

**DEVELOPMENT OF DATABASE AND  
COMPUTATIONAL METHODS FOR DISEASE  
DETECTION AND DRUG DISCOVERY**

**HAN BUCONG**

*(M.Sc, B.Sc, Xiamen Univ.)*

**A THESIS SUBMITTED**

**FOR THE DEGREE OF DOCTOR OF PHILOSOPHY**

**IN COMPUTATION AND SYSTEMS BIOLOGY (CSB)**

**SINGAPORE-MIT ALLIANCE**

**NATIONAL UNIVERSITY OF SINGAPORE**

**2013**

---

## DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety.

I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

HAN BUCONG

Han Bucong  
25 January 2013

---

## **ACKNOWLEDGEMENTS**

First and foremost, I would like to present my sincere gratitude to my Singapore supervisor, Professor Chen Yu Zong, who provides me with excellent guidance, invaluable advices and suggestions throughout my Ph.D study. I have tremendously benefited from his profound knowledge, expertise in scientific research, as well as his enormous support, which will inspire and motivate me to go further in my future professional career. I was delighted to interact with Professor Bruce Tidor by having him as my MIT supervisor. His insights, knowledge and great efforts form the strong support to my adventure in computational biology.

I would also like to thank our present and previous BIDD group members for their insight suggestions and collaborations in my research work. In particulars, I would like to thank Dr. Pankaj Kumar, Dr. Liu Xianghui, Dr. Ma Xiaohua, Dr. Jia jia, Dr. Zhu Feng, Dr. Shi Zhe, Ms Liu Xin, Mr. Zhang Jiangxian, Ms Wei Xiaona etc. and other previous research staffs. BIDD is like a big family and I really enjoy the close friendship among us.

Last, but not the least, I am grateful to my parents and my wife for their encouragement and accompany.

---

## TABLE OF CONTENTS

|   |      |
|---|------|
| DECLARATION.....  | I    |
| ACKNOWLEDGEMENTS.....   | II   |
| TABLE OF CONTENTS .....   | III  |
| SUMMARY .....   | VIII |
| LIST OF TABLES .....  | X    |
| LIST OF FIGURES .....   | XII  |
| LIST OF ACRONYMS .....  | XIV  |
| Chapter 1 Introduction .....  | 1    |
| 1.1 Overview of pathogen detection .....                                    | 1    |
| 1.1.1 Application areas requiring pathogen detection. ....                  | 1    |
| 1.1.2 Brief introduction to pathogens induced infectious diseases.....      | 2    |
| 1.1.3 Conventional pathogen detection methods .....                         | 6    |
| 1.1.4 Molecular pathogen detection methods .....                            | 7    |
| 1.2 Bioinformatics and cheminformatics in drug discovery .....              | 9    |
| 1.3 Introduction of bioinformatics and cheminformatics database development | 11   |
| 1.4 Overview of virtual screening in drug discovery.....                    | 15   |

---

|                             |  |    |
|-----------------------------|--|----|
| 1.5                         | Objective and outline of this thesis .....                 | 27 |
| Chapter 2 Methodology ..... |  | 29 |
| 2.1                         | Database development .....                                 | 29 |
| 2.1.1                       | Database model and rational schema design .....            | 29 |
| 2.1.2                       | Data collection .....                                      | 31 |
| 2.1.3                       | Data integration and organization.....                     | 33 |
| 2.1.4                       | Database management system .....                           | 35 |
| 2.1.5                       | User Interface.....  | 36 |
| 2.2                         | Dataset collection and preprocess for building models..... | 38 |
| 2.2.1                       | Dataset resource .....                                     | 38 |
| 2.2.2                       | Dataset quality .....                                      | 39 |
| 2.2.3                       | Dataset structural diversity .....                         | 40 |
| 2.3                         | Molecular descriptor .....                                 | 41 |
| 2.4                         | Scaling of molecular descriptors .....                     | 45 |
| 2.5                         | Machine learning classification methods .....              | 46 |
| 2.5.1                       | Support vector machine (SVM).....                          | 48 |
| 2.5.2                       | k-nearest neighbors (kNN).....                             | 52 |
| 2.5.3                       | Probabilistic neural network (PNN).....                    | 54 |

---

|   |  |    |
|---|--|----|
| 2.5.4   | Tanimoto similarity searching method .....   | 58 |
| 2.5.5   | Generation of putative negatives .....   | 58 |
| 2.6   | Virtual screening model optimization, validation and performance measurements..... | 62 |
| 2.6.1   | Model optimization and validation .....  | 62 |
| 2.6.2   | Performance evaluation .....   | 63 |
| 2.6.3   | Overfitting problem and its detection .....  | 65 |
| Chapter 3 Development of MicrobPad MD: microbial pathogen diagnostic methods database.....  |  | 66 |
| 3.1   | Introduction .....   | 66 |
| 3.2   | Database construction .....  | 68 |
| 3.3   | Data collection and access.....  | 69 |
| 3.4   | Database usage and validation .....  | 78 |
| 3.5   | Concluding remarks .....   | 80 |
| Chapter 4 Development of TTD: therapeutic target database .....   |  | 82 |
| 4.1   | Introduction .....   | 82 |
| 4.2   | Target and drug data collection and access .....                                   | 84 |
| 4.3   | Ways to access therapeutic targets database .....                                  | 86 |
| 4.4   | Target and drug similarity searching.....  | 93 |
| Chapter 5 Development and experimental test of support vector machines virtual screening method for searching Src inhibitors from large compound libraries..... |  | 97 |
| 5.1   | Introduction .....   | 97 |

---

|   |   |     |
|---|---|-----|
| 5.2   | Materials and methods .....   | 101 |
| 5.2.1   | Compound collections and construction of training and testing datasets..                                    | 101 |
| 5.3   | Results and discussion.....   | 104 |
| 5.3.1   | Performance of SVM, kNN and PNN identification of Src inhibitors based on 5-fold cross validation test..... | 104 |
| 5.3.2   | Virtual screening performance of SVM in searching Src inhibitors from large compound libraries.....         | 108 |
| 5.3.3   | Experimental test of a SVM identified virtual-hit .....   | 111 |
| 5.3.4   | Evaluation of SVM identified MDDR virtual-hits .....  | 112 |
| 5.3.5   | Comparison of virtual screening performance of SVM with those of other virtual screening methods .....      | 115 |
| 5.3.6   | Does SVM select Src inhibitors or membership of compound families?..  | 118 |
| 5.4   | Conclusions .....   | 118 |
| Chapter 6 Support vector machines virtual screening of VEGFR-2 Inhibitors from large compound libraries: model development and experimental test..... |   | 120 |
| 6.1   | Background .....  | 120 |
| 6.2   | Materials and methods .....   | 123 |
| 6.2.1   | Compound collections and construction of training and testing datasets..                                    | 123 |
| 6.3   | Results and Discussion.....   | 127 |
| 6.3.1   | VEGFR-2 Inhibitor prediction Performance of SVM, kNN and PNN evaluated by 5-fold cross validation test..... | 127 |
| 6.3.2   | Virtual screening performance of SVM in searching VEGFR-2 inhibitors from large compound libraries .....    | 132 |
| 6.3.3   | Experimental test of a SVM identified virtual-hit .....   | 135 |

---

|                                    |   |     |
|------------------------------------|---|-----|
| 6.3.4                              | Evaluation of SVM identified MDDR virtual-hits .....  | 136 |
| 6.3.5                              | Comparison of virtual screening performance of SVM with<br>tanimoto-based similarity searching method ..... | 140 |
| 6.3.6                              | Does SVM select VEGFR inhibitors or membership of compound<br>families?.....                                | 142 |
| 6.4                                | Concluding remarks .....  | 142 |
| Chapter 7 Concluding remarks ..... |   | 144 |
| 7.1                                | Major findings and merits .....   | 144 |
| 7.1.1                              | Merits of the development of MicrobPad MD: microbial pathogen<br>diagnostic methods database.....           | 144 |
| 7.1.2                              | Merits of the updates of TTD in facilitating multi-target drug discovery .                                  | 145 |
| 7.1.3                              | Merits of virtual screening model for Src inhibitors.....   | 146 |
| 7.1.4                              | Merits of virtual screening model for VEGFR-2 inhibitors .....  | 147 |
| 7.2                                | Limitations and suggestions for future studies.....   | 147 |
| Reference .....                    |   | 151 |
| Appendices.....                    |   | 183 |
| List of publication .....          |   | 195 |



---

## SUMMARY

Drug discovery is an expensive and time-consuming process which requires large amount of financial investment. Efforts in bioinformatics and cheminformatics are extensively explored to increase the efficiency and reduce costs of drug discovery and development. Bioinformatics tools such as database and computational methods such as machine learning method based virtual screening (VS) have been developed for searching novel lead compounds.

Database development is a promising approach which can accelerate drug discovery by systematically managing and providing medicinal chemicals and biomolecules information with a web accessible interface. This information is a useful resource for further drug discovery application besides a data storing pool. VS is known to contribute to discovery of hits and lead compounds and VS has been investigated and explored intensively. Various tools and applications have been developed according to VS. However, there are many issues of many conventional VS tools including insufficiency of compound diversity coverage, slow screening speed of large compound libraries and high false positive rate. It is demanded to overcome these problems and it would be very useful to develop application of VS tools to discover novel compounds by screening large compound libraries rapidly at good yields and low false-hit rates.

---

In this work, several computational approaches for facilitating disease detection and drug discovery are presented. MicrobPad MD: Microbial pathogen diagnostic methods database is built to provide comprehensive information about the molecular detection for pathogens. It may help accurate, sensitive and low-cost detection of medical pathogens and diagnosis of disease. The updated TTD is expected to be a useful resource in complement to other related databases by providing comprehensive information about the primary targets and drug of the approved, clinical trial, and experimental drugs. These database lead to a better understanding of the disease and benefit for drug discovery.

Src promotes tumour invasion and metastasis, and facilitates VEGF-mediated angiogenesis and survival in endothelial cells. Both Src and VEGFR-2 are very important for disease, particularly cancers. To facilitate drug discovery by saving time and cost in developing novel lead, the machine learning methods are used to build screening models for Src and VEGFR-2 inhibitors. It is shown that SVM based VS tools work efficiently in the discovery of Src, VEGFR-2 inhibitors and other active compounds at low false-hit rates. The virtual hits of models have been tested experimentally to further verify the models. These projects facilitate drug discovery by reducing the cost and time in developing novel drug lead.

---

## LIST OF TABLES

|   |     |
|---|-----|
| <b>Table 1-1</b> Four categories of pathogen inducing infectious human disease. Their infection are briefly described. Examples of the types of pathogens are listed, along with the disease they cause. ....   | 3   |
| <b>Table 1-2</b> The top 10 leading cause of death worldwide in 2008 reported by WHO fact sheet. ....   | 5   |
| <b>Table 1-3</b> Three pathogenic diseases mortality rate in 2013.....  | 6   |
| <b>Table 1-4</b> Popular bioinformatics databases. ....   | 12  |
| <b>Table 1-5</b> Popular chemical databases .....   | 14  |
| <b>Table 1-6</b> Comparison of the reported performance of different VS methods in screening large libraries of compounds (adopted from Han et al[114]). ....   | 23  |
| <b>Table 2-1</b> 98 molecular descriptors used in this work. ....   | 43  |
| <b>Table 2-2</b> Websites that contain codes of machine learning methods .....  | 47  |
| <b>Table 5-1</b> Performance of SVM for identifying Src inhibitors and non-inhibitors evaluated by 5-fold cross validation study.....   | 105 |
| <b>Table 5-2</b> Performance of kNN for identifying Src inhibitors and non-inhibitors evaluated by 5-fold cross validation study.....   | 106 |
| <b>Table 5-3</b> Performance of PNN for identifying Src inhibitors and non-inhibitors evaluated by 5-fold cross validation study.....   | 107 |
| <b>Table 5-4</b> Virtual screening performance of support vector machines for identifying Src inhibitors from large compound libraries .....  | 109 |
| <b>Table 5-5</b> MDDR classes that contain higher percentage ( $\geq 3\%$ ) of SVM virtual-hits and the percentage values. Virtual-hits are identified by SVMs in screening 168K MDDR compounds for Src inhibitors. The total number of SVM identified virtual hits is 1,496..... | 113 |
| <b>Table 5-6</b> Comparison of virtual screening performance of SVM with those of other methods .....   | 117 |

---

|   |     |
|---|-----|
| <b>Table 6-1</b> Performance of SVM for identifying VEGFR-2 inhibitors and non-inhibitors evaluated by 5-fold cross validation study.....   | 129 |
| <b>Table 6-2</b> Performance of kNN for identifying VEGFR-2 inhibitors and non-inhibitors evaluated by 5-fold cross validation study.....   | 130 |
| <b>Table 6-3</b> Performance of PNN for identifying VEGFR-2 inhibitors and non-inhibitors evaluated by 5-fold cross validation study.....   | 131 |
| <b>Table 6-4</b> Virtual screening performance of support vector machines for identifying VEGFR-2 inhibitors from large compound libraries.....   | 133 |
| <b>Table 6-5</b> MDDR classes that contain higher percentage ( $\geq 3\%$ ) of SVM virtual-hits and the percentage values. Virtual-hits are identified by SVMs in screening 168K MDDR compounds for VEGFR-2 inhibitors. The total number of SVM identified virtual hits is 2,717..... | 137 |
| <b>Table 6-6</b> Comparison of virtual screening performance of SVM with those of other methods.....  | 141 |

---

## LIST OF FIGURES

|   |    |
|---|----|
| <b>Figure 1-1</b> SBVS and LBVS for drug discovery procedure (adopted from Ref [76]). SBVS is shown on the left and LBVS is shown on the right.....   | 18 |
| <b>Figure 2-1</b> Schematic diagram of the process of the training a prediction model and using it for predicting active compounds of a compound class from their structurally-derived properties (molecular descriptors) by using support vector machines; A, B, E, F and (h <sub>j</sub> , p <sub>j</sub> , v <sub>j</sub> ,...) represents such structural and physicochemical properties as hydrophobicity, volume, polarizability, etc.....                                      | 51 |
| <b>Figure 2-2</b> Schematic diagram illustrating the process of the prediction of compounds of a particular property from their structure by using k-nearest neighbors (kNN). Feature vector (h <sub>j</sub> , p <sub>j</sub> , v <sub>j</sub> ,...) represents such structural and physicochemical properties as hydrophobicity, volume, polarizability, etc; green dots: agents with the property; black box : agents without the property.....                                     | 53 |
| <b>Figure 2-3</b> Schematic diagram illustrating the process of the prediction of compounds of a particular property from their structure by using probabilistic neural networks (PNN). A, B: feature vectors of agents with the property; E, F: feature vectors of agents without the property; feature vector (h <sub>j</sub> , p <sub>j</sub> , v <sub>j</sub> ,...) represents such structural and physicochemical properties as hydrophobicity, volume, polarizability, etc..... | 57 |
| <b>Figure 3-1</b> Home page of MicrobPad MD database .....  | 72 |
| <b>Figure 3-2</b> Customized search page. This page provides search fields of genus name, species name, target name, disease indication and virulence factor. ....  | 73 |
| <b>Figure 3-3</b> List result page. This page provides genus name, species name, virulence factor, target gene, disease indications, and the number of diagnostic methods. ....   | 74 |
| <b>Figure 3-4</b> Related species and diagnostic methods page. This page provides detailed description about the related species and the diagnostic methods.....  | 75 |
| <b>Figure 3-5</b> Data download page of MicrobPad MD database.....  | 76 |
| <b>Figure 3-6</b> Data upload page of MicrobPad MD database.....  | 77 |
| <b>Figure 4-1</b> Home page of TTD 2010 .....   | 87 |

---

|  |     |
|--|-----|
| <b>Figure 4-2</b> Customized search page of TTD 2010 .....   | 88  |
| <b>Figure 4-3</b> Sequence similarity search page of TTD 2010 .....  | 88  |
| <b>Figure 4-4</b> Drug tanimoto similarity search page of TTD 2010 .....   | 89  |
| <b>Figure 4-5</b> Targets list page of “VEGFR” .....   | 90  |
| <b>Figure 4-6</b> TTD target detail information page.....  | 91  |
| <b>Figure 4-7</b> TTD drug detail information page .....   | 92  |
| <b>Figure 5-1</b> The structures of representative c-Src inhibitors. Compound 1:SKI-606<br>IC <sub>50</sub> =0.25μm [144]; Compound 2: AG-1879, IC <sub>50</sub> =0.085μm; Compound 3:<br>Sunitinib, SU 11248, IC <sub>50</sub> =1μm [282]; Compound 4: IC <sub>50</sub> =0.5μm [280]; Compound<br>5: IC <sub>50</sub> =0.26μm [281]; Compound 6: IC <sub>50</sub> =0.001μm [282]..... | 102 |
| <b>Figure 5-2</b> The 5-fold cross-validation studies of Src inhibitors across methods with<br>the averaged sensitivity together with their respective error bars.....   | 108 |
| <b>Figure 5-3</b> Virtual hit inhibiting Src at a moderate rate of 4.85% at 20μM .....   | 112 |
| <b>Figure 6-1</b> The structures of representative VEGFR-2 inhibitors. Compound 1:<br>Sunitinib,IC <sub>50</sub> =0.009μm,; Compound 2:IC <sub>50</sub> =0.032μm [335]; Compound 3:Vatalanb<br>(PTK787), IC <sub>50</sub> =0.037μm; Compound 4: IC <sub>50</sub> =0.012μm [336]; Compound 5:<br>IC <sub>50</sub> =0.004 μm [337]; Compound 6:IC <sub>50</sub> =0.111μm[338].....       | 125 |
| <b>Figure 6-2</b> Performance for identifying VEGFR-2 inhibitors evaluated by 5-fold cross<br>validation study across methods. This figure is illustrating the 5-fold cross validation<br>studies of VEGFR-2 inhibitors across methods with the averaged sensitivity together<br>with their respective error bars. ....  | 132 |
| <b>Figure 6-3</b> The structure of a SVM virtual hit tested to show moderate VEGFR-2<br>inhibitory activity.....   | 136 |

---

## LIST OF ACRONYMS

|                     |  |
|---------------------|--|
| <b>FN</b>           | False negative                                 |
| <b>FP</b>           | False positive                                 |
| <b>HTS</b>          | High throughput screening                      |
| <b>k-NN</b>         | k-nearest neighbors                            |
| <b>LBVS</b>         | Ligand-based Virtual Screening                 |
| <b>Lck</b>          | Lymphocyte-specific protein tyrosine kinase    |
| <b>MCC</b>          | Matthews correlation coefficient               |
| <b>MDDR</b>         | MDL Drug Data Report                           |
| <b>ML</b>           | Machine Learning                               |
| <b>MicrobPad MD</b> | Microbial Pathogen Diagnostic Methods Database |
| <b>PNN</b>          | Probabilistic neural network                   |
| <b>SBVS</b>         | Structure-based Virtual Screening              |
| <b>Src</b>          | Tyrosine-protein kinase Src                    |
| <b>Std Dev</b>      | Standard Deviation                             |
| <b>Std Err</b>      | Standard Error                                 |
| <b>SVM</b>          | Support vector machine                         |
| <b>TN</b>           | True negative                                  |
| <b>TP</b>           | True positive                                  |

---

|                |   |
|----------------|---|
| <b>TTD</b>     | Therapeutic targets database                  |
| <b>VS</b>      | Virtual Screening                             |
| <b>VEGFR-2</b> | Vascular endothelial growth factor receptor 2 |



---

## **Chapter 1 Introduction**

*Disease detection and drug discovery is typically a costly and lengthy process which takes more than 10 years to develop a successful drug from initial design to market. Although a log of efforts have been made for drug discovery, the successful drugs did not increase significantly over the past few decades. Bioinformatics and cheminformatics tools are explored to make drug research and development more efficient and effective. To help achieve this purpose, this work on "Development of Database and Computational Methods for Disease Detection and Drug Discovery" is conducted as one of the strategies illustrated in this chapter. The thesis contains database development of disease detection and therapeutic targets as well as discovery of potential drug lead by silico virtual screening. This introduction chapter includes: (1) conventional and molecular detection methods of pathogen; (2) bioinformatics and cheminformatics in drug discovery; (3) database development; (4) virtual screening of drug discovery; (5) objectives and outlines.*

### **1.1 Overview of pathogen detection**

#### **1.1.1 Application areas requiring pathogen detection.**

The detection of pathogens is the most important procedure for the identification and prevention of health and safety problems. It will cause terrible consequences in some

---

areas especially in clinical diagnostics, environment quality control and food industry where failure to detect pathogens. The pathogen detection has become the critical part in many research areas and application areas including pathology research, disease diagnosis, biodefense, food and water safety and epidemic prevention. Three application areas account for over two thirds of all research in the field of pathogen detection including food industry, water and environment quality control and clinical diagnosis [1-3]. Particularly in European Union, about 275 million pathogen detection of food were conducted in 2011 and this number in 2016 will to get to 350 million [4].

### **1.1.2 Brief introduction to pathogens induced infectious diseases**

A biological agent that cause diseases to its host is known as pathogen. Pathogens are most often used to refer to numerous infectious microorganisms such as bacteria, viruses, fungi and parasites which infect unicellular or multicellular organisms including human, animals and plants by disrupting the normal physiological function [5]. Pathogenic diseases is a term used for the diseases clinically caused by pathogen. Usually there are four kinds of pathogens including bacteria, viruses, fungi and parasites [5, 6]. Brief Description of four categories of pathogen together with the associated diseases are described in Table 1-1.

**Table 1-1** Four categories of pathogen inducing infectious human disease. Their infection are briefly described. Examples of the types of pathogens are listed, along with the disease they cause.

| Type of pathogen | Typically size                   | Description of infection   | Examples of pathogen                                  | Associated diseases                      |
|------------------|----------------------------------|--|---|--|
| Bacteria         | 1-5 $\mu\text{m}$                | Inhibit immune system and released edotoxins, extoxins and toxic factors which will block host protein synthesis, make cell deficient or cause inflammatory reaction.  | <i>Escherichia Coli</i>                               | Food poisoning                           |
|                  |                                  |  | <i>Chlamydia pneumoniae</i> [7]                       | Atherosclerosis                          |
|                  |                                  |  | <i>Helicobacter pylori</i> [8]                        | Psoriasis                                |
|                  |                                  |  | <i>Francisella tularensis</i> [9]                     | Tularemia                                |
|                  |                                  |  | <i>Mycobacterium tuberculosis</i> [10]                | Tuberculosis                             |
|                  |                                  |  | <i>Yersinia pestis</i> [11]                           | Plague                                   |
| Viruses          | 20-300 nm                        | Infection and the severe level of disease symptoms are relied on the virus virulence factors. Receptor typically endocytosed protein are often required on host cells for virus binding. Virus virulence factors can block MHCI processing for host immune system dysfunction.   | <i>Human immunodeficiency virus (HIV)</i> [12]        | AIDS                                     |
|                  |                                  |  | <i>Dengue virus</i> [13]                              | Dengue fever                             |
|                  |                                  |  | <i>SARS coronavirus</i> [14]                          | Severe acute respiratory syndrome (SARS) |
|                  |                                  |  | <i>Ebola virus</i> [15]                               | Ebola hemorrhagic fever                  |
|                  |                                  |  | <i>Coxsackie A virus, Enterovirus 71 (EV-71)</i> [16] | Hand, foot and mouth disease (HFMD)      |
|                  |                                  |  | <i>Influenzavirus A</i> [17]                          | Swine Flu                                |
| Fungi            | Spore size of 1-40 $\mu\text{m}$ | Fungi diseases are induced through host barriers penetration or immunological debilitation by fungi. Fungi infect host through three ways: iatrogenicity, trauma or inhalation [18]. The common fungal diseases include respiratory fungal allergy, immune reconstitution inflammatory syndrome, skin diseases, mucosal infections and | <i>Blastomyces dermatitidis</i> [19]                  | Blastomycosis                            |
|                  |                                  |  | <i>Candida albicans</i> [20]                          | Thrush                                   |
|                  |                                  |  | <i>Histoplasma capsulatum</i> [21]                    | Histoplasmosis                           |

|          |           |   |  |                 |
|----------|-----------|---|--|-----------------|
|          |           | eosinophilia-driven hypersensitivity diseases. Fungi can also induce opportunistic infection in AIDS and cancer patients.   |  |                 |
| Parasite | Up to 1mm | Traditionally, there are more than one host within lifestages of parasite. Parasite can be divided into four types: roundworms, tapeworms, flukes and single celled protozoa. Some parasites can cause diseases by toxins, others directly cause diseases. Parasitic infection can be caused by contamination of soil, water, food, pet and insect. The parasite infection is typically chronic and immunology defection. | <i>Entameba histolytica</i> [22]                             | Amoebiasis      |
|          |           |   | <i>Ascaris lumbricoides</i> [23]                             | Ascariasis      |
|          |           |   | <i>Plasmodium malariae</i> ,<br><i>Plasmodium ovale</i> [24] | Malaria         |
|          |           |   | <i>Schistosoma mansoni</i> [25]                              | Schistosomiasis |

Although medical advances have been made to protect human from pathogen infection, pathogens still threaten human life and difficult for treatment since the variation of the pathogens, particularly viruses, is significant fast. Over the decades, more serious pathogen diseases have been induced by viruses such as human immunodeficiency virus (HIV), hepatitis B, meningococcal disease [26] and some cancer such bladder cancer [27] and cervical cancer [28]. The pathogenic diseases are extremely harmful for human health and life quality. **Table 1-2** shows the top 10 leading cause of death worldwide in 2008 reported by WHO fact sheet [29]. Four

pathogenic diseases involved in the top 10 causes of death worldwide and the total proportion of all the death is up to 15.90%.

**Table 1-2** The top 10 leading cause of death worldwide in 2008 reported by WHO fact sheet.

| World                                    | Deaths in millions | % of deaths |
|--|--------------------|-------------|
| Ischaemic heart disease                  | 7.25               | 12.80       |
| Stroke and other cerebrovascular disease | 6.15               | 10.80       |
| Lower respiratory infections             | 3.46               | 6.10        |
| Chronic obstructive pulmonary disease    | 3.28               | 5.80        |
| Diarrhoeal diseases                      | 2.46               | 4.30        |
| HIV/AIDS                                 | 1.78               | 3.10        |
| Trachea, bronchus, lung cancers          | 1.39               | 2.40        |
| Tuberculosis                             | 1.34               | 2.40        |
| Diabetes mellitus                        | 1.26               | 2.20        |
| Road traffic accidents                   | 1.21               | 2.10        |

Although the many effort have been made to diagnosis and treatment of pathogenic diseases, challenges exist in accurately identifying pathogens rapidly. According to the world health statistics report [30], pathogenic diseases mortality rate is still significant as shown in **Table 1-3**. Some pathogenic diseases e.g. H5N1 influenza [31] induce high mortality rate due to mutations. Some diseases e.g. poliomyelitis [32] cause very bad consequence even with low mortality rate. Therefore, early detection of pathogens to identify the pathogenic sources is extremely important for fast disease diagnosis, proper treatment and pathogenesis processes research. It is desired to enable fast, accurate, sensitive and low-cost diagnosis of pathogens [33-36].

**Table 1-3** Three pathogenic diseases mortality rate in 2013.

| WHO region                   | Pathogenic diseases mortality rate (per 100 000 population) |      |         |  |      |
|------------------------------|---|------|---------|--|------|
|                              | HIV/AIDS  |      | Malaria | Tuberculosis among HIV-negative people |      |
|                              | 2001  | 2011 | 2010    | 2000                                   | 2011 |
| African Region               | 219   | 139  | 72      | 37                                     | 26   |
| Region of the Americas       | 12  | 9    | 0.2     | 3.6                                    | 2.2  |
| South-East Asia Region       | 14  | 12   | 2.4     | 43                                     | 26   |
| European Region              | 5   | 11   | NA      | 8                                      | 5    |
| Eastern Mediterranean Region | 4.8   | 7.7  | 3.5     | 29                                     | 16   |
| Western Pacific Region       | 2.4   | 4.4  | 0.2     | 12                                     | 6.9  |

### 1.1.3 Conventional pathogen detection methods

Traditionally, microbial morphology and growth variables are the predominant characteristics using for microorganisms identification and differentiation through morphologic features, growth variables, and biochemical utilization of organic substrates [37]. In addition to phenotypic approaches based on various medium, other methods have been developed and used for pathogen detection over decades. For instance, immunological methods using antigen and antibody, rapid microscopic

---

smear analysis and manually or semi-automated biochemical testing for characterization of pathogens have been widely applied.

However, there are significant drawbacks existing in these conventional methods because they highly depend on traditional microbiology characteristics and chemical profiles monitoring approaches which are time-consuming, high cost, low sensitivity, high manpower cost and require labile natural products. Due to the cultivation time of microorganisms, high expense, false positives and causative agent, it is difficult to conduct high-throughput screening for environmental and clinical samples. Moreover, these techniques that are routinely established for pathogen identification but do not directly identify virulence factors [38]. These methods cannot provide important information of the identified pathogens about the potential pathogenesis and virulence factors for further research. In summary, to overcome the problems of conventional identification methods, more reliable, rapid and accurate tools for pathogen determination have been developed.

#### **1.1.4 Molecular pathogen detection methods**

There are numerous molecular techniques have developed to detect pathogens with the advantage of speed along with the relative simplicity, specific and sensitive detection [39]. Molecular detection methods mainly refer to nucleic acid based

---

molecular detection technology. It plays a key role when great efforts made to development of pathogen detection. Nucleic acid based methods rely on the premise that unique DNA or RNA sequences marker of an organism is specific and different from other species. The unique sequence can be used as a detection target of a pathogen. The nucleic acid based molecular methods include several kinds of nucleic acid amplification techniques such as polymerase chain reaction (PCR), reverse transcriptase polymerase chain reaction (RT-PCR) and quantitative PCR (Q-PCR), molecular beacon technology [40], fluorescent in situ hybridization (FISH) [41], microarray based strategies. Among these molecular detection methods, microarray-based detection is considered as the technique of high sensitivity, specificity and throughput because it can integrate nucleic acid amplification and high throughput screening technique. Nucleic acid probes using in microarray technique are able to identify pathogen organisms at, above, and below the species level [37]. However, microarray based detections cost a lot and require plenty of PCR reactions which are complex for arrangement. These disadvantages of microarray application have severely impeded the utilization and further development of this technique.

Molecular detection methods are much safer using in laboratory than conventional methods. Some pathogens such as *Mycobacterium tuberculosis*, *Influenza A virus* and *SARS virus* causing serious fevers and symptoms are laboratory hazards and risks [42,



---

43]. These organisms have severe risks for laboratory worker and may contribute to severe diseases or mortality.

## **1.2 Bioinformatics and cheminformatics in drug discovery**

The combination of random screening and rational drug design have played an important role in drug discovery [44]. The traditional drug discovery process comprise seven basic steps including disease selection, target selection, lead compound identification, lead optimization, preclinical trial evaluation, clinical trials and drug manufacturing [45]. Drug discovery process costs typically 10-17 years and \$800 million totally, but success rate is still less than 10%. Definitely, it is a time-consuming, expensive and low success rate procedure [46]. Target, efficacy and safety are three major problems of current drug discovery strategy. Current drugs design aims at a few know targets, but many targets for existing diseases and new diseases are still unknown. Novel targets are demanded to be investigated to treat new disease or overcome drug resistances problems. Some drug candidates may lose efficacy or cause safety issues such as severe side effects during the clinical trials phage.

New techniques have been utilized to drug discovery to make it more effective and efficient especially in early stage of drug discovery such as target selection, lead compound identification and optimization. Since the development of molecular

---

biology and genomics for comprehensive understanding disease mechanism and therapeutic intervention, new techniques including microarray, genomic DNA sequencing, RNA-seq, Chip-seq, and high throughput screening have been applied and shown great potential for solving current problems. For instance, genomic DNA sequencing and next generation sequencing combining bioinformatics may help identify up to 10,000 new molecular targets [47]. On the chemistry side, HTS and cheminformatics may help discover new leads from large compound library. Bioinformatics, an interdisciplinary of biology, mathematics and computer science, mostly refers to informatics processes in biotic systems [48]. Bioinformatics can develop computational methods and tools to obtain and analysis data as well as generate biological knowledge [48]. Cheminformatics aims to use computer and informational techniques to solve a serious of chemistry problems [49, 50].

Bioinformatics and cheminformatics tools are developed which are able to congregate all the required information regarding potential targets like nucleotide and protein sequencing, homologue mapping [51, 52], function prediction [53, 54], pathway information [55], structural information [56] and disease associations [57], chemistry information. The availability of that information can help pharmaceutical companies in saving time and money on target identification and validation.

---

### **1.3 Introduction of bioinformatics and cheminformatics database development**

Since the biological and chemistry data increase rapidly due to the new technology such as HTS and nucleotide sequencing, it is necessary to collect, store and manage data effectively to assist research on disease mechanisms and drug candidates. However, some data may lack of organization and standard format from different resource. Further process of validation and analysis are needed for the data to extract useful information. Implementation of tools is also required to provide an easy and powerful way for data access. Database is such a technique can meet these requirements by providing latest information and data that related to disease mechanism studies, pharmaceutical research and drug development. They provide interdisciplinary data of different areas such as biological information, chemistry information, bioinformatics and cheminformatics data, system model of pathway, bimolecular interaction data and so on. Databases store data in various formats such as relational database tables, XML files, YML files, flat files, protein structure 3D object files.

Databases have been developed over decades. There are many public bioinformatics cheminformatics databases available which are listed in **Table 1-4** and **Table 1-5**. In this work, we focus on the development of web accessible databases for pathogen

detection and therapeutic targets and drugs. Owing to the effort of target discovery, hundreds of success targets and more than 1000 research targets have been identified [58-61]. There are several well known target and drug databases available such as SuperTarget [62], BindingDB and DrugBank.

**Table 1-4** Popular bioinformatics databases.

| Database  | Description  | Web link  |
|---|--|---|
| National Center for Biotechnology Information (NCBI)      | Biomedical and genomic resource funded by U.S. government. | <a href="http://www.ncbi.nlm.nih.gov/">http://www.ncbi.nlm.nih.gov/</a>   |
| EMBL-EBI  | Integrated bioinformatics research and services            | <a href="http://www.ebi.ac.uk/services">http://www.ebi.ac.uk/services</a>   |
| NCBI GenBank  | Publicly available annotated DNA sequences                 | <a href="http://www.ncbi.nlm.nih.gov/genbank/">http://www.ncbi.nlm.nih.gov/genbank/</a>                                 |
| DNA Data Bank of Japan (DDBJ)                             | Sole nucleotide sequence data in Asia and other countries  | <a href="http://www.ddbj.nig.ac.jp/">http://www.ddbj.nig.ac.jp/</a>   |
| European Nucleotide Archive (ENA)                         | Worldwide nucleotide sequencing information                | <a href="http://www.ebi.ac.uk/ena/">http://www.ebi.ac.uk/ena/</a>   |
| NCBI Genome   | Sequence and map data of the whole genomes                 | <a href="http://www.ncbi.nlm.nih.gov/genome/">http://www.ncbi.nlm.nih.gov/genome/</a>                                   |
| Genomes OnLine Database (GOLD)                            | Worldwide genome and metagenome sequencing                 | <a href="http://www.genomesonline.org/cgi-bin/GOLD/index.cgi">http://www.genomesonline.org/cgi-bin/GOLD/index.cgi</a>   |
| TIGR  | Plant Transcript Assemblies                                | <a href="http://plantta.jcvi.org/">http://plantta.jcvi.org/</a>   |
| PEDANT  | Genomes protein analysis tools                             | <a href="http://pedant.gsf.de/index.jsp">http://pedant.gsf.de/index.jsp</a>   |
| Comprehensive Microbial Resource (CMR)                    | Publicly available prokaryotic genomes                     | <a href="http://cmr.jcvi.org/tigr-scripts/CMR/CMRHomePage.cgi">http://cmr.jcvi.org/tigr-scripts/CMR/CMRHomePage.cgi</a> |
| Microbial Genome Database for Comparative Analysis (MBGD) | Comparative analysis of microbial genomes                  | <a href="http://mbgd.genome.ad.jp/">http://mbgd.genome.ad.jp/</a>   |
| Clusters of Orthologous Groups of proteins (COG)          | Comparative protein sequences based on complete genomes    | <a href="http://www.ncbi.nlm.nih.gov/COG/">http://www.ncbi.nlm.nih.gov/COG/</a>   |
| Mitomap   | human mitochondrial genome                                 | <a href="http://www.mitomap.org/MITOMAP">http://www.mitomap.org/MITOMAP</a>   |
| Uniprot   | Protein knowledgebase                                      | <a href="http://www.uniprot.org/">http://www.uniprot.org/</a>   |

|  |   |   |
|--|---|---|
|  | (UniProtKB/Swiss-Prot: manually annotated and reviewed;<br>UniProtKB/TrEMBL: automatically annotated and not reviewed |   |
| BRENDA   | Collection of enzyme data   | <a href="http://www.brenda-enzymes.org/">http://www.brenda-enzymes.org/</a>   |
| ExPASy -ENZYME                                     | Enzymes information   | <a href="http://enzyme.expasy.org/">http://enzyme.expasy.org/</a>   |
| CAZy   | Carbohydrate-Active enZymes   | <a href="http://www.cazy.org/">http://www.cazy.org/</a>   |
| Pfam   | Collection of protein families  | <a href="http://pfam.sanger.ac.uk/">http://pfam.sanger.ac.uk/</a>   |
| TIGRFAMs   | Resource of protein sequence classification   | <a href="http://www.jcvi.org/cgi-bin/tigrfams/index.cgi">http://www.jcvi.org/cgi-bin/tigrfams/index.cgi</a>                 |
| SUPFAM   | Homologous protein domain families  | <a href="http://supfam.mbu.iisc.ernet.in/">http://supfam.mbu.iisc.ernet.in/</a>   |
| ExPASy - PROSITE                                   | Patterns and profiles of protein domains, families and functional sites   | <a href="http://prosite.expasy.org/">http://prosite.expasy.org/</a>   |
| CATH:Protein Structure Classification Database     | Hierarchical domain classification of PDB protein structures  | <a href="http://www.cathdb.info/">http://www.cathdb.info/</a>   |
| PRINTS   | Protein fingerprints  | <a href="http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/index.php">http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/index.php</a> |
| SCOP: Structural Classification of Proteins        | Detailed structural and evolutionary relationships of PDB protein   | <a href="http://scop.mrc-lmb.cam.ac.uk/scop/">http://scop.mrc-lmb.cam.ac.uk/scop/</a>                                       |
| Protein Data Bank                                  | Structures of proteins, nucleic acids, and Complex  | <a href="http://www.rcsb.org/pdb/home/home.do">http://www.rcsb.org/pdb/home/home.do</a>                                     |
| MMDB   | Protein structure at NCBI   | <a href="http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml">http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml</a>   |
| BIND The Biomolecular Interaction Network Database | Molecular interactions  | <a href="http://bond.unleashedinformatics.com/">http://bond.unleashedinformatics.com/</a>                                   |
| Database of Interacting Proteins (DIP)             | Experimentally verified protein interaction   | <a href="http://dip.doe-mbi.ucla.edu/dip/Main.cgi">http://dip.doe-mbi.ucla.edu/dip/Main.cgi</a>                             |
| MINT   | Experimentally verified protein interaction   | <a href="http://mint.bio.uniroma2.it/mint/Welcome.do">http://mint.bio.uniroma2.it/mint/Welcome.do</a>                       |
| KEGG   | Collection and manually drawn pathway maps  | <a href="http://www.genome.jp/kegg/pathway.html">http://www.genome.jp/kegg/pathway.html</a>                                 |
| Signaling Pathway Database (SPAD)                  | Signal transduction and genetic information   | <a href="http://www.grt.kyushu-u.ac.jp/spad/">http://www.grt.kyushu-u.ac.jp/spad/</a>                                       |

|                              |                                       |   |
|------------------------------|---------------------------------------|---|
| BioCarta                     | Dynamic graphical pathway map         | <a href="http://www.biocarta.com/">http://www.biocarta.com/</a>                           |
| cPath                        | Pathway collection and software suite | <a href="http://cbio.mskcc.org/software/cpath/">http://cbio.mskcc.org/software/cpath/</a> |
| ExPASy -Biochemical Pathways | Biochemical Pathways                  | <a href="http://web.expasy.org/pathways/">http://web.expasy.org/pathways/</a>             |

**Table 1-5** Popular chemical databases

| Database                                 | Description  | Web link  |
|--|--|---|
| BindingDB                                | Binding affinities of proteins and protein ligand              | <a href="http://www.bindingdb.org/bind/index.jsp">http://www.bindingdb.org/bind/index.jsp</a>   |
| MDDR                                     | Information of biologically active molecules                   | <a href="http://accelrys.com/products/databases/bioactivity/mddr.html">http://accelrys.com/products/databases/bioactivity/mddr.html</a> |
| PubChem                                  | Information on the biological active molecules                 | <a href="http://pubchem.ncbi.nlm.nih.gov/">http://pubchem.ncbi.nlm.nih.gov/</a>   |
| ZINC                                     | Commercially-available compounds for virtual screening         | <a href="http://zinc.docking.org/">http://zinc.docking.org/</a>   |
| ChEMBL                                   | Bioactive molecules  | <a href="http://www.ebi.ac.uk/chembl/">http://www.ebi.ac.uk/chembl/</a>   |
| DrugBank                                 | Drug data with drug target                                     | <a href="http://www.drugbank.ca/">http://www.drugbank.ca/</a>   |
| eMolecules                               | Chemical molecules commercial available                        | <a href="http://www.emolecules.com/">http://www.emolecules.com/</a>   |
| WOMBAT                                   | Chemogenomics with bioactivity annotations                     | <a href="http://www.sunsetmolecular.com">http://www.sunsetmolecular.com</a>   |
| 4SC                                      | Targeted small molecule drugs                                  | <a href="http://www.4sc.de">www.4sc.de</a>  |
| chemspider                               | Over 28 million free chemical structure                        | <a href="http://www.chemspider.com/">http://www.chemspider.com/</a>   |
| NIST Chemistry WebBook                   | Thermochemical, thermophysical, and ion energetics properties  | <a href="http://webbook.nist.gov/chemistry/">http://webbook.nist.gov/chemistry/</a>   |
| chemexper                                | Over 200,000 different chemicals with physical characteristics | <a href="http://www.chemexper.com/">http://www.chemexper.com/</a>   |
| Chemical Structure Lookup Service (CSLS) | 46 million unique structures                                   | <a href="http://cholla.chemnavigator.com/cgi-bin/lookup/search">http://cholla.chemnavigator.com/cgi-bin/lookup/search</a>               |
| ChemDB                                   | Nearly 5 million small molecules                               | <a href="http://cdb.ics.uci.edu/CHEM/">http://cdb.ics.uci.edu/CHEM/</a>   |
| AffinDB                                  | Affinity data of PDB protein-ligand                            | <a href="http://pc1664.pharmazie.uni-marburg.de/affinity/index.php">http://pc1664.pharmazie.uni-marburg.de/affinity/index.php</a>       |
| ChemBank                                 | Small molecules library with over 36,000 biological assays     | <a href="http://chembank.broadinstitute.org/">http://chembank.broadinstitute.org/</a>   |

|                               |  |   |
|-------------------------------|--|---|
| ChemIDplus                    | Free 350000 chemical compounds   | <a href="http://chem.sis.nlm.nih.gov/chemidplus/">http://chem.sis.nlm.nih.gov/chemidplus/</a>                                     |
| ACB Blocks                    | 90,000 combinatorial chemistry   | <a href="http://www.acbblocks.com/content/view/page/services">http://www.acbblocks.com/content/view/page/services</a>             |
| Advanced ChemTech             | Manufacturer of amino acids, chemicals and reagents                                      | <a href="https://www.advancedchemtech.com/">https://www.advancedchemtech.com/</a>   |
| Asinex                        | Libraries of medicinal chemistry, including biodesign, synergy, medchem building blocks. | <a href="http://www.asinex.com/">http://www.asinex.com/</a>   |
| COMBI-BLOCKS                  | Combinatorial building blocks, organics and chemicals                                    | <a href="http://www.combi-blocks.com">http://www.combi-blocks.com</a>   |
| ComGenex                      | Freely accessible chemicals catalog  | <a href="http://www.rdchemicals.com/index.html">http://www.rdchemicals.com/index.html</a>   |
| EMC microcollection           | Organic chemical and Biochemicals  | <a href="http://www.microcollections.de/">http://www.microcollections.de/</a>   |
| InterBioScreen                | Biologically active natural organic compounds  | <a href="http://www.ibscreen.com/">http://www.ibscreen.com/</a>   |
| Maybridge                     | Chemical Building Blocks and Screening Compounds,  | <a href="http://www.maybridge.com/">http://www.maybridge.com/</a>   |
| MicroSource Discovery Systems | Biocompatible compounds and pure natural products  | <a href="http://www.msdiscovery.com">http://www.msdiscovery.com</a>   |
| Polyphor                      | Innovative pharmaceutical compound   | <a href="http://www.polyphor.com">http://www.polyphor.com</a>   |
| Sigma-Aldrich                 | Drug-like compounds  | <a href="http://www.sigmaaldrich.com/chemistry/drug-discovery.html">http://www.sigmaaldrich.com/chemistry/drug-discovery.html</a> |
| Specs.net                     | Over 240,000 true novel compounds  | <a href="http://www.specs.net/snpage.php?snpageid=home">http://www.specs.net/snpage.php?snpageid=home</a>                         |
| TimeTec                       | Over 1,000,000 Chemical Structure  | <a href="http://www.timtec.net/">http://www.timtec.net/</a>   |
| Tripos                        | Over 50,000 compounds of biological activity   | <a href="http://leadquest.tripos.com/">http://leadquest.tripos.com/</a>   |
| ChemBridge                    | Over 900,000 diverse compounds and 14,000 chemical building blocks                       | <a href="http://www.chembridge.com/index.php">http://www.chembridge.com/index.php</a>   |
| ChemDiv                       | Drug discovery compounds   | <a href="http://eu.chemdiv.com/">http://eu.chemdiv.com/</a>   |

## 1.4 Overview of virtual screening in drug discovery

High throughput screening (HTS) known as a fast test large amount of chemicals tool has been used extensively in pharmaceutical industry. However, HTS has problems of over-reliance, no assurance of success and high cost. Without assay development,

---

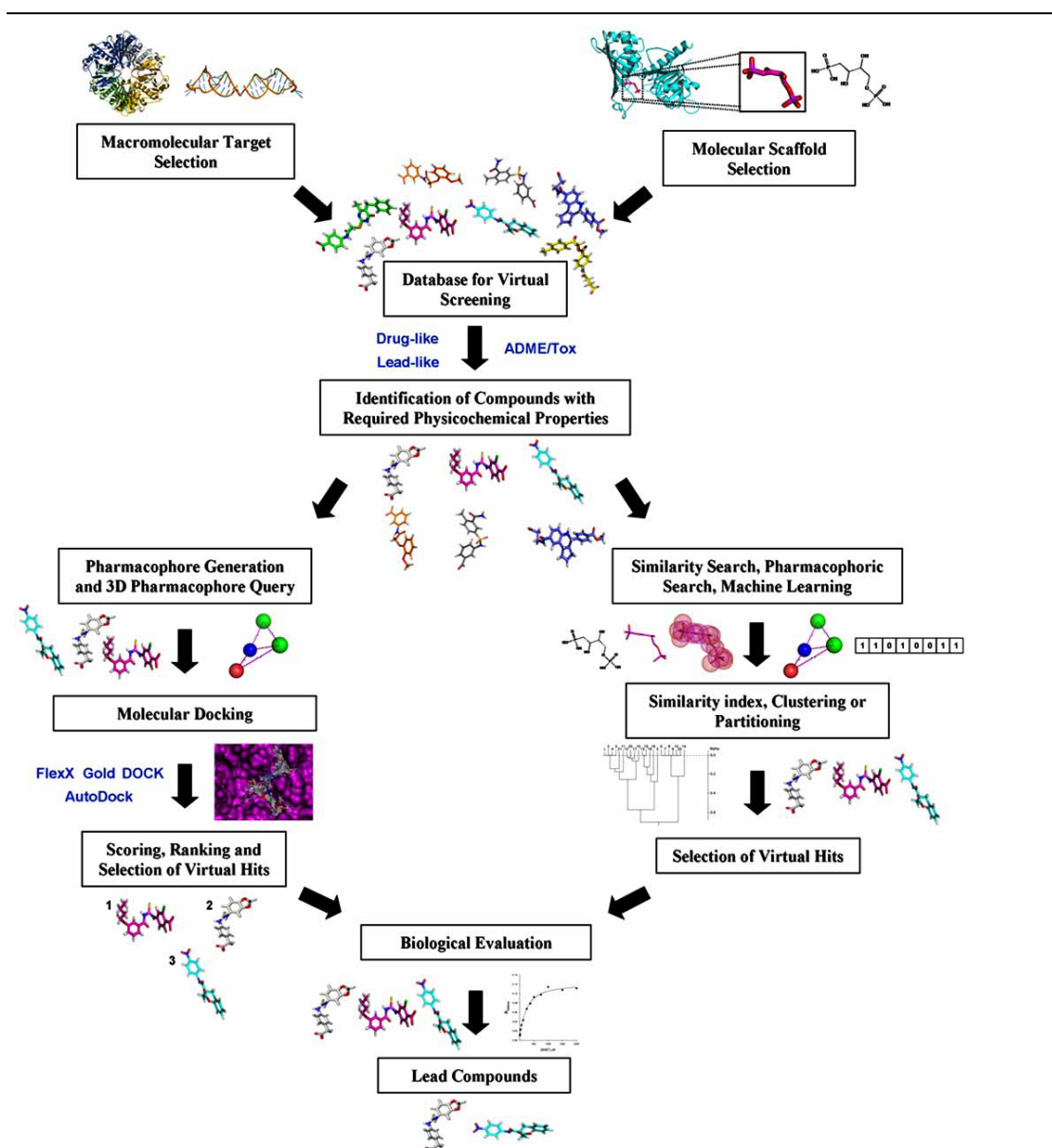
HTS process still costs approximately US \$75,000.[63]. The cost will get much higher if expensive assay is integrated. Moreover, the entire chemical space is too huge to be covered by natural and synthesized compounds which only occupy limited proportion of the chemical space [64, 65]. Even if only drug-like compounds are considered, chemical space of drug-like compounds is still magnitude larger than that of pharmaceutical industry screening collection [66]. Considering these drawbacks, it is necessary to explore technologies to complement HTS assay and synthesis.

Virtual screening (VS) is a computational technique used in drug discovery research. It involves rapid *in silico* assessment of large libraries of chemical structures in order to identify those structures that are most likely to bind to a drug target, typically a protein receptor or enzyme [67, 68]. VS has been used to describe a process of computationally analyzing large compound collections in order to prioritize compounds for synthesis or assay. During the last decade, a broad range of computational techniques have been applied to search for novel bioactive compounds for many targets. VS has been extensively explored for facilitating lead discovery [69-72], identifying agents of desirable pharmacokinetic and toxicological properties [73, 74] and other areas. There are two broad categories of screening techniques: structure-based virtual screening (SBVS) and ligand-based virtual screening (LBVS) [75]. SBVS involves the virtual docking of candidate ligands into a protein target



---

followed by the estimation of the probability of the high affinity binding between them calculated by a scoring function. LBVS methods, such as pharmacophore methods and chemical similarity analysis methods, require the ligand structure information, they focus on discovery the new drug hits by analyzing the physical and chemical similarities of known compound pools by computational means. **Figure 1-1** shows the general procedure used in SBVS and LBVS.



**Figure 1-1** SBVS and LBVS for drug discovery procedure (adopted from Ref [76]). SBVS is shown on the left and LBVS is shown on the right.

Structure-based virtual screening (SBVS) starts with a 3-D structure of a target protein and a database of the 3-D structures of ligands as the screening pool. It is usually applied when the 3D structure of a protein target, derived either from experimental data (X-ray or NMR spectroscopy) or from homology modeling, when

---

available. The SBVS procedure consists of docking and scoring. Docking is most straightforward VS method and it is preferred by the chemists. The docking algorithms [77, 78] are designed to predict the ligand conformation and orientation within the targeted active site of the target. The scoring methods are empirically or semi-empirically derived to attempt [79] to estimate the binding tightness of the ligand and the protein in bound complexes. Docking and scoring algorithms are combined to detect the compounds with higher affinity against a target by predicting their binding mode (by docking) and affinity (by scoring), and retrieving those with the highest scores. To date, more than 60 docking programs and 30 scoring functions have been reported [80, 81]. The major drawback with SBVS is the unavailability of appropriate scoring functions to differentiate between correct and incorrect poses of bound ligands and identifying false negative and positive hits. Some of the key challenges encountered by SBVS include the appropriate treatment of ionization, tautomerization of ligand and protein residues, target/ligand flexibility, choice of force fields, solvation effects, dielectric constants, exploration of multiple binding modes and, most importantly, the approximations in the scoring functions that lead to false-positives and miss true-hits. Moreover, most docking algorithms and scoring functions are tuned towards high throughput, which requires a compromise between the speed and accuracy of binding mode and energy prediction. The hit enrichment is defined as the fraction of true active compounds in, for example, the upper 1% of the

---

ranked VS hit list compared with the average fraction of active compounds in the search space. The performance of a docking program is difficult to evaluate in advance, and depends on the nature and quality of the target structure [80-82]. Despite all optimization efforts, the available scoring functions do not provide reliable estimates of free binding energies, and are not able to rank-order compounds according to affinity [81, 83]. The published comparisons of docking programs have been critically reviewed [84-86].

As compared with structure-based methods, LBVS methods including pharmacophore methods and chemical similarity analysis methods have shown better performance in terms of speed, yield and enrichment factor. Hit rate is defined as the relation between the number of true hits found in the hit list respect to the total number of compounds in the hit list; and the enrichment factor (EF) is the hit rate divided by the total number of hits in the full database relative to the total number of compounds in the database. To improve the coverage, performance and speed of VS tools, machine learning (ML) methods, including SVM, neural network and etc, have been used for developing LBVS tools [87-94] to complement or to be combined with SBVS [69, 95-106] and other LBVS [70, 107-110] tools. ML methods have been used as part of the efforts to overcome several problems that have impeded progress in more extensive applications of SBVS and LBVS tools [69, 111]. These problems include

---

the vastness and sparse nature of chemical space needs to be searched, limited availability of target structures (only 15% of known proteins have known 3D structures), complexity and flexibility of target structures, and difficulties in computing binding affinity and solvation effects. ML methods have been explored for developing such alternative VS tools [87-89] because of their high speed [112] and capability for covering highly diverse spectrum of compounds [113]. Han et al [114] did a comparative study for reported performance of different VS methods in screening large libraries of compounds as shown in **Table 1-6**. ML methods show good potential for a better performance at VS of extremely large libraries with over 1M compounds. The reported yield, hit-rate and enrichment factor of ML tools are in the range of 55%~81%, 0.2%~0.7% and 110~795 respectively [88, 91, 93], compared to those of 62%~95%, 0.65%~35% and 20~1,200 by SBVS tools [98, 99]. Moreover, he also developed a new putative negative generation method in which negatives were generated from 3M PubChem compounds. With this method he significantly improved yield, hit-rate and enrichment factor to 52.4%~78.0%, 4.7%~73.8%, and 214~10,543 respectively in screening libraries of over 1 million compounds. For SBVS methods, approaches of using additional filters are often required in order to further minimize the false positives. One approach is the selection of top-ranked hits, which has been extensively used in LBVS [88, 89, 93, 94, 115, 116] and SBVS [98, 100-102, 117, 118]. The second approach is the elimination of potentially

---

unpromising hits in pre-screening stage by using such filters as Lipinski's rule of five [119] [99], and recognition of pharmacophore [101] and specific chemical groups or interaction patterns [98, 100, 104, 120]. The last one is the combination of LBVS and SBVS methods. All these approaches take quite some time. However, they are not required for SVM based approaches which already have a low false positives rate.

**Table 1-6** Comparison of the reported performance of different VS methods in screening large libraries of compounds (adopted from Han et al[114]).

| Type of VS method and size of compound libraries screened                   | VS method (number of studies) [references]    | Compounds screened |                  |                       | Virtual hits selected by VS method       |  | Known hits selected by VS method |          |            |                   |
|---|---|--------------------|------------------|-----------------------|--|--|----------------------------------|----------|------------|-------------------|
|   |   | No of compounds    | No of known hits | Percent of known hits | No of compounds selected as virtual hits | Percent of screened compounds selected as virtual hits | No of known hits selected        | Yield    | Hit rates  | Enrichment factor |
| Structure-based VS, extremely large libraries ( $\geq 1M$ )                 | Docking + pre-screening filter (2) [98, 99]   | 1M~2M              | 355~630          | ~0.03%                | 1K~60K                                   | 0.08%~3%   | 340~390                          | 62%~ 95% | 0.65%~ 35% | 20~1200           |
| Structure-based VS, large libraries   | Docking + pre-screening filter (11) [100-106] | 134K~400K          | 100~1016         | 0.12%~0.76%           | 375~4.5K                                 | 0.28%~3%   | 5~231                            | 2%~ 30%  | 0.11%~ 17% | 4~66              |
| Ligand-based VS (machine learning), extremely large libraries ( $\geq 1M$ ) | Machine learning - SVM (2)[88, 91, 93]        | 2.5M               | 22~46            | 0.0009%~0.0018%       | 2.5K~11K                                 | 0.1%~0.45%   | 18~25                            | 55%~ 81% | 0.2%~ 0.7% | 110~795           |
| Ligand-based VS (machine learning), large libraries                         | Machine learning - SVM (2)[89]                | 172K               | 118~128          | ~0.07%                | 1.7K                                     | 1%   | 26~70                            | 22%~ 55% | 1.5%~ 4.1% | 22~55             |
|   | Machine learning - SVM (11)[92]               | 98.4K              | 259~1146         | 0.26%~1.16%           | 984                                      | 1%   | 131~710                          | 44%~ 69% | 14%~ 72%   | 44~69             |

|   |  |                |              |                     |                |            |         |          |                    |           |
|---|--|----------------|--------------|---------------------|----------------|------------|---------|----------|--------------------|-----------|
|   | Machine learning<br>– BKD (12)[89,<br>91, 93, 94]                  | 101K~1<br>03K  | 259~<br>1166 | 0.25%~<br>1.2%      | 5.1K           | 5%         | 65~972  | 14%~ 94% | 1.2%~ 18.9%        | 3~19      |
|   | Machine learning<br>– LMNB (1)[91,<br>93]                          | 172K           | 118          | 0.069%              | 1.7K           | 1%         | 19      | 16%      | 1%                 | 15        |
|   | Machine learning<br>– CKD (18)[92]                                 | 98.4K          | 259~<br>1211 | 0.26%~<br>1.23%     | 984            | 1%         | 132~960 | 34%~ 94% | 13%~ 98%           | 53~94     |
| Ligand-based VS<br>(clustering), large<br>libraries | Hierarchical<br>k-means (5)[108]                                   | 344.5K         | 91~155<br>6  | 0.026%<br>~0.45%    | 3750~2128<br>5 | 1.1%~6.2%  | 27~761  | 23% ~55% | 0.72%~5%           | 7.97~31.2 |
|   | NIPALSTREE<br>(5)[108]   | 344.5K         | 91~155<br>6  | 0.026%<br>~0.45%    | 3469~2812<br>5 | 1.0%~8.2%  | 17~625  | 18% ~50% | 0.49%~ 2.8%        | 3.51~18.7 |
|   | Hierarchical<br>k-means +<br>NIPALSTREE<br>disjunction<br>(5)[108] | 344.5K         | 91~155<br>6  | 0.026%<br>~0.45%    | 7317~4316<br>5 | 2.1%~12.3% | 30~980  | 33% ~72% | 0.41% ~2.9%        | 4.86~17.6 |
|   | Hierarchical<br>k-means +<br>NIPALSTREE<br>conjunction<br>(5)[108] | 344.5K         | 91~155<br>6  | 0.026%<br>~0.45%    | 538~6692       | 0.16%~1.9% | 14~406  | 6% ~32%  | 1.1% ~10.2%        | 7.77~98   |
| Ligand-based VS<br>(structural signatures),         | Pharmacophore<br>(3)[109, 121,<br>.8M                              | 1.77M~3<br>.8M | 55~144       | 0.0014%<br>~0.0081% | 20K~1M         | 1.15%~26%  | 6~39    | 11% ~70% | 0.0039%~<br>0.084% | 3~10.3    |



|  |                        |        |      |         |      |         |      |       |       |       |
|--|------------------------|--------|------|---------|------|---------|------|-------|-------|-------|
| extremely large libraries ( $\geq 1M$ )  | 122]                   |        |      |         |      |         |      |       |       |       |
| Ligand-based VS (structural signatures), large libraries   | Pharmacophore (1)[110] | 380K   | 30   | 0.0079% | 6917 | 1.82%   | 23   | 76.7% | 0.33  | 41.8  |
| Ligand-based VS, extremely large libraries ( $\geq 1M$ ) for HIV protease, inhibitors DHFR inhibitors, Dopamine antagonists, CNS active agents | SVM[114]               | 2.986M | 2351 | 0.076%  | 8157 | 0.27%   | 1833 | 78.0% | 22.5% | 296   |
|  | SVM[114]               | 2.986M | 225  | 0.007%  | 160  | 0.0054% | 118  | 52.4% | 73.8% | 10543 |
|  | SVM[114]               | 2.986M | 37   | 0.0012% | 299  | 0.01%   | 23   | 62.2% | 7.7%  | 6417  |
|  | SVM[114]               | 2.986M | 664  | 0.022%  | 9502 | 0.32%   | 442  | 66.6% | 4.7%  | 214   |

---

As it is common for the pharmaceutical industry to screen >1 million compounds per high-throughput screening campaign [123]. A small rise in the hit rate will lead to hundreds or thousands compounds to test. Improvement in screening performance is therefore very significant. We want to further improve SVM based VS as a well accepted VS method like docking. Current models were generated by using two-tier supervised classification SVM methods [87-89, 91-94, 111]. The inactive compounds in these models have been collected from up to a few hundred known inactive compounds or/and putative inactive compounds from up to a few dozen biological target classes in MDDR database [87-89, 91-94, 111], which may not always be sufficient to fully represent inactive compounds in the vast chemical space, thereby making it difficult to optimally minimize false hit prediction rate of ML models. Han et al[114] has demonstrated the potential of putative negatives generation method in helping to increase the performance of SVM based VS methods. We will carry on the study to further improve the method to generate more diverse negatives for training. Besides SVM, some other common ML methods include artificial neural network (ANN), probabilistic neural network (PNN), k nearest neighbor (kNN), C4.5 decision tree (C4.5DT), linear discriminate analysis (LDA) and logistic regression (LR) were used. Some of these methods will be explained in Chapter 2 and attempted for comparison. Several types of pharmaceutical agents, including Src kinase inhibitors, VEGFR-2 inhibitors will be investigated. Moreover, our SVM based VS system is

---

also evaluated in terms of prediction on novel types structures because it is also one goal of VS [75].

## **1.5 Objective and outline of this thesis**

The ultimate goal of this thesis is to develop comprehensive databases to facilitate disease detection and drug discovery for the disease using computation methods.

Overall, there are three major objectives for this work:

1. To develop and update databases with enhanced storage, management, integration and provide the customized biological and chemistry information data for pathogen detection, disease diagnosis, therapeutic targets and drugs.
- 2 To develop SVM based virtual screening method for prediction of potential Src and VEGFR-2 inhibitors from large compound libraries and test the model experimentally.
- 3 To compare the virtual screening performances of several the machine learning methods SVM, kNN, PNN and similarity searching in identification of inhibitors.

The complete outline of this thesis is as follows:

**Chapter 1** describes pathogen induced diseases and their detection methods, and introduces background of cheminformatics and bioinformatics. Then the introduction of virtual screening methods is given.

---

**Chapter 2** shows methods used in this work, including database development method and procedure of the application of VS tools. In particular, data collection, theoretical backgrounds of machine learning methods, virtual screening model validation and performance measurements.

**Chapter 3 and Chapter 4** elaborate the development of MicrobPad MD: Microbial pathogen diagnostic methods database and update of therapeutic targets database.

**Chapter 5 and Chapter 6** are devoted to the application of our SVM based VS system for pharmaceutical agents Src and VEGFR-2 inhibitors from large compound libraries. In these chapters, SVM based VS system combined with a novel putative negative generation method is evaluated as a highly efficient VS tool. The comparison between kNN and PNN based VS model are described.

**Chapter 7** summarizes major findings and contributions of current work and also rationalizes the limitations and suggestions for future studies.

---

## **Chapter 2 Methodology**

*This chapter includes methods of database development and virtual screening for drug agents. The database development methods is usually consisted of the following four components: (1) database design; (2) data collection; (3) data integration and organization; (4) user interface. Methods of virtual screening include (1) Datasets collection and quality analysis; (2) Molecular descriptors calculation; (3) Machine learning methods; (4) Machine learning methods model development and evaluations;*

### **2.1 Database development**

A database is a well-organized data collection of information and their supporting data structures, typically in digital form. It involves the data and their supporting data structures. Database development is comprehensive and time consuming involving collecting relevant data, designing reasonable database scheme, integrating data from various resource, designing database interface and implementing database function.

#### **2.1.1 Database model and rational schema design**

A database model is a theoretical foundation of a database and fundamentally determines in which manner data can be stored, organized,

---

and manipulated in a database system. There are several different basic ways of constructing databases including flat file model, hierarchical model, network model, relational model, dimensional model, multi-value model and object-oriented model. The relational model has been extensively used in biological database development. Relational database comprises multiple tables of data, related to each other by primary keys and foreign keys. Each table is a collection of records and each record in a table has the same attributes. Relational database is the predominant form of database in use today, especially in biological research field. In this study, the relational model was applied in the database development. After relational model was selected, a rational schema is important for the database construction.

A rational schema is designed before the construction of the database to help define the scope of the database and focus on relevant problem. Information need to be collected to pave the way for the information collection stage. The database performance, the ease with which users retrieve information, the search engine coding and other database function implementations are greatly influenced by the schema design. Therefore, schema design is the fundamental step in planning a new database is to identify and design the

---

tables to be included in the database, specify their contents, and define the relationships among them.

Use MicroPad MD for an example, as described in Chapter 1, medical pathogens of bacterial, fungal, and viral species induce infections, disease and sometimes serious medical conditions in the infected hosts. Fast, accurate, sensitive and low-cost diagnosis of medical pathogens is important and desired for proper treatment and investigation of pathogenesis processes. To facilitate the development of diagnostic methods and device, MicrobPad MD was designed to provide comprehensive information about diagnostic technique, targets, and primers/probes for the known bacterial, fungal and viral pathogens. Based on this preliminary architecture, the more details schema including several tables and their relationships was built.

### **2.1.2 Data collection**

Generally, a database is supposed to provide enough domain knowledge around a specific subject together with information of related subjects. For instance, MicrobPad MD provides integrated molecular diagnostic information including molecular diagnostic techniques, targets, primers/probes, virulence factors, disease, etc. Data collection of these

---

information can be done by various ways. Data can be captured from literature, books, experiments or software output, customized data collected programmatically from other databases locally or over the internet, text mining by programs, and so on. Literature is typically on unstructured data source. Names of the subjects that are stored in different synonymous terms, various abbreviations, or totally different expressions, are difficult to be recognized by automatic language processing. It is very difficult to invent a fully automated literature information extraction system to gather useful information from literature efficiently. Manual data collection from literature or manual curation of collected data is considered be one of the most feasible ways for information data collection. However, it is too time consuming and expensive [124]. 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 [125].

In this work, automatic data retrieval methods with manual curation process was combined to ensure good quality. It is useful to have program parser to filter the data since the amount of biological data is generally very huge. Automated text retrieval programs developed in PERL with efficient use of regular expression were implemented in retrieving information from



---

literatures that contained the key word related to searching the subject via local Medline packages [126]. The useful subject information was selected manually and the full literature was referred to facilitate information searching. Meanwhile, the detail biological information of subject and cross-links were automatically extracted from some general or specific biological databases, such NCBI genome, SwissProt and UniProt. Moreover, an html parser was also written to parse some html pages with unstructured data. The information obtained by the program were extracted and verified manually.

### **2.1.3 Data integration and organization**

Data is still often available in an unstructured manner even when it does have a strong internal structure. Data integration is a necessary procedure when data from different sources needs to be standardized before implementation. The integration of biological and chemical data coming from various source sometimes become a big challenge. Improper integration can lead to missing of some part of data or even can induce mistakes. The correct way of data integration for biological databases can generally be divided into two parts: syntactic integration and semantic integration. In syntactic integration, data from different sources and of different file formats

---

are standardized to have single file format. In semantic integration, 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 (extensive markup languages). The structure of XML is such that it can hold data of various types such as simple plain table, 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 and database software. In this work, the powerful feature of XML has been utilized for various purposes e.g. collection of PubMed extracts for the medical pathogen as keywords using NCBI E-utilities. On the other hand, semantic data integration gives flexibility to mix complex biological data in semi structured way when it is difficult to standardize a part of data to the convention of unified single file format. In addition, data can be integrated manually. Information curation process is time consuming and tedious but sometimes it becomes indispensable to ensure data achieves high quality. Manual data integration is also achieved through scripting languages like PERL or Python which are handy to use yet very powerful. Scripts can help manipulate database tables by integrating plain unformatted text taken from literature or web page. Relied on the power of programming languages, major public

---

databases hosted by NCBI and EMBL provide data access service through user written program. As an example, E-utilities provides many example scripts to obtain customized data by constructing user defined pipeline over its database.

After data integration, standardized data will be organized based on the schema for efficient and effective data creation, storage and manipulation. Formal definitions of all the included information in a database, a couple of relational tables of relevant data and the relationship among them will be established. In these relation tables, certain fields may be designated as keys, by which the separated tables can be linked together for facilitating to search specific values of that field. Primary key uniquely identifies each record in the table. Foreign key matching the primary key of other tables can be used to cross-reference tables.

#### **2.1.4 Database management system**

After database construction, the database should be organized and managed effectively by Database Management System (DBMS). DBMS is a couple of programs and tools that used to store, maintain, and extract information from a database [127]. There are several DBMS software available e.g. Oracle,

---

Microsoft SQL Server, Access, MySQL and PostgreSQL. In this work, Oracle and Access based relational database management systems have been built to manage involving define, create and modify the various information as well as privilege. By using Structured Query Language (SQL) queries, all entry data from the related tables can therefore be retrieved together for display and output.

### **2.1.5 User Interface**

A web based user interface allows the user to submit query, obtain data and interact with database. Without user interface, the database is less useful to end user even if the database is complete, well organized and maintained. A good database interface can help the user get information stored in database quickly and convenient as well as multi-level search capability. A bad interface will cause difficulty in searching, locating and displaying data. There are two main categories of web based user interface: static web pages and dynamic web pages.

Static web pages are also called flat page or stationary page. In contrast to dynamic webpage, static web pages are delivered to user exactly as stored and the same information to all users. They are usually HTML document

---

files stored in the file system and are available through the HTTP web service. Dynamic pages present various content to different users according to the parameters provided by them. Dynamic pages are often produced by Common Gateway Interface (CGI) with the assistant of server-side languages such as Active Server Pages (ASP), Java Server Pages (JSP), Perl, Hypertext Preprocessor (PHP) and other languages. The client side dynamic web page creation is generally achieved through JavaScript or ActionScript.

In this work, ASP technology is used for server side dynamic web page creation and JavaScript is used for client side dynamic web page creation. Server side dynamic web page generation over database includes submission of user customized query to web server which further interacts with DBMS such as MySQL and Oracle. The client side technology is based on Internet browsers with support of JavaScript e.g. Internet Explorer, Mozilla Firefox and Google Chrome to extract and display the data. The client side dynamic web page is generally used to present content more friendly and convenient such as change in color or short string tips when mouse is on or off some part of the content.

---

## 2.2 Dataset collection and preprocess for building models

### 2.2.1 Dataset resource

Currently, massive amount of data about small molecules and their related annotation information have been accumulated in scientific literatures and cheminformatics databases. **Table 1-5** lists some of the widely known small molecule databases. For instance, BindingDB is a public, web-accessible database of measured binding affinities, focusing chiefly on the interactions of protein targets with small drug-like molecules. DrugBank is also a unique bioinformatics and cheminformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information. MDDR is a database covering the patent literature, journals, meetings and congresses produced by Symyx and Prous Science.

The datasets used in this work mainly are retrieved from the following two types of sources. First, we collected small molecular data from credible journals such as Bioorganic & Medicinal Chemistry Letters, Bioorganic & Medicinal Chemistry, European Journal of Medicinal Chemistry, European Journal of Organic Chemistry and Journal of Medicinal Chemistry, etc. Second, we use cheminformatics databases that contain accurate and reliable data such as PubChem and ChEMBL [128].

---

### **2.2.2 Dataset quality**

The sufficient and high quality data with low experimental errors is important for the development of reliable machine learning classification model which depends on the quality and quantity of data. Many factors such as quality, size and relevance of the dataset can affect machine learning process greatly. The data quality is usually assessed during the produce of data collection. The data collected from less credit resource will lead to faulty models which will induce weak predictive power or even wrong prediction. Ideally, the measurements of pharmacological data properties should be conducted with a same protocol so that there is a common ground to compare different compounds with each other. However, some pharmacological properties measurements have been used only for a limited number of compounds and most pharmacological properties measurements are rarely determined by the same protocol. Thus the collected data consist of compound data measured by different protocols and the incorporation of additional experimental information. To maintain the stability of data quality, in this work, several methods are adopted to ensure that inter-laboratory variations caused by different experimental protocols do not significantly affect the quality of the training sets. The pharmacological property measurements for data were investigated and the ones that contain large

---

variations in experimental protocols compared to the majority of the data are filtered. It is estimated that the most common range of the pharmacological properties measurements for the compounds investigated in more than one source was used to select compounds for the different classes [129].

In this work, the data were collected from varied sources. This approach can enrich the diversity in the datasets and reduce the potential bias that may arise from a monotonic due to the preferences of the researchers. However, since the data are presented by independent researchers who don't share pre-existing agreement on their individual data collection. It is likely that there is a certain level of redundancy between the datasets from different sources. The redundancy could contrarily deduce diversity in the datasets. Therefore, 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.

### **2.2.3 Dataset structural diversity**



---

Diversity Index (DI) is applied to evaluate the structural diversity of a collection of compounds. It is defined as the average value of the similarity between pairs of compounds in a dataset [130],

$$DI = \frac{\sum_{i,j \in D \wedge i \neq j} sim(i, j)}{|D|(|D| - 1)} \quad (1)$$

where  $sim(i, j)$  is a measure of similarity between compounds  $i$  and  $j$ ,  $D$  is the dataset and  $|D|$  is set cardinality (number of elements of the set). The dataset is more diverse when DI approaches 0. Tanimoto coefficient [131] were used to compute  $sim(i, j)$  in this study,

$$sim(i, j) = \frac{\sum_{d=1}^k x_{d_i} x_{d_j}}{\sum_{d=1}^k (x_{d_i})^2 + \sum_{d=1}^k (x_{d_j})^2 - \sum_{d=1}^k x_{d_i} x_{d_j}} \quad (2)$$

where  $k$  is the number of descriptors calculated for the compounds in the datasets. A compound  $i$  is considered to be similar to a known active  $j$  in the active dataset if the corresponding  $sim(i, j)$  value is greater than a cut-off value.

### 2.3 Molecular descriptor

Molecular descriptors are frequently used to quantitatively represent various physicochemical or structural properties of molecules for many

---

computational studies small molecules. A descriptor is the final result of a logical and mathematical procedure which transforms chemical information encoded within a symbolic representation of a compound into a useful number or the result of some standardized experiment. Molecular descriptors have been extensively used in deriving structure-activity relationships [132, 133], quantitative structure activity relationships [134, 135], and machine learning prediction models for pharmaceutical agents [136-143]. They represent compounds in the form of mathematical vectors. This transformation enables the statistical analysis of chemical compounds.

A number of programs e.g. DRAGON[144], Molconn-Z[145], MODEL[146], Chemistry Development Kit(CDK) [147, 148], JOELib [149], and Xue descriptor set [140] are available to calculate chemical descriptors. These methods can be used for deriving >3,000 molecular descriptors including constitutional descriptors, topological descriptors, RDF descriptors [150], molecular walk counts [151], 3D-MoRSE descriptors [152], BCUT descriptors [153], WHIM descriptors [154], Galvez topological charge indices and charge descriptors [155], GETAWAY descriptors [156], 2D autocorrelations, functional groups, atom-centred descriptors, aromaticity indices [157], Randic molecular profiles [158], electrotopological state

descriptors [159], linear solvation energy relationship descriptors [160], and other empirical and molecular properties. Not all of the available descriptors are needed for representing features of a particular class of compounds. Moreover, without properly selecting the appropriate set of descriptors, the performance of a developed machine learning VS tool may be affected to some degrees due to the noise arising from the high redundancy and overlapping of the available descriptors.

In this work, 98 1D and 2D descriptors are computed which are widely used in machine learning based virtual screening models. These 98 descriptors were selected from the descriptors derived from MODEL program by discarding those that were redundant and unrelated to the problem studied here. These 98 descriptors are showed in **Table 2-1**.

**Table 2-1** 98 molecular descriptors used in this work.

| <b>Descriptor Class</b>     | <b>No of Descriptors in Class</b> | <b>Descriptors</b>   |
|-----------------------------|-----------------------------------|--|
| Simple molecular properties | 18                                | Number of C,N,O,P,S, Number of total atoms, Number of rings, Number of bonds, Number of non-H bonds, Molecular weight,, Number of rotatable bonds, number of H-bond donors, number of H-bond acceptors, Number of 5-member aromatic rings, Number of 6-member aromatic rings, Number of N heterocyclic rings, Number |

|                                  |    |   |
|----------------------------------|----|---|
|                                  |    | of O heterocyclic rings, Number of S heterocyclic rings.  |
| Chemical properties              | 3  | Sanderson electronegativity, Molecular polarizability, ALogp  |
| Molecular Connectivity and shape | 35 | Schultz molecular topological index, Gutman molecular topological index, Wiener index, Harary index, Gravitational topological index, Molecular path count of length 1-6, Total path count, Balaban Index J, 0-2th valence connectivity index, 0-2th order delta chi index, Pogliani index, 0-2th Solvation connectivity index, 1-3th order Kier shape index, 1-3th order Kappa alpha shape index, Kier Molecular Flexibility Index, Topological radius, Graph-theoretical shape coefficient, Eccentricity, Centralization, Logp from connectivity. |
| Electro-topological state        | 42 | Sum of Estate of atom type sCH3, dCH2, ssCH2, dsCH, aaCH, sssCH, dssC, aasC, aaaC, sssC, sNH3, sNH2, ssNH2, dNH, ssNH, aaNH, dsN, aaN, sssN, ddsN, aOH, sOH, ssO, sSH; Sum of Estate of all heavy atoms, all C atoms, all hetero atoms, Sum of Estate of H-bond acceptors, Sum of H Estate of atom type HsOH, HdNH, HsSH, HsNH2, HssNH, HaaNH, HtCH, HdCH2, HdsCH, HaaCH, HCsat, HCsat, Havin, Sum of H Estate of H-bond donors   |

In this work, the 2D structure of each of the compounds was generated by using ChemDraw [106] or downloaded from databases such as PubChem and BindingDB [161]. Then they were subsequently converted into 3D structure by using CORINA [162]. The 3D structure of each compound was manually inspected to ensure the proper chirality of each chiral agent. All salts and elements, such as sodium or calcium, were removed prior to descriptor

---

calculation. The optimization of generated geometries was conducted without symmetry restrictions. The 3D structures of the compounds then were used to compute the molecular descriptors by the in-house programs and scripts.

## 2.4 Scaling of molecular descriptors

Molecular descriptors are usually scaled before they can be employed for machine learning methods. Scaling of chemical descriptors ensures that each of descriptor have unbiased contribution in constructing the prediction models[163]. Scaling can be done by various of ways e.g. auto-scaling, range scaling, Pareto scaling [164], and feature weighting [165, 166]. 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:

$$d_{ij}^{scaled} = \frac{d_{ij} - d_{j,min}}{d_{j,max} - d_{j,min}} \quad (3)$$

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.

---

## **2.5 Machine learning classification methods**

A machine learning (ML) method takes a training set of objects that have previously been classified into two or more classes as input.

Machine learning classification methods employ computational and statistical methods to construct mathematical models from a training set of objects 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 divided into two types: Supervised and Unsupervised. Supervised machine learning, as the name indicates, generally needs feeding which generally involve already labeled or classified training data. Example of supervised machine learning includes SVM, ANN, 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 SVM, PNN, kNN. They are explained below in subsequent sub sections. For a comparative study, Tanimoto similarity searching method is

also introduced. Websites for codes of some machine learning methods are given in **Table 2-2**.

**Table 2-2** Websites that contain codes of machine learning methods

| <b>Decision Tree</b>                        |   |
|---|---|
| PrecisionTree                               | <a href="http://www.palisade.com.au/precisiontree/">http://www.palisade.com.au/precisiontree/</a>   |
| DecisionPro                                 | <a href="http://www.vanguardsw.com/decisionpro/jdtree.htm">http://www.vanguardsw.com/decisionpro/jdtree.htm</a>   |
| C4.5  | <a href="http://www2.cs.uregina.ca/~hamilton/courses/831/notes/ml/dtrees/c4.5/tutorial.html">http://www2.cs.uregina.ca/~hamilton/courses/831/notes/ml/dtrees/c4.5/tutorial.html</a> |
| C5.0  | <a href="http://www.rulequest.com/download.html">http://www.rulequest.com/download.html</a>   |
| <b>KNN</b>                                  |   |
| k Nearest Neighbor                          | <a href="http://www.cs.cmu.edu/~zhuxj/courseproject/knndemo/KNN.html">http://www.cs.cmu.edu/~zhuxj/courseproject/knndemo/KNN.html</a>   |
| PERL Module for KNN                         | <a href="http://aspn.activestate.com/ASPN/CodeDoc/AI-Categorize/AI/Categorize/kNN.html">http://aspn.activestate.com/ASPN/CodeDoc/AI-Categorize/AI/Categorize/kNN.html</a>           |
| Java class for KNN                          | <a href="http://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/classify/old/KNN.html">http://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/classify/old/KNN.html</a> |
| <b>LDA</b>                                  |   |
| DTREG                                       | <a href="http://www.dtreg.com/lda.htm">http://www.dtreg.com/lda.htm</a>   |
| <b>LR</b>                                   |   |
| Paul Komarek's Logistic Regression Software | <a href="http://komarix.org/ac/lr/lrtrirls">http://komarix.org/ac/lr/lrtrirls</a>   |
| Web-based logistic regression calculator    | <a href="http://statpages.org/logistic.html">http://statpages.org/logistic.html</a>   |
| <b>Neural Network</b>                       |   |
| BrainMaker                                  | <a href="http://www.calsci.com/">http://www.calsci.com/</a>   |
| Libneural                                   | <a href="http://pochat.online.fr/webus/tutorial/BPN_tutorial7.html">http://pochat.online.fr/webus/tutorial/BPN_tutorial7.html</a>   |
| fann  | <a href="http://leenissen.dk/fann/">http://leenissen.dk/fann/</a>   |
| NeuralWorks Predict                         | <a href="http://www.neuralware.com/products.jsp">http://www.neuralware.com/products.jsp</a>   |
| NeuroShell Predictor                        | <a href="http://www.mbaware.com/neurpred.html">http://www.mbaware.com/neurpred.html</a>   |
| <b>SVM</b>                                  |   |
| SVM light                                   | <a href="http://svmlight.joachims.org/">http://svmlight.joachims.org/</a>   |
| LIBSVM                                      | <a href="http://www.csie.ntu.edu.tw/~cjlin/libsvm/">http://www.csie.ntu.edu.tw/~cjlin/libsvm/</a>   |
| mySVM                                       | <a href="http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/index.html">http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/index.html</a>   |
| BSVM  | <a href="http://www.csie.ntu.edu.tw/~cjlin/bsvm/">http://www.csie.ntu.edu.tw/~cjlin/bsvm/</a>   |
| SVMTorch                                    | <a href="http://www.idiap.ch/learning/SVMTorch.html">http://www.idiap.ch/learning/SVMTorch.html</a>   |

---

### 2.5.1 Support vector machine (SVM)

Support vector machine (SVM) is based on the structural risk minimization principle of statistical learning theory [167, 168], which consistently shows outstanding classification performance, is less penalized by sample redundancy, and has lower risk for over-fitting [169, 170].

In linearly separable cases, SVM constructs a hyper-plane to separate active and inactive classes of compounds with a maximum margin. A compound is represented by a vector  $\mathbf{x}_i$  composed of its molecular descriptors. The hyper-plane is constructed by finding another vector  $\mathbf{w}$  and a parameter  $b$  that minimizes  $\|\mathbf{w}\|^2$  and satisfies the following conditions:

$$\mathbf{w} \cdot \mathbf{x}_i + b \geq +1, \text{ for } y_i = +1 \text{ Class 1 (active)} \quad (4)$$

$$\mathbf{w} \cdot \mathbf{x}_i + b \leq -1, \text{ for } y_i = -1 \text{ Class 2 (inactive)} \quad (5)$$

where  $y_i$  is the class index,  $\mathbf{w}$  is a vector normal to the hyperplane,  $|b|/\|\mathbf{w}\|$  is the perpendicular distance from the hyperplane to the origin and  $\|\mathbf{w}\|^2$  is the Euclidean norm of  $\mathbf{w}$ . Based on  $\mathbf{w}$  and  $b$ , a given vector  $\mathbf{x}$  can be classified by  $f(\mathbf{x}) = \text{sign}[(\mathbf{w} \cdot \mathbf{x}) + b]$ . A positive or negative  $f(\mathbf{x})$  value indicates that the vector  $\mathbf{x}$  belongs to the active or inactive class respectively.

In nonlinearly separable cases, which almost always occur in classifying compounds of diverse structures [111, 171-177], SVM maps the input



---

vectors into a higher dimensional feature space by using a kernel function  $K(\mathbf{x}_i, \mathbf{x}_j)$ . We used RBF kernel  $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|\mathbf{x}_j - \mathbf{x}_i\|^2 / 2\sigma^2}$  which has been extensively used and consistently shown better performance than other kernel functions [178-180]. Linear SVM can then applied to this feature space based on the following decision function:

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^l \alpha_i^0 y_i K(\mathbf{x}, \mathbf{x}_i) + b\right),$$

where the coefficients  $\alpha_i^0$  and  $b$  are

determined by maximizing the following Langrangian expression:

$$\sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

under the conditions  $\alpha_i \geq 0$  and

$$\sum_{i=1}^l \alpha_i y_i = 0.$$

A positive or negative  $f(\mathbf{x})$  value indicates that the vector  $\mathbf{x}$

belongs to the active or inactive class respectively. For a given training set of instance-label pairs  $(x_i, y_i)$ ,  $i=1, \dots, l$  where  $x_i \in R^n$  and  $y_i \in \{1, -1\}$  in  $l$ , in SVM, the task of finding the hyper-plane which is able to separate active and inactive classes with a maximum margin, in essence, is to look for the solution of the following optimization problem:

$$\min_{w, b, \xi} \frac{1}{2} w^T w + C \sum_{i=1}^l \xi_i$$

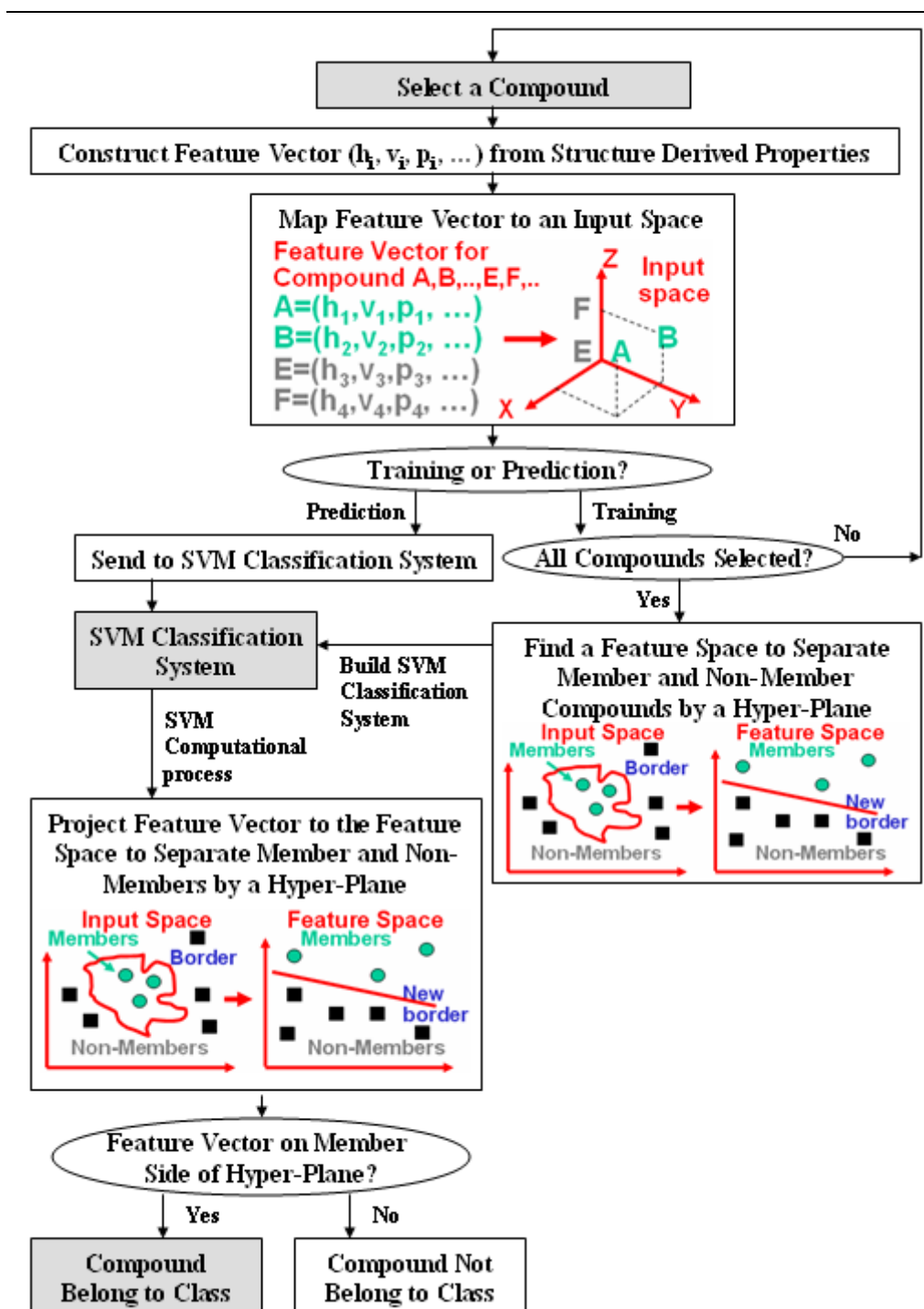
$$\text{subject to } y_i(w^T \Phi(x_i) + b) \geq 1 - \xi_i,$$

$$\xi_i \geq 0.$$

---

In developing our SVM VS tool, a hard margin  $C=100,000$  was used. The margin parameter  $C$  is penalty parameter that controls the trade-off between the training errors and sample separation. Increasing  $C$  imposes a higher penalty for training errors. Our chosen value corresponds to a very high penalty. Software LibSVM [181], an integrated software for support vector classification, regression and distribution estimation, was chosen to do the machine learning in this work.

The process of training and using a SVM VS model for screening compounds based on their molecular descriptors is schematically illustrated in **Figure 2-1**.



**Figure 2-1** Schematic diagram of the process of the training a prediction model and using it for predicting active compounds of a compound class from their structurally-derived properties (molecular descriptors) by using support vector machines; A, B, E, F and  $(h_j, p_j, v_j, \dots)$  represents such structural and physicochemical properties as hydrophobicity, volume, polarizability, etc.

---

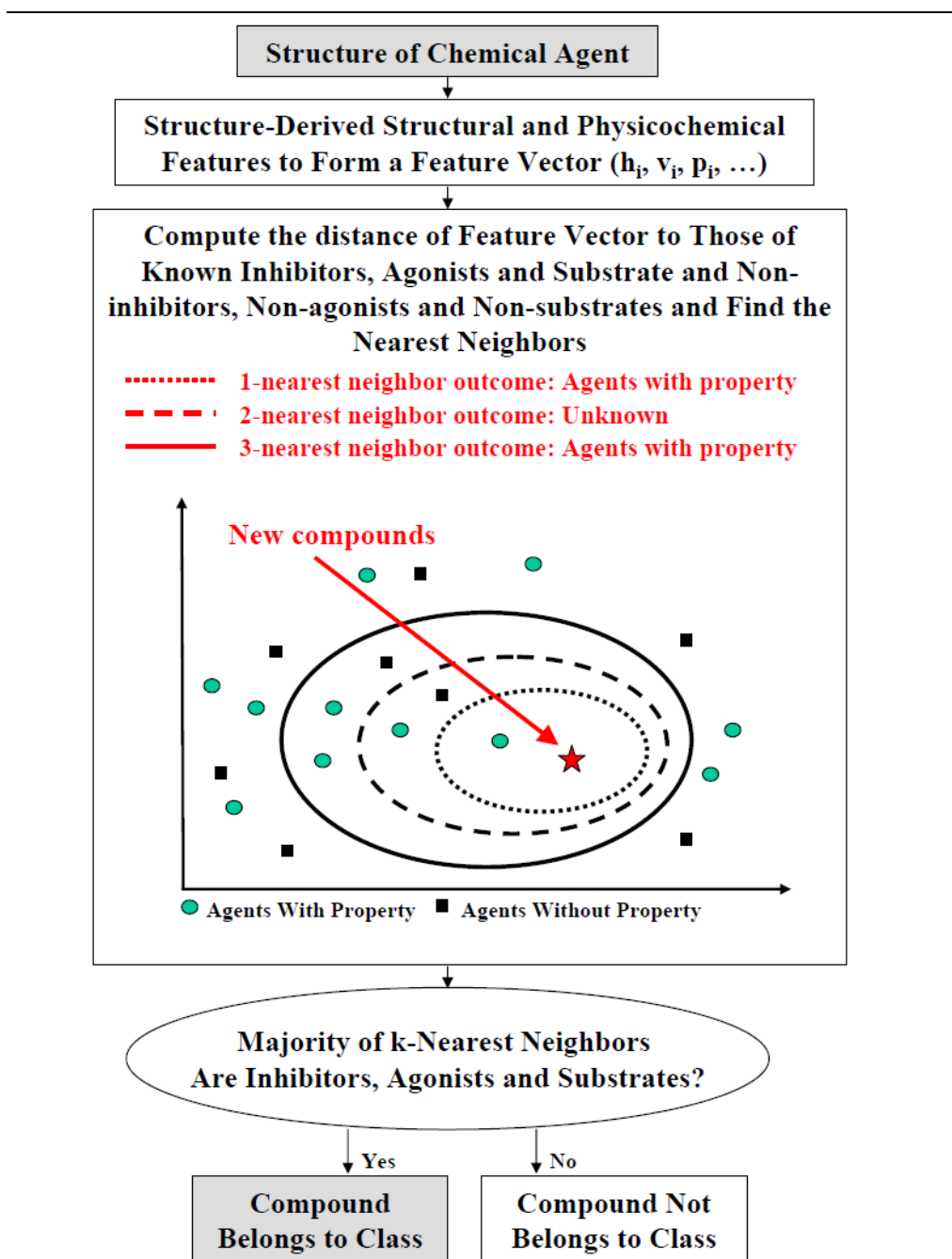
## 2.5.2 k-nearest neighbors (kNN)

kNN measures the Euclidean distance  $D = \sqrt{\|\mathbf{x} - \mathbf{x}_i\|^2}$  between a compound  $\mathbf{x}$  and each individual inhibitor or non-inhibitor  $\mathbf{x}_i$  in the training set [182, 183].

A total of  $k$  number of vectors nearest to the vector  $\mathbf{x}$  are used to determine the decision function  $f(\mathbf{x})$ :

$$\hat{f}(\mathbf{x}) \leftarrow \arg \max_{v \in V} \sum_{i=1}^k \delta(v, f(\mathbf{x}_i)) \quad (6)$$

where  $\delta(a, b) = 1$  if  $a = b$  and  $\delta(a, b) = 0$  if  $a \neq b$ ,  $\arg \max$  is the maximum of the function,  $V$  is a finite set of vectors  $\{v_1, \dots, v_s\}$  and  $\hat{f}(\mathbf{x})$  is an estimate of  $f(\mathbf{x})$ . Here estimate refers to the class of the majority compound group (i.e. inhibitors or non-inhibitors) of the  $k$  nearest neighbors. The procedure of kNN is illustrated in **Figure 2-2**.



**Figure 2-2** Schematic diagram illustrating the process of the prediction of compounds of a particular property from their structure by using k-nearest neighbors (kNN). Feature vector ( $h_j, p_j, v_j, \dots$ ) represents such structural and physicochemical properties as hydrophobicity, volume, polarizability, etc; green dots: agents with the property; black box : agents without the property.

---

### 2.5.3 Probabilistic neural network (PNN)

Probabilistic neural network (PNN) belongs to the neural network methods. It is designed for classification through the use of Bayes' optimal decision rule [129]:  $h_i c_i f_i(\mathbf{x}) > h_j c_j f_j(\mathbf{x})$ , where  $h_i$  and  $h_j$  are the prior probabilities,  $c_i$  and  $c_j$  are the costs of misclassification and  $f_i(x)$  and  $f_j(x)$  are the probability density function for class  $i$  and  $j$  respectively. An unclassified vector  $\mathbf{x}$  is classified into population  $i$  if the product of all the three terms is greater for class  $i$  than for any other class  $j$  (not equal to  $i$ ). In most applications, the prior probabilities and costs of misclassifications are treated as being equal. The probability density function for each class for a univariate case can be estimated by using the Parzen's nonparametric estimator[184],

$$g(\mathbf{x}) = \frac{1}{n\sigma} \sum_{i=1}^n W\left(\frac{\mathbf{x} - \mathbf{x}_i}{\sigma}\right) \quad (7)$$

where  $n$  is the sample size,  $\sigma$  is a scaling parameter which defines the width of the bell curve that surrounds each sample point,  $W(d)$  is a weight function which has its largest value at  $d = 0$  and  $(\mathbf{x} - \mathbf{x}_i)$  is the distance between the unknown vector and a vector in the training set. The Parzen's nonparametric estimator was later expanded by Cacoullos for the multivariate case.

$$g(x_1, \dots, x_p) = \frac{1}{n\sigma_1 \dots \sigma_p} \sum_{i=1}^n W\left(\frac{x_1 - x_{1,i}}{\sigma_1}, \dots, \frac{x_p - x_{p,i}}{\sigma_p}\right) \quad (8)$$

---

The Gaussian function is frequently used as the weight function because it is well behaved, easily calculated and satisfies the conditions required by Parzen's estimator. Thus the probability density function for the multivariate case becomes

$$g(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \exp\left(-\sum_{j=1}^p \left(\frac{x_j - x_{ij}}{\sigma_j}\right)^2\right) \quad (9)$$

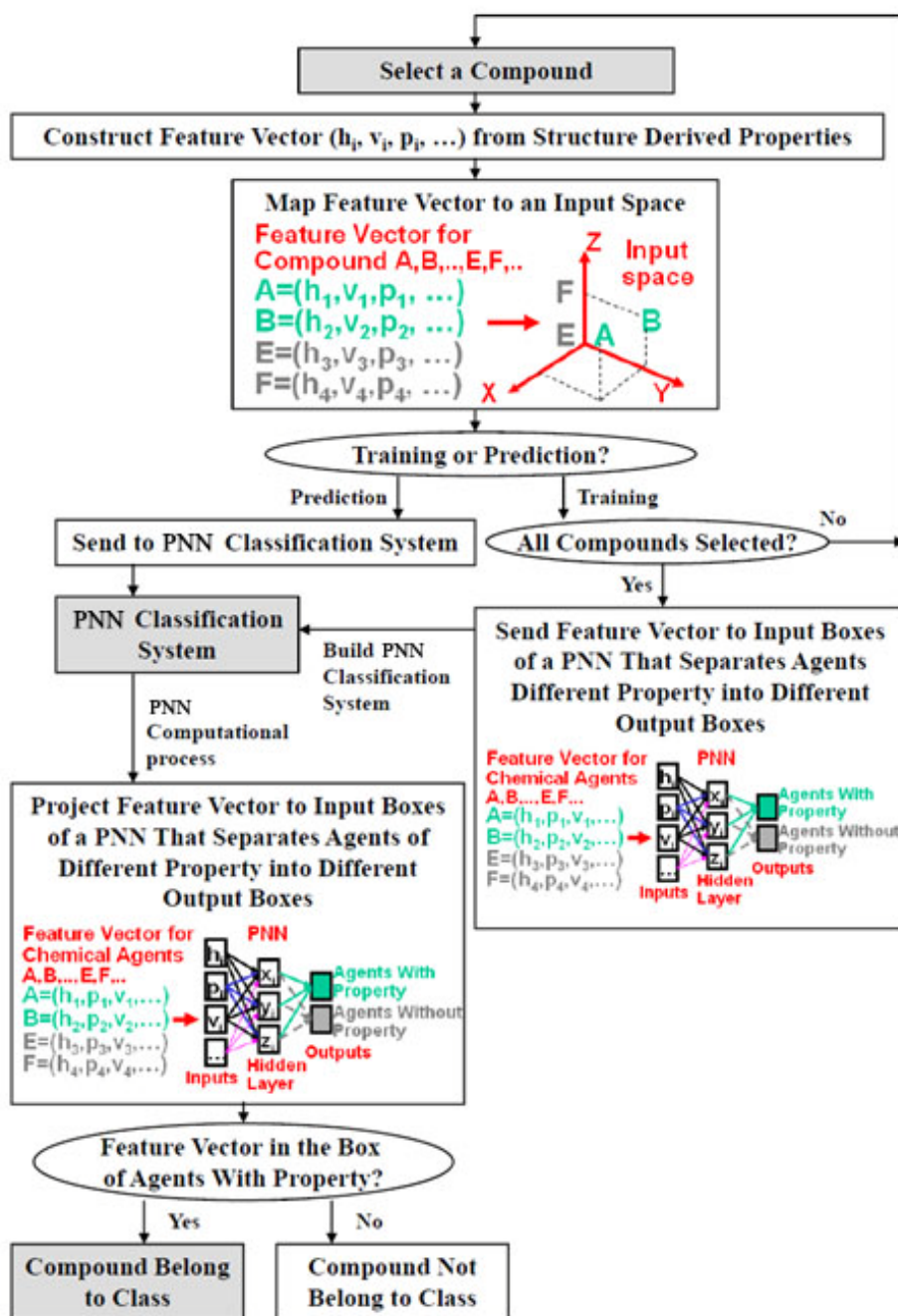
The network architectures of PNN are determined by the number of compounds and descriptors in the training set. PNN are constituted of four layers, the input layer, the pattern layer, the summation layer and the output layer. The input layer provides input values to all neurons in the pattern layer and has as many neurons as the number of descriptors in the training set. The number of pattern neurons is determined by the total number of compounds in the training set. Each pattern neuron computes a distance measure between the input and the training case represented by that neuron and then subjects the distance measure to the Parzen's nonparametric estimator. The summation layer has a neuron for each class and the neurons sum all the pattern neurons' output corresponding to members of that summation neuron's class to obtain the estimated probability density function for that class. Finally, the single neuron in the output layer then estimates the class of the unknown vector  $\mathbf{x}$  by comparing all the probability density function from the summation neurons

---

and choosing the class with the highest probability density function. **Figure**

**2-3** illustrates the procedure of PNN method.





**Figure 2-3** Schematic diagram illustrating the process of the prediction of compounds of a particular property from their structure by using probabilistic neural networks (PNN). A, B: feature vectors of agents with the property; E, F: feature vectors of agents without the property; feature vector  $(h_j, p_j, v_j, \dots)$  represents such structural and physicochemical properties as hydrophobicity, volume, polarizability, etc.

---

## 2.5.4 Tanimoto similarity searching method

Determining if two compounds are similar to each other or not in a training dataset can be conducted by using the Tanimoto coefficient  $sim(i,j)$  [131]

$$sim(i, j) = \frac{\sum_{d=1}^l x_{di} x_{dj}}{\sum_{d=1}^l (x_{di})^2 + \sum_{d=1}^l (x_{dj})^2 - \sum_{d=1}^l x_{di} x_{dj}} \quad (10)$$

where  $l$  is the number of molecular descriptors. A compound  $i$  is considered to be similar to a known active  $j$  in the active dataset if the corresponding  $sim(i,j)$  value is greater than a cut-off value. In this work, in computing  $sim(i,j)$ , the molecular descriptor vectors  $\mathbf{x}_i$ s were scaled with respect to all of the MDDR. The cut-off values for similarity compounds are typically in the range of 0.8 to 0.9 [185, 186]. A stricter cut-off value of 0.9 was used in this work.

## 2.5.5 Generation of putative negatives

Both positive data (e.g. active compounds) and negative data (e.g. inactive compounds) are compulsory for building machine learning prediction models. As for prediction of compound inhibitors, positives can be formed from known active compounds but negatives are usually lacking. Previous studies have used known inactive compounds and active compounds of other biological target classes as putative inactive compounds [87, 111, 171-174, 187, 188]. In our group a new approach extensively used for generating

---

inactive proteins in SVM classification of various functional classes of proteins [189-191] has been attempted for generating putative inactive compounds[114]. An advantage of this approach is its independence on the knowledge of known inactive compounds and active compounds of other biological target classes, which enables more expanded coverage of the “inactive” chemical space in cases of limited knowledge of inactive compounds and compounds of other biological classes. A drawback of this approach is the possible inclusion of some yet-to-be-discovered active compounds in the “inactive” class, which may affect the capability of SVM for identifying novel active compounds. As has been demonstrated in an earlier study[114], such an adverse effect is expected to be relatively small for many biological target classes. In applying this approach to proteins, all known proteins are clustered into ~8,933 protein domain families in based on the clustering of their amino acid sequences [149], and a set of putative inactive proteins can be tentatively extracted from a few representative proteins in those families without a single known active protein. Undiscovered active proteins of a specific functional class typically cover no more than a few hundred families, which gives a maximum possible “wrong” family representation rate of <10.2% even when all of the undiscovered active proteins are misplaced into the inactive class [192]. Importantly,

---

inclusion of the representative of a “wrong” family into the inactive class does not preclude other active family members from being classified as active. Statistically, a substantial percentage of active members can be classified by ML methods as active even if its family representative is in the inactive class [114, 192]. Therefore, in principle, a reasonably good SVM classification model can be derived from these putative inactive samples, which has been confirmed by a number of studies of proteins [189-192].

In a similar manner, known compounds can be grouped into compound families by clustering them in the chemical space defined by their molecular descriptors [193, 194]. As SVM predict compound activities based on their molecular descriptors, in developing SVM VS tools, it makes sense to cluster as well as to represent compounds in terms of molecular descriptors. By using a K-means method [193, 194] and molecular descriptors computed from our own software [195], we generated 8,423 compound families from the 13.56M compounds in the PUBCHEM and MDDR databases that we were able to compute the molecular descriptors, which is consistent with the 12,800 compound-occupying neurons (regions of topologically close structures) for 26.4 million compounds of up to 11 atoms [65], and the 2,851 clusters for 171,045 natural products [196].

---

The number of compound inhibitors of a specific target is usually around 1000 and distributed in several hundred families respectively. Because of the extensive effort in searching the known compound libraries for identifying active compounds in these target classes, the number of undiscovered “active” families in PUBCHEM database is expected to be relatively small, most likely no more than several hundred families. The ratio of the discovered and undiscovered “active” families (hundreds) and the families that contain no known active compound (~8423 based on the current versions of PUBCHEM and MDDR) for these and possibly many other target classes is expected to be <15%. Therefore, putative inactive training datasets can be generated by extracting a few representative compounds of those families that contain no known active compound in the active training set, with a maximum possible “wrong” family representation rate of <15% even when all of the undiscovered active compounds are misplaced into the inactive class, and with the expectation that a substantial percentage of active members in the putative “inactive” families can be classified as active despite of their family representatives are placed into the inactive training sets. As has been shown in a study of SVM VS tools, a substantial percentage of identified virtual hits are from these “inactive” families [114].

---

## **2.6 Virtual screening model optimization, validation and performance measurements**

### **2.6.1 Model optimization and validation**

*In-silico* modeling offers the prediction of the pharmacological properties of compounds which have not been clinically or biologically tested. Therefore it is important to estimate and validate the predicting ability of the pharmacological-data-derived models by their performances with the compounds that are not present in the training set. In this work, 5-fold cross-validation and independent validation datasets were used for this purpose. In 5-fold cross-validation, compounds are randomly divided into five subsets of approximately equal size. Four subsets are used as the training set for developing a model; the remaining one is used as a testing set for evaluating the prediction performance of the model. This procedure is repeated five times such that every subset is used as a testing set once. Through this procedure, the optimal parameter can be obtained. Models have types of parameters that must be optimized. In this work, SVM is trained by using a radial basis function kernel which has an adjustable parameter  $\sigma$ . For PNN, the only parameter to be optimized is a scaling parameter  $\sigma$ . In kNN, the optimum number of nearest neighbors,  $k$ , needs to be derived for each training set. Optimization of the parameter for each of these methods is conducted by

---

scanning the parameter through a range of values. The average accuracy of the five time models is seen as the accuracy predicting capability of the model constructed with the machine learning method. Five-fold cross-validation can reflect the average performance of a model, however, it has the tendency of underestimating the prediction capability of a classification model, especially if important molecular features happen to be contained only in a minority of the compounds in the training set [197, 198]. Hence if a model has relatively low cross-validation accuracy, it can still be predictive [197]. Therefore, cross-validation alone is not decisive to the performance of a model. To complement cross-validation, independent validation datasets are used. They may provide a more reliable estimation of the prediction capability of a pharmacological property prediction model [199, 200]. The independent validation dataset should be strictly independent from the training.

### **2.6.2 Performance evaluation**

Measurements such as sensitivity, specificity and the overall prediction accuracy are employed to quantitatively assess the performance of virtual screening models. They are defined in terms of true positives TP (pharmaceutical agents possessing a specific pharmacological property), true negatives TN (pharmaceutical agents not possessing a specific

---

pharmacological property), false positives FP (pharmaceutical agents not possessing a specific pharmacological property but predicted as agents possessing the specific pharmacological property) and false negatives FN (pharmaceutical agents possessing a specific pharmacological property but predicted as agents not possessing the specific pharmacological property). Sensitivity and specificity are the measurement of prediction accuracy for pharmaceutical agents possessing a specific pharmacological property and agents not possessing that pharmacological property respectively. The overall prediction accuracy (Q) and Matthews correlation coefficient (MCC) [201] are used to measure the overall prediction performance. They are defined as follows:

$$SE = \frac{TP}{TP + FN} \quad (11)$$

$$SP = \frac{TN}{TN + FP} \quad (12)$$

$$Q = \frac{TP + TN}{TP + TN + FP + FN} \quad (13)$$

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \quad (14)$$

The typical measurements of a model performance in screening large libraries include [202] yield (percentage of known positives predicted as virtual hits), hit-rate (percentage of virtual hits that are known positives), false hit-rate (percentage of virtual hits that are known negatives) and enrichment factor EF (magnitude of hit-rate improvement over random selection):



---

$$\text{Yield} = SE \quad (15)$$

$$\text{Hit-rate} = TP/(TP+FP) \quad (16)$$

$$\text{False hit-rate} = FP/(TP+FP) \quad (17)$$

$$\text{Enrichment factor EF} = \text{hit-rate} / (TP+FN)/(TP+FN+TN+FP) \quad (18)$$

### 2.6.3 Overfitting problem and its detection

Overfitting is a major concern in machine learning classification methods. It happens when a model that agrees well with the observed data but has no predictive ability, which means it does not have any value to unseen or future data. There are two main types of overfitting situations: (1) a model more flexible than it needs to be and (2) a model including irrelevant descriptors [198]. An over-fitted classification system tends to obtain much higher prediction accuracies in the cross-validation sets than in the independent validation sets. Hence frequently used method for checking whether a model is overfitted is to compare the prediction accuracies in the cross-validation procedure with those found in testing independent validation sets [198].

---

## **Chapter 3 Development of MicrobPad MD: microbial pathogen diagnostic methods database**

### **3.1 Introduction**

Medical pathogens of bacterial, fungal, and viral species induce infections, illnesses and sometimes serious medical conditions in the infected hosts [35, 36, 203, 204]. Diagnosis of these pathogens is important for proper treatment and investigation of pathogenesis processes, and extensive efforts have been made for developing molecular techniques that enable fast, accurate, sensitive and low-cost diagnosis of these pathogens [33-36]. Based on these molecular techniques, advanced diagnostic devices have been developed for a number of medical pathogens [35, 205]. More devices are needed for comprehensive coverage and faster diagnosis of medical pathogens, and for direct detection of multiple species [33, 205, 206].

Several databases have been developed and explored for providing the information and tools about the molecular diagnostic methods of specific classes of pathogenic species. For instance, the RIDOM website provides medical micro-organism differentiation services based on the analysis of small subunit ribosomal 16S rDNA sequences [207]. An expanded MicroSeq 500 16S rDNA sequence library database [208] and an integrated database network system [209] are useful for the identification

---

of *nocardia* species. The fourth international spoligotyping database [210] has been explored for the identification of *mycobacterium* species. A three-locus DNA sequence database is useful for the identification of the 69 *Fusarium* species associated with human or animal mycoses [211]. The 16SpathDB database supports automated identification of medically important bacteria by 16S rRNA gene sequencing [212]. Another database provides pulsed-field gel electrophoresis patterns of epidemic-type oxacillin-resistant *Staphylococcus aureus* strains [213]. GenoBASE-pylori is useful for genotype searching of the human gastric pathogen *Helicobacter pylori* [214]. TrED is a relational database that provides integrated access to various expression data of *Trichophyton rubrum* for developing effective diagnostic and treatment strategies [215].

These databases and web-tools are highly useful for the development of diagnostic devices of specific classes of medical pathogens. To facilitate the development of diagnostic devices for more diverse groups of medical pathogens, a database with integrated information about diagnostic methods, targets, and primers/probes for the known bacterial, fungal and viral pathogens is needed. Therefore, we developed the Microbial pathogen diagnostic methods database, MicrobPad MD (<http://bidd.nus.edu.sg/group/MicrobPad/MicrobPad.asp> or <http://pha-bidd.nus.edu.sg/group/MicrobPad/MicrobPad.asp>), to provide

---

comprehensive information about the molecular diagnostic techniques, targets, primers/probes, detection procedures and conditions, and tested diagnostic accuracies and limit of diagnosis for 314 bacterial, fungal and viral species from 61 genera.

### **3.2 Database construction**

MicrobPad MD is intended as a comprehensive resource for facilitating the research, development, and evaluation of molecular diagnostic methods for faster detection of pathogens that conventional diagnostic methods are inadequate to meet the treatment demand. For instance, more than half of the *Tuberculous meningitis* (TB) cases cannot be confirmed microbiologically and the conventional diagnostic method CSF takes over two weeks time for the test outcome, resulting in many patients being treated on the basis of clinical suspicion before the diagnosis is confirmed [216]. Partly for dealing with this problem, two PCR-based molecular diagnostic devices, TB Amplikor and E-MTD, have been developed and approved by the FDA for diagnosis of TB from clinical specimens [217].

In addition to the development of MicrobPad MD as a resource for microbial pathogen diagnostic methods, we also aim to provide additional information useful for understanding the characteristics and mechanisms of the microbial pathogens. These include pathogen strains and hosts, tissue distribution or habitats, cultivation methods,

---

biochemical characteristics, virulence factors, morphology, diseases, symptoms, treatment and prevention methods are provided for facilitating the study of the molecular mechanisms of medical pathogens. Cross-links to the NCBI genome and SwissProt/UniProt databases are provided.

MicrobPad MD is a freely accessible public online database and the full version of the database in text format can also be downed from the download page. Users are recommended to use the web-version because of its user friendly format. Our database continues to be regularly updated and supported. Queries and suggestions are welcome and can be sent via email link provided in the MicrobPad MD webpage. Users are also welcome to send their new data via email or the new data upload page.

### **3.3 Data collection and access**

The relevant data were collected from the literature searched from the Pubmed database [218] by using keyword combinations of “diagnosis”, “diagnostic”, “detection”, “detect”, “bacterial”, “fungal” and “microbial”, “viral”, “pathogen”, and “pathogenetic”, and from the information described in such review journals as Expert Rev Mol Diagn, Nature Rev Microbiology, and Trends in Biotechnology (The journal titles were listed in the **Appendix A**). A total of 382 papers were collected, which report or describe the molecular diagnostic methods, gene targets, primers/probes,

---

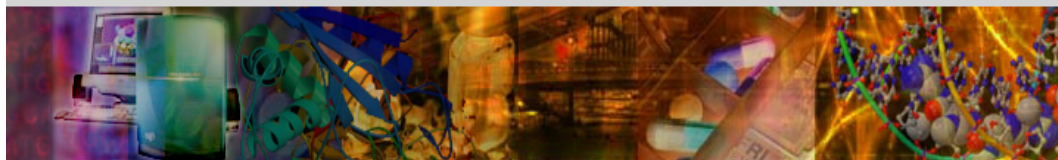
detection procedures and conditions, and the tested diagnostic accuracies and limit of diagnosis for 205 bacterial species from 25 bacteria genera, 17 fungal species from 6 fungal genera, and 92 viral species from 30 virus genera. We further extracted the information from these papers and additional literature searched from Pubmed [218] for finding the pathogen strains and hosts, tissue distribution or habitats, cultivation methods, biochemical characteristics, virulence factors, morphology, diseases, symptoms, treatment and prevention methods for each species.

The MicrobPad MD data can be accessed by keyword or customized search. The keyword search is case insensitive and wildcards are supported. In a query, a user can specify full name or any part of the name in a text field. Wild characters of '\*' and '?' are allowed in text field. Here, '?' represents any one character and '\*' represents a string of characters of any length. For example, input of 'toxin' in the query field finds entries containing 'toxin' in their names, such as alpha toxin, beta toxin, epsilon toxin, RTX toxin, enterotoxins, etc. On the other hand, input of 'Clostridium\*' finds all the species start their genus names with 'Clostridium'. In this case, '\*' represents 'perfringens A' , 'perfringens B', 'septicum', 'difficile', etc.

**Figure 3-1** shows the home page of the database. Customized search (**Figure 3-2**) fields include genus name, species name, target name, disease indication and

---

virulence factor. The result of a search is illustrated in **Figure 3-3**, in which all entries that satisfy the search criteria are listed. This list includes the MicrobPad entry ID, genus name, species name, virulence factor, target gene, disease indications, and the number of diagnostic methods. The related species and diagnosis method page (**Figure 3-4**) can be obtained by clicking the “MicrobPad ID” link of a selected MicrobPad entry. The page of species and diagnostic method contains two sections. The first and second section provides detailed description about the medical species and the diagnostic methods respectively. Further information about the genome of the species, target genes and virulence factors can be accessed via crosslink to NCBI genome databases [218] and SwissProt/UniProt database [219]. The whole MicrobPad methods data can be downloaded via the download link as showed in **Figure 3-5**. It also allow users to contribute to the database by uploading data in certain format illustrated in **Figure 3-6**.



HOME

BROWSE

DOWNLOAD

HELP

#### Search Whole Database

Search

Reset

Examples: Bacillus; toxin; brucellosis; haemolysin ...

Read more about MePad [Query Methods](#)

#### Information of MePad Data

Microbial Pathogen Diagnostic Methods Database MicrobPad MD provides comprehensive information about the molecular diagnostic techniques, targets, primers/probes, detection procedures and conditions, and tested diagnostic accuracies and limit of diagnosis for 314 bacterial, fungal and viral species from 61 genera. While available, additional information such as pathogen strains and hosts, tissue distribution or habitats, cultivation methods, biochemical characteristics, virulence factors, morphology, diseases, symptoms, treatment and prevention methods are provided.

#### Statistics of this database

Currently this database covers 242 gene targets, 700 primers/ probes, 340 virulence factors, and 261 diseases. It contains 205 bacterial species from 25 bacteria genera, 17 fungal species from 6 fungal genera, and 92 viral species from 30 virus genera.

#### Upload new entry into Mepad database

Please click [here](#) to upload new entry.

#### Database Version

V2 Sep 2012

Figure 3-1 Home page of MicrobPad MD database





[HOME](#)

[BROWSE](#)

[DOWNLOAD](#)

[HELP](#)

| Field Name   | Match Text   |
|--|--|
| <b>Species Name</b>  | Genus Name : <input type="text" value="Please Select a Genus Name"/> |
|  | Species Name: <input type="text" value="ALL"/>                       |
| <input type="button" value="Submit"/> <input type="button" value="Reset"/> |  |

| Field Name   | Match Text  |
|--|---|
| <b>Species Name</b>  | Species Name: <input type="text" value="Please select a species name"/> |
| <input type="button" value="Submit"/> <input type="button" value="Reset"/> |   |

| Field Name   | Match Text  |
|--|---|
| <b>Target Name</b>   | Target Name: <input type="text" value="Please Select a Target Name"/> |
| <input type="button" value="Submit"/> <input type="button" value="Reset"/> |   |

| Field Name   | Match Text  |
|--|---|
| <b>Disease Indication</b>  | Disease Indication: <input type="text" value="Please Select a Disease Name"/> |
| <input type="button" value="Submit"/> <input type="button" value="Reset"/> |   |

| Field Name   | Match Text   |
|--|--|
| <b>Virulence Factor</b>  | Virulence Factor: <input type="text" value="Please Select a Virulence Factor Name"/> |
| <input type="button" value="Submit"/> <input type="button" value="Reset"/> |  |

**Figure 3-2** Customized search page. This page provides search fields of genus name, species name, target name, disease indication and virulence factor.

# MicrobPad: Microbial Pathogen Diagnostic Methods Database



[HOME](#)

[BROWSE](#)

[DOWNLOAD](#)

[HELP](#)

You are searching for:

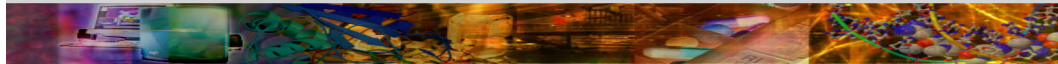
'B. anthracis'

<<First <Previous Page 1 of 1 Next> Last>>

| Microb Pad ID           | Genus Name      | Species Name       | Virulence Factor | Target  | Disease Indication          | No. of Methods |
|-------------------------|-----------------|--------------------|------------------|---|-----------------------------|----------------|
| <a href="#">BN01100</a> | <b>Bacillus</b> | Bacillus anthracis | pX01;pX02        | <a href="#">sspE;</a> <a href="#">cya;</a> <a href="#">capB;</a> <a href="#">plcR;</a> <a href="#">capA;</a> <a href="#">capB;</a> <a href="#">capC;</a> <a href="#">lef;</a> <a href="#">pag;</a> <a href="#">16s rRNA;</a> <a href="#">rpoB;</a> <a href="#">gyrB</a> | Anthrax; emetic syndromesnd | 9              |

<<First <Previous Page 1 of 1 Next> Last>>

**Figure 3-3** List result page. This page provides genus name, species name, virulence factor, target gene, disease indications, and the number of diagnostic methods.



| HOME   | BROWSE  | DOWNLOAD | HELP |
|--|---|----------|------|
| <b>MicrobPad ID: BN01115</b>   |   |          |      |
| <b>Species Information</b>   |   |          |      |
| <b>Genus Name</b>  | Bordetella  |          |      |
| <b>Species Name</b>  | B. holmesii   |          |      |
| <b>Scientific Name</b>   | Bordetella holmesii   |          |      |
| <b>Morphology</b>  | Shape: Bacillus; rod-shaped; Length: 0.2~0.7µm; Gram stain: (-); Motility: (-)  |          |      |
| <b>Cultivation Method</b>  | Bordet Gengou (BG) selective media; Columbia agar with nalidixic acid and 5% sheep blood (all from bioMérieux, Marcy-l'Étoile, France), and CHROMagar Candida (BBL, Becton, Dickinson and Company, Le Pont de Claix, France). |          |      |
| <b>Biochemical Characteristics</b>   | Oxygen requirement: aerobes; Glycolysis of glucose: (+); oxidase(-); persist inside their hosts); oxidase (-); catalase (-); urea(-); beta lactamase(-)   |          |      |
| <b>Virulence Factor</b>  | FHA; Fimbriae; Pertactin; TcfA; Endotoxin LPS; <i>BvgAS</i> ; Secretion system TTSS; Brk; Toxin Cya; Toxin Dnt; Toxin Ptx; Toxin TCT  |          |      |
| <b>Tissue Distribution/Habitats</b>  | The reservoir of this bacterium is unknown  |          |      |
| <b>Disease Indication</b>  | septicemia  |          |      |
| <b>Symptoms</b>  | Fever; Headach; Chill; Vomitin; Cough; Shortness of breath  |          |      |
| <b>Treatment Method</b>  | Penicillin; Cefotaxime; Ceftriaxon; Ampicilli; Amoxicilli   |          |      |
| <b>Prevention Method</b>   | NA  |          |      |
| <b>Molecular Diagnostic Method</b>   |   |          |      |
| <b>Diagnostic Technique 1</b>  | Real-Time PCR    |          |      |
| <b>Diagnostic Target 1</b>   | IS481   |          |      |
| <b>Primer 1</b>  | F: GATTCAATAGGTTGTATGCATGGTT; R: GGACACAAACTTGATGGCGA   |          |      |
| <b>Probe 1</b>   | FAM-CGGACCTTCCTACGTGCGCCTCGAAATGGTCCG-BHQ   |          |      |
| <b>Size of PCR product (bp) 1</b>  | 177   |          |      |
| <b>Procedure and Condition 1</b>   | 95°C for 30s, 45 cycles at 94°C for 1s, 58°C for 15s, and 72°C for 15s  |          |      |
| <b>Diagnostic Limit 1</b>  | 0.02- 0.2 CFU   |          |      |
| <b>Diagnostic Accuracy 1</b>   | NA  |          |      |
| <b>Pathogen strains and hosts 1</b>  | clinical isolates   |          |      |
| <b>Diagnostic Technique 2</b>  | Real-Time PCR    |          |      |
| <b>Diagnostic Target 2</b>   | IS481   |          |      |
| <b>Primer 2</b>  | F: TCAATAGGTTGTATGCATGG; R: GATCAATTGCTGGACCATT   |          |      |
| <b>Probe 2</b>   | FAM-GCAGGCGGCCGGATGAACACCCATAAGCCTGC-Dabcyl   |          |      |
| <b>Size of PCR product (bp) 2</b>  | 154   |          |      |
| <b>Procedure and Condition 2</b>   | 95°C for 15min, 50 cycles at 95°C for 30s, at 55°C for 30s, at 72°C for 30s.  |          |      |
| <b>Diagnostic Limit 2</b>  | 1 to 10 CFU/ml  |          |      |
| <b>Diagnostic Accuracy 2</b>   | NA  |          |      |
| <b>Pathogen strains and hosts 2</b>  | patients  |          |      |
| <b>Diagnostic Technique 3</b>  | Real-Time PCR    |          |      |
| <b>Diagnostic Target 3</b>   | IS1001  |          |      |
| <b>Primer 3</b>  | F: CCATGTCGTGGCCAAGTA; R: TGGTTGGCTTGCAGCAA   |          |      |
| <b>Probe 3</b>   | Texas red-GCAGGCGCTGGCTRCTGCTGCGCAAGCCTGC-BHQ2  |          |      |
| <b>Size of PCR product (bp) 3</b>  | 440   |          |      |
| <b>Procedure and Condition 3</b>   | 95°C for 15min, 50 cycles at 95°C for 30s, at 55°C for 30s, at 72°C for 30s.  |          |      |
| <b>Diagnostic Limit 3</b>  | 10 CFU/ml   |          |      |
| <b>Diagnostic Accuracy 3</b>   | NA  |          |      |
| <b>Pathogen strains and hosts 3</b>  | patients  |          |      |
| <b>Reference</b>   |   |          |      |
| 1. Poddar, S.K. (2003) Detection and discrimination of B pertussis and B holmesii by real-time PCR targeting IS481 using a beacon probe and probe-target melting analysis. Mol Cell Probes, 17, 91-98. Pubmed: <a href="#">12788030</a>  |   |          |      |
| 2. Roorda, L., Buitenwerf, J., Ossewaarde, J.M. and van der Zee, A. (2011) A real-time PCR assay with improved specificity for detection and discrimination of all clinically relevant Bordetella species by the presence and distribution of three Insertion Sequence elements. BMC Res Notes, 4, 11. Pubmed: <a href="#">21255383</a>  |   |          |      |
| 3. Templeton, K.E., Scheltinga, S.A., van der Zee, A., Diederer, B.M., van Kruijssen, A.M., Goossens, H., Kuijper, E. and Claas, E.C. (2003) Evaluation of real-time PCR for detection of and discrimination between Bordetella pertussis, Bordetella parapertussis, and Bordetella holmesii for clinical diagnosis. J Clin Microbiol, 41, 4121-4126. Pubmed: <a href="#">12958235</a> |   |          |      |

**Figure 3-4** Related species and diagnostic methods page. This page provides detailed description about the related species and the diagnostic methods.



HOME

BROWSE

DOWNLOAD

HELP

### MicrobPad Database Downloads

|  |                               |
|--|-------------------------------|
| Download all species information                           | <a href="#">Click to save</a> |
| Download Molecular Diagnostic Primer/Probe for all species | <a href="#">Click to save</a> |
| Download Target Name data for all species                  | <a href="#">Click to save</a> |
| Download Disease Name data for all species                 | <a href="#">Click to save</a> |
| Download Virulence Factor data for all species             | <a href="#">Click to save</a> |

### Database Version

Figure 3-5 Data download page of MicrobPad MD database



HOME

BROWSE

DOWNLOAD

HELP

**MicrobPad Upload**

**Species Information**

|                              |                      |
|------------------------------|----------------------|
| Genus Name *                 | <input type="text"/> |
| Species Name *               | <input type="text"/> |
| Scientific Name *            | <input type="text"/> |
| Morphology                   | <input type="text"/> |
| Cultivation Method           | <input type="text"/> |
| Biochemical Characteristics  | <input type="text"/> |
| Virulence Factor             | <input type="text"/> |
| Tissue Distribution/Habitats | <input type="text"/> |
| Disease Indication           | <input type="text"/> |
| Symptoms                     | <input type="text"/> |
| Treatment Method             | <input type="text"/> |
| Prevention Method            | <input type="text"/> |

**Molecular Diagnostic Method**

|                            |                      |
|----------------------------|----------------------|
| Diagnostic Technique *     | <input type="text"/> |
| Diagnostic Target          | <input type="text"/> |
| Primer/Probe               | <input type="text"/> |
| Procedure and Conduction   | <input type="text"/> |
| Diagnostic Limit           | <input type="text"/> |
| Diagnostic Accuracy        | <input type="text"/> |
| Pathogen strains and hosts | <input type="text"/> |
| Other information          | <input type="text"/> |
| Contact Email              | <input type="text"/> |

**Reference**

|                      |
|----------------------|
| <input type="text"/> |
|----------------------|

Submit

\* Compulsory field

Figure 3-6 Data upload page of MicrobPad MD database

---

### **3.4 Database usage and validation**

Users of MicrobPad MD are expected to have basic knowledge about the popular molecular diagnostic techniques such as PCR, Multiplex PCR, real-time PCR, the diagnostic markers of microbial pathogens including the targets, primers, and probes, and the commonly used detection procedures. To facilitate the users for studying the relevant techniques, all the techniques used in the diagnostic methods described in the MicrobPad MD are provided in the help page. For searching MicrobPad MD, users are also expected to have the knowledge of at least one of the following items: pathogen genus name, pathogen species name, virulence factor, detection target name, and disease indication. Users can use both keyword search and browsing facilities (with pull-down manuals of pathogen, Target Name, Disease Indication and Virulence Factor lists) for selecting the relevant diagnostic method. Keyword search function supports incomplete word search such that all items that partially match the input keywords are displayed for user to select appropriate entries. Our database is built based on IIS HTTP server, ASP (Microsoft's server-side script engine for dynamically generated web pages) and Access (Microsoft's database manage system). There is no special requirement for client users. It can be easily accessed on various operation systems by common Internet Browsers such as Internet Explorer, Chrome, Firefox and Safari.

---

As an illustrative example, in order to find the diagnostic method for detecting the disease “Brucellosis” from clinical samples, the keyword “Brucellosis” can be entered into the MicrobPad MD keyword search field, and the search leads to the list of species and corresponding molecular diagnostic method, specifically the 8 species and the 17 methods. After obtaining the relevant information, users are expected to prepare the primer/probe, DNA polymerase and other necessary reagents, and use PCR amplifier and other relevant equipments for developing diagnostic tools.

Validation study was conducted on a molecule diagnostic method for detecting *Mycobacterium tuberculosis*. It has been reported that the *Mycobacterium tuberculosis* specific probe KY172-T3 (59-GGTGGAAAGCGCTTTAGCGGT-39) has been selected from a hypervariable region within the 16S rRNA gene that are conserved among mycobacterial species [220]. The sequence of these probe and sequence was used for searching all similarity sequences in 16S ribosomal RNA sequences (Bacteria and Archaea) database by using NCBI blast (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>) The sequence of *Mycobacterium tuberculosis* strain NCTC 7416 H37Rv 16S ribosomal RNA is one of the top-6 hits among 123 Blast Hits with max score 42.1, total score 62.4, query coverage 100%, e value 2e-05 and max ident 100%. Base on this probe, Roche AMPLICOR for *Mycobacterium tuberculosis* PCR test (TB AMPLICOR) has been developed which is a rapid

---

diagnostic test and has been shown in clinical tests to have a sensitivity of 66.7% and a specificity of 99.6% [221].

### **3.5 Concluding remarks**

Extensive studies of medical pathogens have led to the identification of high numbers of molecular diagnosis signatures and their recognition techniques useful for the development of fast and low cost pathogen diagnostic tools [35, 36, 203-205]. While significant progress has been made in developing molecular diagnostic devices [35, 205], new methods of rapid diagnosis of infectious pathogens are still in urgently demand, particularly for such serious infections such as septic shock wherein the survival rates decrease on hourly basis if appropriate treatment is delayed [206]. General databases such as MicrobPad MD and the specialized databases such as RIDOM [207], MicroSeq 500 16S rDNA sequence library [208], the fourth international spoligotyping [210], 16SpathDB [212], GenoBASE-pylori [214], and TrED [215] are useful resources and tools for facilitating the development of new diagnostic devices. New technologies, such as mass spectrometry [222], next-generation sequencing [223] and single-molecule detection methods [224, 225], in combination with existing diagnostic technologies [35, 36, 203-205] and knowledge of antimicrobial resistances [226], are expected to further improve the speed and precision of the identification of infectious organisms and the



---

determination of their sensitivities to antimicrobial agents [205, 206, 227]. The new methods and data can be added into MicrobPad MD and other databases to facilitate the development of new diagnostic devices for comprehensive sets of medical pathogens.

---

## **Chapter 4 Development of TTD: therapeutic target database**

### **4.1 Introduction**

Pharmaceutical drugs or agents generally exert their therapeutic effects by binding to and subsequently modulating the activity of particular protein, nucleic acid or other molecular (such as membrane) targets [228, 229]. Target discovery efforts have led to the discovery of hundreds of successful targets (targeted by at least one approved/marketed drug), several hundred clinical trial targets (targeted by drug in clinical trial but not any approved/marketed drug) and more than 1,000 research targets (targeted only by experimental drugs only) [58-61]. Rapid advances in genomic, proteomic, structural, functional and systems studies of the known targets and other disease proteins [230-236] enable the discovery of drugs, multi-target agents, combination therapies and new drug targets [58, 61, 230, 237, 238], analysis of on-target toxicity [239] and pharmacogenetic responses [240], and development of discovery tools [241-244].

To facilitate the access of therapeutic targets information, publicly accessible databases such as Drugbank [245], Potential Drug Target Database (PDTD)

---

[246] and our own Therapeutic Target Database (TTD) [247] have been developed. These databases complement each other to provide target and drug profiles. DrugBank is an excellent source for comprehensive drug data with information about drug actions and multiple targets [245]. PDTD contains active-sites as well as functional information for the potential targets with available 3D structures [246] in PDB. TTD provides information about the primary therapeutic targets of a comprehensive set of both approved and experimental drugs [247].

While drugs and agents typically modulate the activities of multiple proteins [248] and up to 14,000 drug-targeted-proteins have been published [249], the reported number of primary targets directly related to the therapeutic actions of approved drugs is limited to 324 [60]. Information about the primary targets of more comprehensive sets of approved, clinical trial and experimental drugs is highly useful for facilitating focused investigations and discovery efforts against the most relevant and proven targets [61, 230, 237, 239, 240, 243]. Therefore, we updated TTD by significantly expanding the target data to include 348 successful, 292 clinical trial, and 1,254 research targets, and added drug data for 1,514 approved, 1,212 clinical trial and 2,302 experimental drugs linked to their primary targets (3,382 small molecule and 649 antisense drugs with available structure and sequence).

---

We collected a slightly higher number of successful targets than the reported number of 320 targets [60] due to the identification of protein subtypes as the targets of some approved drugs and the inclusion of multiple drug targets of approved multi-target drugs and non-protein/nucleic acid targets of anti-infectious drugs (e.g. bacterial cell wall and membrane components). Clinical trial drugs are based on reports since 2005 with the majority since 2008, their corresponding clinical trial phase is specified. We also added new features for data access by drug mode of action, sequence and tanimoto similarity search of targets and drugs, customized and whole data download, and standardized target ID. TTD is now available online, and can be accessed at <http://bidd.nus.edu.sg/group/cjttd/TTD.asp>.

## **4.2 Target and drug data collection and access**

Additional information about the approved, clinical trial and experimental drugs and their primary targets were collected by comprehensive search of literatures, FDA Drugs of FDA webpage (<http://www.accessdata.fda.gov/scripts/cder/drugsatfda/>) with data about FDA approved drugs, latest reports from 17 pharmaceutical companies that describe clinical trial and other pipeline drugs (*Astrazeneca, Bayer, Boehringer Ingelheim, Genentech, GSK, Idenix, Incyte, ISIS, Merck, Novartis, Pfizer, Roche, Sanofi Aventis, Schering-Plough, Spectrum, Takeda* and *Teva*).

---

Literature search was conducted by combinational searching the PubMed database by using keyword “therapeutic” and “target”, “drug” and “target”, “clinical trial” and “drug”, “clinical trial” and “target”, and by searching reputable review journal like *Nature Reviews Drug Discovery*, *Drug Discovery Today*, *Current Opinion in Pharmacology*, *Current Drug Targets*, *Current Topics in Medicinal Chemistry*, *Science*, *Mini-Reviews in Medicinal Chemistry*, *Anti-Cancer Agents in Medicinal Chemistry*, and so on (The journal titles were listed in the **Appendix B**). In the meantime, we also extracted data from 2008 Report of Medicines in Development biotechnology, and 2008 Report of Medicines in Development for HIV/AIDS, cancer, children, diabetes, neurological disorders, women, and rare diseases, which explicitly mentioned the targets and their corresponding drugs. In particular, these searches identified 198 recent papers reporting approved and clinical trial drugs and their targets. As many of the experimental antisense drugs are described in US patents, we specifically searched US patent databases to identify 745 antisense drugs targeting 104 targets. Primary targets of 211 drugs and drug binding modes of 79 drugs are not specified in our collected documents. Further literature search was conducted to find the relevant information for these drugs. The criteria for identifying the primary target of a drug or targets of a multi-target drug is based on the developer or literature

---

reported cell-based or *in vivo* evidence that links the target to the therapeutic effect of the drug. These searched documents are listed in the respective target or drug entry page of TTD and many cross links are provided for the respective PubMed abstracts, US patents, or developer web-page.

However, in order to double check and have an overall understanding on the status of these targets, we have searched from the literature of reported IC50/EC50 values against the target/targets and cell-lines and the reports of *in vivo* studies to confirm that the reported primary targets are accurate.

### **4.3 Ways to access therapeutic targets database**

TTD data can be accessed by both whole database (**Figure 4-1**) and customized (**Figure 4-2**) keyword search, and by target sequence similarity (**Figure 4-3**) and drug tanimoto similarity search (**Figure 4-4**). Full TTD data download is also provided. Two optional whole database searches are provided: one is to search by target name, and another is by drug name. Different whole database search options will list search results in different manners, which is designed to facilitate users with different initial searching information. Customized search fields include target name, drug name, disease indication, target biochemical class, drug mode of action, and drug therapeutic class. In current TTD, 112 disease indications, 61 target biochemical classes,

20 drug mode of actions, and 157 drug therapeutic classes are available for customized selection.

**Therapeutic Targets Database** BIDD Bioinformatics and Drug Design group

HOME Customized Search Target Similarity Search Drug Similarity Search Download

**Search Whole Database**

List search results by drugs:

Search Reset

List search results by targets:

Search Reset

Examples: Oseltamivir, Alzheimer's disease, MAPK pathway, Muscarinic acetylcholine receptor ...  
Read more about TTD [Query Methods](#)

**Therapeutic Target Database**

A database to provide information about the known and explored therapeutic protein and nucleic acid targets, the targeted disease, pathway information and the corresponding drugs directed at each of these targets. Also included in this database are links to relevant databases containing information about target function, sequence, 3D structure, ligand binding properties, enzyme nomenclature and drug structure, therapeutic class, clinical development status. All information provided are fully referenced.

**Statistics of this database**

This database currently contains **1,894** targets, including **348** successful, **292** clinical trial and **1,254** research targets, and **5,126** drugs, including **1,515** approved, **1,279** clinical trial and **2,332** experimental drugs (**3,257** small molecules and **652** antisense drugs with available structure or oligonucleotide sequence). Targets and drugs in this database cover **61** protein biochemical class and **140** drug therapeutic classes respectively.

**How to cite our database**

Zhu F, Han BC, Pankaj Kumar, Liu XH, Ma XH, Wei XN, Huang L, Guo YF, Han LY, Zheng CJ, Chen YZ. Update of TTD: Therapeutic Target Database. *Nucleic Acids Res.* **2009** [PubMed](#)

Chen X, Ji ZL, Chen YZ. TTD: Therapeutic Target Database. *Nucleic Acids Res.* **2002** [PubMed](#)

**Last update by**

March 19th, 2010

Figure 4-1 Home page of TTD 2010

**Therapeutic Targets Database** 





HOME Customized Search Target Similarity Search Drug Similarity Search Download

| Field Name               | Match Text  |
|--------------------------|---|
| Target Name              | <input type="text"/><br><input checked="" type="radio"/> All <input type="radio"/> Successful <input type="radio"/> Clinical Trial <input type="radio"/> Research |
| Drug Name                | <input type="text"/><br><input checked="" type="radio"/> All <input type="radio"/> Approved <input type="radio"/> Clinical Trial                                  |
| Disease Indication       | Please Select a Disease Name <input type="text" value="Please Select a Disease Name"/>  |
| Target BioChemical Class | Please Select a Target BioChemical Class <input type="text" value="Please Select a Target BioChemical Class"/>  |
| Drug Mode of Action      | Please Select a Drug Mode of Action <input type="text" value="Please Select a Drug Mode of Action"/>  |
| Drug Therapeutic Class   | Please Select a Drug Therapeutic Class <input type="text" value="Please Select a Drug Therapeutic Class"/>  |

Submit Reset

Figure 4-2 Customized search page of TTD 2010

**Therapeutic Targets Database** 



HOME Customized Search Target Similarity Search Drug Similarity Search Download

**Input your protein sequence in FASTA format (example)**

```

MSLPNSSCLEDKMCENKTTMASPQLMPLVVVLSITICLVTVGLNLLVLYAVRSEKRLHT
VGNLYIVSLSVADLIVGAVVMPMNILYLLMSKWSLGRPLCLFWLSMDYVASTASIFSVEI
LCIDRYRSVQQPLRYLKYRTKTRASATILGAWFLSFLWVIPILGWNHFMQQTSVRREDKC
ETDFYDVTWFKVMTAIFNYLPTLLMLWFIYAKIYKAVRQHCQHRELINRSLPSFSEIKLR
PENPKGDARKPKGKESPWEVLKRRPKDAGGGSVLKSQSPQPKEMKSPVVFVSEQEDDREVDKL
YCFPLDIVHMQAAAEGSSRDYVAVNRSHGQLKTDQGLNTHGASEISEDQMLGDSQSFSSR
TDSDTTITETAPGKGLRSGSNTGLDYIKFTWKRLRSHSRQYVSGLHMNRERKAAKQLGFI
MAAFILCWIPYFIFFMVIAFCRNCNEHLHMFITWLGYNSTLNPLIYPLCNENFKKTFK
RILHIRS
  
```

Search Reset

Figure 4-3 Sequence similarity search page of TTD 2010





**Figure 4-4** Drug tanimoto similarity search page of TTD 2010

After input keywords and search in TTD database, the intermediate searching results will be displayed for user to choose from. For example, “VEGFR” was used into the search box–“List search results by targets” at the home page. The intermediate search results page (**Figure 4-5**) will display Vascular endothelial growth factor receptor 1, Vascular endothelial growth factor receptor 2, Vascular endothelial growth factor receptor 3 and mRNA of VEGFR1 for users to make further selection.

<<First <Previous Page 1 of 1 Next> Last>>

| TTD ID  | Search Result |   |
|---|---------------|---|
| <b>TTDS00007</b><br><a href="#">Target Info</a> | Target Name   | <b>Vascular endothelial growth factor receptor 1</b>                          |
|   | Target type   | Successful target   |
|   | Disease       | <a href="#">Angiogenesis</a> ; <a href="#">Inflammatory diseases</a>          |
|   | Drugs         | <a href="#">Telbermin Drug Info</a> ; <a href="#">Sorafenib Drug Info</a> ... |
| <b>TTDS00008</b><br><a href="#">Target Info</a> | Target Name   | <b>Vascular endothelial growth factor receptor 2</b>                          |
|   | Target type   | Successful target   |
|   | Disease       | <a href="#">Cancers</a> ; <a href="#">Angiogenesis</a>                        |
|   | Drugs         | <a href="#">XL999 Drug Info</a> ; <a href="#">XL880 Drug Info</a> ...         |
| <b>TTDC00290</b><br><a href="#">Target Info</a> | Target Name   | <b>Vascular endothelial growth factor receptor 3</b>                          |
|   | Target type   | Clinical trial target   |
|   | Disease       | <a href="#">Angiogenesis in atherosclerotic processes</a> ...                 |
|   | Drugs         | <a href="#">SU-14813 Drug Info</a> ; <a href="#">Pazopanib Drug Info</a> ...  |
| <b>TTDC00334</b><br><a href="#">Target Info</a> | Target Name   | <b>mRNA of VEGFR1</b>   |
|   | Target type   | Clinical trial target   |
|   | Disease       | <a href="#">Age-Related Macular Degeneration</a>                              |
|   | Drugs         | <a href="#">Sirna-027 Drug Info</a>   |

<<First <Previous Page 1 of 1 Next> Last>>

**Figure 4-5** Targets list page of “VEGFR”

Target detail information page (**Figure 4-6**) lists target name, target status (successful, clinical trial and research), synonyms, disease, corresponding drugs, target bio-chemical class, pathway involved, target uniprot accession number, PDB structure, protein function, sequence information, US patents, drug mode of action, references, and so on. Moreover, further information about each target can be accessed via crosslink to external databases, like SwissProt/UniProt, PDB, KEGG, OMID, and Brenda database.

**TTD Target ID: TTDS00007**

| Target Information       |  |                           |                          |  |
|--------------------------|--|---------------------------|--------------------------|--|
| <b>Name</b>              | Vascular endothelial growth factor receptor 1  |                           |                          |  |
| <b>Type of target</b>    | Successful target  |                           |                          |  |
| <b>Synonyms</b>          | Flt-1  |                           |                          |  |
|                          | Fms-like tyrosine kinase 1   |                           |                          |  |
|                          | Tyrosine-protein kinase FRT  |                           |                          |  |
|                          | Tyrosine-protein kinase receptor FLT   |                           |                          |  |
|                          | VEGFR-1  |                           |                          |  |
| <b>Disease</b>           | Vascular permeability factor receptor  |                           |                          |  |
|                          | Angiogenesis [1]   |                           |                          |  |
|                          | Angiogenesis in metastatic and atherosclerotic processes [2]   |                           |                          |  |
| <b>Drug(s)</b>           | Inflammatory diseases [1]  |                           |                          |  |
|                          | Ranibizumab  | <a href="#">Drug Info</a> | Approved                 | Age-related macular degeneration [3]   |
|                          | Sorafenib  | <a href="#">Drug Info</a> | Launched                 | Advanced renal cell carcinoma [4][5][6]  |
|                          | Ranibizumab  | <a href="#">Drug Info</a> | Phase III                | Diabetic macular edema and retinal vein occlusion [3][7]   |
|                          | Sorafenib  | <a href="#">Drug Info</a> | Phase III                | Hepatocellular carcinoma, NSCLC, melanoma [4][5][6]  |
|                          | Sorafenib  | <a href="#">Drug Info</a> | Phase II                 | Myelodysplastic syndrome, AML, head & neck cancer, breast, colon, ovarian, pancreatic cancer [4][5][6] |
|                          | Telbermin  | <a href="#">Drug Info</a> | Discontinued in Phase II | Diabetic foot ulcers [8]   |
| <b>BioChemical Class</b> | Transferases transferring phosphorus-containing groups   |                           |                          |  |
| <b>EC Number</b>         | EC 2.7.1.112   |                           |                          |  |
| <b>Pathway</b>           | <a href="#">Cytokine-cytokine receptor interaction</a>   |                           |                          |  |
|                          | <a href="#">Focal adhesion</a>   |                           |                          |  |
| <b>UniProt ID</b>        | P17948   |                           |                          |  |
| <b>PDB Structure</b>     | 1FLT; 1QSV; 1QSZ; 1QTY; 1RV6.  |                           |                          |  |
| <b>Function</b>          | Receptor for VEGF, VEGFB and PGF, and has a tyrosine-protein kinase activity. The VEGF-kinase ligand/receptor signaling system plays a key role in vascular development and regulation of vascular permeability.   |                           |                          |  |
| <b>Sequence</b>          | <p>MVSYWDTGVLLCALLSCLLLTGSSSSGSKLDPPELSLKGTHIMQAGQTLHLQCRGEAAHK<br/> WSPPEMVSKSEERLSITKACGRRNGKQFCSTLLTNTAQANHTGFYSCRYLAVPTSKKKEE<br/> ESAIYIFISDTGRFVEMVSEIPELIHMTGRELVIKCVTSSEMITVTLKKEPLDLEID<br/> GKRIIWDSRKGFIIISNATYKEIGLLTCEATVNGHLYKTNVLTHTQNTIIVDVQISTPRPV<br/> KLLRGHTLVLNCTATTPLNTRVQMTWSYPDEKNKRASVRRRIQDSNSHANIYFVSLTIDK<br/> MQNKDKGLYTCRVRSGPFSKSVNTSVHIYDKAFITVKHRRKQVLETVAGKRSYRLSMKVK<br/> AFPSPEVVWLKDLGDPATEKSARYLTRYSLIIKDVTEEDAGNYTILLSIKQSNVKNLTA<br/> TLIVNVKPIYKAVSFFDPPALYPLGSRQILTCTAYGIPQPTIKWFHPCNNHNSHSEARC<br/> DFCSNNEESFILDADSNMGNRIEITQRMALIEGKNKMASTLVVADSRISGIYICIASNK<br/> VSTVGRNISFYITDVEGSHVNLKEMTEGDEKLSCTVWKFVFRDVTWILLRVTNNRIM<br/> HYSISKQKMAITKEHSITLNLTIMNVSLQDSGTACRARNVYTGEEILQKKEITIRDQEA<br/> PYLLRNLSDHTVAISSSTLTDCHANGVPEPQITWFKNNHKKIQQEPGIIILPGSSSTLFIER<br/> VTEDEGCVYHCKATNKGKSVESAYLTVQGTSDKSNLELITLCTCVAATLFWLLLTFLFI<br/> RKMKRSSSEIKTDYLSIIMDPDEVPLDEQCCERLPYDASKWEFARERLKLKGLSLGRGAFGK<br/> VVQASAFGIKKSPTCRTVAVKMLKEGATASEYKALMTELKILTHIGHHLNVNLLGACTK<br/> QGGFLMVIYCYKYGNSLNLKSKRDLEFLNKDAALHMEPKKERMEGLGCKKRLDSV<br/> TSSVFASISGFQEDKLSLSDVEEEDSDGFYKEPIITMEDLSYSFQVARGMFLSSRKCIIH<br/> RDLAARNILLSENNIVKICDFGLARDIYKNDPVVRKGDTRLPKWMAPESIFDKIYSTKS<br/> DVWSYGVLLWEIFSLGSSPYPGVQMDDEDFCSRLREGMRRAPEYSTPEIYQIMLDCWHRD<br/> PKERPRFAELVEKLGDLQANVQDQGDYIPIINALTGNSGFTYSTPAFSEDFFKESISA<br/> PKFNSGSSDDVRYVNAFKFMSLERIKTFEELLPNATSMFDDYQGDSSSTLLASPMLKRFV<br/> TDSKPKASLKIDLRVTSKSKESGLSDVSRPFSFCHSSCGHVSEGKRRFTYDHAELERKIAC<br/> CSPFPDYNSVLYSTFPI</p> |                           |                          |  |
| <b>Inhibitor</b>         | AAL-993  | <a href="#">Drug Info</a> | [9]                      |  |
|                          | Ranibizumab  | <a href="#">Drug Info</a> | [3]                      |  |
|                          | Ranibizumab  | <a href="#">Drug Info</a> | [3][7]                   |  |
|                          | SEMAXINIB  | <a href="#">Drug Info</a> | [9]                      |  |
|                          | SU-11652   | <a href="#">Drug Info</a> | [10]                     |  |
|                          | SU-5416  | <a href="#">Drug Info</a> | [11]                     |  |
| <b>Activator</b>         | Sorafenib  | <a href="#">Drug Info</a> | [4][5][6]                |  |
| <b>Activator</b>         | Telbermin  | <a href="#">Drug Info</a> | [8]                      |  |
| <b>Ref 1</b>             | Placental growth factor and its receptor, vascular endothelial growth factor receptor-1: novel targets for stimulation of ischemic tissue revascularization and inhibition of angiogenic and inflammatory disorders. J Thromb Haemost. 2003 Jul;1(7):1356-70. <a href="#">To Reference</a>   |                           |                          |  |
| <b>Ref 2</b>             | Inhibition of vascular endothelial growth factor (VEGF) as a novel approach for cancer therapy. Medicina (B Aires). 2000;60 Suppl 2:41-7. <a href="#">To Reference</a>   |                           |                          |  |
| <b>Ref 3</b>             | The Effect of Intravitreal Ranibizumab on the Fellow Untreated Eye with Subfoveal Scarring due to Exudative Age-Related Macular Degeneration. Ophthalmologica. 2009 Jul 15;223(6):383-389. [Epub ahead of print] <a href="#">To Reference</a>  |                           |                          |  |
| <b>Ref 4</b>             | A comparison of physicochemical property profiles of marketed oral drugs and orally bioavailable anti-cancer protein kinase inhibitors in clinical development. Curr Top Med Chem. 2007;7(14):1408-22. <a href="#">To Reference</a>  |                           |                          |  |
| <b>Ref 5</b>             | Pituitary tumors. Curr Treat Options Neurol. 2009 Jul;11(4):287-96. <a href="#">To Reference</a>   |                           |                          |  |
| <b>Ref 6</b>             | Multi-target therapeutics: when the whole is greater than the sum of the parts. Drug Discov Today. 2007 Jan;12(1-2):34-42. Epub 2006 Nov 28. <a href="#">To Reference</a>  |                           |                          |  |
| <b>Ref 7</b>             | Roche. Product Development Pipeline. July 29 2009. <a href="#">To Reference</a>  |                           |                          |  |
| <b>Ref 8</b>             | Emerging drugs for diabetic foot ulcers. Expert Opin Emerg Drugs. 2006 Nov;11(4):709-24. <a href="#">To Reference</a>  |                           |                          |  |
| <b>Ref 9</b>             | J Med Chem. 2002 Dec 19;45(26):5687-93. Anthranilic acid amides: a novel class of antiangiogenic VEGF receptor kinase inhibitors. <a href="#">To Reference</a>   |                           |                          |  |
| <b>Ref 10</b>            | J Med Chem. 2003 Mar 27;46(7):1116-9. Discovery of 5-[5-fluoro-2-oxo-1,2-dihydroindol-(3Z)-ylidene-methyl]-2,4-dimethyl-1H-pyrrole-3-carboxylic acid (2-diethylaminoethyl)amide, a novel tyrosine kinase inhibitor targeting vascular endothelial and platelet-derived growth factor receptor tyrosine kinase. <a href="#">To Reference</a>  |                           |                          |  |
| <b>Ref 11</b>            | Antiangiogenic effect by SU5416 is partly attributable to inhibition of Flt-1 receptor signaling. Mol Cancer Ther. 2002 Mar;1(5):295-302. <a href="#">To Reference</a>   |                           |                          |  |

Figure 4-6 TTD target detail information page

Drug detail information page (**Figure 4-7**) lists drug name, drug synonyms, trade name, company information, disease indication, 3D drug structure displayed, 2D&3D structural MOL files for download, target therapeutic class, CAS number, formula, PubChem ID, ChEBI ID, SuperDrug ATC & CAS IDs, primary therapeutic target(s), references, and so on. Furthermore, further drug information can be accessed via cross links to the external databases, such as PubChem, DrugBank, SuperDrug, and ChEBI.

| TTD Drug ID: DAP001260      |   |                             |                      |
|-----------------------------|---|-----------------------------|----------------------|
| Drug Information            |   |                             |                      |
| <b>Name</b>                 | Ranibizumab   |                             |                      |
| <b>Synonyms</b>             | Ranibizumab (genetical recombination) (JAN); Ranibizumab (USAN/INN); D05697; Lucentis; 347396-82-1; Ranibizumab (genetical recombination); Ranibizumab; Lucentis (TN)   |                             |                      |
| <b>Trade Name</b>           | Lucentis  |                             |                      |
| <b>Company</b>              | Roche & Genentech   |                             |                      |
| <b>Indication</b>           | Age-related macular degeneration  | Approved                    | [1]                  |
|                             | Diabetic macular edema and retinal vein occlusion   | Phase III                   | [1]                  |
| <b>CAS Number</b>           | <a href="#">CAS 347396-82-1</a>   |                             |                      |
| <b>Formular</b>             | C2158H3282N562O681S12   |                             |                      |
| <b>PubChem Substance ID</b> | <a href="#">SID 47207358</a> .  |                             |                      |
| <b>Target</b>               | Vascular endothelial growth factor 1  | <a href="#">Target Info</a> | Inhibitor [2]        |
|                             | Vascular endothelial growth factor 1  | <a href="#">Target Info</a> | Inhibitor [2]<br>[3] |
| <b>Ref 1</b>                | Future pharmacological treatment options for nonexudative and exudative age-related macular degeneration. Expert Opin Emerg Drugs. 2005 Feb;10(1):119-35. <a href="#">To Reference</a>  |                             |                      |
| <b>Ref 2</b>                | The Effect of Intravitreal Ranibizumab on the Fellow Untreated Eye with Subfoveal Scarring due to Exudative Age-Related Macular Degeneration. Ophthalmologica. 2009 Jul 15;223(6):383-389. [Epub ahead of print] <a href="#">To Reference</a> |                             |                      |
| <b>Ref 3</b>                | Roche. Product Development Pipeline. July 29 2009. <a href="#">To Reference</a>   |                             |                      |

Figure 4-7 TTD drug detail information page

---

Related target or drug entries can be recursively searched by clicking a disease or drug name. Similarity targets of an input protein sequence in FASTA format can be searched by using the NCBI BLAST sequence alignment tool [250]. Similarity drugs of an input drug structure can be searched by using molecular descriptor based tanimoto similarity searching method[251, 252]. Target and drug entries are assigned standardized TTD IDs for easy identification, analysis and linkage to other related databases. The whole TTD data, target sequences along with Swissprot and Entrez gene IDs, and drug structures can be downloaded via the download link. A separate downloadable file contains the list of TTD drug ID, drug name and the corresponding IDs in other cross-matching database PubChem, DrugBank, SuperDrug, and ChEBI. The corresponding HGNC name and Swissprot and Entrez gene ID of each target is provided in the target page. The SMILES and InCHI of each drug is provided in the drug page.

#### **4.4 Target and drug similarity searching**

Target similarity search is based on BLAST [250] algorithm to determine the similarity level between the sequence of an input protein and the sequence of each of the TTD target entries. The NCBI website (<ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/>) is used for downloading

---

BLAST program. The result of similarity targets searched out are ranked by E-value and BLAST score [250]. E-value has been reported to give reliable predictions of the homologous relationships [253] and a cutoff of 0.001 can be used to find 16% more structural relationships in the SCOP database than when using a standard sequence similarity with a 40% sequence-identity threshold[254]. The majority of protein pairs sharing ~50% (or higher) sequence-identity differ by  $< 1 \text{ \AA}$  RMS deviation[255, 256]. A larger structural deviation alters drug-binding properties probably.

Drug similarity search is based on the tanimoto similarity search method [251]. An input compound structure in MOL or SDF format is converted into a vector composed of molecular descriptor by using MODEL[161]. These molecular descriptors are quantitative representations of structural and physicochemical features of molecules, which have been extensively used in deriving structure-activity relationships, quantitative structure-activity relationship and virtual screening tool for drug discovery[146, 202]. Based on the results of our earlier studies[47], a total of 98 1D and 2D descriptors were used as the components of the compound vector, which include 18 descriptors in the class of simple molecular property, 3 descriptors in chemical property, 35 descriptors in molecular connectivity and shape, and 42 descriptors in

---

electro-topological state. The vector of an input compound  $i$  is then compared to drug  $j$  in TTD by using the Tanimoto coefficient  $sim(i,j)$ [251]:

$$sim(i, j) = \frac{\sum_{d=1}^l x_{di}x_{dj}}{\sum_{d=1}^l (x_{di})^2 + \sum_{d=1}^l (x_{dj})^2 - \sum_{d=1}^l x_{di}x_{dj}}$$

where  $l$  is the total number of molecular descriptors. Tanimoto coefficient of similarity compounds are typically in the range of 0.8 to 0.9[185, 186]. Hence compound  $i$  is considered to be very similar, similar, moderately similar, or un-similar to drug  $j$  if  $sim(i,j) > 0.9$ ,  $0.85 < sim(i,j) < 0.9$ ,  $0.75 < sim(i,j) < 0.85$ , or  $sim(i,j) < 0.75$  respectively.

In conclusion, TTD 2010 update is intended to be a more useful resource in complement to other related databases by providing comprehensive information to the primary targets and other drug data for the approved, clinical trial, and experimental drugs. In addition to the continuous update of new target and drug information, efforts will be devoted to the incorporation of more features into TTD. Increasing amounts of data about the genomic, proteomic, structural, functional and systems profiles of therapeutic targets have been and are being generated[230-236]. Apart from establishing crosslink to the emerging sources, some of the profiles extracted or derived from the relevant data[58] may be further incorporated into TTD. Target data

---

has been used for developing target discovery methods[241-243], some of these methods may be included in TTD in addition to the BLAST tool for similarity target searching. As in the case of PDTD[246], some of the virtual screening methods and datasets may also be included in TTD for facilitating target oriented drug lead discovery.



---

## **Chapter 5 Development and experimental test of support vector machines virtual screening method for searching Src inhibitors from large compound libraries**

### **5.1 Introduction**

Src promotes tumour invasion and metastasis, facilitates VEGF-mediated angiogenesis and survival in endothelial cells, and enhances growth factor driven proliferation in fibroblasts [257]. Src is known to modulate cell proliferation and cancers through several signaling pathways such as STAT3 pathway, the PI3K pathway and the MAPK pathway [258]. Src consists of 6 functional domain: homology domain 4 (SH4), unique domain, homology domain (SH3), homology domain (SH2), catalytic domain (SH1), and C-terminal regulatory tail [259]. Src activation can be altered by many different cell processes through upstream kinases or phosphatases. Src associated pathway, its structure and activity regulation [260] have been explored.

Src is one of the multiple kinase targets of a number of multi-target kinase inhibitors effective in the clinical treatment of leukemia and in clinical trials of other cancers [261-263]. The successes and problems of these inhibitors have raised significant interest and efforts in discovering new Src inhibitors [264-266]. Several *in-silico* methods have been used for facilitating the search

---

and design of Src inhibitors, which include pharmacophore [267], Quantitative Structure Activity Relationship (QSAR) [268], and molecular docking [265].

While these *in-silico* methods have shown impressive capability in the identification of potential Src inhibitors, their applications may be affected by such problems as the vastness and sparse nature of chemical space needing to be searched, complexity and flexibility of target structures, difficulties in accurately estimating binding affinity and solvation effects on molecular binding, and limited representativeness of training active compounds [69, 269, 270]. It is desirable to explore other *in-silico* methods that complement these methods by expanded coverage of chemical space, increased screening speed, and reduced false-hit rates without necessarily relying on the modelling of target structural flexibility, binding affinity and solvation effects.

Support vector machines (SVM) has recently been explored as a promising ligand-based virtual screening (VS) method that produces high yields and low false-hit rates in searching active agents of single and multiple mechanisms from large compound libraries [114] and in identifying active agents of diverse structures [114, 171-174]. Good VS performance can also be achieved by SVM trained from sparsely distributed active compounds [252]. SVM classifies active compounds based on the separation of active and inactive compounds in a hyperspace constructed by their physicochemical properties rather than structural similarity to active compounds *per se*, which has the advantage of not relying on the accurate computation of structural flexibility,

---

activity-related features, binding affinity and solvation effects. Moreover, the fast speed of SVM enables efficient search of vast chemical space. Therefore, SVM may be a potentially useful VS tool to complement other *in-silico* methods for searching Src inhibitors from large libraries.

In this work, we developed a SVM VS model for identifying Src inhibitors, and evaluated its performance by both 5-fold cross validation test and large compound database screening test. In 5-fold cross validation test, a dataset of Src inhibitors and non-inhibitors was randomly divided into 5 groups of approximately equal size, with 4 groups used for training a SVM VS tool and 1 group used for testing it, and the test process is repeated for all 5 possible compositions to derive an average VS performance. In large database screening test, a SVM VS tool was developed by using Src inhibitors published before 2011, its yield (percent of known inhibitors identified as virtual-hits) was estimated by using Src inhibitors reported since 2011 and not included in the training datasets, virtual-hit rate and false-hit rate in searching large libraries were evaluated by using 13.56M PubChem and 168K MDDR compounds, and an additional set of 9,305 MDDR compounds similar in structural and physicochemical properties to the known Src inhibitors.

Moreover, VS performance of SVM was compared to those of two similarity-based VS methods, Tanimoto similarity searching and k nearest neighbour (kNN), and an alternative but equally popularly used machine learning method, probabilistic neural network (PNN) method, based on the

---

same training and testing datasets (same sets of PubChem and MDDR compounds) and molecular descriptors. In a study that compares the performance of SVM to 16 classification methods and 9 regression methods, it has been reported that SVMs shows mostly good performances both on classification and regression tasks, but other methods proved to be very competitive [271]. Therefore, it is useful to evaluate the VS performance of SVM in searching large compound libraries by comparison with those of both similarity-based approaches and other typical machine learning method.

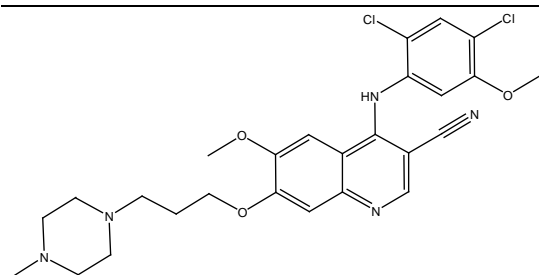
PubChem and MDDR contain high percentages of inactive compounds significantly different from the known Src inhibitors, and the easily distinguishable features may make VS enrichments artificially good [272]. Therefore, VS performance may be more strictly tested by using subsets of compounds that resemble the physicochemical properties of the known Src inhibitors so that enrichment is not simply a separation of trivial physicochemical features [186]. To further evaluate whether our SVM VS tool predict Src inhibitors and non-inhibitors rather than membership of certain compound families, distribution of the predicted active and inactive compounds in the compound families were analyzed.

---

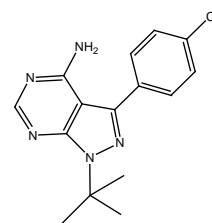
## 5.2 Materials and methods

### 5.2.1 Compound collections and construction of training and testing datasets

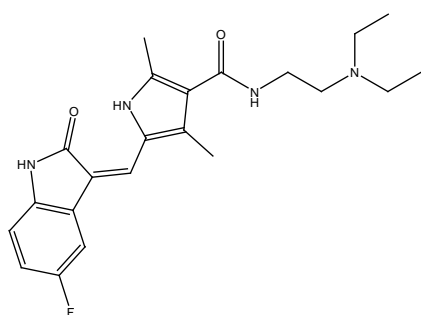
We collected 1,703 Src inhibitors reported before 2011, with  $IC_{50} < 10\mu M$ , from the literatures [273-277] and the BindingDB database [161]. The inhibitor selection criterion of  $IC_{50} < 10\mu M$  was used because it covers most of the reported HTS and VS hits [278, 279]. The structures of representative Src inhibitors are shown in **Figure 5-1**. As few non-inhibitors have been reported, putative non-inhibitors were generated by using our method for generating putative inactive compounds [246, 252]. This method requires no knowledge of known inactive compounds and active compounds of other target classes, which enables more expanded coverage of the “non-inhibitor” chemical space. Although the yet-to-be-discovered inhibitors are likely distributed in some of these “non-inhibitor” families, a substantial percentage of these inhibitors are expected to be identified as inhibitors rather than non-inhibitors even-though representatives of their families are putatively assigned as non-inhibitors [246]. 13.56M PubChem and 168K MDDR compounds were grouped into 8,423 compound families by clustering them in the chemical space defined by their molecular descriptors [193, 194]. The number of generated families is consistent with the 12,800 compound-occupying neurons (regions of topologically close structures) for 26.4 million compounds of up to 11 atoms [65], and the 2,851 clusters for 171,045 natural products [196].



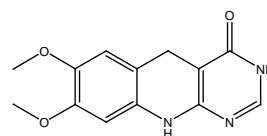
compound 1



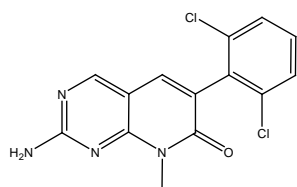
compound 2



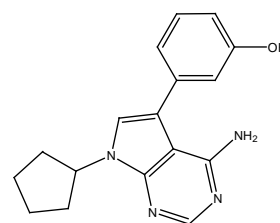
compound 3



compound 4



compound 5



compound 6

**Figure 5-1** The structures of representative c-Src inhibitors. Compound 1:SKI-606  $IC_{50}=0.25\mu\text{m}$  [144]; Compound 2: AG-1879,  $IC_{50}=0.085\mu\text{m}$ ; Compound 3: Sunitinib, SU 11248,  $IC_{50}=1\mu\text{m}$  [282]; Compound 4:  $IC_{50}=0.5\mu\text{m}$  [280]; Compound 5:  $IC_{50}=0.26\mu\text{m}$  [281]; Compound 6:  $IC_{50}=0.001\mu\text{m}$  [282].

Our collected Src inhibitors are distributed in 493 families. Because of the extensive efforts in searching kinase inhibitors from known compound libraries, the number of undiscovered Src inhibitor families in PubChem and

---

MDDR databases is expected to be relatively small, most likely no more than several hundred families. The ratio of the discovered and undiscovered inhibitor families (hundreds) and the families that contain no known Src inhibitor (8,423 based on the current versions of PubChem and MDDR) is expected to be <15%. Therefore, putative non-inhibitor training dataset can be generated by extracting a few representative compounds from each of those families that contain no known inhibitor, with a maximum possible “wrong” classification rate of <15% even when all of the undiscovered inhibitors are misplaced into the non-inhibitor class. The noise level generated by up to 15% “wrong” negative family representation is expected to be substantially smaller than the maximum 50% false-negative noise level tolerated by SVM [172]. Based on earlier studies [246, 252] and this work, it is expected that a substantial percentage of the un-discovered inhibitors in the putative “non-inhibitor” families can be classified as inhibitor despite their family representatives are placed into the non-inhibitor training sets.

In the database screening test, 60.1% of the families that contain Src inhibitors reported since 2011 [283-288] are not covered by the Src inhibitor training dataset (inhibitors reported before 2011). The representative compounds of these families, none of which happen to be Src inhibitor, were deliberately placed into the inactive training sets because the inhibitors in these families are not supposed to be known in our study. As shown in earlier studies [246, 252] and in this work, a substantial percentage of the inhibitors in these misplaced inhibitor-containing “non-inhibitor” families were predicted as

---

inhibitors by our SVM VS tool. Moreover, a small percentage of the compounds in these putative non-inhibitor datasets are expected to be un-reported and un-discovered inhibitors, their presence in these datasets is not expected to significantly affect the estimated false hit rate of SVM.

## **5.3 Results and discussion**

### **5.3.1 Performance of SVM, kNN and PNN identification of Src inhibitors based on 5-fold cross validation test**

The parameters of our SVM, kNN and PNN models were determined by 5-fold cross-validation studies of Src inhibitors and non-inhibitors. The results of these tests for SVM, kNN and PNN are shown in **Tables 5-1, 5-2, 5-3** and **Figure 5-2** respectively. Overall, the sensitivity of SVM, k-NN and PNN is in the range of 93.53%~95.01%, 88.56%~92.94% and 93.53%~97.06%, the specificity in the range of 99.81%~99.90%, 99.57%~99.77% and 97.76%~98.03%, and overall accuracy Q in the range of 99.67%~99.76%, 99.35%~99.48% and 97.69%~97.91% respectively. The inhibitor accuracies of our SVM are comparable to or slightly better than the reported accuracies of 58.3%~67.3% for protein kinase C inhibitors by SVM-RBF and CKD methods [192], 83% for Lck inhibitors by SVM method [289], and 74%~87% for inhibitors of any of the 8 kinases (3 Ser/Thr and 5 Tyr kinases) by SVM, ANN, GA/kNN, and RP methods [290]. The non-inhibitor accuracies are comparable to the value of 99.9% for Lck inhibitors [289] and substantially



better than the typical values of 77%~96% of other studies [192, 290]. Caution needs to be exercised about straightforward comparison of these results, which might be misleading because the outcome of VS strongly depends on the datasets and molecular descriptors used. Based on these rough comparisons, SVM appears to show good capability in identifying Src inhibitors at low false-hit rates.

**Table 5-1** Performance of SVM for identifying Src inhibitors and non-inhibitors evaluated by 5-fold cross validation study

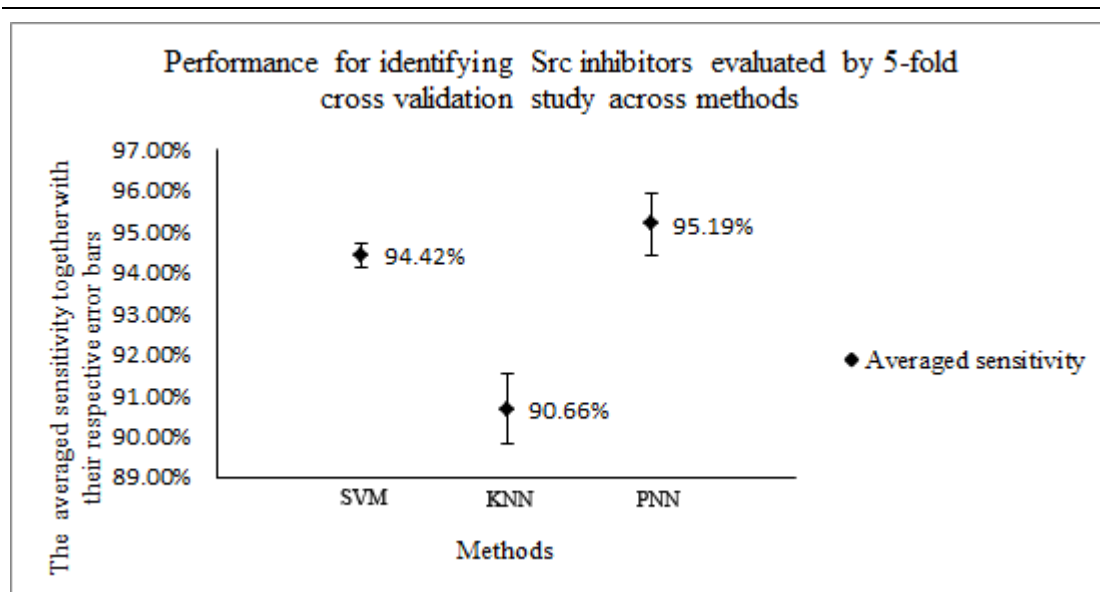
| Cross-Validation | Src inhibitors                    |     |     |        | Src non-inhibitors                    |       |    |        | Q      | C      |
|------------------|-----------------------------------|-----|-----|--------|---------------------------------------|-------|----|--------|--------|--------|
|                  | No of training/testing inhibitors | TP  | F N | SEN    | No of training/testing non-inhibitors | TN    | FP | SP     |        |        |
| 1                | 1362/341                          | 320 | 21  | 93.84% | 50654/12664                           | 12651 | 13 | 99.90% | 99.74% | 0.948  |
| 2                | 1362/341                          | 324 | 17  | 95.01% | 50654/12664                           | 12650 | 14 | 99.89% | 99.76% | 0.953  |
| 3                | 1362/341                          | 324 | 17  | 95.01% | 50654/12664                           | 12640 | 24 | 99.81% | 99.68% | 0.939  |
| 4                | 1363/340                          | 318 | 22  | 93.53% | 50655/12663                           | 12642 | 21 | 99.83% | 99.67% | 0.935  |
| 5                | 1363/340                          | 322 | 18  | 94.71% | 50655/12663                           | 12643 | 20 | 99.84% | 99.71% | 0.943  |
| Average          |                                   |     |     | 94.42% |                                       |       |    | 99.85% | 99.71% | 0.944  |
| Std Dev          |                                   |     |     | 0.0069 |                                       |       |    | 0.0004 | 0.0004 | 0.0072 |
| Std Err          |                                   |     |     | 0.0031 |                                       |       |    | 0.0002 | 0.0002 | 0.0032 |

**Table 5-2** Performance of kNN for identifying Src inhibitors and non-inhibitors evaluated by 5-fold cross validation study

| Cross<br>-Validation | Src inhibitors                              |     |    |        | Src non-inhibitors                              |       |    |        | Q      | C      |
|----------------------|---|-----|----|--------|---|-------|----|--------|--------|--------|
|                      | No of<br>training/<br>testing<br>inhibitors | TP  | FN | SEN    | No of<br>training/<br>testing<br>non-inhibitors | TN    | FP | SP     |        |        |
| 1                    | 1362/341                                    | 302 | 39 | 88.56% | 50654/12664                                     | 12635 | 29 | 99.77% | 99.48% | 0.896  |
| 2                    | 1362/341                                    | 313 | 28 | 91.79% | 50654/12664                                     | 12620 | 44 | 99.65% | 99.45% | 0.894  |
| 3                    | 1362/341                                    | 311 | 30 | 91.20% | 50654/12664                                     | 12610 | 54 | 99.57% | 99.35% | 0.878  |
| 4                    | 1363/340                                    | 316 | 24 | 92.94% | 50655/12663                                     | 12619 | 44 | 99.65% | 99.48% | 0.901  |
| 5                    | 1363/340                                    | 302 | 38 | 88.82% | 50655/12663                                     | 12632 | 31 | 99.76% | 99.47% | 0.895  |
| Average              |   |     |    | 90.66% |   |       |    | 99.68% | 99.44% | 0.893  |
| Std Dev              |   |     |    | 0.0191 |   |       |    | 0.0008 | 0.0005 | 0.0085 |
| Std Err              |   |     |    | 0.0085 |   |       |    | 0.0004 | 0.0002 | 0.0038 |

**Table 5-3** Performance of PNN for identifying Src inhibitors and non-inhibitors evaluated by 5-fold cross validation study

| Cross<br>-Validation | Src inhibitors                              |     |    |        | Src non-inhibitors                              |       |     |        | Q      | C      |
|----------------------|---|-----|----|--------|---|-------|-----|--------|--------|--------|
|                      | No of<br>training/<br>testing<br>inhibitors | TP  | FN | SEN    | No of<br>training/<br>testing<br>non-inhibitors | TN    | FP  | SP     |        |        |
| 1                    | 1362/341                                    | 319 | 22 | 93.55% | 50654/12664                                     | 12413 | 251 | 98.02% | 97.90% | 0.715  |
| 2                    | 1362/341                                    | 324 | 17 | 95.01% | 50654/12664                                     | 12380 | 284 | 97.76% | 97.69% | 0.702  |
| 3                    | 1362/341                                    | 330 | 11 | 96.77% | 50654/12664                                     | 12395 | 269 | 97.88% | 97.85% | 0.722  |
| 4                    | 1363/340                                    | 330 | 10 | 97.06% | 50655/12663                                     | 12389 | 274 | 97.84% | 97.82% | 0.720  |
| 5                    | 1363/340                                    | 318 | 22 | 93.53% | 50655/12663                                     | 12413 | 250 | 98.03% | 97.91% | 0.715  |
| Average              |   |     |    | 95.19% |   |       |     | 97.90% | 97.83% | 0.715  |
| Std Dev              |   |     |    | 0.0169 |   |       |     | 0.0012 | 0.0009 | 0.0075 |
| Std Err              |   |     |    | 0.0076 |   |       |     | 0.0005 | 0.0004 | 0.0034 |



**Figure 5-2** The 5-fold cross-validation studies of Src inhibitors across methods with the averaged sensitivity together with their respective error bars.

### 5.3.2 Virtual screening performance of SVM in searching Src inhibitors from large compound libraries

As outlined in the methods section, we developed a SVM VS tool for searching Src inhibitors from large were developed by using Src kinases reported before 2011. The VS performance of SVM in identifying Src inhibitors reported since 2011 and in searching MDDR and PubChem databases is summarised in **Table 5-4**. The yield in searching Src inhibitors reported since 2011 is 70.45%, which is comparable to the reported 50%~94% yields of various VS tools [291]. Strictly speaking, direct comparison of the reported performances of these VS tools is inappropriate because of the differences in the type, composition and diversity of compounds screened, and in

the molecular descriptors, VS tools and their parameters used. The comparison cannot go beyond the statistics of accuracies.

**Table 5-4** Virtual screening performance of support vector machines for identifying Src inhibitors from large compound libraries

|                               |   |                   |
|-------------------------------|---|-------------------|
| Inhibitors in Training Set    | Number of Inhibitors  | 1703              |
|                               | Number of Chemical Families Covered by Inhibitors   | 493               |
| Inhibitors in Testing Set     | Number of Inhibitors  | 44                |
|                               | Number of Chemical Families Covered by Inhibitors   | 35                |
|                               | Percent of Inhibitors in Chemical Families Covered by Inhibitors in Training Set                        | 51.43%            |
| Virtual Screening Performance | Yield   | 70.45%            |
|                               | Number and Percent of Identified True Inhibitors Outside Training Chemical Families                     | 15(34.1%)         |
|                               | Number and Percent of 13.56M PubChemCompounds Identified as Inhibitors                                  | 44,843<br>(0.33%) |
|                               | Number and Percent of the 168K MDDR Compounds Identified as Inhibitors                                  | 1,496<br>(0.89%)  |
|                               | Number and Percent of the 9,305 MDDR Compounds Similar to the Known Inhibitors Identified as Inhibitors | 719 (7.73%)       |

---

We also evaluated virtual-hit rates and false-hit rates of SVM in screening compounds that resemble the structural and physicochemical properties of the known Src inhibitors by using 9,305 MDDR compounds similar to an Src inhibitor in the training dataset. Similarity was defined by Tanimoto similarity coefficient  $\geq 0.9$  between a MDDR compound and its closest inhibitor [252]. This stricter similarity metric was used for conducting a stricter test of our SVM model. SVM identified 719 virtual-hits from these 9,305 MDDR similarity compounds (virtual-hit rate 7.73%), which suggests that SVM has some level of capability in distinguishing Src inhibitors from non-inhibitor similarity compounds. Significantly lower virtual-hit rates and thus false-hit rates were found in screening large libraries of 168K MDDR and 13.56M PubChem compounds. The numbers of virtual-hits and virtual-hit rates in screening 168K MDDR compounds are 1,496 and 0.89% respectively. The numbers of virtual-hits and virtual-hit rates in screening 13.56M PubChem compounds are 44,843 and 0.33% respectively.

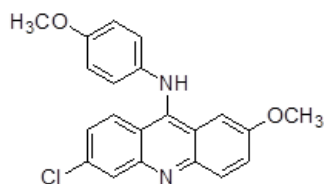
Substantial percentages of the MDDR virtual-hits belong to the classes of antineoplastic, tyrosine-specific protein kinase inhibitors, signal transduction inhibitors, antiangiogenic, and antiarthritic (**Table 5-5**, details in next section). As some of these virtual-hits may be true Src inhibitors, the false-hit rate of our SVM is at most equal to and likely less than the virtual-hit rate. Hence the false-hit rate is

---

<7.73% in screening 9,305 MDDR similarity compounds, <0.89% in screening 168K MDDR compounds, and <0.33% in screening 13.56M PubChem compounds, which are comparable and in some cases better than the reported false-hit rates of 0.0054%~8.3% of SVM [246, 252], 0.08%~3% of structure-based methods, 0.1%~5% by other machine learning methods, 0.16%~8.2% by clustering methods, and 1.15%~26% by pharmacophore models [291].

### **5.3.3 Experimental test of a SVM identified virtual-hit**

Three virtual hits of the same novel scaffold from in-house libraries not found in the known the Src inhibitor were evaluated for inhibitory activity against Src. Src kinase was incubated with substrates, compounds and ATP in a final buffer of 25mM HEPES (pH 7.4), 10mM MgCl<sub>2</sub>, 0.01% Triton X-100, 100μg/mL BSA, 2.5mM DTT in 384-well plate with the total volume of 10μl. The assay plate was incubated at 30°C for 1h and stopped with the addition of equal volume of kinase glo plus reagent. The luminescence was read at envision. The signal was correlated with the amount of ATP present in the reaction and was inversely correlated with the kinase activity. One of three virtual hits showing in Figure 5-3 was found to inhibit Src at a moderate rate of 4.85% at 20μM.



**Figure 5-3** Virtual hit inhibiting Src at a moderate rate of 4.85% at 20 $\mu$ M

### 5.3.4 Evaluation of SVM identified MDDR virtual-hits

SVM identified MDDR virtual-hits were evaluated based on the known biological or therapeutic target classes specified in MDDR. **Table 5-5** gives the MDDR classes that contain higher percentage ( $\geq 3\%$ ) of SVM virtual -hits and the percentage values. We found that 623 (41.6%) of the 1,496 virtual-hits belong to the antineoplastic class, which represent 2.9% of the 21,557 MDDR compounds in the class. In particular, 231 (15.4%) of the virtual-hits belong to the tyrosine-specific protein kinase inhibitor class, which represent 19.6% of the 1,181 MDDR compounds in the class. Moreover, 194 (13.0%) and 75 (5.0%) of the virtual-hits belong to the signal transduction inhibitor and antiangiogenic classes, representing 9.5% and 4.6% of the 2,037 and 1,629 members in these classes respectively. Therefore, many of the SVM virtual-hits are antineoplastic compounds that inhibit tyrosine kinases and possibly other kinases involved in signal transduction and angiogenesis pathways. While some of these kinase



inhibitors might be true Src inhibitors, a significant percentage of them are expected to arise from false selection of inhibitors of other kinases.

**Table 5-5** MDDR classes that contain higher percentage ( $\geq 3\%$ ) of SVM virtual -hits and the percentage values. Virtual-hits are identified by SVMs in screening 168K MDDR compounds for Src inhibitors. The total number of SVM identified virtual hits is 1,496.

| MDDR Classes that Contain Higher Percentage ( $\geq 3\%$ ) of Virtual Hits | No of Virtual Hits in Class | Percentage of Class Members Selected as Virtual Hits |
|--|-----------------------------|--|
| Antineoplastic   | 623                         | 2.9%   |
| Tyrosine-Specific Protein Kinase Inhibitor                                 | 231                         | 19.6%  |
| Signal Transduction Inhibitor  | 194                         | 9.5%   |
| Antiarthritic  | 176                         | 1.5%   |
| Antiallergic/Antiasthmatic   | 83                          | 0.8%   |
| Antihypertensive   | 76                          | 0.7%   |
| Antiangiogenic   | 75                          | 4.6%   |
| Treatment for Osteoporosis   | 55                          | 2.2%   |
| Antidepressant   | 49                          | 0.8%   |

A total of 176 (11.8%) SVM virtual-hits belong to the antiarthritic class. A primary feature of rheumatoid arthritis in synovial tissues is the abnormal stimulation of fibrin deposition, angiogenesis and proinflammatory processes, which are promoted by

---

thrombin increased IL-6 production via the PAR1 receptor/PI-PLC/PKC alpha/c-Src/NF-kappaB and p300 signaling pathways [292]. Therefore, Src inhibitors may have some effects against arthritis via interference with some of these processes. Moreover, several other kinases have been implicated in arthritis. An Abl inhibitor Gleevec has been reported to be effective in treatment of arthritis, which is probably due to its inhibition of other related kinases such as c-kit and PDGFR [293]. EGFR-like receptor stimulates synovial cells and its elevated activities may be involved in the pathogenesis of rheumatoid arthritis [294]. VEGF has been related to such autoimmune diseases as systemic lupus erythematosus, rheumatoid arthritis, and multiple sclerosis [295]. FGFR may partly mediate osteoarthritis [296]. PDGF-like factors stimulate the proliferative and invasive phenotype of rheumatoid arthritis synovial connective tissue cells [297]. Lck inhibition leads to immunosuppression and has been explored for the treatment of rheumatoid arthritis and asthma [298]. Therefore, some of the SVM virtual-hits in the antiarthritic class may be inhibitors of these kinases or their kinase-like capable of producing antiarthritic activities.

Moreover, 83 (5.5%), 76 (5.1%), 55 (3.7%) and 49 (3.3%) of the SVM virtual hits are in the antiallergic/antiasthmatic, antihypertensive, osteoporosis treatment and antidepressant classes respectively. Src or Src family kinases have been implicated in and the respective inhibitors have shown observable effects against these diseases. For

---

instance, Src family kinases and lipid mediators have been found to partly control allergic inflammation [299]. Inhibition of Src family kinase-dependent signaling cascades in mast cells may exert anti-allergic activity [300]. Up-regulation of Src signaling has been suggested to be important in the profibrotic and proinflammatory actions of aldosterone in a genetic model of hypertension, which can be significantly reduced by mineralocorticoid receptor blocker and Src inhibitor [301]. Src signalling pathways play critical roles in osteoclasts and osteoblasts, and Src inhibitors have been developed as therapeutic agents for bone diseases [302, 303]. Src-family protein tyrosine kinases negatively regulate cerebellar long-term depression, which can be recovered by the application of Src-family protein tyrosine kinase inhibitors [304]. Therefore, some of the SVM virtual hits in these four MDDR classes may be Src inhibitors or Src family kinase inhibitors capable of regulating allergic inflammation, hypertension, osteoporosis and depression respectively.

### **5.3.5 Comparison of virtual screening performance of SVM with those of other virtual screening methods**

To evaluate the level of performance of SVM and whether the performance is due to the SVM classification models or to the molecular descriptors used, SVM results were compared with those of three other VS methods based on the same molecular descriptors, training dataset of Src inhibitors reported before 2011, and the testing

---

dataset of Src inhibitors reported since 2011 and 168K MDDR compounds. The three other VS methods include two similarity-based methods, Tanimoto-based similarity searching and kNN methods, and an alternative machine learning method PNN. As shown in **Table 5-6**, the yield and maximum possible false-hit rate of the Tanimoto-based similarity searching, kNN and PNN methods are 36.84% and 5.54%, 38.64% and 2.49%, and 50.00% and 2.60% respectively. Compared to these results, the yield of SVM is better than these similarity-based VS method, and the false-hit rate of SVM is significantly reduced by 6.22, 2.80, and 2.92 fold respectively. These suggests that SVM performance is due primarily to the SVM classification models rather than the molecular descriptors used, and SVM is capable of achieving comparable yield at substantially reduced false-hit rate as compared to both similarity-based approach and alternative machine learning method. Our results are consistent with the report that SVM shows mostly good performances both on classification and regression tasks, but other classification and regression methods proved to be very competitive [271].

**Table 5-6** Comparison of virtual screening performance of SVM with those of other methods

| Method                       | Inhibitors in Training Set |   | Inhibitors in Testing Set |   |  | Virtual Screening Performance |   |  |   |
|------------------------------|----------------------------|---|---------------------------|---|--|-------------------------------|---|--|---|
|                              | No of Inhibitors           | No of Chemical Families Covered by Inhibitors | No of Inhibitors          | No of Chemical Families Covered by Inhibitors | Percent of Inhibitors in Chemical Families Covered by Inhibitors in Training Set | Yield                         | No and Percent of Identified True Inhibitors Outside Training Chemical Families | No and Percent of the 168K MDDR Compounds Identified as Inhibitors | No and Percent of the 9,305 MDDR Compounds Similar to the Known Inhibitors Identified as Virtual Inhibitors |
| Support Vector Machines      | 1703                       | 493   | 44                        | 35  | 51.43%   | 70.45%                        | 15(34.1%)   | 1,496 (0.89%)  | 719 (7.73%)   |
| Tanimoto Similarity          |                            |   |                           |   |  | 36.84%                        | 9(20.5%)  | 9,305 (5.54%)  | 9,305 (100%)  |
| K Nearest Neighbour          |                            |   |                           |   |  | 38.64%                        | 10(22.7%)   | 4,182 (2.49%)  | 1,169 (12.57%)  |
| Probabilistic Neural Network |                            |   |                           |   |  | 50.0%                         | 13(29.5%)   | 4,386 (2.60%)  | 1,184 (12.72%)  |

---

### **5.3.6 Does SVM select Src inhibitors or membership of compound families?**

To further evaluate whether SVM identifies Src inhibitors rather than membership of certain compound families, compound family distribution of the identified Src inhibitors and non-inhibitors were analyzed. 34.1% of the identified inhibitors belong to the families that contain no known Src inhibitors. For those families that contain at least one known Src inhibitor, >70% of the compounds (>90% in majority cases) in each of these families were predicted as non-inhibitor by SVM. These results suggest that SVM identify Src inhibitors rather than membership to certain compound families. Some of the identified inhibitors not in the family of known inhibitors may serve as potential “novel” Src inhibitors. Therefore, as in the case shown by earlier studies [114], SVM has certain capacity for identifying novel active compounds from sparse as well as regular-sized active datasets.

## **5.4 Conclusions**

Our study suggested that SVM is capable of identifying Src inhibitors at comparable yield and in many cases substantially lower false-hit rate than those of typical VS tools reported in the literatures. It can be used for searching large compound libraries at sizes comparable to the 13.56M PubChem and 168K MDDR compounds at low false-hit rates. The performance of SVM is substantially improved against several

---

other VS method based on the same datasets and molecular descriptors, suggesting that the VS performance of SVM is primarily due to SVM classification models rather than the molecular descriptors used. Three SVM virtual hits of the same novel scaffold were experimentally tested, one of which showed moderate Src inhibition rate. Because of its high computing speed and generalization capability for covering highly diverse spectrum compounds, SVM can be potentially explored to develop useful VS tools to complement other VS methods or to be used as part of integrated VS tools in facilitating the discovery of Src inhibitors and other active compounds [305-307].

---

## Chapter 6 Support vector machines virtual screening of

### VEGFR-2 Inhibitors from large compound libraries: model development and experimental test

#### 6.1 Background

VEGFR regulates angiogenesis, growth, migration and survival [308]. There are 3 main VEGFR subtypes, VEGFR-2 mediates almost all of the known cellular responses to VEGF, VEGFR-1 modulates VEGFR-2 signaling and acts as a dummy/decoy receptor, and VEGFR-3 mediates lymphangiogenesis in response to VEGF-C and VEGF-D [308]. VEGFR inhibitors have been successfully used for cancer treatments [261, 309]. While increasing number of VEGFR inhibitors have been developed and tested, several problems limit the scope of their practical applications. These problems include increased toxicity partly due to the targeting of multiple kinases, acquired resistances, and reduced tumor responses (VEGFR inhibitors can cause extensive tumor necrosis without a marked decrease in tumor size) [310]. Moreover, on-target toxicity against specific VEGFR subtypes in various tissues is also a significant problem for the applications of VEGFR inhibitors [311]. The successes of VEGFR inhibitors and the encountered problems have led to further efforts for discovering new inhibitors [261, 309].

*In-silico* methods such as pharmacophore [312], QSAR [313-315], fragment-based method [316], molecular docking [317, 318], and their combinations [312, 315] have been used for facilitating the search and design of VEGFR inhibitors, which have shown impressive



---

capability in the identification of potential VEGFR inhibitors, but their applications may be affected by such problems as the vastness and sparse nature of chemical space needing to be searched, complexity and flexibility of target structures, difficulties in accurately estimating binding affinity and solvation effects on molecular binding, and limited representativeness of training active compounds [69, 269, 270]. Therefore, it is desirable to explore other *in-silico* methods that complement these methods by expanded coverage of chemical space, increased screening speed, and reduced false-hit rates without necessarily relying on the modelling of target structural flexibility, binding affinity and solvation effects.

Support vector machines (SVM) has been explored as such a VS method capable of producing high yields and low false-hit rates in searching active agents of single and multiple mechanisms from large compound libraries [319] and in identifying active agents of diverse structures [171-174, 319]. Good VS performance can also be achieved by SVM trained from sparsely distributed active compounds [252]. SVM classifies active compounds based on the separation of active and inactive compounds in a hyperspace constructed by their physicochemical properties rather than structural similarity to active compounds *per se*, which has the advantage of not necessarily relying on the modeling of target structural flexibility and the computation of activity-related features, binding affinity and solvation effects. Moreover, the fast speed of SVM enables efficient search of vast chemical space. Therefore, SVM may be a potentially useful VS tool to complement other *in-silico* methods for searching VEGFR inhibitors from large libraries.

---

In this work, SVM was tested for its capability in searching VEGFR-2 inhibitors from large compound libraries. Our focus on inhibitors of VEGFR-2 subtype was based on the availability of reported inhibitors of the subtype and the consideration that VEGFR-2 mediates almost all of the known cellular responses to VEGF [308]. The performance of SVM was evaluated by both 5-fold cross validation test and large database screening test. In 5-fold cross validation test, VEGFR-2 inhibitors and non-inhibitors was randomly divided into 5 groups of approximately equal size, with 4 groups used for training a SVM VS tool and 1 group used for testing it, and the test process is repeated for all 5 possible compositions to derive an average VS performance. In large database screening test, SVM was developed by using VEGFR-2 inhibitors published before 2012, its yield (percent of known inhibitors identified as virtual-hits) was estimated by using VEGFR-2 inhibitors reported since 2012 and not included in the training datasets, virtual-hit rate and false-hit rate of the SVM in searching large libraries were evaluated by using 13.56M PubChem, and 168K MDDR, and an additional set of 13,872 MDDR compounds similar in structural and physicochemical properties to the known VEGFR-2 inhibitors.

Moreover, VS performance of SVM was compared to those of two similarity-based VS methods, Tanimoto similarity searching and k nearest neighbour (kNN), and an alternative but equally popularly used machine learning method, probabilistic neural network (PNN) method, based on the same training and testing datasets (same sets of PubChem and MDDR compounds) and molecular descriptors. In a study that compares the performance of SVM to 16 classification methods and 9 regression methods, it has been reported that SVMs shows mostly good performances both on classification and regression tasks, but other methods proved to be very competitive [271]. Therefore, it is useful to evaluate the VS

---

performance of SVM in searching large compound libraries by comparison with those of both similarity-based approaches and other typical machine learning method.

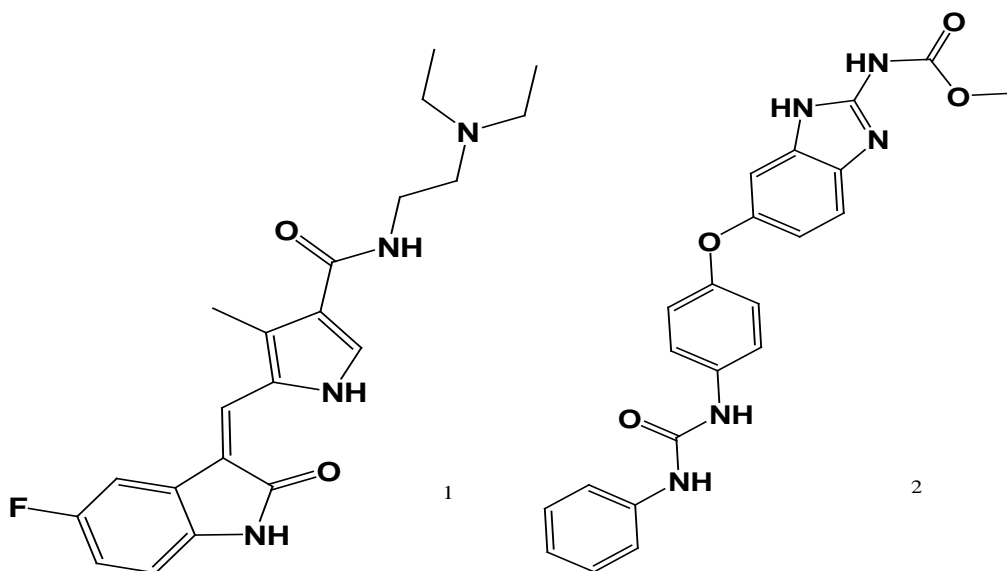
Databases such as PubChem and MDDR contain high percentages of inactive compounds significantly different from VEGFR-2 inhibitors, and the easily distinguishable features may make VS enrichments artificially good [272]. Therefore, VS performance may be more strictly tested by using subsets of compounds that resemble the physicochemical properties of the known VEGFR-2 inhibitors so that enrichment is not simply a separation of trivial physicochemical features [186]. To further evaluate whether SVM predict VEGFR-2 inhibitors and non-inhibitors rather than membership of certain compound families, distribution of the predicted active and inactive compounds in the compound families were analyzed. Moreover, VS performance of SVM for screening MDDR compounds was compared with that of Tanimoto similarity search method on the same molecular descriptors, training dataset to determine whether the performance of SVM is due to the SVM classification models or to the molecular descriptors used.

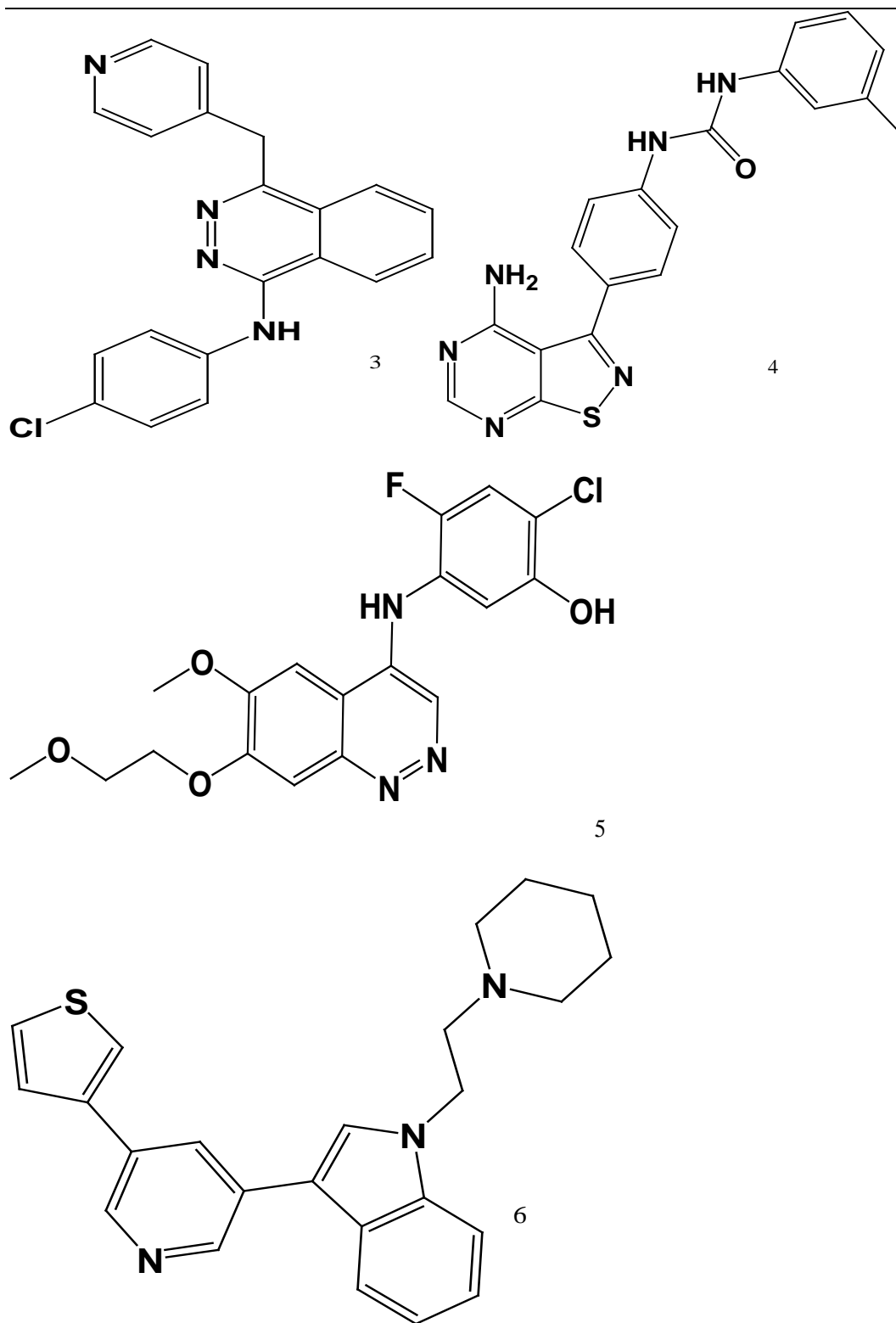
## **6.2 Materials and methods**

### **6.2.1 Compound collections and construction of training and testing datasets**

Using the inhibitor selection criterion of  $IC_{50} < 10 \mu M$ , which covers most of the reported HTS and VS hits [278, 279], we collected 3,653 VEGFR-2 inhibitors regardless of their activities against other VEGFR subtypes from the literature reported before 2012 [319-334] and the BindingDB database [161]. The structures of representative VEGFR-2 inhibitors

are shown in **Figure 6-1**. As few non-inhibitors have been reported, putative non-inhibitors were generated by using our method for generating putative inactive compounds [246, 252]. This method requires no knowledge of known inactive compounds and active compounds of other target classes, which enables more expanded coverage of the “non-inhibitor” chemical space. Although the yet-to-be-discovered inhibitors are likely distributed in some of these “non-inhibitor” families, a substantial percentage of these inhibitors are expected to be identified as inhibitors rather than non-inhibitors even-though representatives of their families are putatively assigned as non-inhibitors [246]. 13.56M PubChem and 168K MDDR compounds were grouped into 8,423 compound families by clustering them in the chemical space defined by their molecular descriptors [193, 194]. The number of generated families is consistent with the 12,800 compound-occupying neurons (regions of topologically close structures) for 26.4 million compounds of up to 11 atoms [65], and the 2,851 clusters for 171,045 natural products [196].





**Figure 6-1** The structures of representative VEGFR-2 inhibitors. Compound 1: Sunitinib,  $IC_{50}=0.009\mu\text{m}$ .; Compound 2:  $IC_{50}=0.032\mu\text{m}$  [335]; Compound 3: Vatalanb (PTK787),  $IC_{50}=0.037\mu\text{m}$ ; Compound 4:  $IC_{50}=0.012\mu\text{m}$  [336]; Compound 5:  $IC_{50}=0.004\mu\text{m}$  [337]; Compound 6:  $IC_{50}=0.111\mu\text{m}$  [338].

---

Our collected VEGFR-2 inhibitors are distributed in 845 families. Because of the extensive efforts in searching kinase inhibitors from known compound libraries, the number of undiscovered VEGFR-2 inhibitor families in PubChem and MDDR databases is expected to be relatively small, most likely no more than several hundred families. The ratio of the discovered and undiscovered inhibitor families (hundreds) and the families that contain no known inhibitor of each kinase (8,423 based on the current versions of PubChem and MDDR) is expected to be <15%. Therefore, putative non-inhibitor training dataset can be generated by extracting a few representative compounds from each of those families that contain no known inhibitor, with a maximum possible “wrong” classification rate of <15% even when all of the undiscovered inhibitors are misplaced into the non-inhibitor class. The noise level generated by up to 15% “wrong” negative family representation is expected to be substantially smaller than the maximum 50% false-negative noise level tolerated by SVM [172]. Based on earlier studies [246, 252] and this work, it is expected that a substantial percentage of the un-discovered inhibitors in the putative “non-inhibitor” families can be classified as inhibitor despite their family representatives are placed into the non-inhibitor training sets.

In conducting large database screening test, 3,653 VEGFR-2 inhibitors reported before 2012 were used as a training dataset for developing SVM and 92 VEGFR-2 inhibitors reported since 2012 [339-345] were used as an independent testing dataset for testing SVM. Only 28.57% of the families that contain VEGFR-2 inhibitors reported since 2012 are covered in the families that contain at least one VEGFR-2 inhibitor reported before 2012, and the representative compounds of these families, none of which happen to be VEGFR-2 inhibitor, were deliberately placed into the inactive training sets because the inhibitors in

---

these families are not supposed to be known in our study. As shown in earlier studies [246, 252] and in this work, a substantial percentage of the inhibitors in these misplaced inhibitor-containing “non-inhibitor” families were predicted as inhibitors by SVM. Moreover, a small percentage of the compounds in these putative non-inhibitor datasets are expected to be un-reported and un-discovered inhibitors, their presence in these datasets is not expected to significantly affect the estimated false hit rate of SVM.

## **6.3 Results and Discussion**

### **6.3.1 VEGFR-2 Inhibitor prediction Performance of SVM, kNN and PNN evaluated by 5-fold cross validation test**

**Table 6-1, 6-2 and 6-3** give the 5-fold cross validation test results of SVM, kNN and PNN models in identifying VEGFR-2 inhibitors and non-inhibitors. **Figure 6-2** shows 5-fold cross validation performance for identifying VEGFR-2 inhibitors across methods with the averaged sensitivity together with their respective error bars. Overall, the sensitivity of SVM, kNN and PNN is in the range of 93.98%~95.89%, 88.10%~90.00% and 91.79%~93.01%, the specificity in the range of 99.53%~99.70%, 98.65%~98.72% and 97.81%~98.01%, and overall accuracy Q in the range of 99.24%~99.45%, 98.10%~98.27% and 97.53%~97.69% respectively. The inhibitor accuracies of our SVM are comparable to or better than the reported accuracies of 58.3%~67.3% for protein kinase C inhibitors by SVM-RBF and CKD methods [192], 83% for Lck inhibitors by SVM method [289], and 74%~87% for inhibitors of any of the 8 kinases (3 Ser/Thr and 5 Tyr kinases) by SVM, ANN, GA/kNN, and RP methods [290]. The non-inhibitor accuracies are comparable to the

---

value of 99.9% for Lck inhibitors [289] and substantially better than the typical values of 77%~96% of other studies [192, 290]. These are consistent with the result of a study of the comparison of SVM with 16 classification methods and 9 regression methods, which has shown that SVMs showed mostly good performances both on classification and regression tasks but other methods proved to be very competitive [346]. Caution needs to be raised about straightforward comparison of these results, which might be misleading because the outcome of VS strongly depends on the datasets and molecular descriptors used. Based on these rough comparisons, SVM appears to show good prediction capability in identifying VEGFR-2 inhibitors at low false-hit rates. Similar prediction accuracies are also found from two additional 5-fold cross validation studies conducted by using training-testing sets separately generated from different random number seed parameters.



**Table 6-1** Performance of SVM for identifying VEGFR-2 inhibitors and non-inhibitors evaluated by 5-fold cross validation study

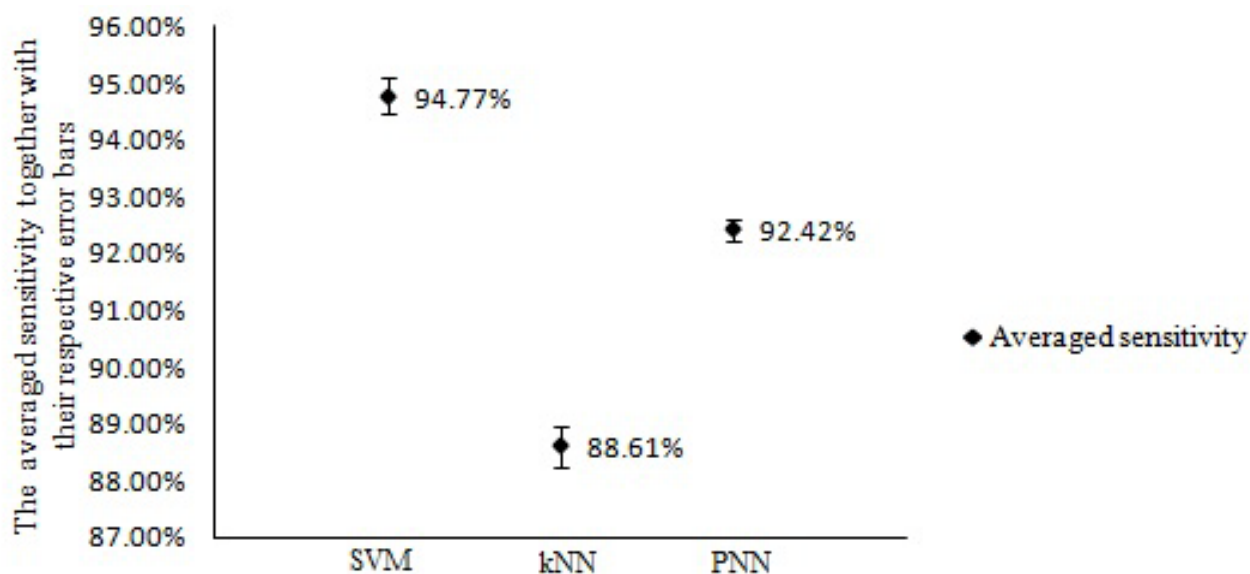
| Cross<br>-Validation | VEGFR-2 inhibitors                          |     |    |        | VEGFR-2 non-inhibitors                          |       |    |        | Q      | C     |
|----------------------|---|-----|----|--------|---|-------|----|--------|--------|-------|
|                      | No of<br>training/<br>testing<br>inhibitors | TP  | FN | SEN    | No of<br>training/<br>testing<br>non-inhibitors | TN    | FP | SP     |        |       |
| 1                    | 2922/731                                    | 692 | 39 | 94.66% | 53585/13397                                     | 13335 | 62 | 99.54% | 99.29% | 0.928 |
| 2                    | 2922/731                                    | 687 | 44 | 93.98% | 53585/13397                                     | 13334 | 63 | 99.53% | 99.24% | 0.924 |
| 3                    | 2922/731                                    | 694 | 37 | 94.94% | 53586/13396                                     | 13356 | 40 | 99.70% | 99.45% | 0.945 |
| 4                    | 2923/730                                    | 689 | 41 | 94.38% | 53586/13396                                     | 13349 | 47 | 99.65% | 99.38% | 0.937 |
| 5                    | 2923/730                                    | 700 | 30 | 95.89% | 53586/13396                                     | 13343 | 53 | 99.60% | 99.41% | 0.941 |
| Average              |   |     |    | 94.77% |   |       |    | 99.60% | 99.35% | 0.935 |
| Std Dev              |   |     |    | 0.0072 |   |       |    | 0.0007 | 0.0009 | 0.009 |
| Std Err              |   |     |    | 0.0032 |   |       |    | 0.0003 | 0.0004 | 0.004 |

**Table 6-2** Performance of kNN for identifying VEGFR-2 inhibitors and non-inhibitors evaluated by 5-fold cross validation study.

| Cross<br>-Validation | VEGFR-2 inhibitors                          |     |    |        | VEGFR-2 non-inhibitors                          |       |     |        | Q      | C     |
|----------------------|---|-----|----|--------|---|-------|-----|--------|--------|-------|
|                      | No of<br>training/<br>testing<br>inhibitors | TP  | FN | SEN    | No of<br>training/<br>testing<br>non-inhibitors | TN    | FP  | SP     |        |       |
| 1                    | 2922/731                                    | 644 | 87 | 88.10% | 53585/13397                                     | 13216 | 181 | 98.65% | 98.10% | 0.819 |
| 2                    | 2922/731                                    | 644 | 87 | 88.10% | 53585/13397                                     | 13216 | 181 | 98.65% | 98.10% | 0.819 |
| 3                    | 2922/731                                    | 646 | 85 | 88.37% | 53586/13396                                     | 13218 | 178 | 98.67% | 98.14% | 0.823 |
| 4                    | 2923/730                                    | 657 | 73 | 90.00% | 53586/13396                                     | 13224 | 172 | 98.72% | 98.27% | 0.836 |
| 5                    | 2923/730                                    | 646 | 84 | 88.49% | 53586/13396                                     | 13217 | 179 | 98.66% | 98.14% | 0.823 |
| Average              |   |     |    | 88.61% |   |       |     | 98.67% | 98.15% | 0.824 |
| Std Dev              |   |     |    | 0.0079 |   |       |     | 0.0003 | 0.0007 | 0.007 |
| Std Err              |   |     |    | 0.0036 |   |       |     | 0.0001 | 0.0003 | 0.003 |

**Table 6-3** Performance of PNN for identifying VEGFR-2 inhibitors and non-inhibitors evaluated by 5-fold cross validation study.

| Cross<br>-Validation | VEGFR-2 inhibitors                          |     |    |        | VEGFR-2 non-inhibitors                          |       |     |        | Q      | C     |
|----------------------|---|-----|----|--------|---|-------|-----|--------|--------|-------|
|                      | No of<br>training/<br>testing<br>inhibitors | TP  | FN | SEN    | No of<br>training/<br>testing<br>non-inhibitors | TN    | FP  | SP     |        |       |
| 1                    | 2922/731                                    | 671 | 60 | 91.79% | 53585/13397                                     | 13131 | 266 | 98.01% | 97.69% | 0.799 |
| 2                    | 2922/731                                    | 676 | 55 | 92.48% | 53585/13397                                     | 13110 | 287 | 97.86% | 97.58% | 0.794 |
| 3                    | 2922/731                                    | 675 | 56 | 92.34% | 53586/13396                                     | 13117 | 279 | 97.92% | 97.63% | 0.797 |
| 4                    | 2923/730                                    | 675 | 55 | 92.47% | 53586/13396                                     | 13102 | 294 | 97.81% | 97.53% | 0.791 |
| 5                    | 2923/730                                    | 679 | 51 | 93.01% | 53586/13396                                     | 13110 | 286 | 97.87% | 97.61% | 0.797 |
| Average              |   |     |    | 92.42% |   |       |     | 97.89% | 97.61% | 0.796 |
| Std Dev              |   |     |    | 0.0044 |   |       |     | 0.0008 | 0.0006 | 0.003 |
| Std Err              |   |     |    | 0.0019 |   |       |     | 0.0004 | 0.0003 | 0.002 |



**Figure 6-2** Performance for identifying VEGFR-2 inhibitors evaluated by 5-fold cross validation study across methods. This figure is illustrating the 5-fold cross validation studies of VEGFR-2 inhibitors across methods with the averaged sensitivity together with their respective error bars.

### 6.3.2 Virtual screening performance of SVM in searching

#### VEGFR-2 inhibitors from large compound libraries

A SVM in searching VEGFR-2 inhibitors from large libraries was developed by using VEGFR-2 inhibitors reported before 2012. The VS performance of this SVM in identifying VEGFR-2 inhibitors reported since 2012 and in searching MDDR and PubChem databases is summarised in **Table 6-4**. The yield in searching VEGFR-2 inhibitors reported since 2012 is 85.87%, which is comparable to the reported

50%~94% yields of various VS tools [291]. Strictly speaking, direct comparison of the reported performances of these VS tools is inappropriate because of the differences in the type, composition and diversity of compounds screened, and in the molecular descriptors, VS tools and their parameters used. The comparison cannot go beyond the statistics of accuracies as the reports are not detailed enough to address questions of whether all methods detect the same hit.

**Table 6-4** Virtual screening performance of support vector machines for identifying VEGFR-2 inhibitors from large compound libraries

|                                   |  |                   |
|-----------------------------------|--|-------------------|
| Inhibitors in Training<br>Dataset | No of Inhibitors   | 3653              |
|                                   | No of Chemical Families Covered by Inhibitors                                    | 845               |
| Inhibitors in Testing<br>Dataset  | No of Inhibitors   | 92                |
|                                   | No of Chemical Families Covered by Inhibitors                                    | 35                |
|                                   | Percent of Inhibitors in Chemical Families Covered by Inhibitors in Training Set | 56.52%            |
| Virtual Screening<br>Performance  | Yield  | 85.87%            |
|                                   | No and Percent of Identified True Inhibitors Outside Training Chemical Families  | 31 (39.24%)       |
|                                   | No and Percent of 13.56M PubChemCompounds Identified as Inhibitors               | 31,624<br>(0.23%) |

|  |  |                |
|--|--|----------------|
|  | No and Percent of the 168K MDDR Compounds Identified as Inhibitors                                   | 2,717 (1.62%)  |
|  | No and Percent of the 13,872 MDDR Compounds Similar to the Known Inhibitors Identified as Inhibitors | 1,714 (12.36%) |

Virtual-hit rates and false-hit rates of SVM in screening compounds that resemble the structural and physicochemical properties of the VEGFR-2 inhibitors were evaluated by using 13,872 MDDR compounds similar to a VEGFR-2 inhibitor in the training dataset. Similarity was defined by Tanimoto similarity coefficient  $\geq 0.9$  between a MDDR compound and its closest dual-inhibitor [252]. This stricter similarity metric was used for conducting a stricter test of our SVM model. SVM identified 1,714 virtual-hits from these 13,872 MDDR similarity compounds (virtual-hit rate 12.36%), which suggests that SVM has some level of capability in distinguishing VEGFR-2 inhibitors from similarity non-inhibitors. Significantly lower virtual-hit rates and thus false-hit rates were found in screening large libraries of 168K MDDR and 13.56M PubChem compounds. The numbers of virtual-hits and virtual-hit rates in screening 168K MDDR compounds are 2,717 and 1.62% respectively. The numbers of virtual-hits and virtual-hit rates in screening 13.56M PubChem compounds are 31,624 and 0.23% respectively.

---

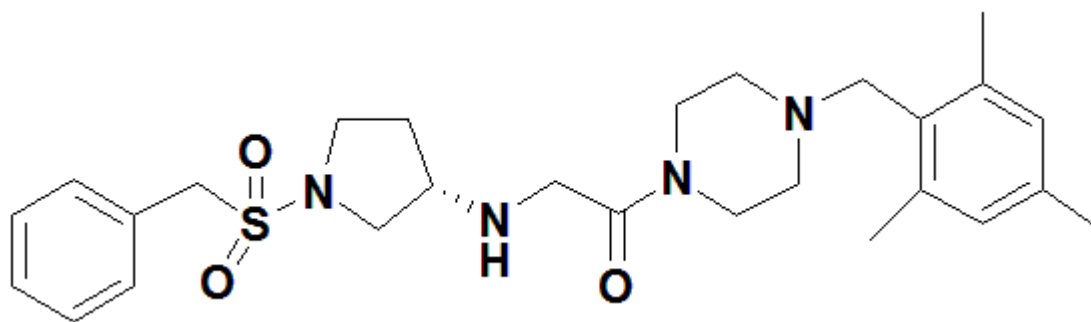
Many of the 2,717 MDDR virtual-hits belong to the classes of antineoplastic (45.3%), tyrosine-specific protein kinase inhibitor (12.7%), signal transduction inhibitor (12.7%), antiarthritic (11.0%), and antiangiogenic (9.3%), antihypertensive (5.1%), antiallergic/antiasthmatic (4.3%), and antidepressant (3.4%) (**Table 6-5**, details in next section). As some of these virtual-hits may be true VEGFR inhibitors, the false-hit rate of our SVM is at most equal to and likely less than the virtual-hit rate. Hence the false-hit rate is 12.36% in screening 13,872 MDDR similarity compounds,  $\leq 1.62\%$  in screening 168K MDDR compounds, and  $\leq 0.23\%$  in screening 13.56M PubChem compounds, which are comparable and in some cases better than the reported false-hit rates of 0.0054%~8.3% of SVM [246, 252], 0.08%~3% of structure-based methods, 0.1%~5% by other machine learning methods, 0.16%~8.2% by clustering methods, and 1.15%~26% by pharmacophore models [291].

### **6.3.3 Experimental test of a SVM identified virtual-hit**

Three virtual hits of the same novel scaffold from in-house libraries not found in the known the VEGFR-2 inhibitor were evaluated for inhibitory activity against VEGFR-2. VEGFR-2 kinase was incubated with substrates, compounds and ATP in a final buffer of 25mM HEPES (pH 7.4), 10mM MgCl<sub>2</sub>, 0.01% Triton X-100, 100 $\mu$ g/mL BSA, 2.5mM DTT in 384-well plate with the total volume of 10 $\mu$ l. The

---

assay plate was incubated at 30°C for 1h and stopped with the addition of equal volume of kinase glo plus reagent. The luminescence was read at envision. The signal was correlated with the amount of ATP present in the reaction and was inversely correlated with the kinase activity. One of three virtual hits shown in **Figure 6-3** was found to inhibit VEGFR-2 at a moderate rate of 4.54% at 20 $\mu$ M.



**Figure 6-3** The structure of a SVM virtual hit tested to show moderate VEGFR-2 inhibitory activity.

#### 6.3.4 Evaluation of SVM identified MDDR virtual-hits

SVM identified MDDR virtual-hits were evaluated based on the known biological or therapeutic target classes specified in MDDR. **Table 6-5** gives the MDDR classes that contain higher percentage ( $\geq 3\%$ ) of SVM virtual-hits and the percentage values. We found that 1,230 or 45.3% of the 2,717 virtual-hits belong to the antineoplastic class, which represent 5.7% of the 21,557 MDDR compounds in the class. In particular, 346 or 12.7% of the virtual-hits belong to the tyrosine-specific protein kinase inhibitor



class, which represent 29.3% of the 1,181 MDDR compounds in the class. Moreover, 12.7% and 9.4% of the virtual-hits belong to the signal transduction inhibitor and antiangiogenic classes, representing 16.9% and 15.7% of the 2,037 and 1,629 members in the two classes respectively. Therefore, many of the SVM virtual-hits are antineoplastic compounds that inhibit tyrosine kinases and possibly other kinases involved in signal transduction, angiogenesis and other cancer-related pathways. Some of these SVM selected kinase inhibitors might have VEGFR inhibitory activities, and others were expectedly selected due to false selection of inhibitors of other kinases (at  $\leq 1.62\%$ ~ $12.36\%$  false-hit rates).

**Table 6-5** MDDR classes that contain higher percentage ( $\geq 3\%$ ) of SVM virtual -hits and the percentage values. Virtual-hits are identified by SVMs in screening 168K MDDR compounds for VEGFR-2 inhibitors. The total number of SVM identified virtual hits is 2,717.

| MDDR Classes that Contain Higher Percentage (>3%) of Virtual Hits | No and Percentage of Virtual Hits in Class | Percentage of Class Members Selected as Virtual Hits |
|---|--|--|
| Antineoplastic  | 1230 (45.3%)                               | 5.7%   |
| Tyrosine-Specific Protein Kinase Inhibitor                        | 346 (12.7%)                                | 29.3%  |
| Signal Transduction Inhibitor                                     | 345 (12.7%)                                | 16.9%  |
| Antiarthritic   | 300 (11.0%)                                | 2.6%   |

|                            |            |       |
|----------------------------|------------|-------|
| Antiangiogenic             | 256 (9.3%) | 15.7% |
| Antihypertensive           | 139 (5.1%) | 1.3%  |
| Antiallergic/Antiasthmatic | 118 (4.3%) | 1.1%  |
| Antidepressant             | 93 (3.4%)  | 1.5%  |

Substantial percentages of the SVM virtual-hits belong to the antiarthritic (11.0%), antihypertensive (5.1%), and antiallergic/antiasthmatic (4.3%) therapeutic classes. Some VEGFR inhibitors have been reported to show respective therapeutic effects. VEGF has been related to such autoimmune diseases as systemic lupus erythematosus, rheumatoid arthritis, and multiple sclerosis [295]. Both VEGFR-1 and VEGFR-2 are expressed in human osteoarthritic cartilage [347]. VEGFR-2 and VEGFR-3 are present in most of the sublining blood vessels in arthritic synovium [348]. A VEGFR-2 inhibitor, PTK787/ZK222584, has been reported to cause significant anti-arthritic effects in models of rheumatoid arthritis via anti-angiogenic actions [349]. Hypertension is characterized by the development of a hyperdynamic circulation which can be markedly inhibited by EGFR-2 inhibitor (e.g. SU5416) blockade of the VEGF signaling pathway, leading to the consideration of modulation of angiogenesis for the treatment of hypertension [350]. VEGFR-2 and VEGFR-1 have been shown to be involved in the pathogenesis of the contact hypersensitivity reaction, and both the induction and elicitation phases of contact hypersensitivity can

---

be inhibited by VEGFR inhibitor PTK787/ZK 222584 [351]. Therefore, some of the SVM virtual-hits in the antiarthritic, antihypertensive, and antiallergic/antiasthmatic classes may be VEGFR inhibitors capable of producing the respective therapeutic effects.

Moreover, 93 (3.4%) of the SVM virtual hits are in the antidepressant class. It has been reported that depressive episodes in the context of borderline personality disorder may be accompanied by increased serum concentrations of VEGF and FGF-2 [352]. VEGF has been implicated in neuronal survival, neuroprotection, regeneration, growth, differentiation, and axonal outgrowth, which is involved in the pathophysiology of major depressive disorder and the higher expression levels of VEGF in the peripheral leukocytes are associated with the depressive state [353]. Therefore, there is a possibility that inhibition of VEGFR signalling may have some level of antidepressant effect or act as enhancer of other antidepressant agents [354], and some of the SVM virtual hits in the antidepressant class may be possible VEGFR inhibitors that partly explain their antidepressant activities.

---

### 6.3.5 Comparison of virtual screening performance of SVM with Tanimoto-based similarity searching method

To evaluate whether the performance of SVM is due to the SVM classification models or to the molecular descriptors used, SVM results were compared with those of three other VS methods based on the same molecular descriptors, training dataset of VEGFR-2 inhibitors reported before 2012, and the testing dataset of VEGFR-2 inhibitors reported since 2012 and 168K MDDR compounds. The three other VS methods include two similarity-based methods, Tanimoto-based similarity searching and kNN methods, and an alternative machine learning method PNN. As shown in **Table 6-6**, the yield and maximum possible false-hit rate of the Tanimoto-based similarity searching, kNN and PNN methods are 73.91% and 8.26%, 54.35% and 2.48% , and 54.35% and 3.30% respectively.

Compared to these results, the yield of SVM is significantly improved and the false-hit rate of SVM is substantially reduced. This suggests that SVM performance is due primarily to the SVM classification models rather than the molecular descriptors used, and SVM is capable of achieving comparable yield at substantially reduced false-hit rate as compared to both similarity-based approach and alternative machine learning method.

**Table 6-6** Comparison of virtual screening performance of SVM with those of other methods

| Method                       | Inhibitors in Training Set |   | Inhibitors in Testing Set |   |  | Virtual Screening Performance |   |  |   |
|------------------------------|----------------------------|---|---------------------------|---|--|-------------------------------|---|--|---|
|                              | No of Inhibitors           | No of Chemical Families Covered by Inhibitors | No of Inhibitors          | No of Chemical Families Covered by Inhibitors | Percent of Inhibitors in Chemical Families Covered by Inhibitors in Training Set | Yield                         | No and Percent of Identified True Inhibitors Outside Training Chemical Families | No and Percent of the 168K MDDR Compounds Identified as Inhibitors | No and Percent of the 13,872MDDR Compounds Similar to the Known Inhibitors Identified as Virtual Inhibitors |
| Support Vector Machines      | 3653                       | 845   | 92                        | 35  | 56.52%   | 85.87%                        | 31 (39.24%)   | 2,717 (1.62%)  | 1,714 (12.36%)  |
| Tanimoto Similarity          |                            |   |                           |   |  | 73.91%                        | 32 (47.06%)   | 13,872 (8.26%)   | 13,872 (100%)   |
| K Nearest Neighbour          |                            |   |                           |   |  | 54.35%                        | 12 (24.00%)   | 4164 (2.48%)   | 2,689 (19.38%)  |
| Probabilistic Neural Network |                            |   |                           |   |  | 54.35%                        | 12 (24.00%)   | 5552 (3.30%)   | 2738 (19.74%)   |

---

### **6.3.6 Does SVM select VEGFR inhibitors or membership of compound families?**

To further evaluate whether SVM identifies VEGFR-2 inhibitors rather than membership of certain compound families, Compound family distribution of the identified VEGFR-2 inhibitors and non-inhibitors were analyzed. A total of 39.24% of the identified VEGFR-2 inhibitors belong to the families that contain no known VEGFR-2 inhibitors. For those families that contain at least one known inhibitor, >70% of the compounds (>90% in majority cases) in each of these families were predicted as non-inhibitor by SVM. These results suggest that SVM identifies VEGFR-2 inhibitors rather than membership to certain compound families. Some of the identified inhibitors not in the family of known inhibitors may serve as potential “novel” VEGFR-2 inhibitors. Therefore, as in the case shown by earlier studies [319], SVM has certain capacity for identifying novel active compounds from sparse as well as regular-sized active datasets.

## **6.4 Concluding remarks**

By using training dataset of more diverse spectrum of inactive compounds as well as substantial number of literature-reported VEGFR-2 inhibitors, SVM shows substantial capability in identifying VEGFR-2 inhibitors at comparable yield and in many cases

---

substantially lower false-hit rate than those of typical VS tools reported in the literatures. It is capable of searching large compound libraries at sizes comparable to the 13.56M PubChem and 168K MDDR compounds at low false-hit rates. The performance of SVM is significantly better than that of Tanimoto-based similarity search method based on the same datasets and molecular descriptors, suggesting that the VS performance of SVM is primarily due to SVM classification models rather than the molecular descriptors used. Because of their high computing speed and generalization capability for covering highly diverse spectrum compounds, SVM can be potentially explored to develop useful VS tools to complement other VS methods or to be used as part of integrated VS tools in facilitating the discovery of VEGFR inhibitors and other active compounds [305-307].

---

## **Chapter 7 Concluding remarks**

### **7.1 Major findings and merits**

#### **7.1.1 Merits of the development of MicrobPad MD: microbial pathogen diagnostic methods database**

In this work, we developed the microbial pathogen diagnostic methods database MicrobPad to provide comprehensive information about the molecular diagnostic techniques, targets, primers/probes, detection procedures and conditions, and tested diagnostic accuracies and limit of diagnosis for 314 bacterial, fungal and viral species from 61 genera. While available, additional information such as pathogen strains and hosts, tissue distribution or habitats, cultivation methods, biochemical characteristics, virulence factors, morphology, diseases, symptoms, treatment and prevention methods are provided. Our Database covers 242 gene targets, 700 primers/ probes, 340 virulence factors, and 261 diseases. Cross-links to the NCBI genome and SwissProt/UniProt databases are provided. This work can facilitate accurate, sensitive and low-cost diagnosis of medical pathogens and also boost the development of diagnosis devices of comprehensive coverage of medical pathogens.



---

## **7.1.2 Merits of the updates of TTD in facilitating multi-target drug discovery**

TTD providing pharmaceutical information on therapeutic target is a reliable knowledge hub for established therapeutic target since it is developed. However, the profile of drugs under clinical developing keeps changing in the past decade, and many new drugs have been approved for acting on some new targets. Moreover, many drugs in previous TTD did not indicate their primary target, and there is no information of drugs in clinical trial provided. TTD 2010 update takes these challenges and tries to offer a most comprehensive map of drug targets for the modern pharmaceutical era. In this updated version, TTD significantly expanding target data to 348 successful, 292 clinical trial, and 1,254 research targets, and 560 diseases, and added drug data for 1,514 approved, 1,212 clinical trial and 2,302 experimental drugs linked to their primary targets (3382 small molecule and 649 antisense drugs with available structure and sequence). Other features which add additional credits to TTD 2010 include: (1) collection of information of antisense, aptamer and siRNA based drugs; (2) allowance of customized target search by disease indications, target biochemical classes, drug mode of actions, drug therapeutic classes, and so on; (3) allowance of target search by BLAST; (4) allowance of drug search by tanimoto similarity; and (5) user friendly interface and full data download. Comprehensive data

---

integrated, primary targets identified, detail clinical trial stage for both drugs and targets labeled, and functional features added guarantee this version of TTD a reliable, informative, useful, multifunctional and convenient source of drug target information.

### **7.1.3 Merits of virtual screening model for Src inhibitors**

We evaluated support vector machines (SVM) as virtual screening tools for searching Src inhibitors from large compound libraries. SVM trained and tested by 1,703 inhibitors and 63,318 putative non-inhibitors correctly identified 93.53%~ 95.01% inhibitors and 99.81%~ 99.90% non-inhibitors in 5-fold cross validation studies. SVM trained by 1,703 inhibitors reported before 2011 and 63,318 putative non-inhibitors correctly identified 70.45% of the 44 inhibitors reported since 2011, and predicted as inhibitors 44,843 (0.33%) of 13.56M PubChem, 1,496 (0.89%) of 168K MDDR, and 719 (7.73%) of 9,305 MDDR compounds similar to the known inhibitors. We also compared SVM models with other machine learning methods including kNN, PNN and Tanimoto similarity searching method with the same dataset. SVM showed comparable yield and reduced false hit rates in searching large compound libraries compared to the similarity-based and other machine-learning VS methods developed from the same set of training compounds and molecular descriptors. We tested three virtual hits of the same novel scaffold from in-house chemical libraries not reported as Src inhibitor, one of which showed moderate

---

activity. SVM may be potentially explored for searching Src inhibitors from large compound libraries at low false-hit rates.

#### **7.1.4 Merits of virtual screening model for VEGFR-2 inhibitors**

Approach for identification of VEGFR-2 inhibitors from large compound libraries by SVM virtual screening model was constructed. SVM trained and tested by 3,653 inhibitors and 66,982 putative non-inhibitors correctly identified 93.98%~95.89% inhibitors and 99.53%~99.70% non-inhibitors in 5-fold cross validation studies. SVM trained by 3,653 inhibitors reported before 2012 and 66,982 putative non-inhibitors correctly identified 85.87% of the 92 inhibitors reported since 2012, and predicted as inhibitors 31,624 (0.23%) of 13.56M PubChem, 2,717 (1.62%) of 168K MDDR, and 1,714 (12.36%) of 13,872 MDDR compounds similar to the known inhibitors. Based on this model, one of three virtual hits was experimental found to inhibit VEGFR-2 at a moderate rate of 4.54% at 20 $\mu$ M. In summary, SVM showed substantial capability in searching VEGFR-2 inhibitors from large compound libraries at low false-hit rates.

#### **7.2 Limitations and suggestions for future studies**

MicrobPad MD consists of data of medical pathogens of bacterial, fungal, and viral species. There are innumerable medical pathogens in the nature and more and more

---

are recognized. The number and the diversity of species need to be expanded. New techniques develop quickly, more effective pipeline are needed to process increasing data quickly. More functions are also necessary to added such as portal of data transfer for further applications and utilization of controlled vocabulary space for large scale of data.

Current TTD provides information of targets and drugs on clinical trial phase o which is one of most crucial characteristic. However, the clinical trial status keeps changing for our modern pharmacology is a dynamically moving process, it is difficult to update it manually. A solution to this may be to integrate automatic information system which helps TTD to search latest data update from reliable sources. Other function such as similarly, docking and QSAR model can be implemented.

The compound descriptors of current SVM approach were calculated using our MODEL software. It provides more than 500 diverse types descriptors. However, these still do not cover all the important descriptors. As shown in the study of acute toxicity, some more important descriptors shall be included and evaluated.

The generation of putative negatives was used for the machine learning methods application. This approach requires a classification of the chemical space which has

---

always been a difficult task in chemoinformatics. The classification of the chemical space needs a clustering method, a distance matrix selection and descriptors. K-means clustering method was used in this work. It is not the best clustering method but is suitable and computable for large chemical spaces. In future studies, more advanced clustering algorithm can be developed for improving the accuracy of chemical space clustering. Additionally, the selection of correlation coefficients and other chemical descriptors such as fingerprint can also help the improvement. Another possible drawback associated with the putative negatives generation approach is the possible inclusion of some undiscovered active compounds in the “inactive” class. This may hinder the identification of novel active compounds by machine learning methods. However, such an adverse effect is expected to be relatively small for many biological target classes.

There is no conclusive answer to which VS approach is the best. Both ligand based and structural based methods have their own advantages and drawbacks. Therefore, the choice of one or another depends on the specific case faced by the medicinal chemist. In terms of performance, ligand based methods have the advantage of better enrichment factors and higher speed serving and they are more efficient in removing non active compounds; structure based methods provide a more direct view of the interactions between the ligand and molecular target and it has an advantage for the

---

detecting of novel structures. The VS approaches aims to firstly include less costly approaches, usually ligand based VS, at the first stage and apply the most demanding methods, such as docking, for the last stage when the original large compound library has been reduced to a manageable size after the previous stage

---

## Reference

1. Touron, A., et al., *Detection of Salmonella in environmental water and sediment by a nested-multiplex polymerase chain reaction assay*. Res Microbiol, 2005. **156**(4): p. 541-53.
2. Atlas, R.M., *Legionella: from environmental habitats to disease pathology, detection and control*. Environ Microbiol, 1999. **1**(4): p. 283-93.
3. Gao, L.Y., et al., *A mycobacterial virulence gene cluster extending RD1 is required for cytolysis, bacterial spreading and ESAT-6 secretion*. Mol Microbiol, 2004. **53**(6): p. 1677-93.
4. Arnandis-Chover, T., et al., *Detection of food-borne pathogens with DNA arrays on disk*. Talanta, 2012. **101**: p. 405-12.
5. D.H. Davies, et al., *Infection and Immunity*. 1999: Taylor & Francis (26 Oct 1998). 237.
6. Parham, P., *The Immune System*. second ed. 2005, New York: Garland Science. 431.
7. Prager, M., et al., *Chlamydia pneumoniae in carotid artery atherosclerosis: a comparison of its presence in atherosclerotic plaque, healthy vessels, and circulating leukocytes from the same individuals*. Stroke, 2002. **33**(12): p. 2756-61.
8. Qayoom, S. and Q.M. Ahmad, *Psoriasis and Helicobacter pylori*. Indian J Dermatol Venereol Leprol, 2003. **69**(2): p. 133-4.
9. Dennis, D.T., et al., *Tularemia as a biological weapon*. JAMA: the journal of the American Medical Association, 2001. **285**(21): p. 2763-2773.
10. Kumar, V. and S.L. Robbins, *Robbins basic pathology*. 8th ed. 2007, Philadelphia, PA: Saunders/Elsevier. xiv, 946 p.
11. Morelli, G., et al., *Yersinia pestis genome sequencing identifies patterns of global phylogenetic diversity*. Nat Genet, 2010. **42**(12): p. 1140-3.

- 
12. Freeman, E.E., et al., *Herpes simplex virus 2 infection increases HIV acquisition in men and women: systematic review and meta-analysis of longitudinal studies*. AIDS, 2006. **20**(1): p. 73-83.
  13. Whitehorn, J. and J. Farrar, *Dengue*. Br Med Bull, 2010. **95**: p. 161-73.
  14. Chan-Yeung, M. and R.H. Xu, *SARS: epidemiology*. Respirology, 2003. **8 Suppl**: p. S9-14.
  15. Carette, J.E., et al., *Ebola virus entry requires the cholesterol transporter Niemann-Pick C1*. Nature, 2011. **477**(7364): p. 340-3.
  16. Ho, M., et al., *An epidemic of enterovirus 71 infection in Taiwan. Taiwan Enterovirus Epidemic Working Group*. N Engl J Med, 1999. **341**(13): p. 929-35.
  17. Kobasa, D., et al., *Aberrant innate immune response in lethal infection of macaques with the 1918 influenza virus*. Nature, 2007. **445**(7125): p. 319-23.
  18. Kobayashi, G.S., *Disease of Mechanisms of Fungi*. 1996.
  19. Bariola, J.R., et al., *Detection of Blastomyces dermatitidis antigen in patients with newly diagnosed blastomycosis*. Diagn Microbiol Infect Dis, 2011. **69**(2): p. 187-91.
  20. Naglik, J.R., et al., *Quantitative expression of the Candida albicans secreted aspartyl proteinase gene family in human oral and vaginal candidiasis*. Microbiology, 2008. **154**(Pt 11): p. 3266-80.
  21. Kauffman, C.A., *Histoplasmosis: a clinical and laboratory update*. Clin Microbiol Rev, 2007. **20**(1): p. 115-32.
  22. Othman, N., et al., *Entamoeba histolytica antigenic protein detected in pus aspirates from patients with amoebic liver abscess*. Exp Parasitol, 2013.
  23. Harhay, M.O., J. Horton, and P.L. Olliaro, *Epidemiology and control of human gastrointestinal parasites in children*. Expert Rev Anti Infect Ther, 2010. **8**(2): p. 219-34.
  24. Mueller, I., P.A. Zimmerman, and J.C. Reeder, *Plasmodium malariae and Plasmodium ovale--the "bashful" malaria parasites*. Trends Parasitol, 2007. **23**(6): p. 278-83.



- 
25. Machado-Silva, J.R., et al., *Schistosoma mansoni* Sambon, 1907: comparative morphological studies of some Brazilian strains. Rev Inst Med Trop Sao Paulo, 1995. **37**(5): p. 441-7.
  26. Bilukha, O.O. and N. Rosenstein, *Prevention and control of meningococcal disease. Recommendations of the Advisory Committee on Immunization Practices (ACIP)*. MMWR Recomm Rep, 2005. **54**(RR-7): p. 1-21.
  27. Botelho, M.C., J.C. Machado, and J.M. da Costa, *Schistosoma haematobium and bladder cancer: what lies beneath?* Virulence, 2010. **1**(2): p. 84-7.
  28. Bosch, F.X., et al., *The causal relation between human papillomavirus and cervical cancer*. J Clin Pathol, 2002. **55**(4): p. 244-65.
  29. WHO, *The 10 leading causes of death worldwide (2008)*. WHO Fact Sheets, 2008.
  30. World Health Organization., *World health statistics 2013*. 2013, World Health Organization: Geneva. p. 1 CD-ROM.
  31. Berns, K.I., et al., *Policy: Adaptations of avian flu virus are a cause for concern*. Nature, 2012. **482**(7384): p. 153-4.
  32. He, Y., et al., *Complexes of poliovirus serotypes with their common cellular receptor, CD155*. J Virol, 2003. **77**(8): p. 4827-35.
  33. Tenover, F.C., *Developing molecular amplification methods for rapid diagnosis of respiratory tract infections caused by bacterial pathogens*. Clin Infect Dis, 2011. **52 Suppl 4**: p. S338-45.
  34. Russek-Cohen, E., et al., *FDA perspectives on diagnostic device clinical studies for respiratory infections*. Clin Infect Dis, 2011. **52 Suppl 4**: p. S305-11.
  35. Ginocchio, C.C., *Strengths and weaknesses of FDA-approved/cleared diagnostic devices for the molecular detection of respiratory pathogens*. Clin Infect Dis, 2011. **52 Suppl 4**: p. S312-25.
  36. Joseph, S.J. and T.D. Read, *Bacterial population genomics and infectious disease diagnostics*. Trends Biotechnol, 2010. **28**(12): p. 611-8.

- 
37. Tang, Y.W., G.W. Procop, and D.H. Persing, *Molecular diagnostics of infectious diseases*. Clin Chem, 1997. **43**(11): p. 2021-38.
  38. Velusamy, V., et al., *An overview of foodborne pathogen detection: in the perspective of biosensors*. Biotechnol Adv, 2010. **28**(2): p. 232-54.
  39. Girones, R., et al., *Molecular detection of pathogens in water--the pros and cons of molecular techniques*. Water Res, 2010. **44**(15): p. 4325-39.
  40. Tsourkas, A. and G. Bao, *Shedding light on health and disease using molecular beacons*. Brief Funct Genomic Proteomic, 2003. **1**(4): p. 372-84.
  41. Cerqueira, L., et al., *DNA mimics for the rapid identification of microorganisms by fluorescence in situ hybridization (FISH)*. Int J Mol Sci, 2008. **9**(10): p. 1944-60.
  42. Singh, K., *Laboratory-acquired infections*. Clin Infect Dis, 2009. **49**(1): p. 142-7.
  43. Gao, R., et al., *Human infection with a novel avian-origin influenza A (H7N9) virus*. N Engl J Med, 2013. **368**(20): p. 1888-97.
  44. Drews, J., *Drug discovery: a historical perspective*. Science, 2000. **287**(5460): p. 1960-4.
  45. Augen, J., *The evolving role of information technology in the drug discovery process*. Drug Discov Today, 2002. **7**(5): p. 315-23.
  46. Ashburn, T.T. and K.B. Thor, *Drug repositioning: Identifying and developing new uses for existing drugs*. Nature Reviews Drug Discovery, 2004. **3**(8): p. 673-683.
  47. Newman, D.J., *Natural products as leads to potential drugs: an old process or the new hope for drug discovery?* J Med Chem, 2008. **51**(9): p. 2589-99.
  48. Hogeweg, P., *The roots of bioinformatics in theoretical biology*. PLoS Comput Biol, 2011. **7**(3): p. e1002021.
  49. Brown, F.K., *Chapter 35. Chemoinformatics: What is it and How does it Impact Drug Discovery*. Annual Reports in Med. Chem, 1998. **33**: p. 375.
  50. Brown, F., *Editorial Opinion: Chemoinformatics – a ten year update*. Current Opinion in Drug Discovery & Development, 2005. **8**(3): p. 296–302.

- 
51. Friedberg, I., T. Kaplan, and H. Margalit, *Evaluation of PSI-BLAST alignment accuracy in comparison to structural alignments*. Protein Sci, 2000. **9**(11): p. 2278-84.
  52. Muller, A., R.M. MacCallum, and M.J. Sternberg, *Benchmarking PSI-BLAST in genome annotation*. J Mol Biol, 1999. **293**(5): p. 1257-71.
  53. Chen, C., et al., *Predicting protein structural class based on multi-features fusion*. J Theor Biol, 2008. **253**(2): p. 388-92.
  54. Li, Z.R., et al., *PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence*. Nucleic Acids Res, 2006. **34**(Web Server issue): p. W32-7.
  55. Cerami, E.G., et al., *cPath: open source software for collecting, storing, and querying biological pathways*. BMC Bioinformatics, 2006. **7**.
  56. Cases, I., et al., *CARGO: a web portal to integrate customized biological information*. Nucleic Acids Res, 2007. **35**(Web Server issue): p. W16-20.
  57. Nakazato, T., et al., *BioCompass: a novel functional inference tool that utilizes MeSH hierarchy to analyze groups of genes*. In Silico Biol, 2008. **8**(1): p. 53-61.
  58. Zheng, C.J., et al., *Therapeutic targets: progress of their exploration and investigation of their characteristics*. Pharmacol Rev, 2006. **58**(2): p. 259-79.
  59. Golden, J.B., *Prioritizing the human genome: knowledge management for drug discovery*. Curr Opin Drug Discov Devel, 2003. **6**(3): p. 310-6.
  60. Overington, J.P., B. Al-Lazikani, and A.L. Hopkins, *How many drug targets are there?* Nat Rev Drug Discov, 2006. **5**(12): p. 993-6.
  61. Imming, P., C. Sinning, and A. Meyer, *Drugs, their targets and the nature and number of drug targets*. Nat Rev Drug Discov, 2006. **5**(10): p. 821-34.
  62. Gunther, S., et al., *SuperTarget and Matador: resources for exploring drug-target relationships*. Nucleic Acids Res, 2008. **36**(Database issue): p. D919-22.
  63. Bajorath, J., *Integration of virtual and high-throughput screening*. Nature reviews, 2002. **1**(11): p. 882-94.

- 
64. Bohacek, R.S., C. McMartin, and W.C. Guida, *The art and practice of structure-based drug design: a molecular modeling perspective*. Med Res Rev, 1996. **16**(1): p. 3-50.
  65. Fink, T. and J.L. Reymond, *Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery*. J Chem Inf Model, 2007. **47**(2): p. 342-53.
  66. Rarey, M. and M. Stahl, *Similarity searching in large combinatorial chemistry spaces*. J Comput Aided Mol Des, 2001. **15**(6): p. 497-520.
  67. Rester, U., *From virtuality to reality - Virtual screening in lead discovery and lead optimization: a medicinal chemistry perspective*. Curr Opin Drug Discov Devel, 2008. **11**(4): p. 559-68.
  68. Rollinger, J.M., H. Stuppner, and T. Langer, *Virtual screening for the discovery of bioactive natural products*. Prog Drug Res, 2008. **65**: p. 211, 213-49.
  69. Shoichet, B.K., *Virtual screening of chemical libraries*. Nature, 2004. **432**(7019): p. 862-5.
  70. Lengauer, T., et al., *Novel technologies for virtual screening*. Drug Discov Today, 2004. **9**(1): p. 27-34.
  71. Davies, J.W., M. Glick, and J.L. Jenkins, *Streamlining lead discovery by aligning in silico and high-throughput screening*. Curr Opin Chem Biol, 2006. **10**(4): p. 343-51.
  72. Willett, P., *Similarity-based virtual screening using 2D fingerprints*. Drug Discov Today, 2006. **11**(23-24): p. 1046-53.
  73. van de Waterbeemd, H. and E. Gifford, *ADMET in silico modelling: towards prediction paradise?* Nat Rev Drug Discov, 2003. **2**(3): p. 192-204.
  74. Matthew W. B. Trotter, S.B.H., *Support Vector Machines for ADME Property Classification*. QSAR & Combinatorial Science, 2003. **22**(5): p. 533-548.
  75. Cavasotto, C.N. and A.J. Orry, *Ligand docking and structure-based virtual screening in drug discovery*. Curr Top Med Chem, 2007. **7**(10): p. 1006-14.

- 
76. Guido, R.V., G. Oliva, and A.D. Andricopulo, *Virtual screening and its integration with modern drug design technologies*. *Curr Med Chem*, 2008. **15**(1): p. 37-46.
  77. Brooijmans, N. and I.D. Kuntz, *Molecular recognition and docking algorithms*. *Annu Rev Biophys Biomol Struct*, 2003. **32**: p. 335-73.
  78. Halperin, I., et al., *Principles of docking: An overview of search algorithms and a guide to scoring functions*. *Proteins*, 2002. **47**(4): p. 409-43.
  79. Wang, R., Y. Lu, and S. Wang, *Comparative evaluation of 11 scoring functions for molecular docking*. *J Med Chem*, 2003. **46**(12): p. 2287-303.
  80. Moitessier, N., et al., *Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go*. *Br J Pharmacol*, 2008. **153 Suppl 1**: p. S7-26.
  81. Warren, G.L., et al., *A critical assessment of docking programs and scoring functions*. *J Med Chem*, 2006. **49**(20): p. 5912-31.
  82. Schulz-Gasch, T. and M. Stahl, *Binding site characteristics in structure-based virtual screening: evaluation of current docking tools*. *J Mol Model*, 2003. **9**(1): p. 47-57.
  83. Kim, R. and J. Skolnick, *Assessment of programs for ligand binding affinity prediction*. *J Comput Chem*, 2008. **29**(8): p. 1316-31.
  84. Kirchmair, J., et al., *Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection--what can we learn from earlier mistakes?* *J Comput Aided Mol Des*, 2008. **22**(3-4): p. 213-28.
  85. Sheridan, R.P., G.B. McGaughey, and W.D. Cornell, *Multiple protein structures and multiple ligands: effects on the apparent goodness of virtual screening results*. *J Comput Aided Mol Des*, 2008. **22**(3-4): p. 257-65.
  86. Jain, A.N., *Bias, reporting, and sharing: computational evaluations of docking methods*. *J Comput Aided Mol Des*, 2008. **22**(3-4): p. 201-12.
  87. Harper, G., et al., *Prediction of biological activity for high-throughput screening using binary kernel discrimination*. *J Chem Inf Comput Sci*, 2001. **41**(5): p. 1295-300.

- 
88. Jorissen, R.N. and M.K. Gilson, *Virtual screening of molecular databases using a support vector machine*. J Chem Inf Model, 2005. **45**(3): p. 549-61.
  89. Glick, M., et al., *Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and laplacian-modified naive bayesian classifiers*. J Chem Inf Model, 2006. **46**(1): p. 193-200.
  90. Li, H., et al., *Prediction of estrogen receptor agonists and characterization of associated molecular descriptors by statistical learning methods*. J Mol Graph Model, 2006. **25**(3): p. 313-23.
  91. Lepp, Z., T. Kinoshita, and H. Chuman, *Screening for new antidepressant leads of multiple activities by support vector machines*. J Chem Inf Model, 2006. **46**(1): p. 158-67.
  92. Chen, B., et al., *Evaluation of machine-learning methods for ligand-based virtual screening*. J Comput Aided Mol Des, 2007.
  93. Hert, J., et al., *New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching*. J Chem Inf Model, 2006. **46**(2): p. 462-70.
  94. Franke, L., et al., *Extraction and visualization of potential pharmacophore points using support vector machines: application to ligand-based virtual screening for COX-2 inhibitors*. J Med Chem, 2005. **48**(22): p. 6997-7004.
  95. Ghosh, S., et al., *Structure-based virtual screening of chemical libraries for drug discovery*. Curr Opin Chem Biol, 2006. **10**(3): p. 194-202.
  96. Shoichet, B.K., et al., *Lead discovery using molecular docking*. Curr Opin Chem Biol, 2002. **6**(4): p. 439-46.
  97. Jansen, J.M. and E.J. Martin, *Target-biased scoring approaches and expert systems in structure-based virtual screening*. Curr Opin Chem Biol, 2004. **8**(4): p. 359-64.
  98. Mozziconacci, J.C., et al., *Optimization and validation of a docking-scoring protocol; application to virtual screening for COX-2 inhibitors*. J Med Chem, 2005. **48**(4): p. 1055-68.

- 
99. Vidal, D., M. Thormann, and M. Pons, *A novel search engine for virtual screening of very large databases*. *J Chem Inf Model*, 2006. **46**(2): p. 836-43.
  100. Cummings, M.D., et al., *Comparison of automated docking programs as virtual screening tools*. *J Med Chem*, 2005. **48**(4): p. 962-76.
  101. Evers, A. and T. Klabunde, *Structure-based drug discovery using GPCR homology modeling: successful virtual screening for antagonists of the alpha1A adrenergic receptor*. *J Med Chem*, 2005. **48**(4): p. 1088-97.
  102. Lorber, D.M. and B.K. Shoichet, *Hierarchical docking of databases of multiple ligand conformations*. *Curr Top Med Chem*, 2005. **5**(8): p. 739-49.
  103. Stiefl, N. and A. Zaliani, *A knowledge-based weighting approach to ligand-based virtual screening*. *J Chem Inf Model*, 2006. **46**(2): p. 587-96.
  104. Vangrevelinghe, E., et al., *Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking*. *J Med Chem*, 2003. **46**(13): p. 2656-62.
  105. Doman, T.N., et al., *Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B*. *J Med Chem*, 2002. **45**(11): p. 2213-21.
  106. Enyedy, I.J., et al., *Discovery of small-molecule inhibitors of Bcl-2 through structure-based computer screening*. *J Med Chem*, 2001. **44**(25): p. 4313-24.
  107. Oprea, T.I. and H. Matter, *Integrating virtual screening in lead discovery*. *Curr Opin Chem Biol*, 2004. **8**(4): p. 349-58.
  108. Bocker, A., G. Schneider, and A. Teckentrup, *NIPALSTREE: a new hierarchical clustering approach for large compound libraries and its application to virtual screening*. *J Chem Inf Model*, 2006. **46**(6): p. 2220-9.
  109. Schuster, D., et al., *The discovery of new 11beta-hydroxysteroid dehydrogenase type 1 inhibitors by common feature pharmacophore modeling and virtual screening*. *J Med Chem*, 2006. **49**(12): p. 3454-66.
  110. Steindl, T., C. Laggner, and T. Langer, *Human rhinovirus 3C protease: generation of pharmacophore models for peptidic and nonpeptidic inhibitors and their application in virtual screening*. *J Chem Inf Model*, 2005. **45**(3): p. 716-24.

- 
111. J. Cui, L.Y.H., H.H. Lin, H.L. Zhang, Z.Q. Tang, C.J. Zheng, Z.W. Cao, and Y.Z. Chen, *Prediction of MHC-Binding Peptides of Flexible Lengths from Sequence-Derived Structural and Physicochemical Properties*. Mol. Immunol, 2007. **44**: p. 866-877.
  112. Lepp, Z., T. Kinoshita, and H. Chuman, *Screening for new antidepressant leads of multiple activities by support vector machines*. Journal of Chemical Information and Modeling., 2006. **46**(1): p. 158-167.
  113. Li, H., et al., *Statistical learning approach for predicting specific pharmacodynamic, pharmacokinetic or toxicological properties of pharmaceutical agents*. . Drug Development Research, 2006. **66**(4): p. 245-259.
  114. Han, L.Y., et al., *A support vector machines approach for virtual screening of active compounds of single and multiple mechanisms from large libraries at an improved hit-rate and enrichment factor*. J Mol Graph Model, 2008. **26**(8): p. 1276-86.
  115. Wilton, D.J., et al., *Virtual screening using binary kernel discrimination: analysis of pesticide data*. J Chem Inf Model, 2006. **46**(2): p. 471-7.
  116. Chen, B., et al., *Virtual screening using binary kernel discrimination: effect of noisy training data and the optimization of performance*. J Chem Inf Model, 2006. **46**(2): p. 478-86.
  117. Alvarez, J.C., *High-throughput docking as a source of novel drug leads*. Curr Opin Chem Biol, 2004. **8**(4): p. 365-70.
  118. Schapira, M., et al., *Discovery of diverse thyroid hormone receptor antagonists by high-throughput docking*. Proc Natl Acad Sci U S A, 2003. **100**(12): p. 7354-9.
  119. Lipinski, C.A., et al., *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings*. Adv Drug Deliv Rev, 2001. **46**(1-3): p. 3-26.
  120. Perola, E., *Minimizing false positives in kinase virtual screens*. Proteins, 2006. **64**(2): p. 422-35.



- 
121. Pirard, B., J. Brendel, and S. Peukert, *The discovery of Kv1.5 blockers as a case study for the application of virtual screening approaches*. J Chem Inf Model, 2005. **45**(2): p. 477-85.
  122. Rella, M., et al., *Structure-based pharmacophore design and virtual screening for novel angiotensin converting enzyme 2 inhibitors*. J Chem Inf Model, 2006. **46**(2): p. 708-16.
  123. Lipinski, C. and A. Hopkins, *Navigating chemical space for biology and medicine*. Nature, 2004. **432**(7019): p. 855-61.
  124. Seringhaus, M.R. and M.B. Gerstein, *Publishing perishing? Towards tomorrow's information architecture*. BMC Bioinformatics, 2007. **8**: p. 17.
  125. Baumgartner, W.A., Jr., et al., *Manual curation is not sufficient for annotation of genomic databases*. Bioinformatics, 2007. **23**(13): p. i41-8.
  126. Wheeler, D.L., et al., *Database resources of the National Center for Biotechnology Information*. Nucleic Acids Res, 2006. **34**(Database issue): p. D173-80.
  127. Beynon-Davies, P., *Database systems*. 3rd ed. 2004, Basingstoke: Palgrave Macmillan. xiv, 601 p.
  128. Overington, J., *ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI)*. Interview by Wendy A. Warr. J Comput Aided Mol Des, 2009. **23**(4): p. 195-8.
  129. Susnow, R.G. and S.L. Dixon, *Use of robust classification techniques for the prediction of human cytochrome P450 2D6 inhibition*. J Chem Inf Comput Sci, 2003. **43**(4): p. 1308-15.
  130. Perez, J.J., *Managing molecular diversity*, in *Chemical Society Reviews*. 2005, Royal Society of Chemistry. p. 143-152.
  131. Willett, P., J.M. Barnard, and G.M. Downs, *Chemical Similarity Searching*. J. Chem. Inf. Comput. Sci., 1998. **38**(6): p. 983-996.
  132. Fang, H., et al., *Structure-activity relationships for a large diverse set of natural, synthetic, and environmental estrogens*. Chem Res Toxicol, 2001. **14**(3): p. 280-94.

- 
133. Tong, W., et al., *Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity*. Environ Health Perspect, 2004. **112**(12): p. 1249-54.
  134. Hu, J.Y. and T. Aizawa, *Quantitative structure-activity relationships for estrogen receptor binding affinity of phenolic chemicals*. Water Res, 2003. **37**(6): p. 1213-22.
  135. Jacobs, M.N., *In silico tools to aid risk assessment of endocrine disrupting chemicals*. Toxicology, 2004. **205**(1-2): p. 43-53.
  136. Byvatov, E., et al., *Comparison of support vector machine and artificial neural network systems for drug/nondrug classification*. J Chem Inf Comput Sci, 2003. **43**(6): p. 1882-9.
  137. Doniger, S., T. Hofmann, and J. Yeh, *Predicting CNS permeability of drug molecules: comparison of neural network and support vector machine algorithms*. J Comput Biol, 2002. **9**(6): p. 849-64.
  138. He, L., et al., *Predicting the genotoxicity of polycyclic aromatic compounds from molecular structure with different classifiers*. Chem Res Toxicol, 2003. **16**(12): p. 1567-80.
  139. Snyder, R.D., et al., *Assessment of the sensitivity of the computational programs DEREK, TOPKAT, and MCASE in the prediction of the genotoxicity of pharmaceutical molecules*. Environ Mol Mutagen, 2004. **43**(3): p. 143-58.
  140. Xue, Y., et al., *Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents*. J Chem Inf Comput Sci, 2004. **44**(5): p. 1630-8.
  141. Yap, C.W., et al., *Prediction of torsade-causing potential of drugs by support vector machine approach*. Toxicol Sci, 2004. **79**(1): p. 170-7.
  142. Yap, C.W. and Y.Z. Chen, *Quantitative Structure-Pharmacokinetic Relationships for drug distribution properties by using general regression neural network*. J Pharm Sci, 2005. **94**(1): p. 153-68.
  143. Zernov, V.V., et al., *Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions*. J Chem Inf Comput Sci, 2003. **43**(6): p. 2048-56.

- 
144. Vultur, A., et al., *SKI-606 (bosutinib), a novel Src kinase inhibitor, suppresses migration and invasion of human breast cancer cells*. *Mol Cancer Ther*, 2008. **7**(5): p. 1185-94.
145. Hall LH, K.G., Haney DN, *Molconn-Z*. 2002: eduSoft LC: Ashland VA.
146. Yap, C.W., et al., *Regression methods for developing QSAR and QSPR models to predict compounds of specific pharmacodynamic, pharmacokinetic and toxicological properties*. *Mini Rev Med Chem*, 2007. **7**(11): p. 1097-107.
147. Steinbeck, C., et al., *The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics*. *J Chem Inf Comput Sci*, 2003. **43**(2): p. 493-500.
148. Steinbeck, C., et al., *Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics*. *Curr Pharm Des*, 2006. **12**(17): p. 2111-20.
149. Wegner, J.K., *JOELib/JOELib2*. 2005, Department of Computer Science, University of Tübingen: Germany.
150. Hemmer, M.C., V. Steinhauer, and J. Gasteiger, *Deriving the 3D structure of organic molecules from their infrared spectra*. *Vibrational Spectroscopy*, 1999. **19**(1): p. 151-164.
151. Rücker, G. and C. Rücker, *Counts of all walks as atomic and molecular descriptors*. *Journal of Chemical Information and Computer Sciences*, 1993. **33**(5): p. 683-695.
152. Schuur, J.H., P. Setzer, and J. Gasteiger, *The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity*. *Journal of Chemical Information and Computer Sciences*, 1996. **36**(2): p. 334-344.
153. Pearlman, R.S. and K.M. Smith, *Metric validation and the receptor-relevant subspace concept*. *Journal of Chemical Information and Computer Sciences*, 1999. **39**(1): p. 28-35.
154. Bravi, G., et al., *MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: A comparative 3D QSAR study in a series of steroids*. *Journal of Computer-Aided Molecular Design*, 1997. **11**(1): p. 79-92.

- 
155. Galvez, J., et al., *Charge indexes. New topological descriptors*. Journal of Chemical Information and Computer Sciences, 1994. **34**(3): p. 520-525.
156. Consonni, V., R. Todeschini, and M. Pavan, *Structure/Response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors*. Journal of Chemical Information and Computer Sciences, 2002. **42**(3): p. 682-692.
157. Randic, M., *Graph theoretical approach to local and overall aromaticity of benzenoid hydrocarbons*. Tetrahedron, 1975. **31**(11-12): p. 1477-1481.
158. Randic, M., *Molecular profiles. Novel geometry-dependent molecular descriptors*. New Journal of Chemistry, 1995. **19**: p. 781-791.
159. Kier, L.B. and L.H. Hall, *Molecular structure description: The electrotopological state*. 1999, San Diego: Academic Press.
160. Platts, J.A., et al., *Estimation of molecular free energy relation descriptors using a group contribution approach*. Journal of Chemical Information and Computer Sciences, 1999. **39**(5): p. 835-845.
161. Liu, T., et al., *BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities*. Nucleic Acids Res, 2007. **35**(Database issue): p. D198-201.
162. Sadowski, J., J. Gasteiger, and G. Klebe, *Comparison of Automatic Three-Dimensional Model Builders Using 639 X-Ray Structures*. J. Chem. Inf. Comput. Sci., 1994. **34**: p. 1000-1008.
163. Dutta, D., et al., *Scalable partitioning and exploration of chemical spaces using geometric hashing*. J Chem Inf Model, 2006. **46**(1): p. 321-33.
164. Eriksson, L., et al., *Multi- and megavariable data analysis - Principles and applications*. 2001, Umea, Sweden: Umetrics, AB.
165. Parsons, H.M., et al., *Improved classification accuracy in 1- and 2-dimensional NMR metabolomics data using the variance stabilising generalised logarithm transformation*. BMC Bioinformatics, 2007. **8**: p. 234.
166. van den Berg, R.A., et al., *Centering, scaling, and transformations: improving the biological information content of metabolomics data*. BMC Genomics, 2006. **7**: p. 142.

- 
167. Vapnik, V.N., *The nature of statistical learning theory*. 1995, New York: Springer.
  168. Burges, C.J.C., *A tutorial on support vector machines for pattern recognition*. *Data Mining and Knowledge Discovery*, 1998. **2**(2): p. 127-167.
  169. Pochet, N., et al., *Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction*. *Bioinformatics*, 2004. **20**: p. 3185-3195.
  170. Li, F. and Y. Yang, *Analysis of recursive gene selection approaches from microarray data*. *Bioinformatics*, 2005. **21**: p. 3741-3747.
  171. Jorissen, R.N. and M.K. Gilson, *Virtual screening of molecular databases using a support vector machine*. *J. Chem. Inf. Model*, 2005. **45**(3): p. 549-61.
  172. Glick, M., et al., *Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and laplacian-modified naive bayesian classifiers*. *J. Chem. Inf. Model*, 2006. **46**(1): p. 193-200.
  173. Lepp, Z., T. Kinoshita, and H. Chuman, *Screening for new antidepressant leads of multiple activities by support vector machines*. *J. Chem. Inf. Model*, 2006. **46**(1): p. 158-67.
  174. Hert, J., et al., *New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching*. *J. Chem. Inf. Model*, 2006. **46**(2): p. 462-70.
  175. Yap, C.W. and Y.Z. Chen, *Quantitative Structure-Pharmacokinetic Relationships for drug distribution properties by using general regression neural network*. *J. Pharm. Sci*, 2005. **94**(1): p. 153-68.
  176. Yap, C.W. and Y.Z. Chen, *Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines*. *J. Chem. Inf. Model*, 2005. **45**(4): p. 982-92.
  177. Grover, I.I., I.I. Singh, and I.I. Bakshi, *Quantitative structure-property relationships in pharmaceutical research - Part 2*. *Pharm. Sci. Technol. Today*, 2000. **3**(2): p. 50-57.

- 
178. Trotter, M.W.B., B.F. Buxton, and S.B. Holden, *Support vector machines in combinatorial chemistry*. Meas. Control, 2001. **34**(8): p. 235-239.
179. Burbidge, R., et al., *Drug design by machine learning: support vector machines for pharmaceutical data analysis*. Comput. Chem., 2001. **26**(1): p. 5-14.
180. Czerminski, R., A. Yasri, and D. Hartsough, *Use of support vector machine in pattern classification: Application to QSAR studies*. Quantitative Structure-Activity Relationships, 2001. **20**(3): p. 227-240.
181. Chang, C.C. and C.J. Lin. *LIBSVM : a library for support vector machines*. 2001; Available from: <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
182. Johnson, R.A. and D.W. Wichern, *Applied multivariate statistical analysis*. 1982, Englewood Cliffs, NJ: Prentice Hall.
183. Fix, E. and J.L. Hodges, *Discriminatory analysis: Non-parametric discrimination: Consistency properties*. 1951, Texas: USAF School of Aviation Medicine. 261-279.
184. Fujishima, S. and Y. Takahashi, *Classification of dopamine antagonists using TFS-based artificial neural network*. J Chem Inf Comput Sci, 2004. **44**(3): p. 1006-9.
185. Bostrom, J., A. Hogner, and S. Schmitt, *Do structurally similar ligands bind in a similar fashion?* J. Med. Chem, 2006. **49**(23): p. 6716-25.
186. Huang, N., B.K. Shoichet, and J.J. Irwin, *Benchmarking sets for molecular docking*. J. Med. Chem, 2006. **49**(23): p. 6789-801.
187. Chen, B., et al., *Evaluation of machine-learning methods for ligand-based virtual screening*. J. Comput. Aided Mol. Des., 2007. **21**(1-3): p. 53-62.
188. Franke, L., et al., *Extraction and visualization of potential pharmacophore points using support vector machines: application to ligand-based virtual screening for COX-2 inhibitors*. J. Med. Chem, 2005. **48**(22): p. 6997-7004.
189. Cai, C.Z., et al., *SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence*. Nucleic Acids Res., 2003. **31**(13): p. 3692-7.

- 
190. Han, L.Y., et al., *Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach*. Nucleic Acids Res., 2004. **32**(21): p. 6437-44.
  191. Lin, H.H., et al., *Prediction of transporter family from protein sequence by support vector machine approach*. Proteins, 2006. **62**(1): p. 218-31.
  192. Han, L.Y., et al., *Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness*. Drug Discov. Today, 2007. **12**(7-8): p. 304-13.
  193. Bocker, A., G. Schneider, and A. Teckentrup, *NIPALSTREE: a new hierarchical clustering approach for large compound libraries and its application to virtual screening*. J. Chem. Inf. Model, 2006. **46**(6): p. 2220-9.
  194. Oprea, T.I. and J. Gottfries, *Chemography: the art of navigating in chemical space*. J. Comb. Chem, 2001. **3**(2): p. 157-66.
  195. Xue, Y., et al., *Prediction of P-glycoprotein substrates by a support vector machine approach*. J. Chem. Inf. Comput. Sci, 2004. **44**(4): p. 1497-505.
  196. Koch, M.A., et al., *Charting biologically relevant chemical space: a structural classification of natural products (SCONP)*. Proc. Natl. Acad. Sci. U.S.A., 2005. **102**(48): p. 17272-7.
  197. Mosier, P.D. and P.C. Jurs, *QSAR/QSPR studies using probabilistic neural networks and generalized regression neural networks*. J Chem Inf Comput Sci, 2002. **42**(6): p. 1460-70.
  198. Hawkins, D.M., *The problem of overfitting*. J Chem Inf Comput Sci, 2004. **44**(1): p. 1-12.
  199. Wold, S. and L. Eriksson, *Statistical validation of QSAR results*, in *Chemometric methods in molecular design*, H. Van de Waterbeemd, Editor. 1995, Wiley-VCH: Weinheim; New York. p. 309-318.
  200. Golbraikh, A. and A. Tropsha, *Beware of q<sup>2</sup>!* J Mol Graph Model, 2002. **20**(4): p. 269-76.
  201. Matthews, B., *Comparison of the predicted and observed secondary structure of T4 phage lysozyme*. Biochim Biophys Acta, 1975. **405**(2): p. 442-51.

- 
202. Li, H., et al., *Machine learning approaches for predicting compounds that interact with therapeutic and ADMET related proteins*. J Pharm Sci, 2007. **96**(11): p. 2838-60.
  203. Liljemark, W.F. and C. Bloomquist, *Human oral microbial ecology and dental caries and periodontal diseases*. Crit Rev Oral Biol Med, 1996. **7**(2): p. 180-98.
  204. Preuner, S. and T. Lion, *Towards molecular diagnostics of invasive fungal infections*. Expert Rev Mol Diagn, 2009. **9**(5): p. 397-401.
  205. Endimiani, A., et al., *Are we ready for novel detection methods to treat respiratory pathogens in hospital-acquired pneumonia?* Clin Infect Dis, 2011. **52 Suppl 4**: p. S373-83.
  206. Ecker, D.J., et al., *New technology for rapid molecular diagnosis of bloodstream infections*. Expert Rev Mol Diagn, 2010. **10**(4): p. 399-415.
  207. Harmsen, D., et al., *RIDOM: Ribosomal Differentiation of Medical Micro-organisms Database*. Nucleic Acids Res, 2002. **30**(1): p. 416-7.
  208. Cloud, J.L., et al., *Evaluation of partial 16S ribosomal DNA sequencing for identification of nocardia species by using the MicroSeq 500 system with an expanded database*. J Clin Microbiol, 2004. **42**(2): p. 578-84.
  209. Conville, P.S., P.R. Murray, and A.M. Zelazny, *Evaluation of the integrated database network system (IDNS) SmartGene software for analysis of 16S rRNA gene sequences for identification of Nocardia species*. J Clin Microbiol, 2010. **48**(8): p. 2995-8.
  210. Brudey, K., et al., *Mycobacterium tuberculosis complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology*. BMC Microbiol, 2006. **6**: p. 23.
  211. O'Donnell, K., et al., *Internet-accessible DNA sequence database for identifying fusaria from human and animal infections*. J Clin Microbiol, 2010. **48**(10): p. 3708-18.
  212. Woo, P.C., et al., *Automated identification of medically important bacteria by 16S rRNA gene sequencing using a novel comprehensive database, 16SpathDB*. J Clin Microbiol, 2011. **49**(5): p. 1799-809.



- 
213. McDougal, L.K., et al., *Pulsed-field gel electrophoresis typing of oxacillin-resistant Staphylococcus aureus isolates from the United States: establishing a national database*. J Clin Microbiol, 2003. **41**(11): p. 5113-20.
  214. Ahmed, N., et al., *genoBASE pylori: a genotype search tool and database of the human gastric pathogen Helicobacter pylori*. Infect Genet Evol, 2007. **7**(4): p. 463-8.
  215. Yang, J., et al., *TrED: the Trichophyton rubrum Expression Database*. BMC Genomics, 2007. **8**: p. 250.
  216. Thwaites, G., et al., *Tuberculous meningitis*. J Neurol Neurosurg Psychiatry, 2000. **68**(3): p. 289-99.
  217. Yang, S. and R.E. Rothman, *PCR-based diagnostics for infectious diseases: uses, limitations, and future applications in acute-care settings*. Lancet Infect Dis, 2004. **4**(6): p. 337-48.
  218. Sayers, E.W., et al., *Database resources of the National Center for Biotechnology Information*. Nucleic Acids Res, 2011. **39**(Database issue): p. D38-51.
  219. Consortium, U., *Ongoing and future developments at the Universal Protein Resource*. Nucleic Acids Res, 2011. **39**(Database issue): p. D214-9.
  220. Tevere, V.J., et al., *Detection of Mycobacterium tuberculosis by PCR amplification with pan-Mycobacterium primers and hybridization to an M. tuberculosis-specific probe*. J Clin Microbiol, 1996. **34**(4): p. 918-23.
  221. D'Amato, R.F., et al., *Rapid diagnosis of pulmonary tuberculosis by using Roche AMPLICOR Mycobacterium tuberculosis PCR test*. J Clin Microbiol, 1995. **33**(7): p. 1832-4.
  222. Everley, R.A., et al., *Characterization of Clostridium species utilizing liquid chromatography/mass spectrometry of intact proteins*. J Microbiol Methods, 2009. **77**(2): p. 152-8.
  223. Baum, A., R. Sachidanandam, and A. Garcia-Sastre, *Preference of RIG-I for short viral RNA molecules in infected cells revealed by next-generation sequencing*. Proc Natl Acad Sci U S A, 2010. **107**(37): p. 16303-8.

- 
224. Jarvius, J., et al., *Digital quantification using amplified single-molecule detection*. Nat Methods, 2006. **3**(9): p. 725-7.
225. Xiao, M., et al., *Rapid DNA mapping by fluorescent single molecule detection*. Nucleic Acids Res, 2007. **35**(3): p. e16.
226. Howden, B.P., et al., *Reduced vancomycin susceptibility in Staphylococcus aureus, including vancomycin-intermediate and heterogeneous vancomycin-intermediate strains: resistance mechanisms, laboratory detection, and clinical implications*. Clin Microbiol Rev, 2010. **23**(1): p. 99-139.
227. Wallis, R.S., et al., *Biomarkers and diagnostics for tuberculosis: progress, needs, and translation into practice*. Lancet, 2010. **375**(9729): p. 1920-37.
228. Ohlstein, E.H., R.R. Ruffolo, Jr., and J.D. Elliott, *Drug discovery in the next millennium*. Annu Rev Pharmacol Toxicol, 2000. **40**: p. 177-91.
229. Zambrowicz, B.P. and A.T. Sands, *Knockouts model the 100 best-selling drugs--will they model the next 100?* Nat Rev Drug Discov, 2003. **2**(1): p. 38-51.
230. Lindsay, M.A., *Target discovery*. Nat Rev Drug Discov, 2003. **2**(10): p. 831-8.
231. Edwards, A., *Large-scale structural biology of the human proteome*. Annu Rev Biochem, 2009. **78**: p. 541-68.
232. Lundstrom, K., *Structural genomics: the ultimate approach for rational drug design*. Mol Biotechnol, 2006. **34**(2): p. 205-12.
233. Kramer, R. and D. Cohen, *Functional genomics to new drug targets*. Nat Rev Drug Discov, 2004. **3**(11): p. 965-72.
234. Dey, R., S. Khan, and B. Saha, *A novel functional approach toward identifying definitive drug targets*. Curr Med Chem, 2007. **14**(22): p. 2380-92.
235. Hopkins, A.L., *Network pharmacology: the next paradigm in drug discovery*. Nat Chem Biol, 2008. **4**(11): p. 682-90.
236. Giallourakis, C., et al., *Disease gene discovery through integrative genomics*. Annu Rev Genomics Hum Genet, 2005. **6**: p. 381-406.

- 
237. Zimmermann, G.R., J. Lehar, and C.T. Keith, *Multi-target therapeutics: when the whole is greater than the sum of the parts*. Drug Discov Today, 2007. **12**(1-2): p. 34-42.
238. Jia, J., et al., *Mechanisms of drug combinations: interaction and network perspectives*. Nat Rev Drug Discov, 2009. **8**(2): p. 111-28.
239. Liebler, D.C. and F.P. Guengerich, *Elucidating mechanisms of drug-induced toxicity*. Nat Rev Drug Discov, 2005. **4**(5): p. 410-20.
240. Eichelbaum, M., M. Ingelman-Sundberg, and W.E. Evans, *Pharmacogenomics and individualized drug therapy*. Annu Rev Med, 2006. **57**: p. 119-37.
241. Barcellos, G.B., et al., *Molecular modeling as a tool for drug discovery*. Curr Drug Targets, 2008. **9**(12): p. 1084-91.
242. Lee, G.M. and C.S. Craik, *Trapping moving targets with small molecules*. Science, 2009. **324**(5924): p. 213-5.
243. Zhu, F., et al., *What are next generation innovative therapeutic targets? Clues from genetic, structural, physicochemical, and systems profiles of successful targets*. J Pharmacol Exp Ther, 2009. **330**(1): p. 304-15.
244. Han, L.Y., et al., *Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness*. Drug Discov Today, 2007. **12**(7-8): p. 304-13.
245. Wishart, D.S., et al., *DrugBank: a knowledgebase for drugs, drug actions and drug targets*. Nucleic Acids Res, 2008. **36**(Database issue): p. D901-6.
246. Gao, Z., et al., *PDTD: a web-accessible protein database for drug target identification*. BMC Bioinformatics, 2008. **9**: p. 104.
247. Chen, X., Z.L. Ji, and Y.Z. Chen, *TTD: Therapeutic Target Database*. Nucleic Acids Res, 2002. **30**(1): p. 412-5.
248. Yildirim, M.A., et al., *Drug-target network*. Nat Biotechnol, 2007. **25**(10): p. 1119-26.

- 
249. Wishart, D.S., et al., *DrugBank: a comprehensive resource for in silico drug discovery and exploration*. Nucleic Acids Res, 2006. **34**(Database issue): p. D668-72.
250. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*. Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
251. Willett, P., *Chemical Similarity Searching*. J. Chem. Inf. Comput. Sci, 1998. **38**: p. 983-996.
252. Ma, X.H., et al., *Evaluation of virtual screening performance of support vector machines trained by sparsely distributed active compounds*. J Chem Inf Model, 2008. **48**(6): p. 1227-37.
253. George, R.A. and J. Heringa, *Protein domain identification and improved sequence similarity searching using PSI-BLAST*. Proteins, 2002. **48**(4): p. 672-81.
254. Gerstein, M., *Measurement of the effectiveness of transitive sequence comparison, through a third 'intermediate' sequence*. Bioinformatics, 1998. **14**(8): p. 707-14.
255. Wood, T.C. and W.R. Pearson, *Evolution of protein sequences and structures*. J Mol Biol, 1999. **291**(4): p. 977-95.
256. Koehl, P. and M. Levitt, *Sequence variations within protein families are linearly related to structural variations*. J Mol Biol, 2002. **323**(3): p. 551-62.
257. Brunton, V.G. and M.C. Frame, *Src and focal adhesion kinase as therapeutic targets in cancer*. Curr Opin Pharmacol, 2008. **8**(4): p. 427-32.
258. Bjorge, J.D., et al., *Simultaneous siRNA targeting of Src and downstream signaling molecules inhibit tumor formation and metastasis of a human model breast cancer cell line*. PLoS One, 2011. **6**(4): p. e19309.
259. Tatosyan, A.G. and O.A. Mizenina, *Kinases of the Src family: structure and functions*. Biochemistry (Mosc), 2000. **65**(1): p. 49-58.
260. Belsches-Jablonski, A.P., et al., *The Src pathway as a therapeutic strategy*. Drug Discovery Today: Therapeutic Strategies, 2006. **2**(4): p. 313-321.

- 
261. Gill, A.L., et al., *A comparison of physicochemical property profiles of marketed oral drugs and orally bioavailable anti-cancer protein kinase inhibitors in clinical development*. *Curr Top Med Chem*, 2007. **7**(14): p. 1408-22.
262. Lee, D. and O. Gautschi, *Clinical development of SRC tyrosine kinase inhibitors in lung cancer*. *Clin Lung Cancer*, 2006. **7**(6): p. 381-4.
263. Hiscox, S. and R.I. Nicholson, *Src inhibitors in breast cancer therapy*. *Expert Opin Ther Targets*, 2008. **12**(6): p. 757-67.
264. Lin, L.G., et al., *Naturally occurring homoisoflavonoids function as potent protein tyrosine kinase inhibitors by c-Src-based high-throughput screening*. *J Med Chem*, 2008. **51**(15): p. 4419-29.
265. Lee, K., et al., *Structure-based virtual screening of Src kinase inhibitors*. *Bioorg Med Chem*, 2009. **17**(8): p. 3152-61.
266. Farard, J., et al., *Design, synthesis and evaluation of new 6-substituted-5-benzyloxy-4-oxo-4H-pyran-2-carboxamides as potential Src inhibitors*. *J Enzyme Inhib Med Chem*, 2008. **23**(5): p. 629-40.
267. Alfaro-Lopez, J., et al., *Discovery of a novel series of potent and selective substrate-based inhibitors of p60c-src protein tyrosine kinase: conformational and topographical constraints in peptide design*. *J Med Chem*, 1998. **41**(13): p. 2252-60.
268. Chen, P., et al., *Imidazoquinoxaline Src-family kinase p56Lck inhibitors: SAR, QSAR, and the discovery of (S)-N-(2-chloro-6-methylphenyl)-2-(3-methyl-1-piperazinyl)imidazo-[1,5-a]pyrido[3,2-e]pyrazin-6-amine (BMS-279700) as a potent and orally active inhibitor with excellent in vivo antiinflammatory activity*. *J Med Chem*, 2004. **47**(18): p. 4517-29.
269. Ghosh, S., et al., *Structure-based virtual screening of chemical libraries for drug discovery*. *Curr. Opin. Chem. Biol*, 2006. **10**(3): p. 194-202.
270. Li, H., et al., *Machine learning approaches for predicting compounds that interact with therapeutic and ADMET related proteins*. *J. Pharm. Sci*, 2007. **96**(11): p. 2838-60.

- 
271. Mayer, D., F. Leisch, and K. Hornik, *The support vector machine under test*. Neurocomputing, 2003. **55**(1-2): p. 169-186.
272. Verdonk, M.L., et al., *Virtual screening using protein-ligand docking: avoiding artificial enrichment*. J Chem Inf Comput Sci, 2004. **44**(3): p. 793-806.
273. Altmann, E., et al., *7-Pyrrolidinyl- and 7-piperidinyl-5-aryl-pyrrolo[2,3-d]pyrimidines--potent inhibitors of the tyrosine kinase c-Src*. Bioorg Med Chem Lett, 2001. **11**(6): p. 853-6.
274. Widler, L., et al., *7-Alkyl- and 7-cycloalkyl-5-aryl-pyrrolo[2,3-d]pyrimidines--potent inhibitors of the tyrosine kinase c-Src*. Bioorg Med Chem Lett, 2001. **11**(6): p. 849-52.
275. Missbach, M., et al., *Substituted 5,7-diphenyl-pyrrolo[2,3d]pyrimidines: potent inhibitors of the tyrosine kinase c-Src*. Bioorg Med Chem Lett, 2000. **10**(9): p. 945-9.
276. Klutchko, S.R., et al., *2-Substituted aminopyrido[2,3-d]pyrimidin-7(8H)-ones. structure-activity relationships against selected tyrosine kinases and in vitro and in vivo anticancer activity*. J Med Chem, 1998. **41**(17): p. 3276-92.
277. Noronha, G., et al., *Discovery of [7-(2,6-dichlorophenyl)-5-methylbenzo [1,2,4]triazin-3-yl]-[4-(2-pyrrolidin-1-ylethoxy)phenyl]amine--a potent, orally active Src kinase inhibitor with anti-tumor activity in preclinical assays*. Bioorg Med Chem Lett, 2007. **17**(3): p. 602-8.
278. Keseru, G.M. and G.M. Makara, *The influence of lead discovery strategies on the properties of drug candidates*. Nat Rev Drug Discov, 2009. **8**(3): p. 203-12.
279. Keseru, G.M. and G.M. Makara, *Hit discovery and hit-to-lead approaches*. Drug Discov Today, 2006. **11**(15-16): p. 741-8.
280. Dow, R.L., et al., *Selective inhibition of the tyrosine kinase pp60<sup>src</sup> by analogs of 5, 10-dihydropyrimido [4, 5-b] quinolin-4 (1H)-one*. Bioorganic & Medicinal Chemistry Letters, 1995. **5**(9): p. 1007-1010.
281. Klutchko, S.R., et al., *2-Substituted aminopyrido [2, 3-d] pyrimidin-7 (8 H)-ones. Structure-activity relationships against selected tyrosine kinases and*

- 
- in vitro and in vivo anticancer activity*. Journal of medicinal chemistry, 1998. **41**(17): p. 3276-3292.
282. Altmann, E., et al., *7-Pyrrolidinyl-and 7-piperidinyl-5-aryl-pyrrolo [2, 3-*d*] pyrimidines—potent inhibitors of the tyrosine kinase c-Src*. Bioorganic & Medicinal Chemistry Letters, 2001. **11**(6): p. 853-856.
283. Kinoshita, K., et al., *9-substituted 6,6-dimethyl-11-oxo-6,11-dihydro-5H-benzo[b]carbazoles as highly selective and potent anaplastic lymphoma kinase inhibitors*. J Med Chem, 2011. **54**(18): p. 6286-94.
284. Schmidt, S., et al., *Dual IGF-1R/SRC inhibitors based on a N'-aroyl-2-(1H-indol-3-yl)-2-oxoacetohydrazide structure*. Eur J Med Chem, 2011. **46**(7): p. 2759-69.
285. Crew, A.P., et al., *Imidazo[1,5-a]pyrazines: orally efficacious inhibitors of mTORC1 and mTORC2*. Bioorg Med Chem Lett, 2011. **21**(7): p. 2092-7.
286. Pevet, I., et al., *Synthesis and pharmacological evaluation of thieno[2,3-b]pyridine derivatives as novel c-Src inhibitors*. Bioorg Med Chem, 2011. **19**(8): p. 2517-28.
287. Guagnano, V., et al., *Discovery of 3-(2,6-dichloro-3,5-dimethoxy-phenyl)-1-{6-[4-(4-ethyl-piperazin-1-yl)-phenyl amin o]-pyrimidin-4-yl}-1-methyl-urea (NVP-BGJ398), a potent and selective inhibitor of the fibroblast growth factor receptor family of receptor tyrosine kinase*. J Med Chem, 2011. **54**(20): p. 7066-83.
288. Kumar, A., et al., *Synthesis of 3-phenylpyrazolopyrimidine-1,2,3-triazole conjugates and evaluation of their Src kinase inhibitory and anticancer activities*. Bioorg Med Chem Lett, 2011. **21**(5): p. 1342-6.
289. Liew, C.Y., et al., *SVM Model for Virtual Screening of Lck Inhibitors*. J Chem Inf Model, 2009.
290. Briem, H. and J. Gunther, *Classifying "kinase inhibitor-likeness" by using machine-learning methods*. Chembiochem, 2005. **6**(3): p. 558-66.
291. Ma, X.H., et al., *Comparative analysis of machine learning methods in ligand-based virtual screening of large compound libraries*. Comb Chem High Throughput Screen, 2009. **12**(4): p. 344-57.

- 
292. Chiu, Y.C., et al., *Thrombin-induced IL-6 production in human synovial fibroblasts is mediated by PAR1, phospholipase C, protein kinase C alpha, c-Src, NF-kappa B and p300 pathway*. Mol Immunol, 2008. **45**(6): p. 1587-99.
293. Paniagua, R.T., et al., *Selective tyrosine kinase inhibition by imatinib mesylate for the treatment of autoimmune arthritis*. J Clin Invest, 2006. **116**(10): p. 2633-42.
294. Yamane, S., et al., *Proinflammatory role of amphiregulin, an epidermal growth factor family member whose expression is augmented in rheumatoid arthritis patients*. J Inflamm (Lond), 2008. **5**: p. 5.
295. Carvalho, J.F., M. Blank, and Y. Shoenfeld, *Vascular endothelial growth factor (VEGF) in autoimmune diseases*. J Clin Immunol, 2007. **27**(3): p. 246-56.
296. Daouti, S., et al., *Development of comprehensive functional genomic screens to identify novel mediators of osteoarthritis*. Osteoarthritis Cartilage, 2005. **13**(6): p. 508-18.
297. Remmers, E.F., H. Sano, and R.L. Wilder, *Platelet-derived growth factors and heparin-binding (fibroblast) growth factors in the synovial tissue pathology of rheumatoid arthritis*. Semin Arthritis Rheum, 1991. **21**(3): p. 191-9.
298. Meyn, M.A., 3rd and T.E. Smithgall, *Small molecule inhibitors of Lck: the search for specificity within a kinase family*. Mini Rev Med Chem, 2008. **8**(6): p. 628-37.
299. Rivera, J. and A. Olivera, *Src family kinases and lipid mediators in control of allergic inflammation*. Immunol Rev, 2007. **217**: p. 255-68.
300. Lee, J.H., et al., *Mast cell-mediated allergic response is suppressed by Sophorae flos: inhibition of SRC-family kinase*. Exp Biol Med (Maywood), 2008. **233**(10): p. 1271-9.
301. Callera, G.E., et al., *c-Src-dependent nongenomic signaling responses to aldosterone are increased in vascular myocytes from spontaneously hypertensive rats*. Hypertension, 2005. **46**(4): p. 1032-8.
302. Metcalf, C.A., 3rd, et al., *Targeting protein kinases for bone disease: discovery and development of Src inhibitors*. Curr Pharm Des, 2002. **8**(23): p. 2049-75.



- 
303. Shakespeare, W.C., et al., *SAR of carbon-linked, 2-substituted purines: synthesis and characterization of AP23451 as a novel bone-targeted inhibitor of Src tyrosine kinase with in vivo anti-resorptive activity*. Chem Biol Drug Des, 2008. **71**(2): p. 97-105.
304. Tsuruno, S., S.Y. Kawaguchi, and T. Hirano, *Src-family protein tyrosine kinase negatively regulates cerebellar long-term depression*. Neurosci Res, 2008. **61**(3): p. 329-32.
305. Vidal, D., M. Thormann, and M. Pons, *A novel search engine for virtual screening of very large databases*. J. Chem. Inf. Model, 2006. **46**(2): p. 836-43.
306. Stiefl, N. and A. Zaliani, *A knowledge-based weighting approach to ligand-based virtual screening*. J. Chem. Inf. Model, 2006. **46**(2): p. 587-96.
307. Rella, M., et al., *Structure-based pharmacophore design and virtual screening for novel angiotensin converting enzyme 2 inhibitors*. J. Chem. Inf. Model, 2006. **46**(2): p. 708-16.
308. Hicklin, D.J. and L.M. Ellis, *Role of the vascular endothelial growth factor pathway in tumor growth and angiogenesis*. J Clin Oncol, 2005. **23**(5): p. 1011-27.
309. Zhong, H. and J.P. Bowen, *Molecular design and clinical development of VEGFR kinase inhibitors*. Curr Top Med Chem, 2007. **7**(14): p. 1379-93.
310. van Cruijssen, H., A. van der Veldt, and K. Hoekman, *Tyrosine kinase inhibitors of VEGF receptors: clinical issues and remaining questions*. Front Biosci, 2009. **14**: p. 2248-68.
311. Roodhart, J.M., et al., *The molecular basis of class side effects due to treatment with inhibitors of the VEGF/VEGFR pathway*. Curr Clin Pharmacol, 2008. **3**(2): p. 132-43.
312. Yu, H., et al., *The discovery of novel vascular endothelial growth factor receptor tyrosine kinases inhibitors: pharmacophore modeling, virtual screening and docking studies*. Chem Biol Drug Des, 2007. **69**(3): p. 204-11.
313. Cao, H., et al., *3D QSAR studies on a series of potent and high selective inhibitors for three kinases of RTK family*. J Mol Graph Model, 2007. **26**(1): p. 236-45.

- 
314. Sharma, B.K., et al., *A quantitative structure-activity relationship study of novel, potent, orally active, selective VEGFR-2 and PDGFRalpha tyrosine kinase inhibitors: derivatives of N-phenyl-N'-(4-(4-quinolyloxy)phenyl)urea as antitumor agents*. J Enzyme Inhib Med Chem, 2008. **23**(2): p. 168-73.
315. Du, J., et al., *Molecular modeling studies of vascular endothelial growth factor receptor tyrosine kinase inhibitors using QSAR and docking*. J Mol Graph Model, 2009. **27**(5): p. 642-54.
316. Dakshanamurthy, S., et al., *In-silico fragment-based identification of novel angiogenesis inhibitors*. Bioorg Med Chem Lett, 2007. **17**(16): p. 4551-6.
317. Vieth, M. and D.J. Cummins, *DoMCoSAR: a novel approach for establishing the docking mode that is consistent with the structure-activity relationship. Application to HIV-1 protease inhibitors and VEGF receptor tyrosine kinase inhibitors*. J Med Chem, 2000. **43**(16): p. 3020-32.
318. Usui, T., et al., *Discovery of indenopyrazoles as EGFR and VEGFR-2 tyrosine kinase inhibitors by in silico high-throughput screening*. Bioorg Med Chem Lett, 2008. **18**(1): p. 285-8.
319. Ruel, R., et al., *Discovery and preclinical studies of 5-isopropyl-6-(5-methyl-1,3,4-oxadiazol-2-yl)-N-(2-methyl-1H-pyrrolo[2,3-b]pyridin-5-yl)pyrrolo[2,1-f][1,2,4]triazin-4-amine (BMS-645737), an in vivo active potent VEGFR-2 inhibitor*. Bioorg Med Chem Lett, 2008. **18**(9): p. 2985-9.
320. Kiselyov, A.S., V.V. Semenov, and D. Milligan, *4-(Azolyphenyl)-phthalazin-1-amines: Novel inhibitors of VEGF receptors I and II*. Chem Biol Drug Des, 2006. **68**(6): p. 308-13.
321. Fraley, M.E., et al., *Discovery and evaluation of 3-(5-thien-3-ylpyridin-3-yl)-1H-indoles as a novel class of KDR kinase inhibitors*. Bioorg Med Chem Lett, 2003. **13**(18): p. 2973-6.
322. Kuo, G.H., et al., *Synthesis and structure-activity relationships of pyrazine-pyridine biheteroaryls as novel, potent, and selective vascular endothelial growth factor receptor-2 inhibitors*. J Med Chem, 2005. **48**(15): p. 4892-909.
323. Thompson, A.M., et al., *Synthesis and structure-activity relationships of soluble 7-substituted 3-(3,5-dimethoxyphenyl)-1,6-naphthyridin-2-amines and*

- 
- related ureas as dual inhibitors of the fibroblast growth factor receptor-1 and vascular endothelial growth factor receptor-2 tyrosine kinases.* J Med Chem, 2005. **48**(14): p. 4628-53.
324. Nakamura, H., et al., *Synthesis and biological evaluation of benzamides and benzamidines as selective inhibitors of VEGFR tyrosine kinases.* Bioorg Med Chem Lett, 2006. **16**(19): p. 5127-31.
325. Heyman, H.R., et al., *Thienopyridine urea inhibitors of KDR kinase.* Bioorg Med Chem Lett, 2007. **17**(5): p. 1246-9.
326. Peifer, C., et al., *Design, synthesis, and biological evaluation of novel 3-aryl-4-(1H-indole-3yl)-1,5-dihydro-2H-pyrrole-2-ones as vascular endothelial growth factor receptor (VEGF-R) inhibitors.* J Med Chem, 2008. **51**(13): p. 3814-24.
327. Hennequin, L.F., et al., *Design and structure-activity relationship of a new class of potent VEGF receptor tyrosine kinase inhibitors.* J Med Chem, 1999. **42**(26): p. 5369-89.
328. Hasegawa, M., et al., *Discovery of novel benzimidazoles as potent inhibitors of TIE-2 and VEGFR-2 tyrosine kinase receptors.* J Med Chem, 2007. **50**(18): p. 4453-70.
329. Ji, Z., et al., *Isothiazolopyrimidines and isoxazolopyrimidines as novel multi-targeted inhibitors of receptor tyrosine kinases.* Bioorg Med Chem Lett, 2006. **16**(16): p. 4326-30.
330. Oguro, Y., et al., *N-phenyl-N'-[4-(5H-pyrrolo[3,2-d]pyrimidin-4-yloxy)phenyl]ureas as novel inhibitors of VEGFR and FGFR kinases.* Bioorg Med Chem, 2010. **18**(20): p. 7150-63.
331. Raepfel, S., et al., *Identification of a novel series of potent RON receptor tyrosine kinase inhibitors.* Bioorg Med Chem Lett, 2010. **20**(9): p. 2745-9.
332. Saavedra, O., et al., *N3-arylmalonamides: a new series of thieno[3,2-b]pyridine based inhibitors of c-Met and VEGFR2 tyrosine kinases.* Bioorg Med Chem Lett, 2009. **19**(24): p. 6836-9.
333. Renhowe, P.A., et al., *Design, structure-activity relationships and in vivo characterization of 4-amino-3-benzimidazol-2-ylhydroquinolin-2-ones: a*

- 
- novel class of receptor tyrosine kinase inhibitors.* J Med Chem, 2009. **52**(2): p. 278-92.
334. Raepfel, S., et al., *N-(3-fluoro-4-(2-arylthieno[3,2-b]pyridin-7-yloxy)phenyl)-2-oxo-3-phenylimidazolidine-1-carboxamides: a novel series of dual c-Met/VEGFR2 receptor tyrosine kinase inhibitors.* Bioorg Med Chem Lett, 2009. **19**(5): p. 1323-8.
335. Hasegawa, M., et al., *Discovery of novel benzimidazoles as potent inhibitors of TIE-2 and VEGFR-2 tyrosine kinase receptors.* Journal of medicinal chemistry, 2007. **50**(18): p. 4453-4470.
336. Ji, Z., et al., *Isothiazolopyrimidines and isoxazolopyrimidines as novel multi-targeted inhibitors of receptor tyrosine kinases.* Bioorganic & Medicinal Chemistry Letters, 2006. **16**(16): p. 4326-4330.
337. Hennequin, L.F., et al., *Design and structure-activity relationship of a new class of potent VEGF receptor tyrosine kinase inhibitors.* Journal of medicinal chemistry, 1999. **42**(26): p. 5369-5389.
338. Fraley, M.E., et al., *Discovery and evaluation of 3-(5-Thien-3-ylpyridin-3-yl)-1*H*-indoles as a novel class of KDR kinase inhibitors.* Bioorganic & Medicinal Chemistry Letters, 2003. **13**(18): p. 2973-2976.
339. Duffey, M.O., et al., *Discovery of a potent and orally bioavailable benzolactam-derived inhibitor of Polo-like kinase 1 (MLN0905).* J Med Chem, 2012. **55**(1): p. 197-208.
340. Yu, B., et al., *Design, synthesis and antitumor activity of 4-aminoquinazoline derivatives targeting VEGFR-2 tyrosine kinase.* Bioorg Med Chem Lett, 2012. **22**(1): p. 110-4.
341. Heidary, D.K., et al., *VX-322: a novel dual receptor tyrosine kinase inhibitor for the treatment of acute myelogenous leukemia.* J Med Chem, 2012. **55**(2): p. 725-34.
342. Zambon, A., et al., *Small molecule inhibitors of BRAF in clinical trials.* Bioorg Med Chem Lett, 2012. **22**(2): p. 789-92.

- 
343. Rizvi, S.U., et al., *Discovery and molecular docking of quinolyl-thienyl chalcones as anti-angiogenic agents targeting VEGFR-2 tyrosine kinase*. *Bioorg Med Chem Lett*, 2012. **22**(2): p. 942-4.
344. Gangjee, A., et al., *N(4)-Aryl-6-substitutedphenylmethyl-7H-pyrrolo[2,3-d]pyrimidine-2,4-diamines as receptor tyrosine kinase inhibitors*. *Bioorg Med Chem*, 2012. **20**(2): p. 910-4.
345. Gangjee, A., et al., *N(4)-(3-Bromophenyl)-7-(substituted benzyl) pyrrolo[2,3-d]pyrimidines as potent multiple receptor tyrosine kinase inhibitors: design, synthesis, and in vivo evaluation*. *Bioorg Med Chem*, 2012. **20**(7): p. 2444-54.
346. David, M., L. Friedrich, and H. Kurt, *The support vector machine under test*. *Neurocomputing*, 2003. **55**(1-2): p. 169-186.
347. Enomoto, H., et al., *Vascular endothelial growth factor isoforms and their receptors are expressed in human osteoarthritic cartilage*. *Am J Pathol*, 2003. **162**(1): p. 171-81.
348. Paavonen, K., et al., *Vascular endothelial growth factors C and D and their VEGFR-2 and 3 receptors in blood and lymphatic vessels in healthy and arthritic synovium*. *J Rheumatol*, 2002. **29**(1): p. 39-45.
349. Grosios, K., et al., *Angiogenesis inhibition by the novel VEGF receptor tyrosine kinase inhibitor, PTK787/ZK222584, causes significant anti-arthritic effects in models of rheumatoid arthritis*. *Inflamm Res*, 2004. **53**(4): p. 133-42.
350. Fernandez, M., et al., *Inhibition of VEGF receptor-2 decreases the development of hyperdynamic splanchnic circulation and portal-systemic collateral vessels in portal hypertensive rats*. *J Hepatol*, 2005. **43**(1): p. 98-103.
351. Yamamoto, A., et al., *Vascular endothelial growth factor receptor tyrosine kinase inhibitor PTK787/ZK 222584 inhibits both the induction and elicitation phases of contact hypersensitivity*. *J Dermatol*, 2007. **34**(7): p. 419-29.
352. Kahl, K.G., et al., *Angiogenic factors in patients with current major depressive disorder comorbid with borderline personality disorder*. *Psychoneuroendocrinology*, 2009. **34**(3): p. 353-7.

- 
353. Iga, J., et al., *Gene expression and association analysis of vascular endothelial growth factor in major depressive disorder*. *Prog Neuropsychopharmacol Biol Psychiatry*, 2007. **31**(3): p. 658-63.
354. Warner-Schmidt, J.L. and R.S. Duman, *VEGF as a potential target for therapeutic intervention in depression*. *Curr Opin Pharmacol*, 2008. **8**(1): p. 14-9.

---

## Appendices

**Appendix A:** The journal name list for MicrobPad database construction.

| <b>Journal Name</b>                            | <b>ISSN</b> |
|--|-------------|
| AMERICAN JOURNAL OF VETERINARY RESEARCH        | 0002-9645   |
| ANALYTICAL AND BIOANALYTICAL CHEMISTRY         | 1618-2642   |
| APPLIED AND ENVIRONMENTAL MICROBIOLOGY         | 0099-2240   |
| APPLIED MICROBIOLOGY AND BIOTECHNOLOGY         | 0175-7598   |
| ARCHIVES OF VIROLOGY                           | 0304-8608   |
| AVIAN PATHOLOGY                                | 0307-9457   |
| BIOTECHNIQUES                                  | 0736-6205   |
| BMC BIOINFORMATICS                             | 1471-2105   |
| BMC INFECTIOUS DISEASES                        | 1471-2334   |
| BMC MICROBIOLOGY                               | 1471-2180   |
| CANCER RESEARCH                                | 0008-5472   |
| CLINICAL CHEMISTRY                             | 0009-9147   |
| CURRENT MICROBIOLOGY                           | 0343-8651   |
| DIAGNOSTIC MICROBIOLOGY AND INFECTIOUS DISEASE | 0732-8893   |
| EPIDEMIOLOGY AND INFECTION                     | 0950-2688   |

|   |           |
|---|-----------|
| FEMS MICROBIOLOGY ECOLOGY                 | 0168-6496 |
| FEMS MICROBIOLOGY LETTERS                 | 0378-1097 |
| JOURNAL OF APPLIED MICROBIOLOGY           | 1364-5072 |
| JOURNAL OF FELINE MEDICINE AND SURGERY    | 1098-612X |
| JOURNAL OF MICROBIOLOGY                   | 1225-8873 |
| JOURNAL OF MICROBIOLOGY AND BIOTECHNOLOGY | 1017-7825 |
| JOURNAL OF MOLECULAR DIAGNOSTICS          | 1525-1578 |
| JOURNAL OF VIROLOGY                       | 0022-538X |
| JAPANESE JOURNAL OF INFECTIOUS DISEASES   | 1344-6304 |
| LETTERS IN APPLIED MICROBIOLOGY           | 0266-8254 |
| MICROBIAL PATHOGENESIS                    | 0882-4010 |
| MICROBIOLOGY AND IMMUNOLOGY               | 0385-5600 |
| MODERN PATHOLOGY                          | 0893-3952 |
| MOLECULAR BIOTECHNOLOGY                   | 1073-6085 |
| MOLECULAR AND CELLULAR PROBES             | 0890-8508 |
| NATURE                                    | 0028-0836 |
| New Microbiologica                        | 1121-7138 |
| NEW ZEALAND VETERINARY JOURNAL            | 0048-0169 |
| PLoS One                                  | 1932-6203 |



|   |           |
|---|-----------|
| PLoS Pathogens                            | 1553-7366 |
| RESEARCH IN VETERINARY SCIENCE            | 0034-5288 |
| THERIOGENOLOGY                            | 0093-691X |
| VETERINARY IMMUNOLOGY AND IMMUNOPATHOLOGY | 0165-2427 |
| VETERINARY JOURNAL                        | 1090-0233 |
| VETERINARY MICROBIOLOGY                   | 0378-1135 |
| VETERINARY RECORD                         | 0042-4900 |
| VETERINARY RESEARCH                       | 0928-4249 |

**Appendix B:** The journal name list for TTD database update

| <b>Journal Name</b>  | <b>ISSN</b> |
|--|-------------|
| ACTA PAEDIATRICA   | 0803-5253   |
| ADVANCES IN CANCER RESEARCH                                | 0065-230X   |
| ALLERGY AND ASTHMA PROCEEDINGS                             | 1088-5412   |
| AMERICAN JOURNAL OF PATHOLOGY                              | 0002-9440   |
| AMERICAN JOURNAL OF RESPIRATORY AND CRITICAL CARE MEDICINE | 1073-449X   |
| ANALYTICAL BIOCHEMISTRY                                    | 0003-2697   |
| ANESTHESIOLOGY   | 0003-3022   |
| ANNALS OF THE NEW YORK ACADEMY OF SCIENCES                 | 0077-8923   |

|   |           |
|---|-----------|
| ANNALS OF ONCOLOGY                                  | 0923-7534 |
| ANNALS OF THE RHEUMATIC DISEASES                    | 0003-4967 |
| ANNUAL REVIEW OF PHARMACOLOGY AND TOXICOLOGY        | 0362-1642 |
| ANTICANCER RESEARCH                                 | 0250-7005 |
| ANTIMICROBIAL AGENTS AND CHEMOTHERAPY               | 0066-4804 |
| ARCHIVES OF MICROBIOLOGY                            | 0302-8933 |
| ARCHIVES OF TOXICOLOGY                              | 0340-5761 |
| ARTHRITIS AND RHEUMATISM                            | 0004-3591 |
| BEHAVIORAL NEUROSCIENCE                             | 0735-7044 |
| BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS | 0006-291X |
| BIOCHEMICAL JOURNAL                                 | 0264-6021 |
| BIOCHEMICAL PHARMACOLOGY                            | 0006-2952 |
| BIOCHEMISTRY  | 0006-2960 |
| BIOLOGICAL CHEMISTRY                                | 1431-6730 |
| BIOLOGICAL PSYCHIATRY                               | 0006-3223 |
| BIOPHYSICAL JOURNAL                                 | 0006-3495 |
| BIOPOLYMERS   | 0006-3525 |
| BIORHEOLOGY   | 0006-355X |
| BLOOD   | 0006-4971 |

|                                 |           |
|---------------------------------|-----------|
| BMC CANCER                      | 1471-2407 |
| BRAIN RESEARCH                  | 0006-8993 |
| BRAIN RESEARCH BULLETIN         | 0361-9230 |
| Brain Tumor Pathology           | 1433-7398 |
| BRITISH JOURNAL OF CANCER       | 0007-0920 |
| BRITISH JOURNAL OF PHARMACOLOGY | 0007-1188 |
| BRITISH JOURNAL OF SURGERY      | 0007-1323 |
| CANADIAN JOURNAL OF CARDIOLOGY  | 0828-282X |
| CANCER LETTERS                  | 0304-3835 |
| CANCER RESEARCH                 | 0008-5472 |
| CANCER                          | 0008-543X |
| CARCINOGENESIS                  | 0143-3334 |
| CELL DEATH AND DIFFERENTIATION  | 1350-9047 |
| CHEMBIOCHEM                     | 1439-4227 |
| CHEMICO-BIOLOGICAL INTERACTIONS | 0009-2797 |
| CIRCULATION                     | 0009-7322 |
| CLINICAL CANCER RESEARCH        | 1078-0432 |
| CLINICAL CARDIOLOGY             | 0160-9289 |
| CLINICAL PHARMACOKINETICS       | 0312-5963 |

|   |           |
|---|-----------|
| CRITICAL CARE MEDICINE                                | 0090-3493 |
| CURRENT OPINION IN CARDIOLOGY                         | 0268-4705 |
| CURRENT OPINION IN CHEMICAL BIOLOGY                   | 1367-5931 |
| CURRENT OPINION IN HEMATOLOGY                         | 1065-6251 |
| CURRENT OPINION IN NEPHROLOGY AND HYPERTENSION        | 1062-4821 |
| CURRENT OPINION IN RHEUMATOLOGY                       | 1040-8711 |
| CURRENT PHARMACEUTICAL DESIGN                         | 1381-6128 |
| DIABETES  | 0012-1797 |
| DIGESTIVE DISEASES AND SCIENCES                       | 0163-2116 |
| DRUGS   | 0012-6667 |
| EMBO JOURNAL  | 0261-4189 |
| ENDOCRINOLOGY AND METABOLISM CLINICS OF NORTH AMERICA | 0889-8529 |
| ENDOCRINOLOGY   | 0013-7227 |
| ESSAYS IN BIOCHEMISTRY                                | 0071-1365 |
| EUROPEAN JOURNAL OF CANCER                            | 0959-8049 |
| EUROPEAN JOURNAL OF PHARMACOLOGY                      | 0014-2999 |
| EUROPEAN NEUROPSYCHOPHARMACOLOGY                      | 0924-977X |
| FASEB JOURNAL   | 0892-6638 |
| FEBS LETTERS  | 0014-5793 |

|   |           |
|---|-----------|
| GUT   | 0017-5749 |
| HISTOCHEMISTRY AND CELL BIOLOGY                 | 0948-6143 |
| HORMONE AND METABOLIC RESEARCH                  | 0018-5043 |
| HUMAN MOLECULAR GENETICS                        | 0964-6906 |
| IDRUGS  | 1369-7056 |
| IMMUNOLOGY LETTERS                              | 0165-2478 |
| INFECTION AND IMMUNITY                          | 0019-9567 |
| INTERNATIONAL JOURNAL OF CANCER                 | 0020-7136 |
| INTERNATIONAL JOURNAL OF EXPERIMENTAL PATHOLOGY | 0959-9673 |
| INTERNATIONAL JOURNAL OF IMPOTENCE RESEARCH     | 0955-9930 |
| INTERNATIONAL JOURNAL OF PHARMACEUTICS          | 0378-5173 |
| JOURNAL OF ALLERGY AND CLINICAL IMMUNOLOGY      | 0091-6749 |
| JOURNAL OF ANTIBIOTICS                          | 0021-8820 |
| JOURNAL OF ANTIMICROBIAL CHEMOTHERAPY           | 0305-7453 |
| JOURNAL OF BACTERIOLOGY                         | 0021-9193 |
| JOURNAL OF BIOLOGICAL CHEMISTRY                 | 0021-9258 |
| JOURNAL OF CEREBRAL BLOOD FLOW AND METABOLISM   | 0271-678X |
| JOURNAL OF CLINICAL INVESTIGATION               | 0021-9738 |
| JOURNAL OF CLINICAL ONCOLOGY                    | 0732-183X |

|  |           |
|--|-----------|
| JOURNAL OF EXPERIMENTAL MEDICINE           | 0022-1007 |
| JOURNAL OF GASTROENTEROLOGY AND HEPATOLOGY | 0815-9319 |
| JOURNAL OF GENERAL VIROLOGY                | 0022-1317 |
| JOURNAL OF IMMUNOLOGY                      | 0022-1767 |
| JOURNAL OF IMMUNOTHERAPY                   | 1524-9557 |
| JOURNAL OF INORGANIC BIOCHEMISTRY          | 0162-0134 |
| JOURNAL OF LEUKOCYTE BIOLOGY               | 0741-5400 |
| JOURNAL OF LIPID RESEARCH                  | 0022-2275 |
| JOURNAL OF MEDICINAL CHEMISTRY             | 0022-2623 |
| JOURNAL OF MOLECULAR BIOLOGY               | 0022-2836 |
| JOURNAL OF NATURAL PRODUCTS                | 0163-3864 |
| JOURNAL OF THE NATIONAL CANCER INSTITUTE   | 0027-8874 |
| JOURNAL OF NEUROCHEMISTRY                  | 0022-3042 |
| JOURNAL OF NEUROIMMUNOLOGY                 | 0165-5728 |
| JOURNAL OF NEURO-ONCOLOGY                  | 0167-594X |
| JOURNAL OF NEUROPHYSIOLOGY                 | 0022-3077 |
| JOURNAL OF NEUROSCIENCE RESEARCH           | 0360-4012 |
| JOURNAL OF NUTRITION                       | 0022-3166 |
| JOURNAL OF ORGANIC CHEMISTRY               | 0022-3263 |

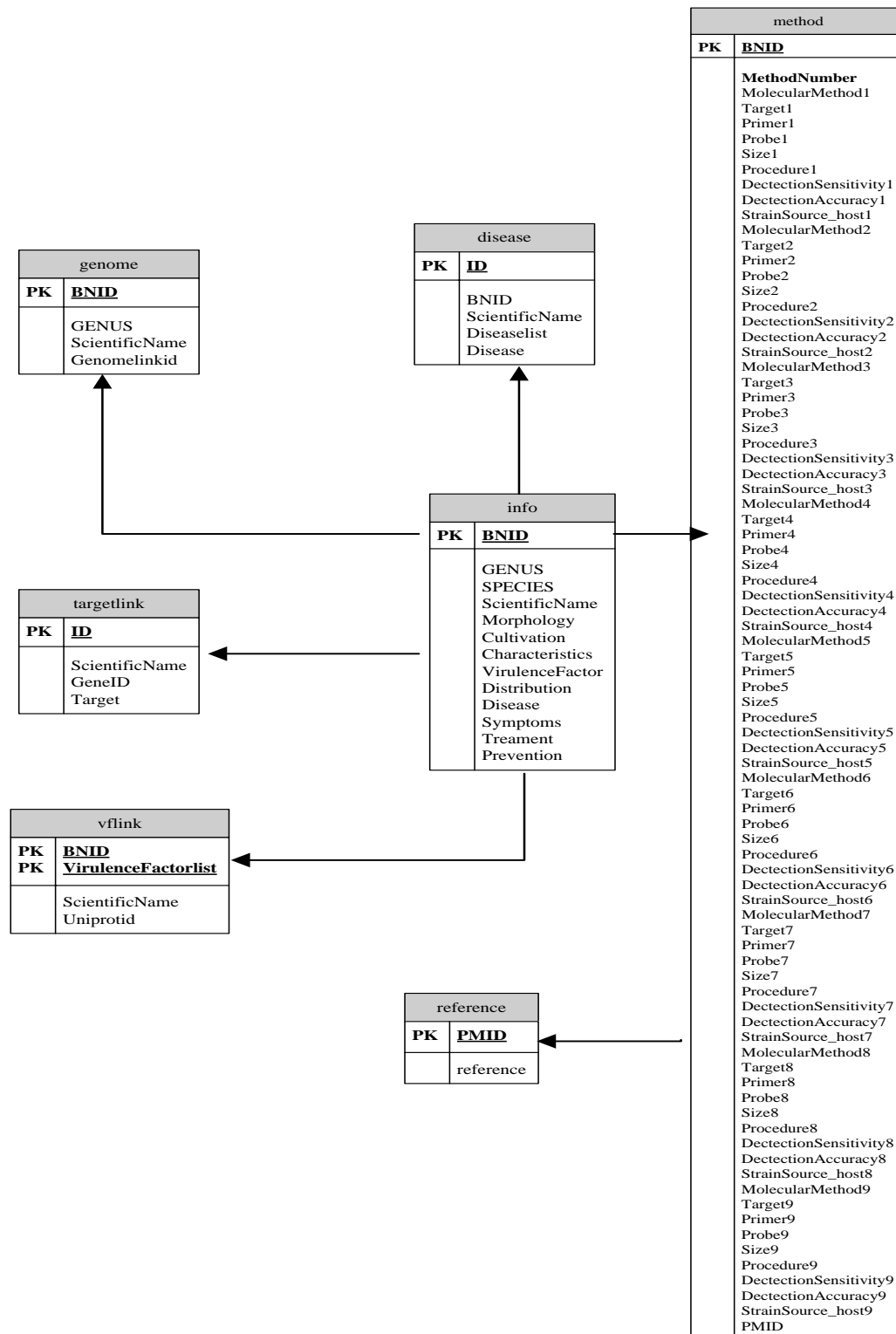
|   |           |
|---|-----------|
| JOURNAL OF PHARMACY AND PHARMACOLOGY                  | 0022-3573 |
| JOURNAL OF PHARMACEUTICAL AND BIOMEDICAL ANALYSIS     | 0731-7085 |
| JOURNAL OF PHARMACOLOGY AND EXPERIMENTAL THERAPEUTICS | 0022-3565 |
| JOURNAL OF REPRODUCTIVE MEDICINE                      | 0024-7758 |
| JOURNAL OF RHEUMATOLOGY                               | 0315-162X |
| JOURNAL OF SURGICAL RESEARCH                          | 0022-4804 |
| JOURNAL OF UROLOGY                                    | 0022-5347 |
| JOURNAL OF VIROLOGY                                   | 0022-538X |
| KIDNEY INTERNATIONAL                                  | 0085-2538 |
| LABORATORY INVESTIGATION                              | 0023-6837 |
| LANCET  | 0140-6736 |
| LEUKEMIA  | 0887-6924 |
| LIFE SCIENCES   | 0024-3205 |
| LUNG CANCER   | 0169-5002 |
| MEDICAL MYCOLOGY                                      | 1369-3786 |
| MEMORIAS DO INSTITUTO OSWALDO CRUZ                    | 0074-0276 |
| MOLECULAR AND BIOCHEMICAL PARASITOLOGY                | 0166-6851 |
| MOLECULAR AND CELLULAR BIOLOGY                        | 0270-7306 |
| MOLECULAR ENDOCRINOLOGY                               | 0888-8809 |

|   |           |
|---|-----------|
| MOLECULAR PHARMACOLOGY                        | 0026-895X |
| MOLECULAR PSYCHIATRY                          | 1359-4184 |
| MOUNT SINAI JOURNAL OF MEDICINE               | 0027-2507 |
| NATURE MEDICINE                               | 1078-8956 |
| NATURE  | 0028-0836 |
| NEUROCHEMICAL RESEARCH                        | 0364-3190 |
| NEUROPHARMACOLOGY                             | 0028-3908 |
| NEUROPSYCHOPHARMACOLOGY                       | 0893-133X |
| NEUROREPORT                                   | 0959-4965 |
| NEUROSCIENCE LETTERS                          | 0304-3940 |
| NEW ENGLAND JOURNAL OF MEDICINE               | 0028-4793 |
| NAUNYN-SCHMIEDEBERGS ARCHIVES OF PHARMACOLOGY | 0028-1298 |
| ONCOGENE                                      | 0950-9232 |
| ONCOLOGIST                                    | 1083-7159 |
| PROGRESS IN LIPID RESEARCH                    | 0163-7827 |
| PROTEOMICS                                    | 1615-9853 |
| PSYCHOPHARMACOLOGY                            | 0033-3158 |
| RHEUMATOLOGY                                  | 1462-0324 |
| SCIENCE                                       | 0036-8075 |



|                                       |           |
|---------------------------------------|-----------|
| SEMINARS IN THROMBOSIS AND HEMOSTASIS | 0094-6176 |
| STEM CELLS                            | 1066-5099 |
| STRUCTURE                             | 0969-2126 |
| SURGERY                               | 0039-6060 |
| TRENDS IN CARDIOVASCULAR MEDICINE     | 1050-1738 |
| TRENDS IN NEUROSCIENCES               | 0166-2236 |
| TRENDS IN PHARMACOLOGICAL SCIENCES    | 0165-6147 |
| VIROLOGY                              | 0042-6822 |

Appendix C: Schema of MicrobPad database.



---

## List of publication

### A. Publication relating to research work from the current thesis

1. **B.C. Han**, X.H. Ma, R. Y. Zhao, J.X. Zhang, X.N. Wei, X.H. Liu, X. Liu, C.L. Zhang, C.Y. Tan, and Y.Y. Jiang, Y. Z. Chen. Development and experimental test of support vector machines virtual screening method for searching Src inhibitors from large compound libraries. *Chem Cent J*:6:139 (2012). doi:10.1186/1752-153X-6-139
2. **B.C. Han**, X.N. Wei, J.X. Zhang, N.Q.T. Truong, C.L. Westgate, R.Y. Zhao, Y.Z. Chen. MicrobPad MD: Microbial pathogen diagnostic methods database. *Infect. Genet. Evol.* 13:261–266 (2012). doi: 10.1016/j.meegid.2012.10.017
3. **B.C. Han** , X.H. Ma , R. Y. Zhao, Z. Shi, C.L. Zhang, C.Y. Tan, and Y. Z. Chen, Y.Y. Jiang. Development and Experimental Test of a Support Vector Machines Virtual Screening Model for Searching VEGFR-2 Inhibitors from Large Compound Libraries. (submitted)
4. 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\*. Update of TTD: Therapeutic Target Database. *Nucleic Acids Res.* 38:D787-91(2010).

### B. Publication from other projects not include in the current thesis

5. Zhang JX, **Han BC**, Wei XN, C.Y. Tan, Y.Y. Jiang, Chen YZ. A two-step Target Binding and Selectivity Support Vector Machines Approach for Virtual Screening of Dopamine Receptor Subtype-selective Ligands. *PLoS ONE* 7(6): e39076. doi:10.1371/journal.pone.0039076 (2012).

- 
6. Zhang JX, J Jia, Ma XH, **Han BC**, Wei XN, C.Y. Tan, Y.Y. Jiang, Chen YZ. Analysis of bypass signaling in EGFR pathway and profiling of bypass genes for predicting response to anticancer EGFR tyrosine kinase inhibitors. *Mol. BioSyst.*, Advance Article, DOI: 10.1039/C2MB25165E. (2012)
  7. F. Zhu, Z. Shi, C. Qin, L. Tao, X. Liu, F. Xu, L. Zhang, Y. Song, X.H. Liu, J.X. Zhang, **B.C. Han**, P. Zhang and Y.Z. Chen\*. Therapeutic Target Database Update 2012: A Resource for Facilitating Target-Oriented Drug Discovery. *Nucleic Acids Res.* *Nucleic Acids Res.* 40(D1):D1128-D1136 (2012).
  8. Wei XN, **Han BC**, Zhang JX, Liu XH, Tan CY, Jiang YY, Low BC, Tidor B, Chen YZ\*. An Integrated Mathematical Model of Thrombin-, Histamine- and VEGF-Mediated Signalling in Endothelial Permeability. *BMC Syst Biol.* Jul 15;5(1):112 (2011).
  9. Pankaj Kumar, X.H. Ma, X.H. Liu, J. Jia, **B.C. Han**, Y. Xue, Z.R. Li, S.Y. Yang, Y.C. Wei and Y.Z. Chen\*. Effect of Training Data Size and Noise Level on Support Vector Machines Virtual Screening of Genotoxic Agents from Large Compound Libraries. *J Comput Aided Mol Des.* 25(5):455-67 (2011)
  10. X.H. Liu, H.Y. Song, J.X. Zhang, **B.C. Han**, X.N. Wei, X.H. Ma, W.K. Chui, Y.Z. Chen\*. Identifying Novel Type ZBGs and Non-hydroxamate HDAC Inhibitors Through a SVM Based Virtual Screening Approach. *Mol Inf.* 29(5): 407-20(2010)
  11. Xiaoxia Liu, Jingxian Zhang, Feng Ni, Xu Dong, **Bucong Han**, Daxiong Han, Zhiliang Ji\* and Yufen Zhao\*. Genome wide exploration of the origin and evolution of amino acids. *BMC Evol Biol.* 2010 Mar 15;10:77

- 
12. P. Kumar, **B.C. Han**, Z. Shi, J. Jia, Y.P. Wang, Y.T. Zhang, L. Liang, Z.L. Ji and Y. Z. Chen\*. Update of KDBI: Kinetic Data of Bio-molecular Interaction Database. *Nucleic Acids Res.* 37: D636-41(2009).