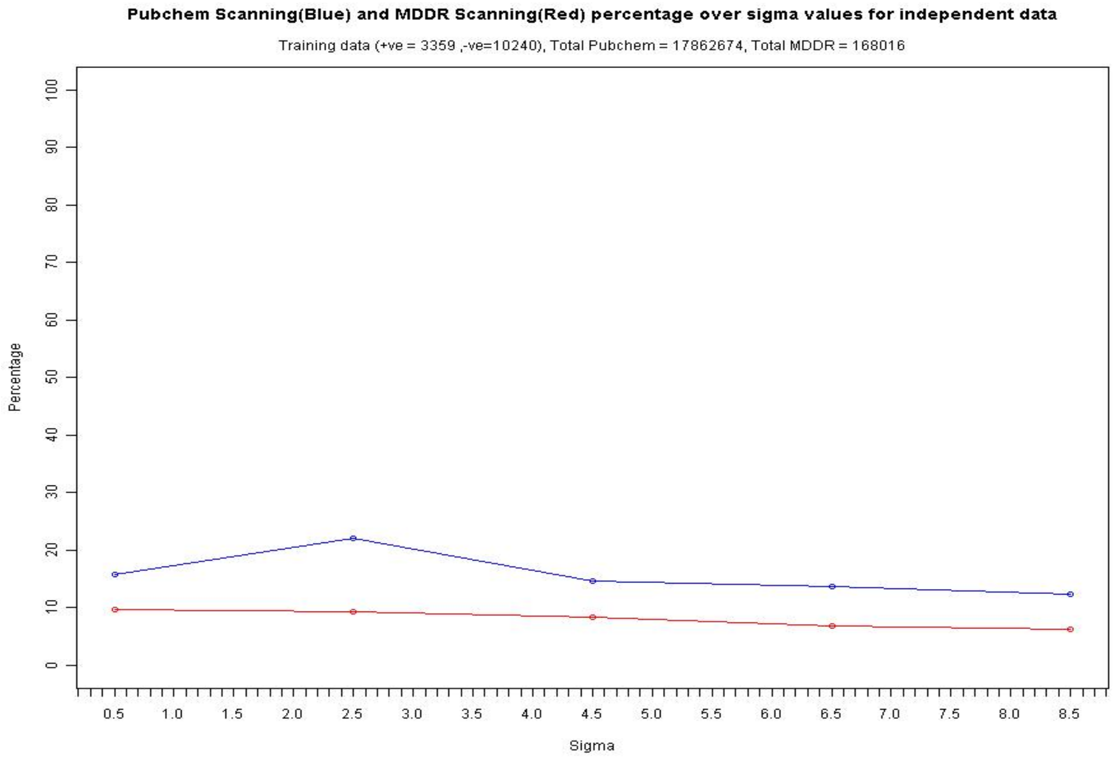


Figure 43: Scanning Pubchem and MDDR (Genotoxicity, SVM, Low diversity low noise (LDLN) (positive in Ames or in vivo) model)

Scanning without clinical trial in the training model

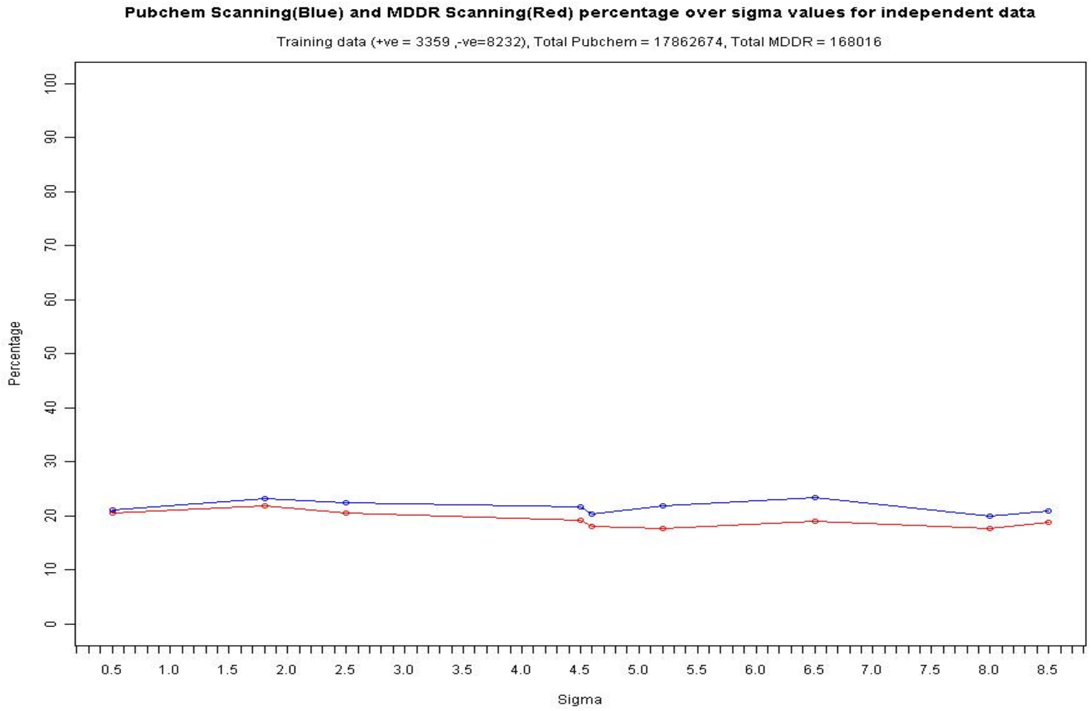


Figure 44: Scanning Pubchem and MDDR (Clinical trial data set excluded while constructing models) (Genotoxicity, SVM, Low diversity low noise (LDLN) (positive in Ames or in vivo) model)

In addition to scanning, analysis of MDDR hits were also done to map compound to the therapeutic class (**Table 13**). The MDDR compound database has the information of therapeutic class for the compounds. The virtual screening hits by models were found to cover 550 therapeutic classes with antineoplastic class having maximum number of hits. This is an agreement with the fact that majority of antineoplastic compounds have the potential for genotoxicity.

Table 13: MDDR classes that contain higher percentage (≥3%) of HDHN SVM model identified virtual GT+ hits in screening 168K MDDR compounds. The total number of SVM identified virtual GT+ hits is 40,257(23.96%)

MDDR Classes that Contain Higher Percentage (>3%) of Virtual Genotoxic Hits	**No and Percentage of Virtual Genotoxic Hits in Class**	**Percentage of Class Members Selected as Virtual Genotoxic Hits**
Antineoplastic	4848(12.04%)	22.47%
Antiallergic/Antiasthmatic	2326(5.78%)	21.68%
Antihypertensive	2095(5.2%)	19.59%
Antiarthritic	1948(4.84%)	25.32%
Cognition Disorders, Agent for	1752(4.35%)	23.02%
Anxiolytic	1363(3.39%)	20.16%
Antidepressant	1232(3.06%)	19.87%
Antiinflammatory	1227(3.05%)	22.04%

5.5 Discussion and Conclusion

The purpose is mentioned accordingly in the thesis as suggested by the examiner. The line has been added which says the average accuracies of fivefold cross validation were relatively over different sigma values. The graphs are also meaningful in the sense of its comparison with other graphs e.g. accuracies of independent data set where positive accuracies over different sigma value are not stable. This gives a better picture of discrepancy of fivefold cross validation and independent data set result. That was the main reason the that sigma value selected for pubchem and MDDR scanning could not be just based on the best parameter selection based on fivefold cross validation result which is the normal protocol in machine learning methods.

The purpose of showing the independent validation results and the scanning of PubChem and MDDR from a series of sigma value is to provide various prediction models. The choice to have different prediction models gives flexibility to end user to choose among the best models.

The existences of different in vivo and in vitro genotoxicity tests are corroborated inference from these test motivate the idea of providing different prediction models. For example, the compounds found positive in standard test battery of ICH guideline (which includes three genotoxicity tests mentioned in the introduction of this chapter) should be further tested extensively for genotoxicity but the compounds found negative in standard test battery are considered safe. Similarly, the compounds found negative by prediction models which were generated at three best sigma values can be considered safer than compounds giving non-consistent result by three prediction models.

The discussion is added to the chapter. The yield about 80% is high enough in my opinion. One cannot expect that all the support vectors will be limited to 79 true genotoxic compounds when the total number of positive genotoxic compounds used in training of SVM models is 2776. The

C value used for training was very high (C= 1000, 000), so the models are not under fitted. The appropriate tanimoto cut-off should be above 0.8, otherwise it generally produce high false positive and is not reliable. The yield at tanimoto cut-off 0.8 and 0.9 are 57% and 48% respectively. So, by tanimoto similarity it is not possible to beat SVM in terms of yield.

The usefulness of machine learning methods, particularly SVM, *k*-NN, and PNN, in facilitating the prediction of GT+ potential of a diverse set of molecules without requiring the intrinsic mechanism knowledge of chemical compounds, has been made possible through this study. The use of a large number of compounds has shown to significantly improve accuracy levels of genotoxicity prediction. HDHN models have better performance than LDLN models which further consolidate the fact that SVM is capable handling some noise when the dataset is large. Virtual Screening can be used for the identification of potential genotoxic compounds in large databases such as the likes of Pubchem and MDDR. The results gained via Virtual Screening can be fruitfully examined by confirmatory wet lab experiments.

Chapter 6 Machine Learning Classification: Prediction of p38 kinase inhibitors

6.1 Introduction of p38 MAPKs

The p38 mitogen-activated protein kinases (MAPK) is a type of mammalian stress activated MAPK. MAPKs belong to the family of serine/threonine kinase which get activated by a conserved mechanism that is phosphorylation of both serine and tyrosine residues. The p38 MAPK gets activated by stress response and has important role in cytokine production. There are four different isoforms of p38MAPK: p38α, p38β, p38γ and p38δ. These isoforms are 60-70% similar in sequence but differ in the size of a lipophilic pocket. Also, the lipophilic pocket of these isoforms which is buried inside the ATP may have different gatekeeper residue. The gatekeeper residue of p38α and p38β is threonine while p38γ and p38δ have methionine as the gatekeeper residue. Out of these four isoforms, p38α isoform has been studied most because of its role in the biosynthesis of inflammatory cytokines interleukin-1β (IL-1β) and tumor necrosis factor alpha (TNFα). The excessive production of IL-1β and TNFα is found to be the cause of many inflammatory diseases(Pettus and Wurz 2008). The therapeutic importance of TNFα and IL-1β in chronic inflammatory diseases has been reported in many studies. Suppression of IL-1β helps in treating cartilage damage and diminishes the cell inflammation, while blocking TNFα alone have shown therapeutic value in treating joint swelling of animals in which Rheumatoid Arthritis have been introduced(Kuiper, Joosten et al. 1998). Furthermore, some of this individual therapeutic approaches are evident in examples such as infliximab, a monoclonal antibody, against TNFα for the treatment of rheumatoid arthritis and Crohn's disease(Maini 2004); adalimumab, a fully humanized antibody, against TNFα; anakinra, against IL-1β receptor for the treatment of rheumatoid arthritis; Remicade, Humira and Enbrel(Goldsmith and Wagstaff 2005; Pettus and Wurz 2008). Although TNFα and IL-1β can be targeted alone for anti-inflammatory actions but the synergistic interaction is illustrated in many studies which becomes the ground for applying combination therapy against these two cytokines (Bendele, Chlipala et al. 2000; Bolos 2005). Moreover the idea of combined therapeutic approach can be consolidated by the fact of drawbacks and individual targeting of TNFα and IL-1β. Some of these drawbacks and limitation of individually targeting TNFα and IL-1β includes short half-life, low oral bioavailability, congestive heart failure, increased risks for infections, possible immune reactions and other malignancies (Palladino, Bahjat et al. 2003). These drawbacks and limitation can be reduced by combined therapeutic approaches in addition to better efficacy. The p38 MAPK signalling pathway shows the path for this combined therapeutic approach. Inhibiting the p38 MAPK will not only suppress TNFα and IL-1β but also the other enzymes like matrix metallonoproteinases and Cyclooxygenase-2 which are also responsible for inflammation(**Figure 45**) (Bolos 2005). The p38 MAP kinase activation of can mediate gene expression as well because of its interaction with many transcription factors .The transcription factors like ATF1, ATF2, ATF-6, SAP1A (Signaling lymphocytic Activation molecule associated Protein-1A), the MEF2A/C (Myocyte Enhance Factor-2A/C), and Elk1 (ETS-domain transcription factor-1) upon interaction with p38 MAP kinase becomes phosphorylated and subsequently becomes activated. The p38 MAPKs also regulate p53 which is a tumor suppressing protein, and NFAT which is an important transcription factor for cell differentiation and embryonic development. In summary, p38 MAPKs has major role in apoptosis pathway, transcriptional regulation, and cytokine production (Ferrer, Blanco et al. 2002; Rasmussen, Iversen et al. 2008).

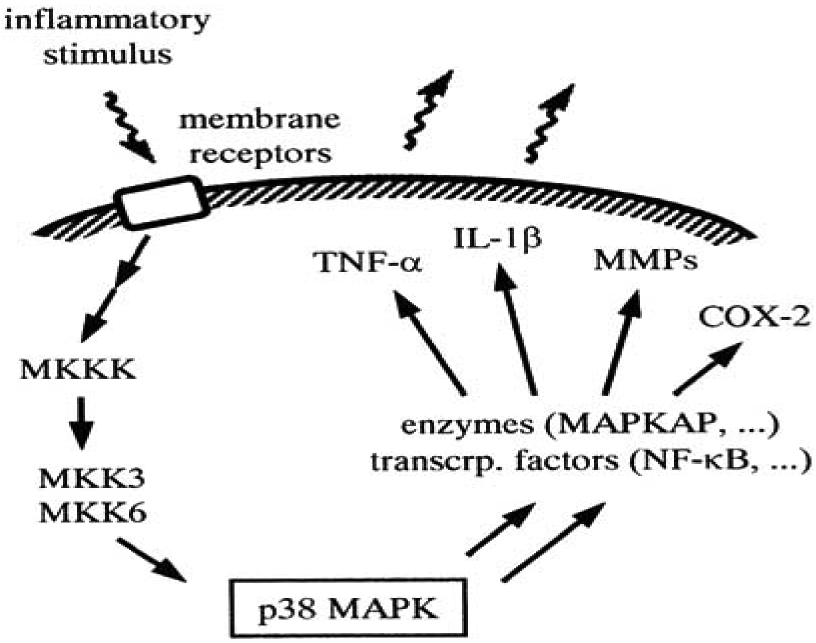


Figure 45: p38 MAPK Signaling

6.2 Methods

The general flowchart for performing machine learning classification method is shown in **Figure 46**. The detailed method for 5-fold cross validation, scaling, virtual screening, and hierarchical clustering is explained in Chapter 2.

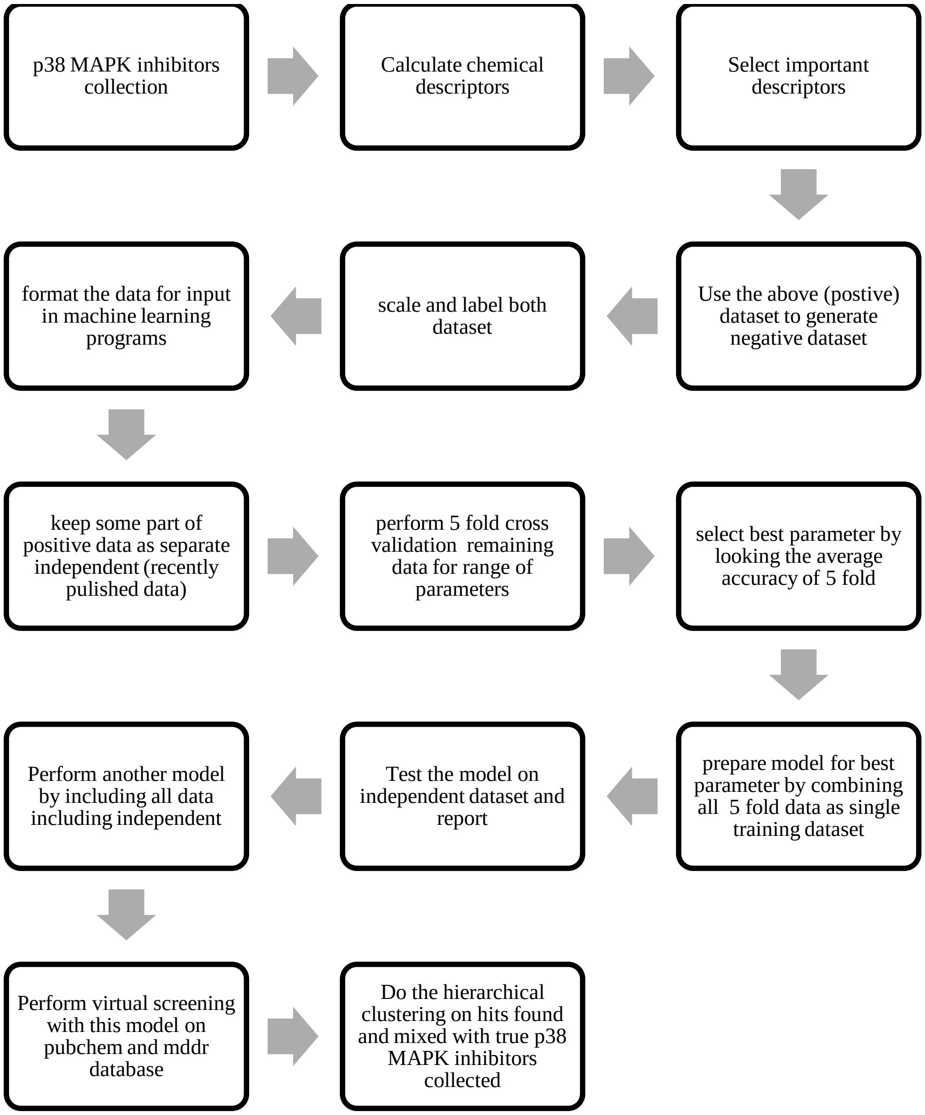

p38 MAPK inhibitors collection
Calculate chemical descriptors
Select important descriptors
Use the above (postive) dataset to generate negative dataset
scale and label both dataset
format the data for input in machine learning programs
keep some part of positive data as separate independent (recently pulished data)
perform 5 fold cross validation remaining data for range of parameters
select best parameter by looking the average accuracy of 5 fold
prepare model for best parameter by combining all 5 fold data as single training dataset
Test the model on independent dataset and report
Perform another model by including all data including independent
Perform virtual screening with this model on pubchem and mddr database
Do the hierarchical clustering on hits found and mixed with true p38 MAPK inhibitors collected

Figure 46: Flowchart for machine learning classification of p38 MAPK inhibitors

6.2.2 Selection of p38 inhibitors and non-inhibitors

A total of 1094 p38 inhibitors were manually collected from literature and drawn using Chemdraw software and subsequently converted to 3d structure using Corina software.

Table 21 in appendix shows the list of journal articles from which p38 inhibitors were collected. Since this study was done in the beginning of year 2008, journal articles are limited till year 2007.

The p38 non-inhibitor was generated by finding complement of p38 inhibitors in chemical space of whole Pubchem database. The Pubchem compounds were divided in 8000 family by k-mean clustering method. The p38 inhibitors were then mapped to these families. The families which were not covered by p38 inhibitors represent the complement family set. From each of these complement family, representative compounds were chosen by starting from centroid of the family to the varying distance in all the side to incorporate diversity. This way a total of 58774 compounds are selected as negative dataset.

6.2.3 Molecular descriptors

A total of 100 important descriptors were chosen from a total of 522 chemical descriptors calculated by our program which were used for generating p38 inhibitor prediction model. The detail about the selected 100 molecular descriptors is shown in **Table 14**. A more detailed description about the descriptors is given in Appendix.

Table 14: Molecular descriptors, selected 100 descriptors out of total 522 descriptors calculated for each compound

Molecular Descriptors			Selected	Total Calculated
Constitutational Descriptors			13	58
Charge Descriptors			6	14
	Electronic-topological descriptors		4	7
Topological descriptors			2	2
	Topological charge index		0	5
	Mean topological charge index		3	10
	Molecular path count		7	7
	Sum of E-State of atom type		28	88
	Sum of H E-State of atom type		16	42
	Moreau-Broto topological autocorrelation		0	0
		Atomic mass weighted Moreau-Broto	0	11
		Electronegativity weighted moreau-Broto	0	11
		VDW radius weighted Moreau-Broto	0	11
		Estate Values weighted Moreau-Broto	0	11
		polarizability weighted Moreau-Broto	0	11
		Van der Waals volume weighted Moreau-Broto	0	11
	Moran topological autocorrelation		0	0
		Atomic mass weighted Moran	0	10
		Electronegativity weighted Moran	0	10
		VDW radius weighted Moran	0	10
		Estate weighted Moran	0	10
		Polarizability weighted Moran	0	10
		VDW volume weighted Moran	0	10
	Geary topological autocorrelation		0	0
		Atomic mass weighted Geary	0	10
		Electronegativity weighted Geary	0	10
		VDW radius weighted Geary	0	10
		E-state weighted Geary	0	10
		Polarizability weighted Geary	0	10
		VDW volume weighted Geary	7	25
	Solvation connectivity index		10	10
	Topological distance related		4	18
		BCUT highest of mass	0	5
		BCUT lowest of mass	0	5
		BCUT highest of electronegativity	0	5
		BCUT lowest of electronegativity	0	5
		BCUT highest of VDW radius	0	5
		BCUT lowest of VDW radius	0	5
		BCUT highest of Estate	0	5
		BCUT lowest of Estate	0	5
		[2.4.52.10] BCUT highest of polarizability	0	5
		BCUT lowest of polarizability	0	5
		BCUT highest of VDW volume	0	5
		BCUT lowest of VDW volume	0	5
Total			100	522

6.3 Results and discussion

6.3.1 Five-fold cross validation and testing on independent dataset

The 5-fold cross validation study was done to see the performance of SVM and to select the best parameters for further testing on independent dataset. **Table 15** shows the 5-fold cross validation result having 95.72% average positive accuracy and 99.82% negative accuracy.

Different machine learning classification methods, other than SVM, were applied to test the performance of prediction capability. The result is shown in **Table 16** and **Table 17**. The test data set in table is randomly selected 300 compounds from journal articles published before year 2006. Similarly, the 15385 negative test data was randomly selected from the total of 58774 negative generated by representative complement of all p38 MAPK inhibitor in chemical space. The independent data consist of 287 compounds collected for journal articles published in year 2006 and 2007.

Table 15: 5-fold cross validation by SVM for p38 MAPK inhibitors. Each fold is comprised of 196 positive labeled (p38 MAPK inhibitor) and 10725 negative labeled compounds (non-inhibitors generated from Pubchem chemical space).

	Accuracy of models (Best models of each fold)			
	Positive Accuracy	Negative Accuracy	Overall Accuracy	MCC
Fold1	95.40	99.80	99.70	0.85
Fold2	94.40	99.80	99.70	0.88
Fold3	97.40	99.90	99.80	0.92
Fold4	94.50	99.80	99.70	0.84
Fold5	96.90	99.80	99.80	0.88
Average	95.72	99.82	99.74	0.87
Max	97.40	99.90	99.80	0.92
STDEV	1.37	0.04	0.05	0.03

Table 16 : Prediction performance of various machine learning methods for test data p38 MAPK inhibitor prediction

Method	Total count	True Positive	True Negative	False Positive	False Negative	Positive Accuracy	Negative Accuracy	Overall Accuracy	MCC
SVM	15685	282	15357	28	18	94	99.81	99.7	0.85
J48 (C4.5)	15685	300	15385	0	0	100	100	100	1
LMT	15685	294	15385	0	6	98	100	99.96	0.98
ADTree	15685	300	15385	0	0	100	100	100	1
BFTree	15685	0	15385	0	300	0	100	98.08	
NBTree	15685	300	15385	0	0	100	100	100	1
Decision Stump	15685	300	15377	8	0	100	99.94	99.94	0.97
Random Forest	15685	300	15385	0	0	100	100	100	1
Random Tree	15685	269	15374	11	31	89.66	99.92	99.73	0.86
REPTree	15685	300	15385	0	0	100	100	100	1
FT	15685	300	15385	0	0	100	100	100	1
J48graft	15685	285	15385	0	15	95	100	99.9	0.95
SimpleCart	15685	0	15385	0	300	0	100	98.08	
NaiveBayes	15685	299	8979	6404	1	99.66	58.36	59.15	0.03
ZeroR	15685	0	15385	0	300	0	100	98.08	
Ibk (KNN)	15685	269	15320	65	31	89.67	99.58	99.39	0.72

Table 17 : Prediction performance of various machine learning methods for independent data in p38 MAPK inhibitor prediction

Method	Total count	True Positive	False Negative	Positive Accuracy
SVM	287	217	70	75.61
J48 (C4.5)	287	190	97	66.20
LMT	287	182	105	63.41
ADTree	287	165	122	57.49
BFTree	287	0	287	0
NBTree	287	154	133	53.66
DecisionStump	287	188	99	65.5
RandomForest	287	177	100	61.67
RandomTree	287	174	113	60.63
REPTree	287	181	106	63.07
FT	287	176	111	63.32
J48graft	287	217	70	75.6
SimpleCart	287	0	287	0
NaiveBayes	287	201	86	70.03
ZeroR	287	0	287	0
Ibk (KNN)	287	175	112	60.97

Performance of different machine learning method varied greatly for p38 MAPK inhibitor classification. Although, the main focus was on SVM from our previous experiences, equal opportunity was given to other decision trees and kNN method. SVM performed very well in prediction accuracy with 75.61 % positive accuracy when tested on independent dataset. J48graft (modified C4.5 algorithm) also showed good performance (75.6 %) in testing on independent dataset. Other methods like Naïve bayes, j48 and Decision stump also showed good performance of 70.03, 66.2 and 65.5 percent respectively in testing on independent dataset.

6.3.2 Virtual screening of Pubchem and MDDR

The performance in scanning MDDR (**Table 18**) did not show good correlation with the percentage obtained except for SVM and kNN. Therefore, SVM and kNN was chosen for further analyses. SVM had clear edge over kNN in terms of positive accuracy for independent dataset. So, SVM was to scan Pubchem database. Also, other methods are very slow for scanning huge database like Pubchem.

Table 18: Machine learning based virtual screening of MDDR database by p38 MAPK inhibitor prediction model

Method	MDDR Total Count	Scanned positive	Percentage
SVM	168016	1221	0.73
J48	168016	33	0.02
LMT	168016	132	0.08
ADTree	168016	54	0.03
BFTree	168016	0	0.00
NBTree	168016	0	0.00
DecisionStump	168016	0	0.00
RandomForest	168016	0	0.00

RandomTree	168016	202	0.12
REPTree	168016	0	0.00
FT	168016	0	0.00
J48graft	168016	54	0.03
SimpleCart	168016	0	0.00
NaiveBayes	168016	123216	73.34
ZeroR	168016	0	0.00
Ibk (KNN)	168016	4372	2.60

Table 19: Pubchem scanning by SVM based p38 MAPK inhibitor prediction model

Method	Pubchem Total Count	Scanned positive	Percentage
SVM	13560720	40464	0.298

Out of total 40464 Pubchem hits found by SVM, 11947 were also found by KNN which was further analysed with hierarchal clustering (**Table 19**).

6.3.3 Hierarchical clustering of Pubchem hits

The hits found after scanning of Pubchem by SVM was further scanned by kNN. Thus, total 11947 Pubchem hits along with 1094 true p38 MAPKs inhibitors in literature were subjected to hierarchical clustering. Hierarchal clustering was performed using WEKA (Frank 2005) class COBWEB(Fisher 1990). **Figure 47** shows the visualization of hierarchal clustering. A total of 106 clusters were formed and the distribution of p38 inhibitors and Pubchem hits are shown in **Figure 48**. The Pubchem hits are well clustered with p38 inhibitors and the distribution is shared, not just segregated set. This shows the potential of p38 MAPKs inhibitor model in finding compound similar to existing p38 inhibitors. However, many clusters does exist which has only Pubchem hits and not the reported p38 inhibitors. This indicates that SVM is capable in finding a pattern out of compound descriptors which was not found by hierarchal clustering. So, the performance of hierarchical clustering by COBWEB in this case is very meaningful in terms of distribution ratio.

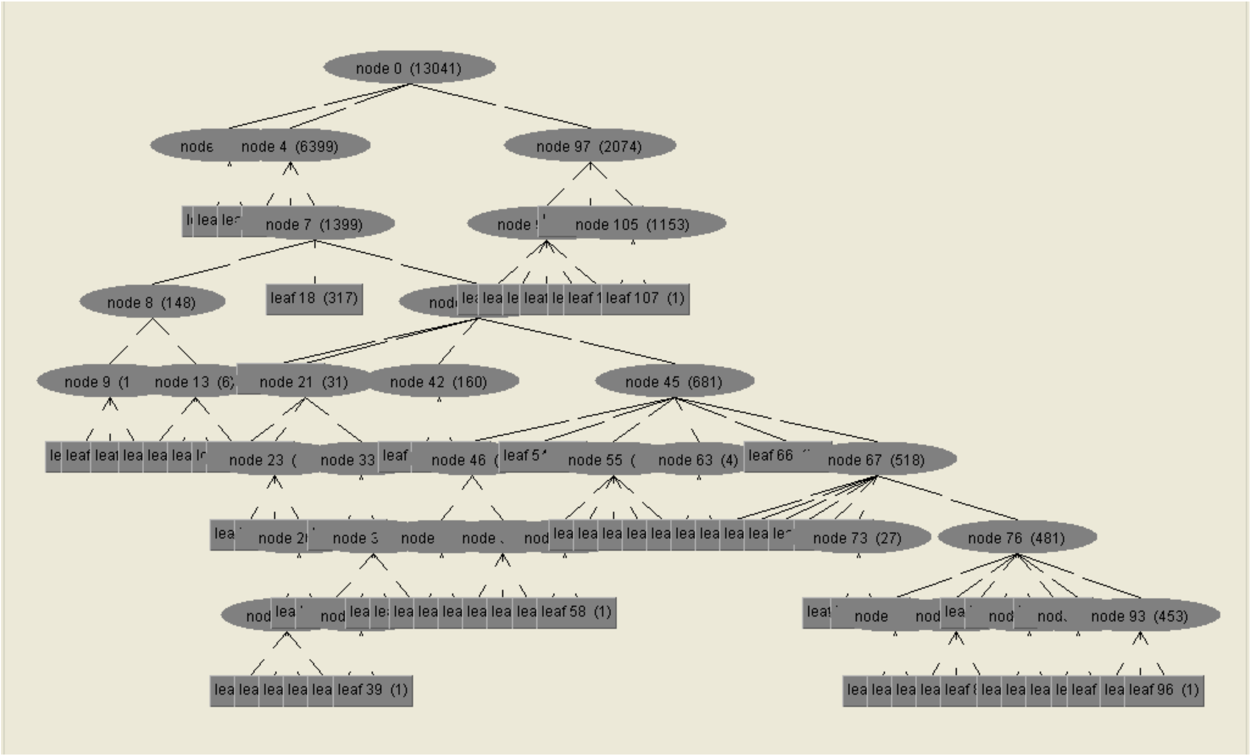


Figure 47: Hierarchal clustering by COBWEB on 13041 compounds (11947 Pubchem hits and 1094 true p38 inhibitors)

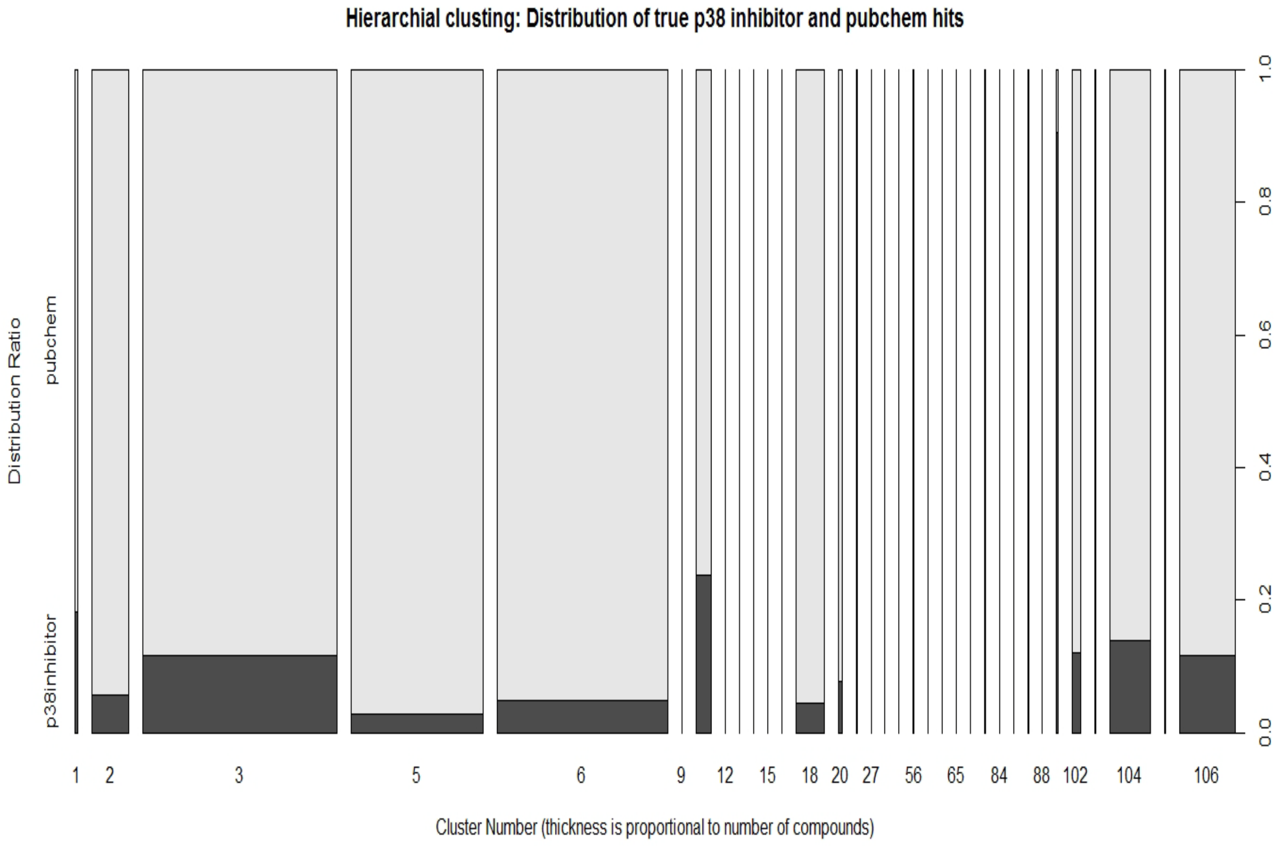


Figure 48: Hierarchal clustering, Distribution ratio of p38 inhibitor and Pubchem hits

6.4 Discussion and Conclusion

Prediction model of p38 MAPKs be very useful for drug discovery of inflammatory diseases. These models can be a handy tool to prediction a potential compound before or after synthesis. This will help in saving time and money. Also, many existing chemical library can be screened and hits can be assessed further in wet lab experiments. In various machine learning classification methods employed, SVM was found to have very good performance in testing on independent data as well as in virtual screening to give nearby the expected percentage of compounds in MDDR and Pubchem database. The good performance of SVM has also been found by other studies as well. This study adds the confidence in SVM for the cheminformatics related work. The other machine learning methods are also useful for comparison. Although, this study shows very good performance of SVM in comparison to other machine learning methods, this cannot guaranteed because of one very important factor which is optimization of parameters. In SVM, through rigorous use and experience the intuition of good parameter range is acquired. Normally, for small molecules with these set of 100 descriptors, sigma value gets optimized in the range of 0.5 to 2. Thus, we build the model at each sigma value starting from 0.1 to 10 with the interval of 0.1. This way we are generally confidant to find best optimized parameter. For other machine learning methods applied through WEKA the parameter optimization were not performed and the default parameter was chosen. However, the same descriptors set and descriptor value were used in every machine learning method. The optimization of parameter for each algorithm can be a subject of future study.

By inspecting incorrectly predicted compounds, it has been observed that some of the compounds are not being fully represented by molecular descriptors used. Such compounds generally contain complex chemical or structural configuration which may include compounds with multi-rings with several heteroatoms such as oxygen, nitrogen, sulphur, chlorine and fluorine. Also, the compounds having large rigid structure along with a flexible hydrophilic tail are sometime incorrectly predicted due to limited coverage of descriptors capable of representing such complexity. A common solution can be the use of all 522 calculated descriptors which was employed in genotoxicity study. Use of entire 522 descriptors will require huge amount of computation especially in the case p38 MAPKs inhibitor prediction model generation since it involve large number (58774) of p38 non-inhibitor generated from Pubchem chemical space by complement method. With such a huge number generating a SVM model may take more than a day for single sigma value. Moreover, comparing to genotoxic compounds p38 MAPKs inhibitors are less diverse. Furthermore, some redundant topological descriptors in 522 descriptor set can introduce noise as well. Therefore, this study was done with selected 100 descriptors only despite of having little error introduced because of that.

In conclusion, the prediction accuracy achieved for p38 MAPKs inhibitor by machine learning method is useful for further research and medicinal chemist and biologist interested in finding novel inhibitor can use this prediction model. Furthermore, machine learning classification method for p38 MAPKs inhibitor can also encourage development for other kind of inhibitors prediction model.

Chapter 7 Concluding remarks

7.1 Findings and Merits

In the process of developing databases, it was found that usefulness of biological databases can be enhanced significantly by including pathway related information. Similarly, many other things which improve the quality of database include manual annotation, addition of critical information needed by other researcher which could significantly increase the speed of their research, presentation and speed of database opening, cross-referencing to other databases, inclusion of newly published data, mechanism to easily update the database. It was also found that biological data format is shifting towards structured file format like XML for easy exchange. It was found that technology employed for database development play a major role in efficiency and speed especially when the database is very huge. For example, handling of protein structures kind of data is very efficient in Oracle or MySQL than Microsoft Access.

In the IHCD development it was found that a bridge is possible to conventional and modern medicine. If the traditional use of herb if could have rationalization in modern mechanistic and system biology based approach, it will be a great help in drug discovery. It was found the mapping of chemical ingredients of Indian herbs to Pubchem can add important information already available in Pubchem database. The merit of IHCD lies in providing diverse information in same window e.g. therapeutic category of chemicals, calculated chemical descriptor and docked complex by INVDOCK wherever possible.

In machine learning classification for medicinal chemicals one common argument is that how efficient it is in finding novel hits. Methods like QSAR generally have their applicability domain. But in SVM, the hyperplane was drawn by the influence of sufficiently large number of positive and negative compounds, and this hyperplane goes till infinity. So, there is no need to impose applicability domain in the SVM method employed in this study and the method is quite capable of finding novel hits as well. This is well support by good performance of SVM on true independent dataset. Various computational related issues and their handling was discovered when virtual screening was performed on very large database like Pubchem. SVM performed very well in terms of computational speed, other methods were quite slow especially lazy learning algorithm like k-NN. This could be because of algorithm and coding language. The SVM code is in C++ which is very near to machine language and can run very fast compared to WEKA which has Java code which runs in Java Virtual Machine (JVM) over the operating system of machine. The speed of calculation for machine learning methods becomes more prominent when the number of dimension is increased. For example, when the number of descriptors used was increased from 100 to 522, the time by SVM to scan 17 billion compounds of Pubchem was 3 days on Linux workstation for a single sigma value, while if the number of descriptors were kept 100 the same scanning could be done for 50 sigma value in same time. Thus, the selection of method can also depend on number of dimension. Computation time can usually be decreased is by parallel processing. The Pubchem database was split into 10 parts to achieve 10 times reduction in time but at the expense of CPU consumption.

7.2 Limitations

This study has few limitations which are basically associated with data availability and methods employed. In IHCD study, involving docking using INVDOCK, only proteins whose structures are deposited in the PDB is used. It is common knowledge that only a small fraction of all proteins have their 3D structures elucidated. This would hamper the widespread use of IHCD/INVDOCK method. One could alleviate this limitation using modeled 3D structure of proteins. In KDBI, some of the important signaling and metabolic pathways were missed due to lack of availability of kinetic parameters. Also, the KDBI server is running on IIS 5.0 web server which has limitation that it can process maximum 10 requests at a time. In KDBI, the SBML file for pathway simulation model is created by Java API of SBML version 2.4. The system biology related software which process SBML file, if upgrade themselves and stop supporting lower version of SBML then the SBML file downloaded from KDBI will not open in that particular software. In these situations, users are advised to edit these SBML file using some SBML editor. Also, pathway simulation parameter set available in KDBI is limited to pathway presented in the referenced article and it may need extra parameter collection if one wants to try the modified or extended pathway simulation.

In genotoxicity study, it was desirable to study the machine learning classification performance based on including *in vivo* genotoxic data alone in positive dataset. But due to lack of sufficient number of such data in literature or databases, the study missed that desired comparison. The machine learning methods employed has their inherent limitations due to their algorithm. Generally, machine learning methods require some minimum number of data points to develop a good prediction model. In addition, machine learning method is greatly influenced by the diversity of data (compounds in this case) for building models. Although, compounds collected in this work are from almost all available sources and can be considered very diverse, still it may be the case that dataset is not representative of certain set of compounds which are yet to be discovered and is very different from any existing compound. Also, the chemical space used in the case of p38 MAPK inhibitor prediction is based on Pubchem database which therefore decides the diversity of chemical compounds. The diversity of Pubchem is undoubtedly very high, thus the limitation associated with this is of little concern.

7.3 Suggestions for future studies

This work has attempted to provide insight to the importance of database development and machine learning classification methods of medicinal chemicals and biomolecules in drug discovery processes. Web accessible databases presented in this work e.g. updating of KDBI and IHCD can be building block for future work. Considering this, KDBI work has already been extended by introducing PIK-BLAST: Web-server of Protein Interaction Kinetic Parameters Estimated from Sequence Similarity. The hypothesis and the introduction are presented in next paragraph.

Knowledge of the kinetics of biomolecular interactions is important for facilitating quantitative study and simulation of biological systems and processes. The limited availability of experimental kinetic parameters is an obstacle of current studies. Literature studies have suggested that the kinetic parameters of interacting protein pairs are roughly correlated with those of protein pairs of similar sequences (Gabdoulline, Stein et al. 2007). With the introduction of a web-server, PIK-BLAST, kinetic parameter can be estimated of a protein pair from the experimental kinetic parameters of the protein pairs with similar sequences. Given the sequences of a protein pair, PIK-BLAST searches a pool of 2628 unique protein pairs (involved in 12896 kinetic reactions and 45 biological pathways) for finding similarity protein pairs and the parameters of the best matched pairs are provided as estimated parameters of the input protein pair. Sequence similarities were conducted by the NCBI BLAST program (Altschul, Madden et al. 1997; Altschul, Wootton et al. 2005) and kinetic data were from KDBI database. PIK-BLAST is publically available at http://bidd.nus.edu.sg/group/kinblast/pikblast.html .

PIK-BLAST work can be studied in detail by incorporating more number of unique protein pairs. By increasing the number of unique protein pairs which have kinetic parameters of interaction, one can improve the BLAST performance for the user specified input sequence. Moreover, the study is needed to establish that, apart from functional correlations, the kinetic parameters of interacting protein pairs are correlated with those of protein pairs of similar sequences. An extensive statistical analysis suggesting this would be more appropriate.

Another database IHCD also can be extended to include herbal formulation. Some tradition Indian herbal formulations have shown promising therapeutic effect in disease like cancer e.g.

triphala a herbal formulation of three different Indian plants namely *Terminalia chebula*, *Terminalia belerica* and *Emblica officinalis* (Deep, Dhiman et al. 2005; Sandhya, Lathika et al. 2006). Future work can be done to link protein targets of chemical ingredients of the herbal ingredients of the herbal formulation and also to provide web interface by incorporating them to IHCD.

Similar to the possible enhancement of database work, machine learning classification for genotoxicity and p38 inhibitors can also be extended. Genotoxicity prediction in this work has been done with the intension of pre-assessment of compounds likeliness to pass in clinical trials. Work can done to create prediction model for post-approval drugs. For p38 MAPKs inhibitor prediction, a web server can be created for online testing of input mol or sdf file based submission.

References

"Friedl,J.E.F. ((1997)) Mastering Regular Expressions. O'Reilly Media, Sebastopol, CA."

"Stubbletine,T. ((2003)) Regular Expression Pocket Reference. O'Reilly Media, Sebastopol, CA. .".

(1999). Indian Herbal Pharmacopoeia. Mumbai, RRL, Jammu and IDMA.

(2004). "Comprehensive Medicinal Chemistry (CMC) (2004) MDL Information Systems Inc., San Leandro, CA (USA) (http://www.mdl.com). The latest version records ~8800 clinically used drugs."

(2004). "MDL Drug Data Report (MDDR) (2004) MDL Information Systems Inc., San Leandro, CA (USA) (http://www.mdl.com). The latest version records ~165,000 drug candidates."

(2007). Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/.

(2009). "The Universal Protein Resource (UniProt) 2009." Nucleic Acids Res **37**(Database issue): D169-74.

Alfarano, C., C. E. Andrade, et al. (2005). "The Biomolecular Interaction Network Database and related tools 2005 update." Nucleic Acids Research **33**: D418-D424.

Altan-Bonnet, G. and R. N. Germain (2005). "Modeling T cell antigen discrimination based on feedback control of digital ERK responses." Plos Biology **3**(11): 1925-1938.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-402.

Altschul, S. F., J. C. Wootton, et al. (2005). "Protein database searches using compositionally adjusted substitution matrices." FEBS J **272**(20): 5101-9.

Alves, R., F. Antunes, et al. (2006). "Tools for kinetic modeling of biochemical networks." Nature Biotechnology **24**(6): 667-672.

Arora, S., K. Kaur, et al. (2003). "Indian medicinal plants as a reservoir of protective phytochemicals." Teratog Carcinog Mutagen **Suppl 1**: 295-300.

Ashburn, T. T. and K. B. Thor (2004). "Drug repositioning: Identifying and developing new uses for existing drugs." Nature Reviews Drug Discovery **3**(8): 673-683.

Badr, G. and B. J. Oommen (2006). "On optimizing syntactic pattern recognition using tries and AI-based heuristic-search strategies." IEEE Trans Syst Man Cybern B Cybern **36**(3): 611-22.

Bairoch, A., R. Apweiler, et al. (2005). "The Universal Protein Resource (UniProt)." Nucleic Acids Res **33**(Database issue): D154-9.

Baumgartner, W. A., Jr., K. B. Cohen, et al. (2007). "Manual curation is not sufficient for annotation of genomic databases." Bioinformatics **23**(13): i41-8.

Bendele, A. M., E. S. Chlipala, et al. (2000). "Combination benefit of treatment with the cytokine inhibitors interleukin-1 receptor antagonist and PEGylated soluble tumor necrosis factor receptor type I in animal models of rheumatoid arthritis." Arthritis Rheum **43**(12): 2648-59.

Berman, H. M., J. Westbrook, et al. (2000). "The Protein Data Bank." Nucleic Acids Res **28**(1): 235-42.

Beyer, A., S. Bandyopadhyay, et al. (2007). "Integrating physical and genetic maps: from genomes to interaction networks." Nature Reviews Genetics **8**(9): 699-710.

Bhaskar, H., D. C. Hoyle, et al. (2006). "Machine learning in bioinformatics: a brief survey and recommendations for practitioners." Comput Biol Med **36**(10): 1104-25.

Birtwistle, M. R., M. Hatakeyama, et al. (2007). "Ligand-dependent responses of the ErbB signaling network: experimental and modeling analyses." Molecular Systems Biology **3**.

Bolos, J. (2005). "Structure-activity relationships of p38 mitogen-activated protein kinase inhibitors." Mini Rev Med Chem **5**(9): 857-68.

Bolzan, A. D. and M. S. Bianchi (2002). "Genotoxicity of streptozotocin." Mutat Res **512**(2-3): 121-34.

Bornstein, B. J., S. M. Keating, et al. (2008). "LibSBML: an API library for SBML." Bioinformatics **24**(6): 880-881.

Bostrom, J., A. Hogner, et al. (2006). "Do structurally similar ligands bind in a similar fashion?" J Med Chem **49**(23): 6716-25.

Bradley, B. P., B. Kalampanayil, et al. (2009). "Protein expression profiling." Methods Mol Biol **519**: 455-68.

Breiman, L. (2001). "Random Forests." Machine Learning **45**(1): 5-32.

Breitkreutz, B. J., C. Stark, et al. (2008). "The BioGRID interaction database: 2008 update." Nucleic Acids Research **36**: D637-D640.

Buurma, N. J. and I. Haq (2007). "Advances in the analysis of isothermal titration calorimetry data for ligand-DNA interactions." Methods **42**(2): 162-172.

Byvatov, E., U. Fechner, et al. (2003). "Comparison of support vector machine and artificial neural network systems for drug/nondrug classification." J Chem Inf Comput Sci **43**(6): 1882-9.

Cases, I., D. G. Pisano, et al. (2007). "CARGO: a web portal to integrate customized biological information." Nucleic Acids Res **35**(Web Server issue): W16-20.

Cash, G. G. (2001). "Prediction of the genotoxicity of aromatic and heteroaromatic amines using electrotopological state indices." Mutat Res **491**(1-2): 31-7.

Cerami, E. G., G. D. Bader, et al. (2006). "cPath: open source software for collecting, storing, and querying biological pathways." Bmc Bioinformatics **7**.

Chen, C., L. X. Chen, et al. (2008). "Predicting protein structural class based on multi-features fusion." J Theor Biol **253**(2): 388-92.

Chen, X., Y. Fang, et al. (2007). "Does drug-target have a likeness?" Methods Inf Med **46**(3): 360-6.

Chen, X., C. Y. Ung, et al. (2003). "Can an in silico drug-target search method be used to probe potential mechanisms of medicinal plant ingredients?" Natural Product Reports **20**(4): 432-444.

Chen, Y. P. and F. Chen (2008). "Identifying targets for drug discovery using bioinformatics." Expert Opin Ther Targets **12**(4): 383-9.

Chen, Y. Z. and C. Y. Ung (2001). "Prediction of potential toxicity and side effect protein targets of a small molecule by a ligand-protein inverse docking approach." Journal of Molecular Graphics & Modelling **20**(3): 199-218.

Chen, Y. Z. and D. G. Zhi (2001). "Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule." Proteins **43**(2): 217-26.

Chin, S. F., Y. Wang, et al. (2006). "Using array-comparative genomic hybridization to define molecular portraits of primary breast cancers." Oncogene.

Cho, C. R., M. Labow, et al. (2006). "The application of systems biology to drug discovery." Curr Opin Chem Biol **10**(4): 294-302.

Chou, K. C. and H. B. Shen (2006). "Large-scale plant protein subcellular location prediction." J Cell Biochem.

Cristianini, N., and Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. New York, Cambridge University Press, Cambridge.

Cruz-Monteagudo, M., M. N. Cordeiro, et al. (2008). "Computational chemistry approach for the early detection of drug-induced idiosyncratic liver toxicity." J Comput Chem **29**(4): 533-49.

Custer, L. L. and K. S. Sweder (2008). "The role of genetic toxicology in drug discovery and optimization." Curr Drug Metab **9**(9): 978-85.

Dandekar, T., B. Snel, et al. (1998). "Conservation of gene order: a fingerprint of proteins that physically interact." Trends Biochem Sci **23**(9): 324-8.

Davidov, E., J. Holland, et al. (2003). "Advancing drug discovery through systems biology." Drug Discov Today **8**(4): 175-83.

Deckard, A., F. T. Bergmann, et al. (2006). "Supporting the SBML layout extension." Bioinformatics **22**(23): 2966-2967.

Deep, G., M. Dhiman, et al. (2005). "Chemopreventive potential of Triphala (a composite Indian drug) on benzo(a)pyrene induced forestomach tumorigenesis in murine tumor model system." J Exp Clin Cancer Res **24**(4): 555-63.

Doniger, S., T. Hofmann, et al. (2002). "Predicting CNS permeability of drug molecules: comparison of neural network and support vector machine algorithms." J Comput Biol **9**(6): 849-64.

Downward, J. (2001). "The ins and outs of signalling." Nature **411**(6839): 759-62.

Drees, B. L., B. Sundin, et al. (2001). "A protein interaction map for cell polarity development." J Cell Biol **154**(3): 549-71.

Dunkel, M., M. Fullbeck, et al. (2006). "SuperNatural: a searchable database of available natural compounds." Nucleic Acids Res **34**(Database issue): D678-83.

Dutta, D., R. Guha, et al. (2006). "Scalable partitioning and exploration of chemical spaces using geometric hashing." J Chem Inf Model **46**(1): 321-33.

Ecker, G. F., T. Stockner, et al. (2008). "Computational models for prediction of interactions with ABC-transporters." Drug Discov Today **13**(7-8): 311-7.

Ekins, S., J. Mestres, et al. (2007). "In silico pharmacology for drug discovery: applications to targets and beyond." Br J Pharmacol **152**(1): 21-37.

Evers, A., G. Hessler, et al. (2005). "Virtual screening of biogenic amine-binding G-protein coupled receptors: comparative evaluation of protein- and ligand-based virtual screening protocols." J Med Chem **48**(17): 5448-65.

Fabrizi, F., S. Bunnapradist, et al. (2003). "Kinetics of hepatitis C virus load during hemodialysis: novel perspectives." J Nephrol **16**(4): 467-75.

Fang, H., W. Tong, et al. (2001). "Structure-activity relationships for a large diverse set of natural, synthetic, and environmental estrogens." Chem Res Toxicol **14**(3): 280-94.

Ferrer, I., R. Blanco, et al. (2002). "Active, phosphorylation-dependent MAP kinases, MAPK/ERK, SAPK/JNK and p38, and specific transcription factor substrates are differentially expressed following systemic administration of kainic acid to the adult rat." Acta Neuropathol **103**(4): 391-407.

Fisher, J. H. G. a. P. L. a. D. (1990). "Models of incremental concept formation." Artificial Intelligence **40**: 11-61.

Franch, T., M. Petersen, et al. (1999). "Antisense RNA regulation in prokaryotes: Rapid RNA/RNA interaction facilitated by a general U-turn loop structure." Journal of Molecular Biology **294**(5): 1115-1125.

Frank, I. H. W. a. E. (2005). Data Mining: Practical machine learning tools and techniques. San Francisco, Morgan Kaufmann

Freund, Y. a. M., L. (1999). "The alternating decision tree learning algorithm." Proceeding of the Sixteenth International Conference on Machine Learning: 124-133.

Friedberg, I., T. Kaplan, et al. (2000). "Evaluation of PSI-BLAST alignment accuracy in comparison to structural alignments." Protein Sci **9**(11): 2278-84.

Funahashi, A., Y. Matsuoka, et al. (2008). "CellDesigner 3.5: A versatile modeling tool for biochemical networks." Proceedings of the Ieee **96**(8): 1254-1265.

Fussenegger, M., J. E. Bailey, et al. (2000). "A mathematical model of caspase function in apoptosis." Nat Biotechnol **18**(7): 768-74.

Gabdoulline, R. R., M. Stein, et al. (2007). "qPIPSA: relating enzymatic kinetic parameters and interaction fields." BMC Bioinformatics **8**: 373.

Galperin, M. Y. and G. R. Cochrane (2009). "Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009." Nucleic Acids Res **37**(Database issue): D1-4.

Gavin, A. C., M. Bosche, et al. (2002). "Functional organization of the yeast proteome by systematic analysis of protein complexes." Nature **415**(6868): 141-7.

Goldsmith, D. R. and A. J. Wagstaff (2005). "Spotlight on etanercept in plaque psoriasis and psoriatic arthritis." BioDrugs **19**(6): 401-3.

Goto, S., Y. Okuno, et al. (2002). "LIGAND: database of chemical compounds and reactions in biological pathways." Nucleic Acids Res **30**(1): 402-4.

Hamelryck, T. (2009). "Probabilistic models and machine learning in structural bioinformatics." Stat Methods Med Res.

Han, L. Y., C. J. Zheng, et al. (2007). "Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness." Drug Discov Today **12**(7-8): 304-13.

Harrill, A. H. and I. Rusyn (2008). "Systems biology and functional genomics approaches for the identification of cellular responses to drug toxicity." Expert Opin Drug Metab Toxicol **4**(11): 1379-89.

Haugh, J. M., A. Wells, et al. (2000). "Mathematical modeling of epidermal growth factor receptor signaling through the phospholipase C pathway: mechanistic insights and predictions for molecular interventions." Biotechnol Bioeng **70**(2): 225-38.

He, L., P. C. Jurs, et al. (2003). "Predicting the genotoxicity of polycyclic aromatic compounds from molecular structure with different classifiers." Chem Res Toxicol **16**(12): 1567-80.

Helma, C., T. Cramer, et al. (2004). "Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds." J Chem Inf Comput Sci **44**(4): 1402-11.

Hoops, S., S. Sahle, et al. (2006). "COPASI- A COmplex PAthway SImulator." Bioinformatics **22**(24): 3067-3074.

Hopkins, A. L. and C. R. Groom (2002). "The druggable genome." Nature Reviews Drug Discovery **1**(9): 727-730.

Hoshino, M., Y. Kawata, et al. (1996). "Interaction of GroEL with conformational states of horse cytochrome c." Journal of Molecular Biology **262**(4): 575-587.

Hu, J. Y. and T. Aizawa (2003). "Quantitative structure-activity relationships for estrogen receptor binding affinity of phenolic chemicals." Water Res **37**(6): 1213-22.

Huang, J., G. Ma, et al. (2007). "Identifying P-glycoprotein substrates using a support vector machine optimized by a particle swarm." J Chem Inf Model **47**(4): 1638-47.

Huang, N., B. K. Shoichet, et al. (2006). "Benchmarking sets for molecular docking." J Med Chem **49**(23): 6789-801.

Huber, W. and F. Mueller (2006). "Biomolecular interaction analysis in drug discovery using surface plasmon resonance technology." Current Pharmaceutical Design **12**(31): 3999-4021.

Hucka, M., A. Finney, et al. (2003). "The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models." Bioinformatics **19**(4): 524-531.

Ichikawa, H., Y. Nakamura, et al. (2007). "Anticancer drugs designed by mother nature: Ancient drugs but modern targets." Current Pharmaceutical Design **13**(33): 3400-3416.

Jacob, L., B. Hoffmann, et al. (2008). "Virtual screening of GPCRs: an in silico chemogenomics approach." Bmc Bioinformatics **9**: 363.

Jacobs, M. N. (2004). "In silico tools to aid risk assessment of endocrine disrupting chemicals." Toxicology **205**(1-2): 43-53.

Jayaraman, K. S. (2006). Break with tradition. Nature. **442:** 342-343.

Jenwitheesuk, E., J. A. Horst, et al. (2008). "Novel paradigms for drug discovery: computational multitarget screening." Trends Pharmacol Sci **29**(2): 62-71.

Ji, H. F., D. X. Kong, et al. (2007). "Distribution patterns of small-molecule ligands in the protein universe and implications for origin of life and drug discovery." Genome Biology **8**(8).

Ji, Z. L., X. Chen, et al. (2003). "KDBI: Kinetic Data of Bio-molecular Interactions database." Nucleic Acids Res **31**(1): 255-7.

Jiao, D. and D. J. Wild (2009). "Extraction of CYP Chemical Interactions from Biomedical Literature Using Natural Language Processing Methods." J Chem Inf Model.

Joshi-Tope, G., M. Gillespie, et al. (2005). "Reactome: a knowledgebase of biological pathways." Nucleic Acids Res **33**(Database issue): D428-32.

Karakoc, E., A. Cherkasov, et al. (2006). "Distance based algorithms for small biomolecule classification and structural similarity search." Bioinformatics **22**(14): e243-51.

Karp, P. D., C. A. Ouzounis, et al. (2005). "Expansion of the BioCyc collection of pathway/genome databases to 160 genomes." Nucleic Acids Research **33**(19): 6083-6089.

Kartal, M. (2007). "Intellectual property protection in the natural product drug discovery, traditional herbal medicine and herbal medicinal products." Phytotherapy Research **21**(2): 113-119.

Kaul, P. N. (1998). "Drug discovery: past, present and future." Prog Drug Res **50**: 9-105.

Kazius, J., R. McGuire, et al. (2005). "Derivation and validation of toxicophores for mutagenicity prediction." J Med Chem **48**(1): 312-20.

Kerrien, S., Y. Alam-Faruque, et al. (2007). "IntAct - open source resource for molecular interaction data." Nucleic Acids Research **35**: D561-D565.

Khare, C. P. (2007). Indian Medicinal Plants: An Illustrated Dictionary. New York, Springer.

Kibler, D. A. a. D. (1991). "Instance-based learning algorithms." Machine Learning **6**: 37-66.

Kirkland, D., M. Aardema, et al. (2005). "Evaluation of the ability of a battery of three in vitro genotoxicity tests to discriminate rodent carcinogens and non-carcinogens I. Sensitivity, specificity and relative predictivity." Mutat Res **584**(1-2): 1-256.

Kitano, H. (2007). "Innovation - A robustness-based approach to systems-oriented drug design." Nature Reviews Drug Discovery **6**(3): 202-210.

Kloppel, S., C. M. Stonnington, et al. (2008). "Accuracy of dementia diagnosis: a direct comparison between radiologists and a computerized method." Brain **131**(Pt 11): 2969-74.

Koehn, F. E. and G. T. Carter (2005). "The evolving role of natural products in drug discovery." Nat Rev Drug Discov **4**(3): 206-220.

Kohavi, R. (1996). "Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid." Second International Conference on Knoledge Discovery and Data Mining: 202-207.

Kong, D. X., X. J. Li, et al. (2009). "Where is the hope for drug discovery? Let history tell the future." Drug Discov Today **14**(3-4): 115-9.

Kononenko, I. (2001). "Machine learning for medical diagnosis: history, state of the art and perspective." Artif Intell Med **23**(1): 89-109.

Korneeva, N. L., B. J. Lamphear, et al. (2001). "Characterization of the two eIF4A-binding sites on human eIF4G-1." Journal of Biological Chemistry **276**(4): 2872-2879.

Kortagere, S., D. Chekmarev, et al. (2008). "New predictive models for blood-brain barrier permeability of drug-like molecules." Pharm Res **25**(8): 1836-45.

Kuhn, M., C. von Mering, et al. (2008). "STITCH: interaction networks of chemicals and proteins." Nucleic Acids Research **36**: D684-D688.

Kuiper, S., L. A. Joosten, et al. (1998). "Different roles of tumour necrosis factor alpha and interleukin 1 in murine streptococcal cell wall arthritis." Cytokine **10**(9): 690-702.

Kumar, P., B. C. Han, et al. (2009). "Update of KDBI: Kinetic Data of Bio-molecular Interaction database." Nucleic Acids Res **37**(Database issue): D636-41.

Larranaga, P., B. Calvo, et al. (2006). "Machine learning in bioinformatics." Brief Bioinform **7**(1): 86-112.

Le Novere, N., B. Bornstein, et al. (2006). "BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems." Nucleic Acids Research **34**: D689-D691.

Legrain, P., J. Wojcik, et al. (2001). "Protein-protein interaction maps: a lead towards cellular functions." Trends in Genetics **17**(6): 346-352.

Lengeler, J. W. (2000). "Metabolic networks: a signal-oriented approach to cellular models." Biol Chem **381**(9-10): 911-20.

Li, H., C. Y. Ung, et al. (2005). "Prediction of genotoxicity of chemical compounds by statistical learning methods." Chem Res Toxicol **18**(6): 1071-80.

Li, Z. R., L. Y. Han, et al. (2007). "MODEL - Molecular descriptor lab: A web-based server for computing structural and physicochemical features of compounds." Biotechnology and Bioengineering **97**(2): 389-396.

Li, Z. R., H. H. Lin, et al. (2006). "PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence." Nucleic Acids Res **34**(Web Server issue): W32-7.

Linding, R., L. J. Jensen, et al. (2008). "NetworKIN: a resource for exploring cellular phosphorylation networks." Nucleic Acids Research **36**: D695-D699.

Lo, S. L., C. Z. Cai, et al. (2005). "Effect of training datasets on support vector machine prediction of protein-protein interactions." Proteomics **5**(4): 876-884.

Ludemann, A., D. Weicht, et al. (2004). "PaVESy: Pathway visualization and editing system." Bioinformatics **20**(16): 2841-2844.

Ma, X. H., R. Wang, et al. (2008). "Advances in machine learning prediction of toxicological properties and adverse drug reactions of pharmaceutical agents." Curr Drug Saf **3**(2): 100-14.

Maini, S. R. (2004). "Infliximab treatment of rheumatoid arthritis." Rheum Dis Clin North Am **30**(2): 329-47, vii.

Mak, H. C., M. Daly, et al. (2007). "CellCircuits: a database of protein network models." Nucleic Acids Research **35**: D538-D545.

Malik, N. N. (2008). "Drug discovery: past, present and future." Drug Discov Today **13**(21-22): 909-12.

Matter, H., K. H. Baringhaus, et al. (2001). "Computational approaches towards the rational design of drug-like compound libraries." Comb Chem High Throughput Screen **4**(6): 453-75.

Mattioni, B. E., G. W. Kauffman, et al. (2003). "Predicting the genotoxicity of secondary and aromatic amines using data subsetting to generate a model ensemble." J Chem Inf Comput Sci **43**(3): 949-63.

McArdle, B. M. and R. J. Quinn (2007). "Identification of protein fold topology shared between different folds inhibited by natural products." Chembiochem **8**(7): 788-798.

McGuffin, M. (2008). "Should Herbal Medicines Be Regulated as Drugs[quest]." Clin Pharmacol Ther **83**(3): 393-395.

Mewes, H. W., D. Frishman, et al. (2002). "MIPS: a database for genomes and protein sequences." Nucleic Acids Res **30**(1): 31-4.

Miglino, O., H. H. Lund, et al. (1995). "Evolving mobile robots in simulated and real environments." Artif Life **2**(4): 417-34.

Mishra, L. C., B. B. Singh, et al. (2001). "Ayurveda: A historical perspective and principles of the traditional healthcare system in India." Alternative Therapies in Health and Medicine **7**(2): 36-42.

Muller, A., R. M. MacCallum, et al. (1999). "Benchmarking PSI-BLAST in genome annotation." J Mol Biol **293**(5): 1257-71.

Muster, W., A. Breidenbach, et al. (2008). "Computational toxicology in drug development." Drug Discov Today **13**(7-8): 303-10.

Nakazato, T., T. Takinaka, et al. (2008). "BioCompass: a novel functional inference tool that utilizes MeSH hierarchy to analyze groups of genes." In Silico Biol **8**(1): 53-61.

Newman, D. J. (2008). "Natural products as leads to potential drugs: an old process or the new hope for drug discovery?" J Med Chem **51**(9): 2589-99.

Nicholson, J. K. and I. D. Wilson (2003). "Opinion: understanding 'global' systems biology: metabonomics and the continuum of metabolism." Nat Rev Drug Discov **2**(8): 668-76.

Nicolas, R., M. Donizelli, et al. (2007). "SBMLeditor: effective creation of models in the Systems Biology Markup Language (SBML)." Bmc Bioinformatics **8**.

Okuda, S., T. Yamada, et al. (2008). "KEGG Atlas mapping for global analysis of metabolic pathways." Nucleic Acids Research **36**: W423-W426.

Oliver, D. E., G. Bhalotia, et al. (2004). "Tools for loading MEDLINE into a local relational database." Bmc Bioinformatics **5**.

Palladino, M. A., F. R. Bahjat, et al. (2003). "Anti-TNF-alpha therapies: the next generation." Nat Rev Drug Discov **2**(9): 736-46.

Paolini, G. V., R. H. B. Shapland, et al. (2006). "Global mapping of pharmacological space." Nature Biotechnology **24**(7): 805-815.

Park, K. and D. Kim (2008). "Binding similarity network of ligand." Proteins: Structure, Function and Genetics **71**(2): 960-971.

Parsons, H. M., C. Ludwig, et al. (2007). "Improved classification accuracy in 1- and 2-dimensional NMR metabolomics data using the variance stabilising generalised logarithm transformation." BMC Bioinformatics **8**: 234.

Pellecchia, M. (2005). "Solution nuclear magnetic resonance spectroscopy techniques for probing intermolecular interactions." Chemistry & Biology **12**(9): 961-971.

Pellegrini, M., E. M. Marcotte, et al. (1999). "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles." Proc Natl Acad Sci U S A **96**(8): 4285-8.

Pestian, J., P. Matykiewicz, et al. (2008). "Using natural language processing to classify suicide notes." AMIA Annu Symp Proc: 1091.

Petrelli, A. and S. Giordano (2008). "From single- to multi-target drugs in cancer therapy: when aspecificity becomes an advantage." Curr Med Chem **15**(5): 422-32.

Pettus, L. H. and R. P. Wurz (2008). "Small molecule p38 MAP kinase inhibitors for the treatment of inflammatory diseases: novel structures and developments during 2006-2008." Curr Top Med Chem **8**(16): 1452-67.

Phizicky, E., P. I. Bastiaens, et al. (2003). "Protein analysis on a proteomic scale." Nature **422**(6928): 208-15.

Qian, J., J. Lin, et al. (2003). "Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data." Bioinformatics **19**(15): 1917-26.

Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. San Francisco, CA, USA Morgan Kaufmann Publishers Inc.

Quinlan, R. (1986). "Induction of decision trees." Machine Learning **1**(1): 81-106.

Quinlan, R. (1993). "C4.5: Programs for Machine Learning." Morgan Kaufmann Publishers San Mateo, CA.

Quinlan, R. J. (1992). "Learning with Continuous Classes." 5th Australian Joint Conference on Artificial Intelligence: 343-348.

Raghavachari, B., A. Tasneem, et al. (2008). "DOMINE: a database of protein domain interactions." Nucleic Acids Research **36**: D656-D661.

Rao, S. (2004). "Database-driven Web sites: a case study with software in biotechnology and bioengineering." Electronic Library **22**(4): 357-361.

Rasmussen, M. K., L. Iversen, et al. (2008). "IL-8 and p53 are inversely regulated through JNK, p38 and NF-kappaB p65 in HepG2 cells during an inflammatory response." Inflamm Res **57**(7): 329-39.

Ratti, E. and D. Trist (2001). "The continuing evolution of the drug discovery process in the pharmaceutical industry." Farmaco **56**(1-2): 13-9.

Rojas, I., M. Golebiewski, et al. (2007). "Storing and annotating of kinetic data." In Silico Biol **7**(2 Suppl): S37-44.

Sahm, H., L. Eggeling, et al. (2000). "Pathway analysis and metabolic engineering in Corynebacterium glutamicum." Biol Chem **381**(9-10): 899-910.

Salwinski, L., C. S. Miller, et al. (2004). "The Database of Interacting Proteins: 2004 update." Nucleic Acids Research **32**: D449-D451.

Sandhya, T., K. M. Lathika, et al. (2006). "Potential of traditional ayurvedic formulation, Triphala, as a novel anticancer drug." Cancer Lett **231**(2): 206-14.

Sasagawa, S., Y. Ozaki, et al. (2005). "Prediction and validation of the distinct dynamics of transient and sustained ERK activation." Nature Cell Biology **7**(4): 365-U31.

Sayers, E. W., T. Barrett, et al. (2009). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Res **37**(Database issue): D5-15.

Schadt, E. E., S. H. Friend, et al. (2009). "A network view of disease and compound screening." Nat Rev Drug Discov **8**(4): 286-95.

Schmidt, H., G. Drews, et al. (2007). "SBML export interface for the systems biology toolbox for MATLAB." Bioinformatics **23**(10): 1297-1298.

Schoeberl, B., C. Eichler-Jonsson, et al. (2002). "Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors." Nature Biotechnology **20**(4): 370-375.

Schomburg, I., A. Chang, et al. (2002). "BRENDA, enzyme data and metabolic information." Nucleic Acids Res **30**(1): 47-9.

Schrattenholz, A. and V. Soskic (2008). "What does systems biology mean for drug development?" Curr Med Chem **15**(15): 1520-8.

Seringhaus, M. R. and M. B. Gerstein (2007). "Publishing perishing? Towards tomorrow's information architecture." Bmc Bioinformatics **8**: 17.

Shacham, S., Y. Marantz, et al. (2004). "PREDICT modeling and in-silico screening for G-protein coupled receptors." Proteins **57**(1): 51-86.

Sivakumaran, S., S. Hariharaputran, et al. (2003). "The Database of Quantitative Cellular Signaling: management and analysis of chemical kinetic models of signaling networks." Bioinformatics **19**(3): 408-15.

Smit, H. F., H. J. Woerdenbag, et al. (1995). "AYURVEDIC HERBAL DRUGS WITH POSSIBLE CYTOSTATIC ACTIVITY." Journal of Ethnopharmacology **47**(2): 75-84.

Snyder, R. D., G. S. Pearl, et al. (2004). "Assessment of the sensitivity of the computational programs DEREK, TOPKAT, and MCASE in the prediction of the genotoxicity of pharmaceutical molecules." Environ Mol Mutagen **43**(3): 143-58.

Sollano, J. A., J. M. Kirsch, et al. (2008). "The economics of drug discovery and the ultimate valuation of pharmacotherapies in the marketplace." Clin Pharmacol Ther **84**(2): 263-6.

Southan, C., P. Varkonyi, et al. (2007). "Complementarity between public and commercial databases: new opportunities in medicinal chemistry informatics." Curr Top Med Chem **7**(15): 1502-8.

Southan, C., P. Varkonyi, et al. (2007). "Complementarity between public and commercial databases: New opportunities in medicinal chemistry informatics." Current Topics in Medicinal Chemistry **7**(15): 1502-1508.

Steinbeck, C., Y. Han, et al. (2003). "The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics." J Chem Inf Comput Sci **43**(2): 493-500.

Steinbeck, C., C. Hoppe, et al. (2006). "Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics." Curr Pharm Des **12**(17): 2111-20.

Stephens, S. M., J. Y. Chen, et al. (2005). "Oracle Database 10g, a platform for BLAST search and Regular Expression pattern matching in life sciences." Nucleic Acids Res **33**(Database issue): D675-9.

Stockinger, H., T. Attwood, et al. (2008). "Experience using web services for biological sequence analysis." Brief Bioinform **9**(6): 493-505.

Suresh, B. C. V., S. M. E. Babar, et al. (2008). "Kinetic analysis of the MAPK and PI3K/Akt signaling pathways." Molecules and Cells **25**(3): 397-406.

Tibshirani, J. F. a. T. H. a. R. (2000). "Additive logistic regression : A statistical view of boosting." Annals of statistics **28**(2): 337-407.

Tong, W., Q. Xie, et al. (2004). "Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity." Environ Health Perspect **112**(12): 1249-54.

Tweats, D. J., D. Blakey, et al. (2007). "Report of the IWGT working group on strategies and interpretation of regulatory in vivo tests I. Increases in micronucleated bone marrow cells in rodents that do not indicate genotoxic hazards." Mutat Res **627**(1): 78-91.

Ung, C. Y., H. Li, et al. (2008). "Simulation of the regulation of EGFR endocytosis and EGFR-ERK signaling by endophilin-mediated RhoA-EGFR crosstalk." Febs Letters **582**(15): 2283-2290.

Valentini, G., R. Tagliaferri, et al. (2009). "Computational intelligence and machine learning in bioinformatics." Artif Intell Med **45**(2-3): 91-6.

van den Berg, R. A., H. C. Hoefsloot, et al. (2006). "Centering, scaling, and transformations: improving the biological information content of metabolomics data." BMC Genomics **7**: 142.

van den Broek, B., M. C. Noom, et al. (2005). "DNA-tension dependence of restriction enzyme activity reveals mechanochemical properties of the reaction pathway." Nucleic Acids Res **33**(8): 2676-84.

Van Gompel, J., F. Woestenborghs, et al. (2005). "An assessment of the utility of the yeast GreenScreen assay in pharmaceutical screening." Mutagenesis **20**(6): 449-54.

Vasanthanathan, P., O. Taboureau, et al. (2009). "Classification of cytochrome P450 1A2 inhibitors and noninhibitors by machine learning techniques." Drug Metab Dispos **37**(3): 658-64.

Vidovszky, T. J., W. Smith, et al. (2006). "Robotic cholecystectomy: learning curve, advantages, and limitations." J Surg Res **136**(2): 172-8.

von Mering, C., L. J. Jensen, et al. (2007). "STRING 7 - recent developments in the integration and prediction of protein interactions." Nucleic Acids Research **35**: D358-D362.

Walters, W. P. and M. A. Murcko (2002). "Prediction of 'drug-likeness'." Adv Drug Deliv Rev **54**(3): 255-71.

Willett, P., J. M. Barnard, et al. (1998). "Chemical similarity searching." Journal of Chemical Information and Computer Sciences **38**(6): 983-996.

Xu, Y., Z. Wang, et al. (2009). "MBA: a literature mining system for extracting biomedical abbreviations." BMC Bioinformatics **10**: 14.

Xue, Y., Z. R. Li, et al. (2004). "Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents." J Chem Inf Comput Sci **44**(5): 1630-8.

Xue, Y., C. W. Yap, et al. (2004). "Prediction of P-glycoprotein substrates by a support vector machine approach." J Chem Inf Comput Sci **44**(4): 1497-505.

Xue, Y., C. W. Yap, et al. (2004). "Prediction of P-glycoprotein substrates by a support vector machine approach." Journal of Chemical Information and Computer Sciences **44**(4): 1497-1505.

Y.Z. Chen, D. G. Z. (2001). "Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule." Proteins: Structure, Function, and Genetics **43**(2): 217-226.

Yang, H., I. Spasic, et al. (2009). "A text mining approach to the prediction of disease status from clinical discharge summaries." J Am Med Inform Assoc **16**(4): 596-600.

Yap, C. W., C. Z. Cai, et al. (2004). "Prediction of torsade-causing potential of drugs by support vector machine approach." Toxicol Sci **79**(1): 170-7.

Yap, C. W. and Y. Z. Chen (2005). "Quantitative Structure-Pharmacokinetic Relationships for drug distribution properties by using general regression neural network." J Pharm Sci **94**(1): 153-68.

Yildirim, M. A., K. I. Goh, et al. (2007). "Drug-target network." Nature Biotechnology **25**(10): 1119-1126.

Yousef, M., L. Showe, et al. (2009). "A study of microRNAs in silico and in vivo: bioinformatics approaches to microRNA discovery and target identification." FEBS J **276**(8): 2150-6.

Yu, X. B., D. K. Xu, et al. (2006). "Label-free detection methods for protein microarrays." Proteomics **6**(20): 5493-5503.

Yuryev, A., Z. Mulyukov, et al. (2006). "Automatic pathway building in biological association networks." BMC Bioinformatics **7**: 171.

Zanzoni, A., L. Montecchi-Palazzi, et al. (2002). "MINT: a Molecular INTeraction database." Febs Letters **513**(1): 135-140.

Zeng, Q., C. L. Teo, et al. (2008). "Collaborative path planning for a robotic wheelchair." Disabil Rehabil Assist Technol **3**(6): 315-24.

Zernov, V. V., K. V. Balakin, et al. (2003). "Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions." J Chem Inf Comput Sci **43**(6): 2048-56.

Zhang, H., Q. Y. Chen, et al. (2008). "In silico prediction of mitochondrial toxicity by using GA-CG-SVM approach." Toxicol In Vitro.

Zhang, H. L., H. H. Lin, et al. (2008). "Prediction of antibiotic resistance proteins from sequence-derived properties irrespective of sequence similarity." Int J Antimicrob Agents **32**(3): 221-6.

Zhang, X., A. Crespo, et al. (2008). "Turning promiscuous kinase inhibitors into safer drugs." Trends Biotechnol **26**(6): 295-301.

Zhang, Y. H., L. H. Fang, et al. (2003). "In vitro inhibitory effects of bergenin and norbergenin on bovine adrenal tyrosine hydroxylase." Phytother Res **17**(8): 967-9.

Zhao, M., P. Rattanatamrong, et al. (2008). "BMI cyberworkstation: enabling dynamic data-driven brain-machine interface research through cyberinfrastructure." Conf Proc IEEE Eng Med Biol Soc **2008**: 646-9.

Zhou, S., E. Chan, et al. (2004). "Therapeutic drugs that behave as mechanism-based inhibitors of cytochrome P450 3A4." Curr Drug Metab **5**(5): 415-42.

Zhou, Y. Y., B. Zhou, et al. (2007). "Large-scale annotation of small-molecule libraries using public databases." Journal of Chemical Information and Modeling **47**(4): 1386-1394.

Zi, Z. and E. Klipp (2006). "SBML-PET: A systems biology markup language-based parameter estimation tool." Bioinformatics **22**(21): 2704-2705.

Appendix

Table A1: Total 522 Molecular descriptors, selected 100 descriptors are highlighted. Machine learning classification studies were performed using either total 522 descriptors or the selected 100 descriptors.

Constitutational Descriptors	
1	**Number of Atoms**
2	Number of Heavy atoms
3	Number of H atoms
4	Number of B atoms
5	**Number of C atoms**
6	**Number of N atoms**
7	**Number of O atoms**
8	Number of F atoms
9	Number of P atoms
10	Number of S atoms
11	Number of Cl atoms
12	Number of Br atoms
13	Number of I atoms
14	**NUmber of Bonds**
15	Number of non-H Bonds
16	**Number of rings**
17	**Molecular weight(MW)**
18	Average molecular weight(AMW)
19	**Number of H-bond donnor**
20	**Number of H-bond acceptor**
21	**Sanderson electronegativity Sum**
22	**Number of rotable bonds**
23	Number of 3-member rings
24	Number of 4-member ings
25	Number of 7-member rings
26	Number of 5-member non-aromatic rings
27	Number of 6-member non-aromatic rings
28	Number of 5-member aromatic rings
29	Number of 6-member aromatic rings
30	Number of heterocyclic rings
31	**Number of N heterocyclic rings**
32	**Number of O heterocyclic rings**
33	Number of S heterocyclic rings
34	Number of Aziridine rings
35	Number of Oxirane rings
36	Number of Thiirane rings
37	Number of Azetidine rings
38	Number of Oxetane rings
39	Number of Thietane rings

40	Number of Pyrrolidine rings
41	Number of Oxolane rings
42	Number of Thiophane rings
43	Number of Pyrrole rings
44	Number of Furane rings
45	Number of Thiophene rings
46	Number of Pyrazole rings
47	Number of Imidazole rings
48	Number of Oxazole rings
49	Number of Isoxazole rings
50	Number of Thiazole rings
51	Number of Isothiazole rings
52	Number of Benzene rings
53	Number of Pyridazine rings
54	Number of Pyrimidine rings
55	Number of Pyrazine rings
56	Number of 1,3,5-trizine rings
57	Number of 1,2,4-trizine rings
58	Number of 1,2,3-trizine rings
Charge Descriptors	
59	**Total absolute atomic charge**
60	Total squared atomic charge
61	**Charge Polarization**
62	**Topological electronic index TE**
63	Topological electronic index CTE
64	Maximum negative charges
65	Maximum positive charges
66	**Local dipol index**
67	Total negative charges
68	Total positive charges
69	Submolecular Polarity Parameter
70	Second-order submolecular polarity parameter
71	**Relative positive charge**
72	**Relative negative charge**
Electronic-topological descriptors	
73	**0th Electronic-topological**
74	**1th Electronic-topological**
75	**2th Electronic-topological**
76	Electron charge density index
77	**Electron charge density connectivity index**
78	Hydrophobic alogp
79	Molecular polarizability
Topological descriptors	
80	**Schultz molecular topological index**

81	**Gutman molecular topological index**
Topological charge index	
82	Topological charge index G1
83	Topological charge index G2
84	Topological charge index G3
85	Topological charge index G4
86	Topological charge index G5
Mean topological charge index	
87	Mean topological charge index J1
88	Mean topological charge index J2
89	Mean topological charge index J3
90	Mean topological charge index J4
91	Mean topological charge index J5
92	Global topological charge index J
93	**Wiener index**
94	Mean Wiener index
95	**Harary index**
96	**Gravitational topological index**
Molecular path count	
97	**Molecular path count of length 1**
98	**Molecular path count of length 2**
99	**Molecular path count of length 3**
100	**Molecular path count of length 4**
101	**Molecular path count of length 5**
102	**Molecular path count of length 6**
103	**Total path count**
Sum of E-State of atom type	
104	Sum of Estate of atom type sLi
105	Sum of Estate of atom type ssBe
106	Sum of Estate of atom type ssssBe
107	Sum of Estate of atom type ssBH
108	Sum of Estate of atom type sssB
109	Sum of Estate of atom type ssssB
110	**Sum of Estate of atom type sCH3**
111	**Sum of Estate of atom type dCH2**
112	**Sum of Estate of atom type ssCH2**
113	Sum of Estate of atom type tCH
114	**Sum of Estate of atom type dsCH**
115	**Sum of Estate of atom type aaCH**
116	**Sum of Estate of atom type sssCH**
117	Sum of Estate of atom type ddC
118	Sum of Estate of atom type tsC
119	**Sum of Estate of atom type dssC**
120	**Sum of Estate of atom type aasC**

121	**Sum of Estate of atom type aaaC**
122	**Sum of Estate of atom type sssC**
123	**Sum of Estate of atom type sNH3**
124	**Sum of Estate of atom type sNH2**
125	**Sum of Estate of atom type ssNH2**
126	**Sum of Estate of atom type dNH**
127	**Sum of Estate of atom type ssNH**
128	**Sum of Estate of atom type aaNH**
129	Sum of Estate of atom type tN
130	Sum of Estate of atom type sssNH
131	**Sum of Estate of atom type dsN**
132	**Sum of Estate of atom type aaN**
133	**Sum of Estate of atom type sssN**
134	**Sum of Estate of atom type ddsN**
135	Sum of Estate of atom type aasN
136	**Sum of Estate of atom type aOH**
137	**Sum of Estate of atom type sOH**
138	Sum of Estate of atom type dO
139	**Sum of Estate of atom type ssO**
140	Sum of Estate of atom type aaO
141	Sum of Estate of atom type F
142	Sum of Estate of atom type ssSiH2
143	Sum of Estate of atom type ssSiH2
144	Sum of Estate of atom type sssSiH
145	Sum of Estate of atom type ssssSi
146	Sum of Estate of atom type sPH2
147	Sum of Estate of atom type ssPH
148	Sum of Estate of atom type sssP
149	Sum of Estate of atom type dsssP
150	Sum of Estate of atom type ssssP
151	**Sum of Estate of atom type sSH**
152	Sum of Estate of atom type dS
153	Sum of Estate of atom type ssS
154	Sum of Estate of atom type aaS
155	Sum of Estate of atom type dssS
156	Sum of Estate of atom type ddssS
157	Sum of Estate of atom type sCl
158	Sum of Estate of atom type sGeH3
159	Sum of Estate of atom type ssGeH2
160	Sum of Estate of atom type sssGeH
161	Sum of Estate of atom type ssssGe
162	Sum of Estate of atom type sAsH2
163	Sum of Estate of atom type ssAsH
164	Sum of Estate of atom type sssAs

165	Sum of Estate of atom type sssdAs
166	Sum of Estate of atom type ssssAs
167	Sum of Estate of atom type sSeH
168	Sum of Estate of atom type dSe
169	Sum of Estate of atom type ssSe
170	Sum of Estate of atom type aaSe
171	Sum of Estate of atom type dssSe
172	Sum of Estate of atom type ddssSe
173	Sum of Estate of atom type sBr
174	Sum of Estate of atom type sSnH3
175	Sum of Estate of atom type ssSnH2
176	Sum of Estate of atom type sssSnH
177	Sum of Estate of atom type ssssSn
178	Sum of Estate of atom type sI
179	Sum of Estate of atom type sPbH3
180	Sum of Estate of atom type ssPbH2
181	Sum of Estate of atom type sssPbH
182	Sum of Estate of atom type ssssPb
183	Sum of Estate of atom type unknown
184	**Sum of Estate of all heavy atoms**
185	**Sum of Estate of all C atoms**
186	Sum of Estate of all halogen atoms
187	**Sum of Estate of all hetero atoms**
188	**Sum of Estate of H-bond acceptors**
189	Average of Estate values
190	Maximum of Estate values
191	Minimum of Estate values
Sum of H E-State of atom type	
192	**Sum of H Estate of atom type HsOH**
193	**Sum of H Estate of atom type HdNH**
194	**Sum of H Estate of atom type HsSH**
195	**Sum of H Estate of atom type HsNH2**
196	**Sum of H Estate of atom type HssNH**
197	**Sum of H Estate of atom type HaaNH**
198	Sum of H Estate of atom type HsNH3p
199	Sum of H Estate of atom type HssNH2p
200	Sum of H Estate of atom type HsssNHp
201	**Sum of H Estate of atom type HtCH**
202	**Sum of H Estate of atom type HdCH2**
203	**Sum of H Estate of atom type HdsCH**
204	**Sum of H Estate of atom type HaaCH**
205	Sum of H Estate of atom type HCHnX
206	**Sum of H Estate of atom type HCsats**
207	Sum of H Estate of atom type HCsatu

208	Sum of H Estate of atom type Havin
209	Sum of H Estate of atom type Hother
210	Sum of H Estate of atom type Hmisc
211	**Sum of H Estate of H-bond donors**
212	Xu index
213	Modified Xu Index
214	**Balaban Index J**
215	Platt Number
216	LOG of superpendentic index
217	First Zagreb Index(M1)
218	Second Zagreb Index(M2)
219	First Modified Zagreb Index
220	Second Modified Zagreb Index
221	Quadratic index(Q)
222	**0th edge connectivity index**
223	**Edge connectivity index**
224	**Extened edge connectivity inndex**
225	2th spectral moment
226	3th spectral moment
227	4th spectral moment
228	5th spectral moment
229	6th spectral moment
230	7th spectral moment
231	8th spectral moment
232	9th spectral moment
233	10th spectral moment
Moreau-Broto topological autocorrelation	
Atomic mass weighted Moreau-Broto	
234	Atomic mass weighted Moreau-Broto lagged 0
235	Atomic mass weighted Moreau-Broto lagged 1
236	Atomic mass weighted Moreau-Broto lagged 2
237	Atomic mass weighted Moreau-Broto lagged 3
238	Atomic mass weighted Moreau-Broto lagged 4
239	Atomic mass weighted Moreau-Broto lagged 5
240	Atomic mass weighted Moreau-Broto lagged 6
241	Atomic mass weighted Moreau-Broto lagged 7
242	Atomic mass weighted Moreau-Broto lagged 8
243	Atomic mass weighted Moreau-Broto lagged 9
244	Atomic mass weighted Moreau-Broto lagged 10
Electronegativity weighted moreau-Broto	
245	Electronegativity weighted Moreau-Broto lagged 0
246	Electronegativity weighted Moreau-Broto lagged 1
247	Electronegativity weighted Moreau-Broto lagged 2
248	Electronegativity weighted Moreau-Broto lagged 3

249	Electronegativity weighted Moreau-Broto lagged 4
250	Electronegativity weighted Moreau-Broto lagged 5
251	Electronegativity weighted Moreau-Broto lagged 6
252	Electronegativity weighted Moreau-Broto lagged 7
253	Electronegativity weighted Moreau-Broto lagged 8
254	Electronegativity weighted Moreau-Broto lagged 9
255	Electronegativity weighted Moreau-Broto lagged 10
VDW radius weighted Moreau-Broto	
256	VDW radius weighted Moreau-Broto lagged 0
257	VDW radius weighted Moreau-Broto lagged 1
258	VDW radius weighted Moreau-Broto lagged 2
259	VDW radius weighted Moreau-Broto lagged 3
260	VDW radius weighted Moreau-Broto lagged 4
261	VDW radius weighted Moreau-Broto lagged 5
262	VDW radius weighted Moreau-Broto lagged 6
263	VDW radius weighted Moreau-Broto lagged 7
264	VDW radius weighted Moreau-Broto lagged 8
265	VDW radius weighted Moreau-Broto lagged 9
266	VDW radius weighted Moreau-Broto lagged 10
Estate Values weighted Moreau-Broto	
267	E-State weighted Moreau-Broto lagged 0
268	E-State weighted Moreau-Broto lagged 1
269	E-State weighted Moreau-Broto lagged 2
270	E-State weighted Moreau-Broto lagged 3
271	E-State weighted Moreau-Broto lagged 4
272	E-State weighted Moreau-Broto lagged 5
273	E-State weighted Moreau-Broto lagged 6
274	E-State weighted Moreau-Broto lagged 7
275	E-State weighted Moreau-Broto lagged 8
276	E-State weighted Moreau-Broto lagged 9
277	E-State weighted Moreau-Broto lagged 10
polarizability weighted Moreau-Broto	
278	Polarizability mass weighted Moreau-Broto lagged 0
279	Polarizability mass weighted Moreau-Broto lagged 1
280	Polarizability mass weighted Moreau-Broto lagged 2
281	Polarizability mass weighted Moreau-Broto lagged 3
282	Polarizability mass weighted Moreau-Broto lagged 4
283	Polarizability mass weighted Moreau-Broto lagged 5
284	Polarizability mass weighted Moreau-Broto lagged 6
285	Polarizability mass weighted Moreau-Broto lagged 7
286	Polarizability mass weighted Moreau-Broto lagged 8
287	Polarizability mass weighted Moreau-Broto lagged 9
288	Polarizability weighted Moreau-Broto lagged 10
Van der Waals volume weighted Moreau-Broto	

289	VDW volume weighted Moreau-Broto lagged 0
290	VDW volume weighted Moreau-Broto lagged 1
291	VDW volume weighted Moreau-Broto lagged 2
292	VDW volume weighted Moreau-Broto lagged 3
293	VDW volume weighted Moreau-Broto lagged 4
294	VDW volume weighted Moreau-Broto lagged 5
295	VDW volume weighted Moreau-Broto lagged 6
296	VDW volume weighted Moreau-Broto lagged 7
297	VDW volume weighted Moreau-Broto lagged 8
298	VDW volume weighted Moreau-Broto lagged 9
299	VDW volume weighted Moreau-Broto lagged 10
Moran topological autocorrelation	
Atomic mass weighted Moran	
300	Atomic mass weighted moran lagged 1
301	Atomic mass weighted moran lagged 2
302	Atomic mass weighted moran lagged 3
303	Atomic mass weighted moran lagged 4
304	Atomic mass weighted moran lagged 5
305	Atomic mass weighted moran lagged 6
306	Atomic mass weighted moran lagged 7
307	Atomic mass weighted moran lagged 8
308	Atomic mass weighted moran lagged 9
309	Atomic mass weighted moran lagged 10
Electronegativity weighted Moran	
310	Electronegativity weighted moran lagged 1
311	Electronegativity weighted moran lagged 2
312	Electronegativity weighted moran lagged 3
313	Electronegativity weighted moran lagged 4
314	Electronegativity weighted moran lagged 5
315	Electronegativity weighted moran lagged 6
316	Electronegativity weighted moran lagged 7
317	Electronegativity weighted moran lagged 8
318	Electronegativity weighted moran lagged 9
319	Electronegativity weighted moran lagged 10
VDW radius weighted Moran	
320	VDW radius weighted moran lagged 1
321	VDW radius weighted moran lagged 2
322	VDW radius weighted moran lagged 3
323	VDW radius weighted moran lagged 4
324	VDW radius weighted moran lagged 5
325	VDW radius weighted moran lagged 6
326	VDW radius weighted moran lagged 7
327	VDW radius weighted moran lagged 8
328	VDW radius weighted moran lagged 9

329	VDW radius weighted moran lagged 10
Estate weighted Moran	
330	E-State weighted moran lagged 1
331	E-State weighted moran lagged 2
332	E-State weighted moran lagged 3
333	E-State weighted moran lagged 4
334	E-State weighted moran lagged 5
335	E-State weighted moran lagged 6
336	E-State weighted moran lagged 7
337	E-State weighted moran lagged 8
338	E-State weighted moran lagged 9
339	E-State weighted moran lagged 10
Polarizability weighted Moran	
340	Polarizability mass weighted moran lagged 1
341	Polarizability mass weighted moran lagged 2
342	Polarizability mass weighted moran lagged 3
343	Polarizability mass weighted moran lagged 4
344	Polarizability mass weighted moran lagged 5
345	Polarizability mass weighted moran lagged 6
346	Polarizability mass weighted moran lagged 7
347	Polarizability mass weighted moran lagged 8
348	Polarizability mass weighted moran lagged 9
349	Polarizability mass weighted moran lagged 10
VDW volume weighted Moran	
350	VDW volume weighted moran lagged 1
351	VDW volume weighted moran lagged 2
352	VDW volume weighted moran lagged 3
353	VDW volume weighted moran lagged 4
354	VDW volume weighted moran lagged 5
355	VDW volume weighted moran lagged 6
356	VDW volume weighted moran lagged 7
357	VDW volume weighted moran lagged 8
358	VDW volume weighted moran lagged 9
359	VDW volume weighted moran lagged 10
Geary topological autocorrelation	
Atomic mass weighted Geary	
360	Atomic mass weighted Geary 1
361	Atomic mass weighted Geary 2
362	Atomic mass weighted Geary 3
363	Atomic mass weighted Geary 4
364	Atomic mass weighted Geary 5
365	Atomic mass weighted Geary 6
366	Atomic mass weighted Geary 7
367	Atomic mass weighted Geary 8

368	Atomic mass weighted Geary 9
369	Atomic mass weighted Geary10
Electronegativity weighted Geary	
370	Electronegativity weighted Geary 1
371	Electronegativity weighted Geary 2
372	Electronegativity weighted Geary 3
373	Electronegativity weighted Geary 4
374	Electronegativity weighted Geary 5
375	Electronegativity weighted Geary 6
376	Electronegativity weighted Geary 7
377	Electronegativity weighted Geary 8
378	Electronegativity weighted Geary 9
379	Electronegativity weighted Geary10
VDW radius weighted Geary	
380	VDW radius weighted Geary 1
381	VDW radius weighted Geary 2
382	VDW radius weighted Geary 3
383	VDW radius weighted Geary 4
384	VDW radius weighted Geary 5
385	VDW radius weighted Geary 6
386	VDW radius weighted Geary 7
387	VDW radius weighted Geary 8
388	VDW radius weighted Geary 9
389	VDW radius weighted Geary10
E-state weighted Geary	
390	Estate weighted Geary 1
391	Estate weighted Geary 2
392	Estate weighted Geary 3
393	Estate weighted Geary 4
394	Estate weighted Geary 5
395	Estate weighted Geary 6
396	Estate weighted Geary 7
397	Estate weighted Geary 8
398	Estate weighted Geary 9
399	Estate weighted Geary10
Polarizability weighted Geary	
400	Polarizability weighted Geary 1
401	Polarizability weighted Geary 2
402	Polarizability weighted Geary 3
403	Polarizability weighted Geary 4
404	Polarizability weighted Geary 5
405	Polarizability weighted Geary 6
406	Polarizability weighted Geary 7
407	Polarizability weighted Geary 8

408	Polarizability weighted Geary 9
409	polarizability weighted Geary10
VDW volume weighted Geary	
410	VDW volume weighted Geary 1
411	VDW volume weighted Geary 2
412	VDW volume weighted Geary 3
413	VDW volume weighted Geary 4
414	VDW volume weighted Geary 5
415	VDW volume weighted Geary 6
416	VDW volume weighted Geary 7
417	VDW volume weighted Geary 8
418	VDW volume weighted Geary 9
419	polarizability weighted Geary10
420	0th Kier-Hall connectivity index
421	1th Kier-Hall connectivity index
422	Mean Randic Connectivity index
423	2th Kier-Hall connectivity index
424	Simple topological index by Narumi
425	Harmonic topological index by Narumi
426	Geometric topological index by Narumi
427	Arithmetic topological index by Narumi
428	**0th valence connectivity index**
429	**1th valence connectivity index**
430	**2th valence connectivity index**
431	**0th order delta chi index**
432	**1th order delta chi index**
433	**2th order delta chi index**
434	**Pogliani index**
Solvation connectivity index	
435	**0th Solvation connectivity index**
436	**1th Solvation connectivity index**
437	**2th Solvation connectivity index**
438	**1th order Kier shape index**
439	**2th order Kier shape index**
440	**3th order Kier shape index**
441	**1th order Kappa alpha shape index**
442	**2th order Kappa alpha shape index**
443	**3th order Kappa alpha shape index**
444	**Kier Molecular Flexibility Index**
Topological distance related	
445	**Topological radius**
446	Topological diameter
447	**Eccentricity**
448	Average atom eccentricity

449	Mean eccentricity deviation
450	Average distance degree
451	Mean distance degree deviation
452	Unipolarity
453	Rouvary index Irouv
454	**Centralization**
455	Variation
456	Dispersion
457	Log of PRS INDEX
458	**Graph-theoretical shape coefficient**
459	RDSQ ondex
460	RDCHI index
461	Optimized 1th connectivity index
462	Logp from connectivity
BCUT highest of mass	
463	BCUT 1th highest of mass
464	BCUT 2th highest of mass
465	BCUT 3th highest of mass
466	BCUT 4th highest of mass
467	BCUT 5th highest of mass
BCUT lowest of mass	
468	BCUT 1th lowest of mass
469	BCUT 2th lowest of mass
470	BCUT 3th lowest of mass
471	BCUT 4th lowest of mass
472	BCUT 5th lowest of mass
BCUT highest of electronegativity	
473	BCUT 1th highest of electronegativity
474	BCUT 2th highest of electronegativity
475	BCUT 3th highest of electronegativity
476	BCUT 4th highest of electronegativity
477	BCUT 5th highest of electronegativity
BCUT lowest of electronegativity	
478	BCUT 1th lowest of electronegativity
479	BCUT 2th lowest of electronegativity
480	BCUT 3th lowest of electronegativity
481	BCUT 4th lowest of electronegativity
482	BCUT 5th lowest of electronegativity
BCUT highest of VDW radius	
483	BCUT 1th highest of VDW radius
484	BCUT 2th highest of VDW radius
485	BCUT 3th highest of VDW radius
486	BCUT 4th highest of VDW radius
487	BCUT 5th highest of VDW radius

BCUT lowest of VDW radius	
488	BCUT 1th lowest of VDW radius
489	BCUT 2th lowest of VDW radius
490	BCUT 3th lowest of VDW radius
491	BCUT 4th lowest of VDW radius
492	BCUT 5th lowest of VDW radius
BCUT highest of Estate	
493	BCUT 1th highest of Estate
494	BCUT 2th highest of Estate
495	BCUT 3th highest of Estate
496	BCUT 4th highest of Estate
497	BCUT 5th highest of Estate
BCUT lowest of Estate	
498	BCUT 1th lowest of Estate
499	BCUT 2th lowest of Estate
500	BCUT 3th lowest of Estate
501	BCUT 4th lowest of Estate
502	BCUT 5th lowest of Estate
BCUT highest of polarizability	
503	BCUT 1th highest of Polarizability
504	BCUT 2th highest of Polarizability
505	BCUT 3th highest of Polarizability
506	BCUT 4th highest of Polarizability
507	BCUT 5th highest of Polarizability
BCUT lowest of polarizability	
508	BCUT 1th lowest of Polarizability
509	BCUT 2th lowest of Polarizability
510	BCUT 3th lowest of Polarizability
511	BCUT 4th lowest of Polarizability
512	BCUT 5th lowest of Polarizability
BCUT highest of VDW volume	
513	BCUT 1th highest of VDW volume
514	BCUT 2th highest of VDW volume
515	BCUT 3th highest of VDW volume
516	BCUT 4th highest of VDW volume
517	BCUT 5th highest of VDW volume
BCUT lowest of VDW volume	
518	BCUT 1th lowest of VDW volume
519	BCUT 2th lowest of VDW volume
520	BCUT 3th lowest of VDW volume
521	BCUT 4th lowest of VDW volume
522	BCUT 5th lowest of VDW volume

Table A2: Literature sources of p38 inhibitors collection

Title: Biphenyl amide p38 kinase inhibitors 2: Optimization and SAR **Journal:** Bioorganic & Medicinal Chemistry Letters (2007)
Title: Molecular modeling studies of phenoxypyrimidinyl imidazoles as p38 kinase inhibitors using QSAR and docking **Journal:** European Journal of Medicinal Chemistry xx (2007) 1-9
Title: Benzimidazoles and Imidazo[4,5-b]pyridines as Potent p38a MAP Kinase Inhibitors with Excellent in vivo Antiinflammatory propertie **Journal:** Bioorganic & Medicinal Chemistry Letters (2007)
Title: Biphenyl amide p38 kinase inhibitors 1: Discovery and binding mode **Journal:** Bioorganic & Medicinal Chemistry Letters (2007)
Title: CoMFA and docking studies on triazolopyridine oxazole derivatives as p38 MAP kinase inhibitors **Journal:** European Journal of Medicinal Chemistry xx (2007) 1-9
Title: Trimethylsilylpyrazoles as novel inhibitors of p38 MAP kinase: A new use of silicon bioisosteres in medicinal chemistry **Journal:** Bioorganic & Medicinal Chemistry Letters, Volume 17, Issue 2, 15 January 2007, Pages 354-357
Title: Synthesis, Crystal Structure, and Activity of Pyrazole-Based Inhibitors of p38 Kinase **Journal:** J. Med. Chem. 2007; 50(23); 5712-5719
Title: Synthesis, Biological Testing, and Binding Mode Prediction of 6,9-Diarylpurin-8-ones as p38 MAP Kinase Inhibitors **Journal:** J. Med. Chem.; (Article); 2007; 50(9); 2060-2066
Title: Design, Synthesis, and Anti-inflammatory Properties of Orally Active 4-(Phenylamino)-pyrrolo[2,1-f][1,2,4]triazine p38a Mitogen-Activated Protein Kinase Inhibitors **Journal:** J. Med. Chem.; 2007; ASAP Article;
Title: Synthesis and Biological Activity of Quinolinone and Dihydroquinolinone p38 MAP Kinase Inhibitors

Journal: Bioorganic & Medicinal Chemistry Letters (2006)

Title: Discovery and design of benzimidazolone based inhibitors of p38 MAP kinase

Journal: Bioorganic & Medicinal Chemistry Letters 16 (2006) 6316-6320

Title: p38 MAP kinase inhibitors. Part 6: 2-Arylpyridazin-3-ones as templates for inhibitor design

Journal: Bioorganic & Medicinal Chemistry Letters 16 (2006) 5809-5813

Title: p38 MAP kinase inhibitors. Part 3: SAR on 3,4-dihydropyrimido-[4,5-d]pyrimidin-2-ones and 3,4-dihydropyrido[4,3-d]-pyrimidin-2-ones

Journal: Bioorganic & Medicinal Chemistry Letters 16 (2006) 4400-4404

Title: Successful Screening of Large Encoded Combinatorial Libraries Leading to the Discovery of Novel p38 MAP Kinase Inhibitors

Journal: Combinatorial Chemistry & High Throughput Screening, 2006, 9, 351-358

Title: New Approaches to the Treatment of Inflammatory Disorders Small Molecule Inhibitors of p38 MAP Kinase

Journal: Current Topics in Medicinal Chemistry, 2006, 6, 113-149

Title: Discovery and Characterization of Triaminotriazine Aniline Amides as Highly Selective p38 Kinase Inhibitors

Journal: THE JOURNAL OF PHARMACOLOGY AND EXPERIMENTAL THERAPEUTICS Vol. 318, No. 2

Title: Inhibitors of unactivated p38 MAP kinase

Journal: Bioorganic & Medicinal Chemistry Letters 16 (2006) 6102-6106

Title: p38 MAP kinase inhibitors. Part 5: Discovery of an orally bio-available and highly efficacious compound based on the 7-amino-naphthyridone scaffol

Journal: Bioorganic & Medicinal Chemistry Letters, Volume 16, Issue 20, 15 October 2006, Pages 5468-5471

Title: Pyrazoloheteroaryls: Novel p38a MAP kinase inhibiting scaffolds with oral activity

Journal: Bioorganic & Medicinal Chemistry Letters, Volume 16, Issue 2, 15 January 2006, Pages 262-266

Title: p38 MAP kinase inhibitors: Metabolically stabilized piperidine-substituted quinolinones and naphthyridinones

Journal: Bioorganic & Medicinal Chemistry Letters, Volume 16, Issue 1, 1 January 2006, Pages

64-68
Title: Design, Synthesis, and Biological Evaluation of Phenylamino-Substituted 6,11-Dihydro-dibenzo[b,e]oxepin-11-ones and Dibenzo[a,d]cycloheptan-5-ones: Novel p38 MAP Kinase Inhibitors **Journal:** J. Med. Chem.; (Brief Article); 2006; 49(26); 7912-7915
Title: Discovery of S-[5-Amino-1-(4-fluorophenyl)-1H-pyrazol-4-yl]-[3-(2,3-dihydroxypropoxy)phenyl]methanone (RO3201195), an Orally Bioavailable and Highly Selective Inhibitor of p38 Map Kinase **Journal:** J. Med. Chem.; (Article); 2006; 49(5); 1562-1575.
Title: Novel 2-Aminopyrimidine Carbamates as Potent and Orally Active Inhibitors of Lck:Synthesis, SAR, and in Vivo Antiinflammatory Activity **Journal:** J. Med. Chem. 2006, 49, 4981-4991
Title: Structure–activity relationships of triazolopyridine oxazole p38 inhibitors: Identification of candidates for clinical development **Journal:** Bioorganic & Medicinal Chemistry Letters 16 (2006) 4339–4344
Title: Design of Potent and Selective 2-Aminobenzimidazole-Based p38r MAP Kinase Inhibitors with Excellent in Vivo Efficacy **Journal:** J. Med. Chem. 2005, 48, 2270-2273
Title: Design of Potent and Selective 2-Aminobenzimidazole-based p38a MAP Kinase Inhibitors with Excellent in vivo Efficacy **Journal:** J. Med. Chem., 2005, 48 (7), pp 2270–2273
Title: Discovery of Highly Selective Inhibitors of p38alpha **Journal:** Current Topics in Medicinal Chemistry, 2005, 5, 941-951
Title: The Discovery of Novel Chemotypes of p38 Kinase Inhibitors **Journal:** Current Topics in Medicinal Chemistry, 2005, 5, 953-965
Title: Small Molecule p38 Inhibitors: Novel Structural Features and Advances from 2002-2005 **Journal:** Current Topics in Medicinal Chemistry, 2005, 5, 967-985
Title: P38 MAP Kinase Inhibitors: Evolution of Imidazole-Based and Pyrido-Pyrimidin-2-One Lead Classes

Journal: Current Topics in Medicinal Chemistry, 2005, 5, 987-1003

Title: Structural Comparison of p38 Inhibitor-Protein Complexes: A Review of Recent p38 Inhibitors Having Unique Binding Interactions

Journal: Current Topics in Medicinal Chemistry, 2005, 5, 1005-1016

Title: Pathway to the Clinic: Inhibition of P38 MAP Kinase. A Review of Ten Chemotypes Selected for Development

Journal: Current Topics in Medicinal Chemistry, 2005, 5, 1017-1029

Title: 5-Cyanopyrimidine Derivatives as a Novel Class of Potent, Selective, and Orally Active Inhibitors of p38r MAP Kinase

Journal: J. Med. Chem. 2005, 48, 6261-6270

Title: Synthesis and Biological Activities of 4-Phenyl-5-pyridyl-1,3-thiazole Derivatives as p38 MAP Kinase Inhibitors

Journal: Chem. Pharm. Bull. 53(4) 410—418 (2005)

Title: Novel Inhibitor of p38 MAP Kinase as an Anti-TNF-r Drug: Discovery of N-[4-[2-Ethyl-4-(3-methylphenyl)-1,3-thiazol-5-yl]-2-pyridyl]benzamide (TAK-715) as a Potent and Orally Active Anti-Rheumatoid Arthritis Agent

Journal: J. Med. Chem. 2005, 48, 5966-5979

Title: Design and synthesis of potent pyridazine inhibitors of p38 MAP kinase

Journal: Bioorganic & Medicinal Chemistry Letters 15 (2005) 2409-2413

Title: Theoretical and Experimental Design of Atypical Kinase Inhibitors: Application to p38 MAP Kinase

Journal: J. Med. Chem.; (Article); 2005; 48(18); 5728-5737

Title: Identification of Novel p38 MAP Kinase Inhibitors Using Fragment-Based Lead Generation

Journal: J. Med. Chem.; (Article); 2005; 48(2); 414-426

Title: Novel p38 inhibitors with potent oral efficacy in several models of rheumatoid arthritis

Journal: Bioorganic & Medicinal Chemistry Letters 14 (2004) 3595-3599

Title: SAR of benzoylpyridines and benzophenones as p38alpha MAP kinase inhibitors with oral activity

Journal: Bioorganic & Medicinal Chemistry Letters 14 (2004) 3601-3605

Title: The Discovery of Orally Active Triaminotriazine Aniline Amides as Inhibitors of p38 MAP Kinase **Journal:** J. Med. Chem. 2004, 47, 6283-6291
Title: Novel, potent and selective anilinoquinazoline and anilinopyrimidine inhibitors of p38 MAP kinase **Journal:** Bioorganic & Medicinal Chemistry Letters 14 (2004) 5389-5394
Title: A novel series of p38 MAP kinase inhibitors for the potential treatment of rheumatoid arthritis **Journal:** Bioorganic & Medicinal Chemistry Letters 14 (2004) 5383-5387
Title: SAR of benzoylpyridines and benzophenones as p38a MAP kinase inhibitors with oral activity **Journal:** Bioorganic & Medicinal Chemistry Letters 14 (2004) 3601-3605
Title: Benzimidazolone p38 inhibitors **Journal:** Bioorganic & Medicinal Chemistry Letters 14 (2004) 919-923
Title: Novel and potent transforming growth factor beta type I receptor kinase domain inhibitor: 7-amino 4-(2-pyridin-2-yl-5,6-dihydro-4H-pyrrolo[1,2-b]pyrazol-3-yl)-quinolines **Journal:** Bioorganic & Medicinal Chemistry Letters, Volume 14, Issue 13, 5 July 2004, Pages 3585-3588
Title: Synthesis and activity of new aryl- and heteroaryl-substituted 5,6-dihydro-4H-pyrrolo[1,2-b]pyrazole inhibitors of the transforming growth factor-ß type I receptor kinase domain **Journal:** Bioorganic & Medicinal Chemistry Letters, Volume 14, Issue 13, 5 July 2004, Pages 3581-3584
Title: The development of new bicyclic pyrazole-based cytokine synthesis inhibitors **Journal:** Bioorganic & Medicinal Chemistry Letters, Volume 14, Issue 19, 4 October 2004, Pages 4945-4948
Title: Indole-Based Heterocyclic Inhibitors of p38 MAP Kinase: Designing a Conformationally Restricted Analogue **Journal:** Bioorganic & Medicinal Chemistry Letters 13 (2003) 3087-3090
Title: p38MAP Kinase Inhibitors. Part 1: Design and Development of a New Class of Potent and Highly Selective Inhibitors Based on 3,4-Dihydropyrido[3,2-d]pyrimidone Scaffold

Journal: Bioorganic & Medicinal Chemistry Letters 13 (2003) 273-276

Title: Design and Synthesis of Potent, Orally Bioavailable Dihydroquinazolinone Inhibitors of p38 MAP Kinase

Journal: Bioorganic & Medicinal Chemistry Letters 13 (2003) 277-280

Title: p38 Inhibitors: Piperidine- and 4-Aminopiperidine-Substituted Naphthyridinones, Quinolinones, and Dihydroquinazolinones

Journal: Bioorganic & Medicinal Chemistry Letters 13 (2003) 467-470

Title: Design and Synthesis of 4-Azaindoles as Inhibitors of p38 MAP Kinase

Journal: J. Med. Chem. 2003, 46, 4702-4713

Title: Imidazopyrimidines, Potent Inhibitors of p38 MAP Kinase

Journal: Bioorganic & MedicinalChemistry Letters 13 (2003) 347-350

Title: Synthesis and Structure-Activity Relationship of Aminobenzophenones. A Novel Class of p38 MAP Kinase Inhibitors with High Antiinflammatory Activity

Journal: J. Med. Chem. 2003, 46, 5651-5662

Title: Thermal Denaturation: A Method to Rank Slow Binding, High-Affinity P38alpha MAP Kinase Inhibitors

Journal: J. Med. Chem. 2003, 46, 4669-4675

Title: Structure-Activity Relationships of the p38r MAP Kinase Inhibitor 1-(5-tert-Butyl-2-p-tolyl-2H-pyrazol-3-yl)-3-[4-(2-morpholin-4-yl-ethoxy)naphnaphthalen-1-yl]urea (BIRB 796)

Journal: J. Med. Chem. 2003, 46, 4676-4686

Title: N-Phenyl-N-purin-6-yl Ureas: The Design and Synthesis of P38 MAP Kinase Inhibitors

Journal: Bioorganic & Medicinal Chemistry Letters 13 (2003) 1191-1194

Title: Indole-Based Heterocyclic Inhibitors of p38 MAP Kinase:Designing a Conformationally Restricted Analogue

Journal: Bioorganic & Medicinal Chemistry Letters 13 (2003) 3087-3090

Title: The Kinetics of Binding to p38MAP Kinase by Analogues of BIRB 796

Journal: Bioorganic & Medicinal Chemistry Letters 13 (2003) 3101-3104

Title: Hybrid-Designed Inhibitors of p38 MAP Kinase Utilizing N-Arylpyridazinones

Journal: J. Med. Chem.; (Letter); 2003; 46(3); 349-352.

Title: The Structure of JNK3 in Complex with Small Molecule Inhibitors: Structural Basis for Potency and Selectivity

Journal: Chemistry & Biology, Vol. 10, 705–712, August, 2003

Title: SAR of 2,6-Diamino-3,5-difluoropyridinyl Substituted Heterocycles as Novel p38MAP Kinase Inhibitors

Journal: Bioorganic & Medicinal Chemistry Letters 12 (2002) 2109-2112

Title: Pyridinylimidazole Based p38 MAP Kinase Inhibitors

Journal: Current Topics in Medicinal Chemistry 2002, 2, 1011-1020

Title: The Non-Diaryl Heterocycle Classes of p38 MAP Kinase Inhibitors

Journal: Current Topics in Medicinal Chemistry 2002, 2, 1021-1035

Title: Pyrazole Urea-Based Inhibitors of p38 MAP Kinase: From Lead Compound to Clinical Candidate

Journal: J. Med. Chem. 2002, 45, 2994-3008

Title: Synthesis and Pharmacological Characterization of a Potent, Orally Active p38 Kinase Inhibitor

Journal: Bioorganic & Medicinal Chemistry Letters 12 (2002) 1559-1562

Title: Pyridazine based inhibitors of p38 MAPK

Journal: Bioorganic & Medicinal Chemistry Letters, Volume 12, Issue 4, 25 February 2002, Pages 689-692

Title: An Algorithm-Directed Two-Component Library Synthesized Via Solid-Phase Methodology Yielding Potent and Orally Bioavailable p38 MAP Kinase Inhibitors

Journal: J. Med. Chem.; (Article); 2002; 45(11); 2173-2184

Title: Phenoxypyrimidine Inhibitors of p38 Kinase: Synthesis and Statistical Evaluation of the p38 Inhibitory Potencies of a Series of 1-(Piperidin-4-yl)-4-(4-fluorophenyl)-5-(2-phenoxypyrimidin-4-yl) Imidazoles

Journal: Bioorganic & Medicinal Chemistry Letters 11 (2001) 1123-1126

Title: The discovery of RPR 200765A, a p38 MAP kinase inhibitor displaying a good oral anti-arthritic efficacy

Journal: Bioorganic & Medicinal Chemistry 9 (2001) 537-554

Title: RPR203494 a pyrimidine analogue of the p38 inhibitor RPR200765A with an improved in vitro potency

Journal: Bioorganic & MedicinalChemistry Letters 11 (2001) 693-696

Title: Pyrimidinylimidazole inhibitors of p38: cyclic N-1 imidazole substituents enhance p38 kinase inhibition and oral activity

Journal: Bioorganic & Medicinal Chemistry Letters, Volume 11, Issue 21, 5 November 2001, Pages 2867-2870

Title: Inhibition of BCRP-mediated drug efflux by fumitremorgin-type indolyl diketopiperazines

Journal: Bioorganic & Medicinal Chemistry Letters, Volume 11, Issue 1, 8 January 2001, Pages 9-12

Title: SAR of 4-Hydroxypiperidine and Hydroxyalkyl Substituted Heterocycles as Novel p38 Map Kinase Inhibitors

Journal: Bioorganic & Medicinal Chemistry Letters 10 (2000) 1261±1264

Title: Discovery of a New Class of p38 Kinase Inhibitors

Journal: Bioorganic & Medicinal Chemistry Letters 10 (2000) 2047±2050

Title: 1-Phenyl-5-pyrazolyl ureas: potent and selective p38 kinase inhibitors

Journal: Bioorganic & Medicinal Chemistry Letters, Volume 10, Issue 18, September 2000, Pages 2051-2054

Title: Design and Synthesis of Potent, Selective, and Orally Bioavailable Tetrasubstituted Imidazole Inhibitors of p38 Mitogen-Activated Protein Kinase

Journal: J. Med. Chem. 1999, 42, 2180-2190 31139

Title: pyrroles and other heterocycles as inhibitors of p38 kinase

Journal: Bioorganic & Medicinal Chemistry Letters 8 (1998) 2689-2694

Title: Potent Inhibitors of The MAP Kinase p38

Journal: Bioorganic & Medicinal Chemistry Letters 8 (1998) 3335-3340

Title: 6-Amino-2-(4-fluorophenyl)-4-methoxy-3-(4-pyridyl)-1H-pyrrolo[2,3-b]pyridine (RWJ 68354): A Potent and Selective p38 Kinase Inhibitor

Journal: J. Med. Chem. 1998, 41, 4196-4198

Title: pyrimidinylimidazole inhibitors of csbp/p38 kinase demonstrating decreased inhibition of hepatic cytochrome p450 enzymes

Journal: Bioorganic & Medicinal Chemistry Letters 8 (1998) 3111-3116

Title: The structural basis for the specificity of pyridinylimidazole inhibitors of p38 MAP kinase

Journal: Chemistry & Biology 1997, Vol 4 No 6

Title: Regulation of Stress-Induced Cytokine Production by Pyridinylimidazoles; Inhibition of CSBP Kinase

Journal: Bioorganic & Medicinal Chemistry, Vol. 5, No. 1, pp 49-64, 1997

Title: 1-Substituted 4-Aryl-5-pyridinylimidazoles: A New Class of Cytokine Suppressive Drugs with Low 5-Lipoxygenase and Cyclooxygenase Inhibitory Potency

Journal: J. Med. Chem. 1996, 39, 3929-3937