

**THERAPEUTIC TARGET ANALYSIS AND DISCOVERY BASED
ON GENETIC, STRUCTURAL, PHYSICOCHEMICAL AND
SYSTEM PROFILES OF SUCCESSFUL TARGETS**

ZHU FENG

(B.Sc. & M.Sc., Beijing Normal University)

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

2010

Acknowledgements

Many people contributed to this dissertation in various ways, and it is my best pleasure to thank them who made this thesis possible.

First and foremost, I would like to present my sincere gratitude to my supervisor, Prof. Chen Yu Zong, for his invaluable guidance on my projects and respectable generosity with his time and energy. His inspiration, enthusiasm and great efforts formed the strongest support to my four years' adventure in bioinformatics. Moreover, He also provided me with encouragement not only for the research project but also for my job-hunting. Again, I would like to express my utmost appreciation, and give my best wishes to him and to his loving family.

I am delighted to interact with Prof. Martti T. Tammi by having him as my co-supervisor. His insights and knowledge always gave me new ideas during our discussion. The most wonderful thing was his innate sense of humor which made every meeting a pleasant journey. Great thanks also go to Prof. YAP Chun Wei, who devoted his time as my Qualifying Examination examiner, wrote recommendation letters for me, and most importantly gave many valuable comments on my research. I would also like to thank Prof. Low Boon Chuan, Prof. Yang Dai Wen and Prof. Tan Tin Wee for their great support and encouragement.

Prof. Chen Xin, Dr. Han Lian Yi, Dr. Zheng Chan Juan and Mr. Xie Bin deserve special thanks as they are pioneers who built up the foundation for target prediction. All results obtained in this thesis are directly or indirectly related to their excellent works on this branch of bioinformatics. It is reasonable to say, without their prior efforts, it would be

really hard for me to obtain results demonstrated in this thesis. Moreover, I also want to present my great thanks to Dr. Lin Hong Huang and his wife Dr. Zhang Hai Lei. Dr. Lin was my guide when I was first in BIDD. Through our collaboration, I learned a lot from his knowledge and research attitude. In my job-hunting, he also gave me valuable advice and help. Best appreciation also goes to former BIDD group members: Ms. Jiang Li, Prof. Li Ze Rong, Dr. Wang Rong, Dr. Cui Juan, Dr. Tang Zhi Qun, Dr. Li Hu, Dr. Ung Choong Yong and Dr. Pankaj Kumar. We shared lots of precious experience and happy time in Singapore, which will be an invaluable treasure for my whole life.

Present BIDD members are the direct sources of my courage and capacity in the past four years, who deserve my most sincere appreciation. I am very grateful to Dr. Liu Xiang Hui for our pleasant collaboration on both TTD and IDAD projects, in which he tried his best to enrich and validate the information even when he was rushing on his thesis. Dr. Jia Jia and Dr. Ma Xiao Hua were enrolled in NUS at the same time as I was. Although I was new to bioinformatics, Jia Jia and Xiao Hua did not hesitate to help me on my project and encouraged me when I was in bad mood. Since all of them has started new career or will leave BIDD soon, I would like to take this chance to thank them, and give my best wishes to their new stage of life and future career. Ms. Liu Xin and Ms. Shi Zhe are two best “Shi Mei” I have ever met, I am really happy that we can have pleasant cooperation experience and good personal friendship. Many thanks also go to Mr. Tao Lin for our friendship, his good temper and his knowledge on gardening, and special appreciation goes to our lovely Shi Mei Ms. Qin Chu who is not only the best collaborator of my research work but also an excellent leader and friend of all our out-door activities. Appreciation also goes to Mr. Zhang Jing Xian, Ms. Huang Lu, Ms. Wei Xiao Na, Mr.

Han Bu Cong, and Mr. Zhang Cheng. Thanks for their time and energy on our collaborative projects, and I think with their intelligence and hard work they will win a lot in their Ph.D. studies.

My most sincere appreciation will never miss my loving friends. This thesis is dedicated to Mr. Zheng Zhong, Ms. Gu Han Lu, and most importantly their cute daughter for their understanding, support, and everything. Ms. Sit Wing Yee, Mr. Tu Wei Min, Mr. Li Nan, Mr. Guo Yang Fan, and Mr. Dong Xuan Chun are my close friends, and our gatherings nearly every week in Boon Lay and Bukit Batok are my most happy and relaxing time in Singapore. Thanks guys! Great appreciation also goes to Mr. Xie Chao, Ms. Hu Yong Li, Mr. Mohammad Asif Khan and Ms. Lim Shen Jean who are my TA partners and give me many supports. I would like to thank Ms. Wang Zhong Li for her support in the past one year. I did enjoy a very happy time with her. Finally, I want to thank Mr. Jiang Jin Wu, Ms. Li Dan, Ms. Ma Wei Li, Ms. Ou Yang Min, Mr. Xu Yang, Ms. Zhang Fan, Ms. Zhang Yan, and Mr. Zhu Jia Ji for their warm support from China.

Last but most importantly, I wish to say “thank you” to my beloved parents, who bore me, raised me, taught me, and loved me. To them I dedicate this thesis.

Zhu Feng

Aug 8th, 2010. Early in the morning

S16, Level 8, Room 08-19, National University of Singapore, Singapore

Table of Contents

Acknowledgements	I
Table of Contents	IV
Summary	VII
List of Figures	IX
List of Tables	XII
List of Abbreviations	XIV
List of Publications	XVI
Chapter 1 Introduction.....	1
1.1 Overview of target discovery in pharmaceutical research.....	2
1.1.1 Drug and target discovery	2
1.1.2 Knowledge of target and target discovery	3
1.1.3 Target identification	4
1.1.4 Target validation.....	7
1.2 Knowledge of established therapeutic targets	10
1.2.1 A review of efforts on evaluating number of successful targets	10
1.2.2 Databases providing therapeutic targets information	12
1.3 Therapeutic target and druggable genome.....	15
1.3.1 Efforts devoted for exploring druggable genome	15
1.3.2 Gap between druggable protein and therapeutic targets.....	16
1.4 Introduction to the prediction of druggable proteins	18
1.4.1 Sequence similarity approach.....	18
1.4.2 Motif based approach	21
1.4.3 Structural analysis approach.....	23
1.4.4 Machine learning methods	25
1.5 Objective and outline of this thesis	28
1.5.1 Objective of this thesis	28
1.5.2 Outline of this thesis.....	29
Chapter 2 Methods used in this thesis	42

2.1 Development of pharmainformatics databases	43
2.1.1 Rational architecture design	43
2.1.2 Information mining for pharmainformatics databases.....	44
2.1.3 Data organization and database structure construction	45
2.2 Methodology for validating therapeutic targets.....	51
2.3 Computational methods for predicting druggable proteins	54
2.3.1 Physicochemical properties of drug targets identified by machine learning methods .	54
2.3.2 Method for analyzing sequence similarity between the drug-binding domain of a studied target and that of a successful target	69
2.3.3 Comparative study of structural fold of the drug-binding domains of studied and successful targets.....	70
2.3.4 Simple system-level druggability rules	71
Chapter 3 Pharmainformatics databases construction	84
3.1 Therapeutic targets database, 2010 update	85
3.1.1 Target and drug data collection and access	86
3.1.2 Ways to access therapeutic targets database.....	88
3.1.3 Target and drug similarity searching	90
3.2 Information of Drug Activity Data.....	93
3.2.1 The data collection of IDAD information	93
3.2.2 The construction of IDAD database	94
3.2.3 Way to accession IDAD database	94
3.3 Therapeutic targets validation database.....	96
3.3.1 Pharmaceutical demands for target validation information.....	96
3.3.2 The data collection of TVD information	97
3.3.3 Explanation on target validation data	98
Chapter 4 Therapeutic targets in clinical trials.....	112
4.1 Trends in the exploration of clinical trial targets.....	113
4.2 Comparison of the characteristics of clinical trial targets with successful targets	117
4.3 The characteristics of clinical trial drugs with respect to approved drugs and drug leads	120

<u>Therapeutic targets analysis and discovery</u>	VI
4.4 Perspectives	123
Chapter 5 Identification of next generation innovative therapeutic targets: an application to clinical trial targets	138
5.1 Summary on materials and methods applied for drug target identification.....	140
5.1.1 Target classification based on characteristics of successful targets detected by a machine learning method	140
5.1.2 Sequence similarity analysis between drug-binding domain of studied target and that of successful target	141
5.1.3 Structural comparison between drug-binding domain of studied target and that of successful target	142
5.1.4 Computation of number of human similarity proteins, number of affiliated human pathways, and number of human tissues of a target	143
5.2 Target identification by collective analysis of sequence, structural, physicochemical, and system profiles of successful targets	144
5.3 Performance of target identification on clinical trial, non-clinical trial, difficult, and non-promising targets	146
Chapter 6 Identification of promising therapeutic targets from influenza genomes	182
6.1 Summary on methods applied for target identification	184
6.2 Target identification results from influenza genomes	185
6.3 Discussion on target identification results	187
Chapter 7 Concluding remarks	196
7.1 Major findings and contributions	196
7.1.1 Merits of TTD in facilitating target discovery.....	196
7.1.2 Merits of collective decision made by four <i>in silico</i> systems in target identification from clinical trial targets	197
7.1.3 Merits of collective decision made by four <i>in silico</i> systems in target identification from influenza genome.....	199
7.2 Limitations and suggestions for future studies	199
Bibliography	202

Summary

Knowledge from established therapeutic targets is expected to be invaluable goldmine for target discovery. To facilitate access to target information, publicly accessible databases have been developed. Information about the primary drug target(s) of comprehensive sets of approved, clinical trial, and experimental drugs is highly useful for facilitating focused investigation and discovery effort. However, none of those databases can accurately provide such data. Thus, a significant update to the Therapeutic Targets Database (TTD) in 2010 was conducted by expanding 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 target(s).

Comprehensive analysis on successful and clinical trial targets is able to reveal their common features. As found, analysis of therapeutic, biochemical, physicochemical, and systems features of clinical trial targets and drugs reveal areas of focuses, progresses and distinguished features. Many new targets, particularly G protein-coupled receptors (GPCRs) and kinases in the upstream signaling pathways are in advanced trial phases against cancer, inflammation, and nervous and circulatory systems diseases. The majority of the clinical trial targets show sequence and system profiles similar to successful targets, but fewer of them show overall sequence, structure, physicochemical, and system features resembling successful ones. Drugs in advanced trial phase show improved potency but increased lipophilicity and molecular weight with respect to approved drugs, and improved potency and lipophilicity but increased molecular weight compared to high throughput screening (HTS) leads. These suggest a need for further improvement in drug-like and target-like features.

Based on information from TTD and other sources, and statistical analysis results on successful and clinical trial targets, a collective approach combining 4 *in silico* methods to identify targets was proposed. These methods include (1) machine learning used for identifying physicochemical properties embedded in target primary structure; (2) sequence similarity in drug-binding domains; (3) 3-D structural fold of drug-binding domains; and (4) simple system level druggability rules. This combination identified 50%, 25%, 10% and 4% of the phase III, II, I, and non-clinical targets as promising, it enriched phase II and III target identification rate by 4.0~6.0 fold over random selection. The phase III targets identified include 7 of the 8 targets with positive phase III results.

Recent emergence of swine and avian influenza A H1N1 and H5N1 outbreaks and various drug-resistant influenza strains underscores the urgent need for developing new anti-influenza drugs. As an application, target discovery approach is used to identify promising targets from the genomes of influenza A (H1N1, H5N1, H2N2, H3N2, H9N2), B and C. The identified promising drug targets are neuraminidase of influenza A and B, polymerase of influenza A, B and C, and matrix protein 2 of influenza A. The identified marginally promising therapeutic targets are haemagglutinin of influenza A and B, and hemagglutinin-esterase of influenza C. The identified promising targets show fair drug discovery productivity level compared to a modest level for the marginally promising targets and low level for unpromising targets. Thus, the results are highly consistent with the current drug discovery productivity levels against these proteins.

List of Figures

Chapter 1

Figure 01- 1 Drug discovery process..... 32

Figure 01- 2 Number of new chemical entities in relation to R&D spending (1992-2006)..... 33

Figure 01- 3 Biochemical class for successful and clinical trial targets in TTD..... 33

Chapter 2

Figure 02- 1 The hierarchical data model..... 74

Figure 02- 2 The network data model 74

Figure 02- 3 The relational data model 75

Figure 02- 4 Logical view of the database 75

Figure 02- 5 Architecture of support vector machines 75

Figure 02- 6 Different hyper planes could be used to separate examples 76

Figure 02- 7 Mapping input space to feature space..... 76

Figure 02- 8 Diagrams of the process for training and predicting targets 77

Figure 02- 9 Illustration of derivation of the feature vector* 78

Chapter 3

Figure 03- 1 Screenshot of home page of TTD 2010 99

Figure 03- 2 Screenshot of customized search page of TTD 2010..... 100

Figure 03- 3 Screenshot of sequence similarity search page of TTD 2010..... 101

Figure 03- 4 Screenshot of drug tanimot similarity search page of TTD 2010..... 102

Figure 03- 5 Screenshot of full database download page of TTD 2010..... 103

Figure 03- 6 Intermediate search results of “dopamine receptor” listed by targets..... 104

Figure 03- 7 Intermediate search results of “influenza virus infection” listed by drugs 105

Figure 03- 8 TTD target main information page 106

Figure 03- 9 TTD drug main information page	107
---	-----

Chapter 4

Figure 04- 1 Top-10 PFAM protein families that contain high number of phase I (yellow), II (green), and III (orange) clinical trial targets along with the number of targets in each family..	129
---	-----

Figure 04- 2 Top-20 KEGG pathways that contain high number of phase I (yellow), II (green), and III (orange), and all clinical trial targets (brown) along with the number of targets in each pathway	129
---	-----

Figure 04- 3 Number of phase I (yellow), II (green), and III (orange) targets distributed in various sub-cellular locations.....	130
--	-----

Figure 04- 4 Top-10 Pfam protein families that contain high number of clinical trial (orange) and successful (red) targets along with the number of targets in each family	130
--	-----

Figure 04- 5 Top-10 clinical trial (orange) and successful (red) targets targeted by phase II clinical trial drugs	131
--	-----

Figure 04- 6 Top-10 clinical trial (orange) and successful (red) targets targeted by phase III clinical trial drugs	131
---	-----

Figure 04- 7 Top-10 clinical trial (orange) and successful (red) targets targeted by all clinical trial drugs	131
---	-----

Figure 04- 8 Distribution of all clinical trial targets (orange) and the innovative successful targets (approved by FDA from 1995 to 2008) (red) by crudely estimated target exploration time	132
---	-----

Figure 04- 9 Distribution of phase I (yellow), phase II (green), and phase III (orange) clinical trial targets by crudely estimated target exploration time	132
---	-----

Figure 04- 10 Distribution of phase I (yellow), phase II (green), and phase III (orange) clinical trial targets and discontinued clinical trial targets (blue) by level of similarity to successful targets*	132
--	-----

Figure 04- 11 Distribution of all clinical trial targets and successful targets with respect to the number of human similarity proteins outside the target family.....	133
--	-----

Figure 04- 12 Distribution of all clinical trial targets and successful targets with respect to the number of human pathways the target is associated with	133
--	-----

Figure 04- 13 Distribution of all clinical trial targets and successful targets with respect to the number of human tissues the target is distributed in	133
Figure 04- 14 Distribution of clinical trial drugs (orange) and approved drugs (red) by potency (IC ₅₀ , EC ₅₀ , Ki etc in units of nM).....	134
Figure 04- 15 Distribution of phase I (yellow), II (green), and III (orange) clinical trial drugs and discontinued clinical trial drugs (blue) by potency (IC ₅₀ , EC ₅₀ , Ki etc in units of nM)	134
Figure 04- 16 Distribution of clinical trial drugs (orange) and approved drugs (red) by molecular weight	135
Figure 04- 17 Distribution of phase I (yellow), II (green), and III (orange) clinical trial drugs by molecular weight	135
Figure 04- 18 Distribution of clinical trial drugs targeting novel clinical trial targets (green), clinical trial targets with protein subtype as successful target (brown), and successful targets (pink) by molecular weight	135
Figure 04- 19 Distribution of clinical trial drugs (orange) and approved drugs (red) by ALogP	136
Figure 04- 20 Distribution of phase I (yellow), II (green), and III (orange) clinical trial drugs and discontinued clinical trial drugs (blue) by ALogP.....	136
Figure 04- 21 Distribution of clinical trial drugs targeting novel clinical trial targets (green), clinical trial targets with protein subtype as successful target (brown), and successful targets (pink) by ALogP	136
Figure 04- 22 Percentage of phase I (yellow), II (green), III (orange) clinical trial drugs and approved drugs (red) obeying Lipinsky's rule of five (dark color), with one violation of rule of five (medium color) and the others (light color). The numbers in this figure refer to number of drugs.	137

List of Tables

Chapter 1

Table 01- 1 Examples of well-known gene expression database	34
Table 01- 2 Brief description, advantages and limitations of loss-of-function target validation technologies.....	36
Table 01- 3 Molecular targets of FDA-approved drugs from Overington's work	38
Table 01- 4 Examples of well-known drug target database	39

Chapter 2

Table 02- 1 Websites that contain freely downloadable codes of machine learning methods	79
Table 02- 2 Division of amino acids into 3 different groups by different physicochemical properties	80
Table 02- 3 List of features for proteins.....	81
Table 02- 4 Characteristic descriptors of cellular tumor antigen p53	82

Chapter 3

Table 03- 1 Main drug-binding databases available online.....	108
Table 03- 2 Potencies of drugs against their efficacy targets CDK2.....	109
Table 03- 3 Potencies of drugs against the disease relevant cell-lines expressing CDK2.....	110
Table 03- 4 Effects of target knock-out in CDK2 sequence, expression and activity in disease models and additional evidences	111

Chapter 4

Table 04- 1 Number of clinical trial targets in different disease classes*	126
Table 04- 2 Distribution of the phase III, II, and I targets that are similar or resemble the properties of successful targets in sequence (A), drug-binding domain structural fold (B), physicochemical features (C), and systems profiles (D)	127
Table 04- 3 Median potency, molecular weight, AlogP, the number of H-bond donor and H-bond acceptor, and the number of rotatable bond of approved, all clinical trial, phase , II and III drugs,	

and clinical trial drugs targeting novel clinical trial targets, clinical trial targets protein subtype as a successful target, and successful targets..... 128

Chapter 5

Table 05- 1 List of phase III targets identified by combinations of at least three of the methods A, B, C and D used in this study 150

Table 05- 2 List of phase II and phase I targets identified by combinations of at least three of the methods A, B, C and D used in this study 153

Table 05- 3 Statistics of promising targets selected from the 1,019 research targets by combinations of methods A, B, C and D, and clinical trial target enrichment factors 157

Table 05- 4 List of phase III targets dropped by combinations of at least three of the methods A, B, C and D used in this study 158

Table 05- 5 List of difficult targets currently discontinued in clinical trials and having no new drug entering clinical trials, and the prediction results..... 160

Table 05- 6 List of unpromising targets failed in HTS campaigns or found non-viable in knockout studies, and the prediction results..... 163

Table 05- 7 Definitions and structures (if available) of drugs and compounds in this chapter ... 166

Chapter 6

Table 06- 1 Target identification results for all encoded proteins in the genomes of the 5 subtypes of influenza A, B and C* 193

List of Abbreviations

ADMET	Absorption, Distribution, Metabolism, Excretion, Toxicity
AI	Artificial Intelligence
BLAST	Basic Local Alignment Search Tool
CLL	Chronic Lymphocytic Leukemia
DBMS	Database Management System
DDMS	Development of Database Management System
ENU	N-ethylnitrosouera
FDA	Food and Drug Administration
FN	False Negatives
FP	False Positives
GO	Gene Ontology
GPCR	G Protein-Coupled Receptor
HDAC	Histone Deacetylase
HMM	Hidden Markov Models
HMMER	Profile Hidden Markov Models
IDAD	Information of Drug Activity Data
MCC	Matthews Correlation Coefficient
NCE	New Chemical Entity
NHL	Non-Hodgkin's Lymphoma
NME	New Molecular Entity
NMR	Nuclear Magnetic Resonance
NSCLC	Non-Small Cell Lung Carcinoma
OODB	Object-Oriented Database
OOPPL	Object-Oriented Programming Language

OSH	Optimal Separating Hyper plane
PDB	Protein Data Bank
PSI-BLAST	Position Specific Iterative BLAST
Q	Overall accuracy
QP	Quadratic Programming
RBF	Radial Basis Function
RNAi	RNA interference
SAM	Sequence Alignment and Modeling
SE	Sensitivity
SP	Specificity
SVM	Support Vector Machine
TN	True Negatives
TP	True Positives
TTD	Therapeutic Targets Database
TVD	Target Validation Database
WHO	World Health Organization

List of Publications

1. **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 and Y.Z. Chen. Update of TTD: Therapeutic Target Database. *Nucleic Acids Res.* 38(Database issue):D787-91(2010).
2. **F. Zhu**, L.Y. Han, C.J. Zheng, B. Xie, M.T. Tammi, S.Y. Yang, Y.Q. Wei and Y.Z. Chen. What are next generation innovative therapeutic targets? Clues from genetic, structural, physicochemical and system profile of successful targets. *J Pharmacol Exp Ther.* 330(1):304-15(2009).
3. **F. Zhu**, L.Y. Han, X. Chen, H.H. Lin, S. Ong, B. Xie, H.L. Zhang and Y.Z. Chen. Homology-Free Prediction of Functional Class of Proteins and Peptides by Support Vector Machines. *Curr. Protein Pept. Sci.* 9:70-95 (2008).
4. **F. Zhu**, C.J. Zheng, L.Y. Han, B. Xie, J. Jia, X. Liu, M.T. Tammi, S.Y. Yang, Y.Q. Wei and Y.Z. Chen. Trends in the Exploration of Anticancer Targets and Strategies in Enhancing the Efficacy of Drug Targeting. *Curr Mol Pharmacol.* 1(3):213-232 (2008).
5. J. Jia, **F. Zhu**, X.H. Ma, Z.W. Cao, Y.X. Li and Y.Z. Chen. Mechanisms of drug combinations from interaction and network perspectives. *Nat. Rev. Drug Discov.* 8(2):111-28 (2009).
6. X.H. Ma, J. Jia, **F. Zhu**, Y. Xue, Z.R. Li and Y.Z. Chen. Comparative analysis of machine learning methods in ligand-based virtual screening of large compound libraries. *Comb. Chem. High Throughput Screen.* 12(4):344-357(2009).
7. R. Li, Y. Chen, L.B. Cui, **F. Zhu**, J. Zhou, D.H. Liu, S. Liu and X.S. Zhang. Effect of number of unit cells of FCC photonic crystal on property of band gaps. *Acta Physica Sinica.* 55(01):0188-04 (2006).

8. L.Y. Han, X.H. Ma, H.H. Lin, J. Jia, **F. Zhu**, Y. Xue, Z.R. Li, Z.W. Cao, Z.L. Ji and Y.Z. Chen. 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. Mod.* 26(8):1276-1286 (2008).
9. L.Y. Han, C.J. Zheng, B. Xie, J. Jia, X.H. Ma, **F. Zhu**, H.H. Lin, X. Chen, and Y.Z. Chen. Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness. *Drug Discov. Today.* 12(7-8): 304-313 (2007).
10. H.H Lin, L.Y. Han, C.W. Yap, Y. Xue, X.H. Liu, **F. Zhu** and Y.Z Chen. Prediction of Factor Xa Inhibitors by Machine Learning Methods. *J. Mol. Graph. Mod.* 26(2):505-518 (2007).

Chapter 1 Introduction

With the advent of post-genomic era, the pharmaceutical industry has been offered with unprecedented opportunities and challenges in drug, specifically target, discovery. On the one hand, the availability of human genome gives us chance to elucidate the genetic basis of human diseases by making overall evaluation on the druggability of all human proteins. On the other hand, huge amount of the genomic data requires the development of high-throughput analysis tools and powerful computational capacity to facilitate data process. In face of these challenges, bioinformatics has evolved many techniques to accelerate the target discovery, which are based on the detection of sequence and functional similarity to established drug targets, motif-based drug-binding domain family affiliation, structural analysis of geometric and energetic features, and statistic machine learning approaches. In **Chapter 1**, I intend to give the audience a brief introduction to these popular methods. In order to make my illustration clear, this chapter has been organized into 5 sections. In **Section 1.1**, an overview of target discovery in current pharmaceutical research is given, which reviews current technologies for both target identification and validation. **Section 1.2** includes a retrospective review of efforts to distinguish established drug targets, and a comprehensive analysis of available drug targets databases. Then, a repetitively exposed concept—“druggable genome” is discussed in **Section 1.3**, together with an explanation of the difference between “druggable protein” and “therapeutic target”. In **Section 1.4**, four bioinformatics methods frequently used in target discovery have been demonstrated. Both their advantages and limitations have been introduced. Finally, the objective and outline of this thesis are presented in the last section of this chapter (**Section 1.5**).

1.1 Overview of target discovery in pharmaceutical research

One of the most serious dilemmas encountered by current biopharmaceutical industry is that the output has not kept pace with the enormous increase in pharmaceutical R&D spending. As the very first step in drug development, target discovery is expected to play an important part in reducing cost and improving efficiency. In this part of my thesis, I intend to have a brief review on strategies currently employed for target discovery. After an overview of drug and target discovery in **Section 1.1.1** and **1.1.2**, I plan to introduce three popular techniques nowadays for identifying target in **Section 1.1.3**. In **Section 1.1.4**, three *in vivo* loss-of-function target validation technologies will be further illustrated. Based on these reviews, we can have some general understanding on the current target discovery process, which will not only provide background knowledge for the main topic of this thesis but also give us some hints on the reasons and strategies of our research conducted for facilitating target discovery.

1.1.1 Drug and target discovery

Drug discovery is a difficult, inefficient, lengthy, and expensive process. As illustrated in **Figure 01-1**, the process of a typical drug discovery involves disease selection, target identification and validation, hit and lead identification, lead optimization, preclinical trial evaluation, and clinical trials. Once a candidate has shown its value in these tests, it will be approved by medical authorities, like Food and Drug Administration (FDA), and then proceed to manufacturing and marketing¹. Despite advances in technology and accumulation of knowledge of biological systems, drug discovery is still time and money

consuming². Currently, the research and development cost for each new molecular entity (NME) is approximately US\$1.8 billion³, while the whole discovery process takes about 10-17 years with less than 10% overall probability of success^{2,4}. **Figure 01-2** shows the number of new chemical entities (NCEs) in relation to pharmaceutical R&D spending since 1992⁵. Therefore, how to increase the efficiency and reduce the cost and time of pharmaceutical research and development is the major task of modern drug discovery.

As the very early stage of drug discovery (**Figure 01-1**), selection and validation of novel molecular targets have become of paramount importance in light of the explosion in the number of new potential therapeutic targets that have emerged from human gene sequencing^{6,7}. Thousands of molecular targets have been cloned and are available as potential novel drug targets for further investigation^{8,9}. According to a brief search in the MEDLINE bibliographic database NCBI (<http://www.ncbi.nlm.nih.gov/pubmed>), a new potential therapeutic approach used for treating a known disease is proposed nearly every week, as a result of the exponential proliferation of novel therapeutic targets. Therefore, with thousands of potential targets available, target selection and validation has become one of the most critical components of drug discovery and will continue to be so in the future. In response to this revolution within the pharmaceutical industry, the development of high-throughput approaches for target discovery has been necessitated¹⁰.

1.1.2 Knowledge of target and target discovery

Before explaining the specific tools and technology used for facilitating modern target discovery, I would like to give a brief introduction first. As illustrated in **Figure 01-1**, the identification and validation of disease-causing target genes is an essential first step in

drug discovery and development. A drug target is typically a key molecule involved in certain metabolic or signaling pathway specific to a disease condition or pathology, or to the infectivity or survival of a microbial pathogen. Drugs are designed to bind onto the active region and inhibit this key molecule, or to enhance normal pathway by promoting specific molecules that may have been affected in the diseased state. In addition, these drugs should also be designed in such a way as not to affect any other important “off-target” that may be similar in appearance to the target molecule, since drug interactions with off-targets may lead to side effects^{11,12}. Target discovery, thus, involves a process to identify key “disease-causing” molecules which can be effectively inhibited or enhanced by their corresponding drugs.

In order to determine the disease-relevance of a therapeutic target to disease of interest and the effectiveness of target inhibition/enhancement by drugs, many key questions should be answered. What is the most popular technology used for determining disease-relevance? How to measure the binding activity of drugs on the targets? If we only know the drug and its corresponding disease, how can we identify its primary target? In **Section 1.1.3** and **1.1.4**, we attempt to answer these questions by illustrating target identification and validation in modern drug discovery.

1.1.3 Target identification

After choosing the disease of interest to study on, the next step is to identify a gene target or a mechanistic pathway which demonstrates correlations with the disease initiation and perpetuation. Target identification is to figure out disease-relevant genes and to uncover additional roles for genes of known functions. Many technologies now are available for

identifying targets, which include: expression profiling genomics, molecular genetics, and proteomics.

1.1.3.1 Expression profiling genomics

Molecular profiling has been proved as powerful tool for analyzing gene expression in disease and normal cells¹³⁻¹⁷. A good example is mRNA expression profiling using DNA microarray for large-scale analysis of cellular transcripts by comparing mRNA expression levels. By integrating knowledge of statistics and bioinformatics, gene expression data have been analyzed using clustering algorithms, and been used for detecting significant changes in gene expression levels.

With the collaborative efforts from researchers in both biology and bioinformatics, the number of gene expression databases and bioinformatics tools has been dramatically increased which offers us new *in silico* strategy to discover therapeutic targets^{13,16}. Numerous gene expression studies can be downloaded from public databases^{15,18-26}. **Table 01-1** lists examples of some well-known gene expression databases, which offer gold mines for target identification. However, one thing we need to keep in mind is that although the *in silico* detection of gene variants turns out to be very effective, it is subjected to the same limitations of all bioinformatics tools in that its results need further experimental validation to avoid false leads derived from noisy data.

Discovering drug targets by analyzing pathways has been proposed as another fruitful approach²⁷. Since pathways are known as genetic networks rather than individual genes, if researchers can identify them as being relevant to disease of interest, it is then possible

to assess the potential druggability of the individual proteins in that pathway¹⁷. Computational methods have been proposed together with mathematical models for gene networks²⁸. These computational methods are able to reflect potential pathway alterations based on the expression data²⁹. Thus, the analysis of pathways after gene knockout or drug treatment plays an important role in identifying target genes.

1.1.3.2 Molecular genetics

Molecular genetics is the field of biology that studies the structure and function of genes at molecular level, and it helps to understand genetic mutations which can cause certain disease. The major advantage of using molecular genetics instead of expression profiling genomics lies in that molecular genetics bridges the gap between genetic variation and disease phenotype³⁰.

One of the most extensively performed technologies available to molecular genetics is the forward genetic screen. The aim of this tool is to identify mutations that produce a certain phenotype. A mutagen *N*-ethylnitrosourea (ENU) is very often used to accelerate random mutations in the genome^{31,32}. For technologies used for forward genetic screen, RNA interference (RNAi) based loss-of-function genetic screen is the most frequently used³³.

Besides forward genetic screen, a more straightforward approach is to determine disease phenotype that results from mutating a given gene. This is called reverse genetics. In some organisms, like yeast and mice, it is possible to induce the deletion of a particular gene, creating a gene knockout. Gene knockout model enables not only the discovery of target function but also possible side effects that result from the affection of the target.

Several known human genes have already been identified with druggability by applying knockout studies^{34,35}.

1.1.3.3 Proteomics

Cellular signaling is coordinated by protein-protein interactions, posttranslational protein modifications, and enzymatic activities that cannot be fully described by mRNA levels. In the meantime, drug targets might be differentially expressed at the protein level that cannot be accurately predicted by mRNA expression either. Therefore, knowledge from protein level should be a necessary complementation to transcript analysis. Proteomics, the large-scale study of the proteins, is a promising technique for identifying novel drug targets³⁶. Among the proteomics techniques, 2D gel electrophoresis, multidimensional liquid chromatography, mass spectrometry, and protein microarray are currently available for drug target identification.

1.1.4 Target validation

Once a potential therapeutic target is identified, the next step is to validate its critical role in disease initiation or perpetuation. Most diseases originate from multiple factors which include acquired or inherited genetic predisposition and environmental causes³⁷⁻⁴². With the rapid accumulation of biological data and increasing understanding of disease mechanisms, the target validation process, however, has become more and more difficult, since many biological systems concerned have certain degrees of complexity⁴³. In other words, any modification on a certain part of the system is quite possible to trigger additional regulation of partners in both upstream and downstream, and consequently

induce effects onto other interconnected pathways. Generally, diagnosis of a disease is based on the occurrence of characteristic pathogenic consequences, which is usually after the initial triggering event. The use of *in vivo* models, therefore, enables investigation of whole-organism complexity. Due to the integration of symptom parameters with target efficacy and side effects evaluation, *in vivo* target validation is essential for providing the most relevant information for exploring effective therapeutics.

Currently, three *in vivo* loss-of-function target validation technologies are frequently used to specifically inactivate mammalian pathways or targets, which include: (1) DNA knockout validation models^{44,45}, (2) mRNA knockdown validation models⁴⁶, and (3) protein knockout models based on vaccination^{47,48}. These three technologies cover the three main biological levels: gene, mRNA and protein, and provide insight into the roles played by the targets in both normal and pathological circumstance.

Table 01-2 illustrates a brief description of these three loss-of-function target validation tools mentioned above, and illustrates their corresponding advantages and limitations. None of these three loss-of-function technologies is capable of answering all questions on complex biological systems. Animal models other than mice with similar biological systems to humans should be used whenever possible^{44,49}, but many of which suffer from absence of genetic models⁵⁰. In this circumstance, siRNA could be helpful as long as the target tissue is accessible via systemic or local delivery⁵⁰. In the meantime, a functional protein-KO could provide a very valuable tool for secreted or receptor target. Therefore, siRNA and protein functional KO technologies can overcome some limitations of gene knockout models. Furthermore, new delivery systems, vaccination and the modulation of

immune response will help expand potential application of these technologies. Nowadays, there is a strong need to combine these techniques because individual gene manipulation is proved to be not enough to understand a pathway and the complex regulation of each biological system involved in the disease⁵⁰.

In summary, drug discovery is a difficult and inefficient process. As the very early step in drug development, target discovery plays a critical role in reducing pharmaceutical R&D spending and improving efficiency for drug development. As we can see, target discovery aims at identifying and validating genes which can be effectively inhibited or enhanced by their corresponding drugs. In order to achieve this goal, many techniques have been applied. Three most popular target identification techniques are: (1) expression profiling genomics, (2) molecular genetics, and (3) proteomics, while three *in vivo* loss-of-function target validation technologies are: (1) DNA knockout validation models, (2) mRNA knockdown validation models, and (3) protein knockout models based on vaccination.

1.2 Knowledge of established therapeutic targets

In contrast to the heavy spending on pharmaceutical industry, there is a surprisingly lack of knowledge of the set of drug targets that modern therapeutics act on. For researchers who try to develop predictive model for identifying new promising molecular targets, the number, characteristics and biological profiles of targets of approved drugs are key data for them to work on. However, the total number of therapeutic targets with at least one drug approved, which we defined here as “successful targets”, has been debated.

1.2.1 A review of efforts on evaluating number of successful targets

In 1996, Drews and Reiser were the first to systematically analyze the existed pool of therapeutic targets, and identified 483 successful targets as “the most fruitful paths for therapeutic development in the past”^{51,52}. Moreover, they categorized these drug targets according to their therapeutic areas. Drug targets that affected synaptic and neuroeffector junction sites, as well as central nervous system drugs, accounted for almost 30% of the total. Almost half of the drug targets were divided more or less equally between drugs that address inflammation, renal and cardiovascular function, infectious disease, or hormone agonists and antagonists. The rest (26%) were targeted by drugs affecting blood diseases, gastrointestinal functions, uterine motility, cancer, immune-modulation, and by vitamins in the role of therapeutics.

Six years later, Hopkins and Groom challenged Drews’ conclusion by proposing “rule-of-five” constrain as new criteria for validating successful targets and suggested that of their set of 399 targets with known rule-of-five-compliant agent and binding affinities below

10 micromole, only 120 proteins had approved or marketed drug. According to comparison between these 120 launched targets and 399 targets with drug-like leads, their overall distributions by biochemical class were similar. For launched targets, enzymes constituted nearly half of them (47%), whereas GPCRs accounted for 30%. The remaining classes included ion channels and nuclear hormone receptors which accounted for less than a quarter of the identified launched targets⁸.

In 2003, Golden reported that all approved drugs acted through 273 proteins^{53,54}, while Wishart *et al.*⁵⁵ proposed 14,000 targets for all approved and experimental drugs. Later, Wishart *et al.* revised the number to 6,000 on the DrugBank database website. In 2006, Imming *et al.* catalogued 218 molecular targets for approved drug substances⁵⁶, whereas Zheng *et al.* disclosed 268 ‘successful’ targets in their 2006 version of the Therapeutic Targets Database^{57,58}.

In late 2006, Overington *et al.*⁵⁹ proposed a consensus number of 324 drug targets for all classes of approved therapeutic drugs (**Table 01-3**). Overington’s work reconciled earlier publications into a comprehensive survey. Analysis of protein family distribution revealed that the majority of (>50%) drugs target primarily on four families: class I GPCRs, nuclear receptors, ligand-gated ion channels and voltage-gated ion channels. The targets with the largest number of drugs were glucocorticoid receptor and histamine H1 receptor.

In 2010, we conducted a comprehensive survey on historical researches and latest reports to identify “successful targets” and its corresponding drugs⁹. In the latest version of Therapeutic Targets Database (TTD, 2010)⁹ (<http://bidd.nus.edu.sg/group/ttd/ttd.asp>), we

collected information of 348 successful, 293 clinical trial and 1254 research targets, 1514 approved, 1212 clinical trial and 2302 experimental drugs linked to their primary targets (3382 small molecule and 649 antisense drugs with available structure and sequence). Our data were consistent with previous report on the number of targets with drug approved. We had added a new category named “clinical trial targets” which referred to therapeutic targets with no drug approved but with drugs in clinical trial. According to the clinical trial stage of the drug, we had further defined targets as “phase III clinical trial targets”, “phase II clinical trial targets”, and “phase I clinical trial targets”. Distribution of successful and clinical trial targets with respect to biochemical classes was given in **Figure 01-3**. Biochemical classes included enzymes, receptors, nuclear receptors, channels and transporters, factors and regulators (factors, hormones, regulators, modulators, and receptor-binding proteins involved in a disease process), antigen and the remaining binding proteins not covered in other classes, structural proteins (non-receptor membrane proteins, adhesion molecules, envelop proteins, capsid proteins, motor proteins, and other structural protein), and nucleic acids. In **Chapter 3** of this thesis, I will illustrate the newly updated Therapeutic Targets Database (2010) in detail.

1.2.2 Databases providing therapeutic targets information

In light of the extensive efforts on exploring established and potential therapeutic targets, many databases have been constructed to provide target information for researchers from various directions, like biomedicine, pharmaceuticals, pharmacogenomics, comparative genomics, and so on.

Table 01-4 lists examples of well-known drug target databases which are currently web-accessible. Each database has their distinguished features, and they are complementary with each other. Since these databases aim to collect target information for different purposes, their size of data varies dramatically. We can use the number of targets as an example. For some databases, such as DengueDT-DB and GTD⁶⁰, the number of targets collected is below 100. This is partly because these databases are focused only on certain diseases, like dengue virus infection and bacterial pathogens, and the genome for these infectious species are relative small. In the other side of the spectrum, there are databases containing huge amount of targets data, such as Binding DB⁶¹ (3,056), DdTargets (4,000), SuperTarget⁶² (2,500), PharmGKB⁶³ (20,000), STITCH⁶⁴ (2.5 million), and TDR⁶⁵ (10,000). The large size of these data is because of their attempts for comprehensively collecting target information, and majority of them do not indicate what percentage of their data are established therapeutic targets.

As illustrated in its website, DrugBank⁶⁶ collected 2,500 proteins “linked to” FDA approved drugs. However, according to analysis in **Section 1.2.1**, this number far exceeds those historical evaluation (300~350). This is because that “link to” may not guarantee that these proteins are the primary therapeutic targets for drugs. In the latest version Therapeutic Targets Database⁹, the total number of targets is around 1,800, with 348 successful, 293 clinical trial and 1254 research targets. Because the number demonstrated in TTD is consistent with the historical exploration records, we choose to use TTD data to appreciate the outstanding properties of established therapeutic targets, and identify common features beneath those properties reflected by successful targets. This will be illustrated in detail in **Chapter 5** and **Chapter 6**.

In conclusion, extensive efforts have been devoted into summarizing the established drug targets. After debates for more than two decade, researchers begin to reach an agreement on 300~350 successful targets established by their approved or marketed drugs. In the meantime, targets in clinical trial have also been identified which can be an invaluable set of data for evaluating the process of current target discovery. Once we get the reliable set of established targets, it is time for us to appreciate their properties which make them outstanding compared to other proteins. With the advent of post-genome era, questions have been frequently asked. How many genes in human genome possess the ability to be targeted by drug-like molecule? How many genes will be established as successful targets? In order to answer these questions, I would like to introduce “druggable genome” first in **Section 1.3**.

1.3 Therapeutic target and druggable genome

The vast majority of successful drugs achieve their activity by binding to, and modifying the activity of, a protein. This limits the number of targets for which commercially viable therapeutics can be developed, thus leading to the concept of “druggable genome”—a subset of the ~30,000 genes in the human genome which express proteins able to bind drug-like molecules⁸. Researchers have been searching through the human genome and trying to identify those which are druggable, and, ideally, determine the size of druggable genome⁶⁷⁻⁶⁹. The estimated size of druggable genome from different research groups varies, because of the diverse sets of successful targets chosen as starting point, various biological hypotheses adopted, and different analysis tools applied.

1.3.1 Efforts devoted for exploring druggable genome

In Drews’ historical works “Genomic sciences and the medicine of tomorrow” published in 1996⁵¹, he was the first to conclude that there could be 5,000~10,000 potential targets on the basis of an estimate of the number of disease-related genes. However, this analysis did not relate the target with its corresponding drugs. As we know, commercially viable molecules possess common properties that can be summarized by Lipinski “rule-of-five”⁷⁰. Since drug targets need to be able to bind compounds with shared properties, it is reasonable to deduce that druggable targets should share some common features. In 2001, Bailey *et al.*⁷¹ introduced methods by assessing the number of ligand-binding domains to measure the number of potential points at which small-molecules could act, and Bailey’s conclusion suggests that the size of druggable genome could be even greater than 10,000.

However, the estimated number shrinks in Hopkins and Groom's publication⁸. A total number of 3,051 proteins have been predicted as druggable based on mapping proteins back to 130 proteins families representing the known drug targets. The estimated number consist ~10% of the whole human genome (30,000 genes)⁷². Hopkins further applied his methods onto *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae*, and estimated the sizes of their druggable genome are 13,601, 18,424, and 6,241 respectively. Their percentages of genome covered by the druggable genes are all around ~10% which is consistent with human genome.

An update on Hopkins and Groom's work was proposed in 2005, which re-estimate the size by using two algorithms: optimistic and conservative. In the optimistic scenario, the number arrives at just over 3000 targets, the same total as Hopkins' reported in 2002. The conservative count yields a total of ~2200 druggable genes¹⁰.

1.3.2 Gap between druggable protein and therapeutic targets

Druggable does not equal therapeutic. The capacity of a protein to bind a small molecule at the required binding affinity might make it druggable, but it does not mean that it is a potential drug target. One reason for this is the protein should also be disease-related, or disease-causing. Researchers have proposed that there are 3,000⁷³ to 10,000⁷⁴ disease-related genes, and large-scale mouse-knockout studies have revealed that only ~10% of all gene knockouts might have the potential to be disease modifying⁷⁵, which is consistent with the lower end of this range. Therefore, the potential therapeutic targets that our pharmaceutical industry should exploit are in the intersection between druggable genome

and disease-causing genes. The number has been suggested as a total of 600-1,500 small molecule drug targets⁸.

In summary, druggable genome has long been a critical issue that attracts broad interests. The size of druggable genome predicted by protein family based affiliation is around 3000. The rapid development on new computational methods has facilitated druggable protein identification. In the next section, I will manage to review these most popular approaches used for predicting druggable proteins.

1.4 Introduction to the prediction of druggable proteins

As illustrated in **Section 1.1.3** and **1.1.4**, various target identification technologies^{75,76} have been developed by analyzing disease relevance, functional roles, expression profiles and loss-of-function genetics between normal and disease states⁷⁷⁻⁸⁴. Computational methods have also been used to predict druggable proteins, the activity of which can be regulated by drug-like molecules⁸, from their genomic, structural and functional information^{8,85,86}—druggable proteins with key roles in a disease can then be explored as therapeutic targets⁸.

New and improved methods⁵⁷ and integrated and systems-based approaches^{77,78,87} have been explored for identifying druggable proteins. These commonly used computational methods have primarily been based on the detection of sequence and functional similarity to known drug targets^{8,85}, motif-based drug-binding domain family affiliation^{8,79}, and structural analysis of geometric and energetic features⁸⁶. On the other hand, machine learning approach takes a different strategy to identify druggability, which will be further illustrated in **Section 1.4.4**.

1.4.1 Sequence similarity approach

The most straightforward method of probing druggable protein from its primary structure is sequence alignment. Sequence alignment aims at measuring similarity to distinguish biologically significant relationship in evolution⁸⁸. The rationale behind this technique is that significant sequence similarity between two genes or proteins is a strong indicator of similar function⁸⁹.

Biological sequence similarity comparison started from the introduction of Needleman and Wunsch's dynamic programming algorithm⁹⁰ in early 1970s, which adopted an iterative matrix method for global alignment of two sequences. Later, Waterman and Smith⁹¹ extended this algorithm for local alignment, in which only sub-segments of two sequences with the highest score were aligned. From then on, many more rigorous algorithms were developed⁹², but their biological meaning was difficult to formulate. All these dynamic programming approaches assigned some sort of penalties to insertions, deletions and replacements of different length and computed an alignment of two sequences to maximize their similarity⁸⁸.

However, Because of their intensive computation requirements, dynamic programming algorithms are impractical for searching large sequence database, which is very common for current biological databases, without using supercomputer⁸⁸. Thus, various heuristic algorithms like FASTA⁹³ with less-cost of computation resources were developed. Unlike dynamics algorithms, heuristic algorithms do not aim for optimal alignments between two sequences, but utilize strategies to find approximate solutions with human heuristics. A significant breakthrough was made by the invention of a heuristic algorithm—BLAST, which gave good balance between computation speed and sensitivity, making it the most popular program for sequence comparison. In order to find distantly related proteins, the PSI-BLAST⁹⁴, allows to iterate BLAST search, with a position-specific score matrix generated from significant alignments found in previous rounds.

Moreover, the correlations between sequence similarity and functional similarity have been tested⁹⁵⁻⁹⁹. Based on the test result, Wilson *et al.*⁹⁶ concluded that for pairs of single-

domain proteins, precise function is usually conserved for sequence identity higher than 40%, and broad functional class is conserved for sequence identity higher than 25%. Thus, 40% identity seems to be an appropriate threshold to transfer the sequence similarity to function similarity.

In 2002, Hopkins and Groom deduced that similar sequence can indicate similar degree of druggability⁸. This would suggest that if one protein was able to bind a drug, other proteins that are substantially similar to it are also able to bind a drug-like molecule. Using this algorithm, Hopkins and Groom predicted 3,051 proteins as potential drug targets.

A real world example of target identification by sequence similarity comparison is the discovery of target candidates SNAIL3^{85,100}, a potent target of pharmacogenomics in the field of oncology and regenerative medicine. This gene was isolated by a similarity search of a known database and the characteristics of the sequence, such as chromosomal location, phylogeny and *in silico* expression analysis, were investigated by BLAST and other bioinformatics tools.

However, in the absence of clear sequence or structural similarities, the criteria for comparison of distantly-related proteins become increasingly difficult to formulate¹⁰¹. In the meantime, the success rate for identifying homologues with a sequence identity in the range of 20~30% is only approximately 50%, and the success of the searches is much lower for identities of less than 20%⁷⁹. Moreover, not all homologous proteins have analogous functions¹⁰². It is thus imperative to find other solutions to assign protein druggability beyond sequence similarity.

1.4.2 Motif based approach

Proteins with similar profiles are likely to be functionally related¹⁰³⁻¹⁰⁵. Thus, detection of common motifs among druggable proteins may provide important clues to targets identification. Motif based methods are usually more sensitive than pair wise comparison at detecting distant relationships between protein sequences. Moreover, motifs are easy to construct and use by biologists who have no training in bioinformatics¹⁰⁶. A number of motif-based databases have been developed to facilitate the identification of short and well-conserved regions, such as ligand-binding sites, enzyme-catalytic sites or post-transcriptional modifications¹⁰⁶. Each is different from others in terms of nomenclature and the approach to pattern recognition¹⁰⁷.

One of the most widely-used motif databases is PROSITE, which consists of a large collection of patterns that describe biologically meaning signatures of protein families¹⁰⁶. PROSITE was developed by manually seeking patterns that best fit particular protein families and functions¹⁰⁸. However, one problem with PROSITE patterns is that they are generally too short, which causes the high false-positive occurrences in unrelated sequences. In addition, there is no way to evaluate the probabilities of variations at a particular position. In order to solve these problems, PRINTS represents protein families through a number of fingerprints, which could be used to characterize features of protein families¹⁰⁶. These fingerprints consist of multiply aligned un-gapped segments derived from the most highly conserved regions in protein family, and they typically cover larger regions of the sequence than PROSITE. Moreover, PRINTS takes into account amino acid substitution matrices, so that it does not require exact matches to a fixed pattern¹⁰⁸.

Beside simple motifs derived directly from protein sequences, a higher level of motifs, called domains, could be used to characterize parts of a protein sequence with a single well-defined function. ProDom database clusters related sequence segments from pairwise sequence comparison into domain families¹⁰⁹, so that a new incoming protein could be compared to the domain database to identify shared domains. Another well known motif database is PFAM database, which collects manually curated multiple sequence alignments for more than 12,000 domain families¹¹⁰, and represents these families through hidden Markov models (HMMs). Each family contains two multiple alignments, one from relatively small number of representative proteins and the other one from full alignment of all members in the database that can be detected. InterPro is another widely used motif database of predictive protein “signature” used for the classification and automatic annotation of proteins and genomes¹¹¹. InterPro classifies sequences at different levels: super-family, family and sub-family, and it is used for predicting the occurrence of functional domains, repeats and important sites.

Motif based approach has been applied in finding out druggable proteins. In Hopkins and Groom’s work⁸, they mapped the sequences of the drug-binding domain of 399 molecular targets into InterPro domains, and identified 130 protein families representing known drug targets. Since proteins with similar profiles are likely to be functionally related¹⁰³⁻¹⁰⁵, those proteins in the 130 protein families are regarded as potentially druggable.

Furthermore, HMM algorithms, HMMER (profile hidden markov models)¹¹² and SAM (sequence alignment and modeling)¹¹³, have been applied for detecting close and remote homologues of gene families that are of specific interest in target discovery⁷⁹. As each

method produces overlapping but non-identical results, both algorithm were used and the results were combined for maximum effectiveness.

But many motifs or domains only reflect positions for post-translational modification or structural signals, without any direct functional implications¹¹⁴. The presence of shared domain within a group of proteins does not necessarily imply that these proteins perform the same function¹¹⁵. Moreover, not all the protein sequences could be covered by currently available motifs and it is reported that the latest version of PFAM (2010) only covers about 75.15% of all protein sequences¹¹⁰.

1.4.3 Structural analysis approach

Besides sequence similarity and motif approaches, structure based approach can also provide valuable insights into protein druggability⁸⁶. In the identification of homolog, structural approach often succeeds when sequence-alone-based methods fail¹¹⁴, since structure is more conserved than sequences. It is well known that proteins can function only when they form 3D structures in their functional environment. Due to structure-function relationship, it is possible to correlate protein structure to function¹¹⁶. Therefore, knowledge of proteins sharing similar structures is a strong indication that they have similar functions^{117,118}, and specifically similar degree of druggability⁸⁶. Now, with the increasing number of protein 3D structures, databases classifying proteins into hierarchical structure families are available, such as CATH^{119,120} and SCOP^{121,122}.

Successful application of structural based approaches onto druggability identification has been reported⁸⁶, which uses hetero-nuclear-NMR (nuclear magnetic resonance)-based

screening data. The relationships between protein druggability and protein binding site properties have been derived. It is found that properties like polar and apolar surface area, surface complexity, and pocket dimensions can correctly classify 94% of the proteins with high-affinity, non-covalent, drug-like leads.

Hajduk *et al.*⁸⁶ demonstrate general procedures applied for identifying disease-modifying and druggable proteins by relying on the 3D structure of the protein target itself. The first step is to identify all possible binding sites on a protein surface, and followed by the assessment of whether the binding site can bind with high affinity and specific small molecule drugs. They have calculated the potential druggability of more than 1,000 non-redundant human proteins, and predicted nearly 35% as containing at least one highly druggable binding site.

The problem of a structure based approach is that only few residues are directly involved in the function of proteins¹²³. Therefore, functions are more likely to be influenced by random mutations than structures. Compared with structure prediction, it is more difficult for computational methods to predict protein function¹²⁴. Another problem with structure based approach is the limitation on available protein structures. The latest version of Protein Data Bank (PDB, 2010)¹²⁵ collected 66,828 protein structures. Comparing with 518,415 unique protein sequences in the latest sequence database¹²⁶, only around 13% of the proteins have available structures. For about 34,000 human protein sequences covered in latest sequence database¹²⁶, only 16,371 have been identified by conducting keyword search “homo sapiens” in PDB, and it is reasonable to expect the real amount of human protein structure will not be larger than this number. Despite progress in structural

genomics, most proteins encoded in newly sequenced genomes are known from their amino acid sequences alone¹¹⁴. Although computational methods like homology modeling¹²⁷⁻¹²⁹ may be used to an extent, however, the number of proteins with structures available is still far less than sequences. Thus, methods that do not require protein structures are needed to predict function of those proteins whose structures are unavailable.

1.4.4 Machine learning methods

Unlike sequence similarity and structure based approaches, machine learning takes a different strategy in predicting druggable proteins. In these methods, proteins are represented by digitalized descriptors instead of direct use of sequences. These descriptors either describe physicochemical properties of the constituent amino acids^{130,131} or capture both local structural motifs and longer conserved regions associated with specific functional properties¹³². Once we have knowledge on druggable proteins and non-druggable proteins, the supervised classification methods—machine learning—can be applied for developing an artificial intelligence system to separate proteins into two classes: druggable and non-druggable.

Machine learning methods have been successfully applied to the prediction of proteins of specific functional class characterized by distinguished biochemical properties or biological activities, including G-protein coupled receptors^{133,134}, nuclear receptors¹³⁵, trans-membrane proteins^{136,137}, lipid-binding proteins¹³⁸, enzymes of various families¹³⁹, and transporters¹³¹. They also show some level of capability in predicting the functional class of proteins that are non-homologous to any protein of known function¹⁴⁰⁻¹⁴³.

Machine learning has been applied to identify druggable protein. Zheng *et al.*⁵⁸ used a total of 1535 successful and research targets to construct the druggable class, and 12,956 representative proteins from 6856 PFAM protein families (with all of the known target-representing families excluded from these families) were used to construct the non-druggable class. Their average prediction accuracy of 5-fold cross validation study is 69.8% for druggable proteins and 99.3% for non-druggable proteins. Similarly, in a later report, Xu *et al.*¹⁴⁴ reached ~72% for sensitivity and ~98% for specificity.

However, successful machine learning process requires accurate and sufficient training data, but ambiguity and typos are very common in biological data. Machine learning methods are not applicable for proteins with insufficient knowledge about their specific functional profile. The searching for information about proteins known to possess a particular profile and for those do not possess that profile is the key to more extensive exploration of machine learning methods for facilitating the study of protein functional profiles.

In conclusion, commonly used computational methods for druggable protein prediction are generally based on the detection of sequence and functional similarity to known drug targets^{8,85}, motif-based drug-binding domain family affiliation^{8,79}, structural analysis of geometric and energetic features⁸⁶, and machine learning approaches^{58,144}. Sequence similarity, motif and structure based methods are less effective in finding targets that exhibit no or low homology to known targets, disease proteins and proteins with available 3D structures. However, such non-homologous and structurally unknown proteins constitute a substantial percentage, ~20-100%, of the open reading frames in many of the

completed genomes and might, therefore, be an untapped source of novel drug targets¹⁴¹. Hence, methods independent of sequence and functional similarity, and structural availability, are highly desirable. Statistical machine learning method, has recently been explored for predicting druggable proteins⁵⁸, anticancer genes¹⁴⁵, proteins in families of high target concentrations^{131,133,135,139,146,147}, as well as proteins of various broadly defined functional and structural classes⁶⁸, from sequence-derived constitutional and physicochemical properties, irrespective of similarity to known proteins. This method is particularly useful for predicting novel targets that exhibit no or low homology to known targets. In **Chapter 2 Section 2.3.1**, I plan to describe the machine learning algorithms into detail.

1.5 Objective and outline of this thesis

1.5.1 Objective of this thesis

The ultimate goal of this thesis is to construct *in silico* target identification approaches to facilitate modern drug target discovery. In order to meet it, this thesis has been divided into three sections, each of which deals with one sub-objective. These three objectives are inter-connected with each other. In the following part of this section, these objectives will be explained step by step, which will then lead to the final goal of this thesis.

The first objective is to construct or update pharmainformatics databases providing data that are pharmaceutically important but no comprehensive data have been provided in the publicly accessible databases. Given the demands on information about the primary drug target(s) of comprehensive sets of approved, clinical trial, and experimental drugs/agents for facilitating focused investigation and discovery, significant expansion of data and careful identification of primary drug target(s) for each drug are conducted in TTD 2010. However, it is difficult to find a place where primary targets information is provided, which needs additional actions on designing procedures of primary drug target(s) validation. Therefore, pharmainformatics databases IDAD and TVD were constructed. IDAD provides the activity information for each drug which acts as an essential factor for primary target validation, and TVD aims at giving *in vitro* and *in vivo* target validation data for targets covered by TTD.

Once the target data are available, well-organized and clearly classified, it is time to reach the second objective: revealing common properties among drug targets by comprehensive

analysis on successful and clinical trial targets. The druggability of proteins varies according to their therapeutic, biochemical, physicochemical, and systems features, and it is reasonable to say that a better understanding of those common features will help to figure out what constitute “druggability”. Thus, analysis on protein family distribution, pathway affiliation, subcellular location, similarity to successful targets, drug potencies, and the Lipinski’ rule of five were planned to reach this objective.

After revealing rules guarding “druggability”, the next question would be: whether is it possible to develop fast approaches for identifying promising drug targets from the whole pool of human genome? In order to answer this question, a collective approach combining 4 *in silico* methods was proposed to facilitate target discovery. With the solid foundation built up by the former two objectives, the third one is reachable. Moreover, as an application, we also planned to extend our target discovery approach to identify promising targets from the genome of influenza A (H1N1, H5N1, H2N2, H3N2, H9N2), B and C. If this real world test is proved to be effective, it would be encouraging to apply the approach further.

In summary, this thesis is to predict therapeutic targets by applying bioinformatics tools. In the first step, pharmainformatics databases with comprehensive and accurate data are constructed. With the analysis on the comprehensive targets data collected, several rules guarded druggability are identified. Based on reliable data and better understanding to drugs targets, it is finally possible to apply combinatorial bioinformatics tools to facilitate drug target discovery.

1.5.2 Outline of this thesis

In **Chapter 1**, an overview of target discovery in current pharmaceutical research is given in the first section, which reviews current technologies for both target identification and validation. Then we have a retrospective review of efforts to distinguish established drug targets, and a comprehensive analysis of available drug targets information. As effective tools for target discovery, popular bioinformatics techniques are then introduced in the final section of this chapter.

Chapter 2 illustrates methods used in this work. In particular, the strategy of developing pharmainformatics database, the methods used for validating primary therapeutic targets for drugs, the machine learning methods, the sequence similarity on drug-binding domain, the comparative study of structural fold on drug-binding domain and system-level rules for druggability are presented in more detail.

In **Chapter 3**, construction of 3 pharmainformatics databases is demonstrated. In the first section, update of therapeutic targets database is shown with detailed illustration on significance and reliability of TTD data and many powerful new features integrated. As byproducts of TTD, IDAD and TVD are introduced in the following section of this chapter.

Chapter 4 systematically analyzes the therapeutic, bio-chemical, physicochemical, and systems features of the targets and drugs in clinical trials. In the first section, an analysis on trends in the exploration of clinical trial targets is given. In the following two sections, two comparisons between characteristics of clinical trial target and that of successful one, and characteristics of clinical trial drug and that of approved drug and drug lead are given.

We conclude in the last section that it necessitates further improvement in drug-like and target-like physicochemical features.

In **Chapter 5**, a collective approach which integrates four *in silico* methods is proposed for target identification. After a brief illustration on these four methods in the first section, the performances of target identification on clinical trial, non-clinical trial, difficult, and non-promising targets are shown.

As an application, **Chapter 6** focuses on identify promising targets from the genomes of influenza A (H1N1, H5N1, H2N2, H3N2, H9N2), B and C by using our target discovery methods. After an introduction on methods used, we illustrate the identification results and make further discussion on them in the rest sections of this chapter.

Finally, in the last chapter, **Chapter 7**, major findings and contributions of current work to modern target discovery are discussed. Limitations and suggestions for future studies are also rationalized in this chapter.

Figure 01- 1 Drug discovery process



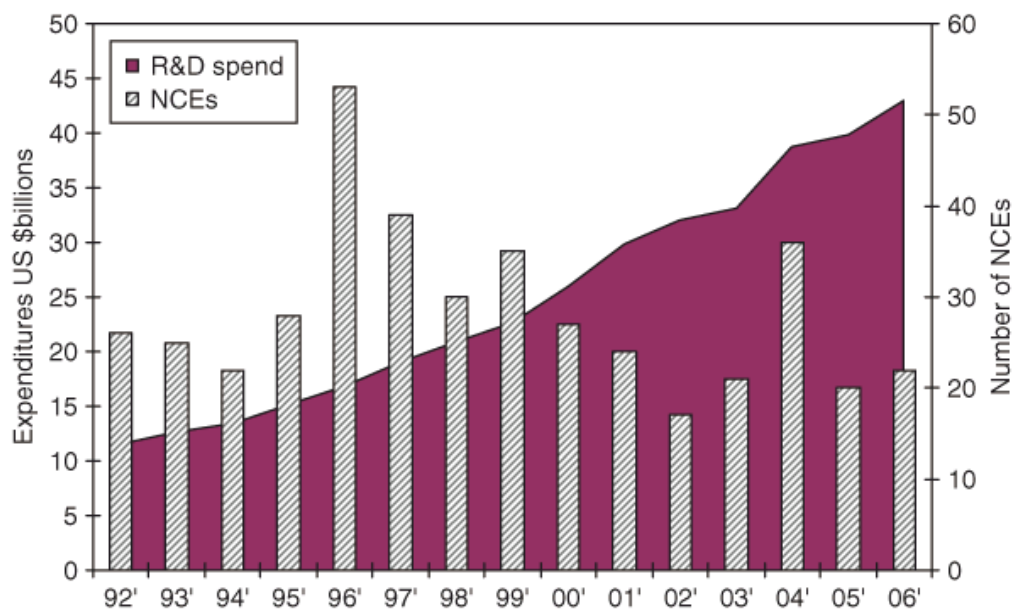
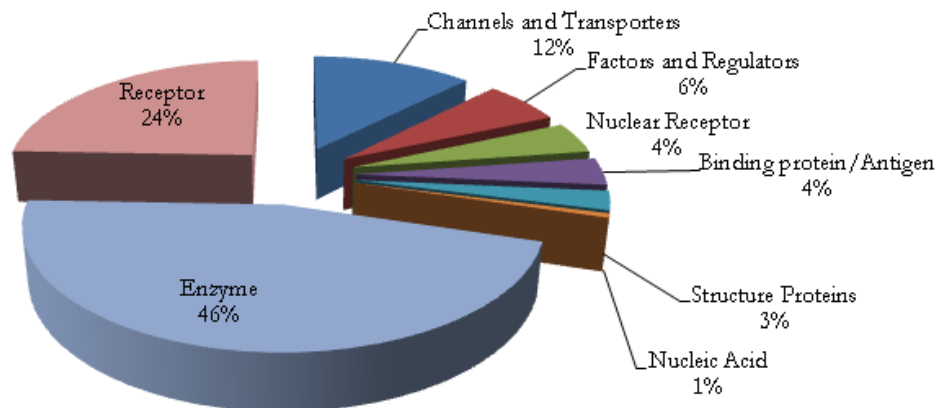
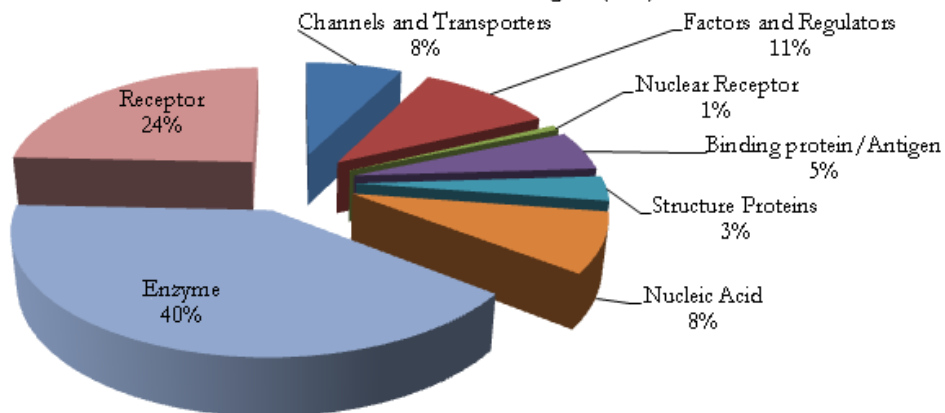
Figure 01- 2 Number of new chemical entities in relation to R&D spending (1992-2006)**Figure 01- 3** Biochemical class for successful and clinical trial targets in TTD**A. BioChemical Class Distribution for Successful Targets (348)****B. BioChemical Class Distribution for Clinical Trial Targets (293)**

Table 01- 1 Examples of well-known gene expression database

Database Name	Brief description of the database and statistics of data stored
ArrayExpress ¹⁸	ArrayExpress Archive is a database of functional genomics experiments including gene expression where you can query and download data collected to MIAME and MINSEQE standards. In 2009, it contains data from over 6,000 experiments comprising approximately 200,000 assays, and doubles in size every 15 months.
BodyMap-Xs ¹⁹	BodyMap is a human and mouse gene expression database that is based on site-directed 3'-expressed sequence tags generated at Osaka University. To date, it contains more than 300,000 tag sequences from 64 human and 39 mouse tissues. For the recent release, the precise anatomical expression patterns for more than half of the human gene entries are generated by introduced amplified fragment length polymorphism (iAFLP), which is a PCR-based high-throughput expression profiling method. The iAFLP data incorporated into BodyMap describes the relative contents of more than 12 000 transcripts across 30 tissue RNAs.
CIBEX ²⁰	CIBEX (Center for Information Biology gene Expression database) is a public database for microarray data, which is aimed at storing MIAME-compliant data in accordance with MGED Society recommendations. In 2009, it contains 94 experiments, 139 arrays and 2488 hybridizations.
DDBJ ²¹	Data releases from DDBJ (DNA Data Bank of Japan) are the complete genome sequence of an endosymbiont within protist cells in the termite gut and Cap Analysis Gene Expression tags for human and mouse deposited from the Functional Annotation of the Mammalian cDNA consortium. In 2010, it contains >120 million gene expression entries and >115 billion bases.
ExpressDB ²²	ExpressDB is a relational database containing yeast and E. coli RNA expression data. Currently, it contains more than 20 million pieces of information loaded from numerous published and in-house expression studies.

GEO²³ GEO (Gene Expression Omnibus) is a public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomic data submitted by the scientific community. In 2009, the database holds over 10,000 experiments comprising 300,000 samples, 16 billion individual abundance measurements, for over 500 organisms, submitted by 5,000 laboratories from around the world.

GXD²⁴ GXD (Gene Expression Database) is a community resource for gene expression information from the laboratory mouse. GXD integrates different types of expression data. In 2007, it contains >56,500 entries covering nearly 12,300 references analyzing nearly 8,700 genes.

HuGE Index²⁵ HuGE Index (Human Gene Expression Index) aims to provide a comprehensive database to aid in understanding the expression of human genes in normal human tissues. Now, it provides the results of 59 microarray experiments conducted using tissue from 19 different human organs from 49 different individuals. All of these experiments are performed using oligo-nucleotide microarrays, which probe for mRNA from approximately 7,000 genes.

SAGEmap¹⁵ SAGEmap is created to provide a central location for depositing, retrieving, and analyzing human gene expression data. This database uses serial analysis of gene expression to quantify transcript levels in both malignant and normal human tissues. Currently, this resource contains over two million tags from 47 SAGE libraries.

SOURCE²⁶ SOURCE (Stanford Microarray Database) is a web-based database that brings together information from a broad range of resources, and provides it in manner particularly useful for genome-scale analyses.

Table 01- 2 Brief description, advantages and limitations of loss-of-function target validation technologies

Loss-of-function technologies	Strategies of technologies	Advantages	Limitations
DNA knockout validation models	Genes are deleted or disrupted to halt their expression ¹⁴⁸	<ul style="list-style-type: none"> (1) Enable the generation of relevant animal disease models and the determination of the biological functions of new targets^{34,149}; (2) Can elucidate complex interactions in both normal and disease pathways^{150,151}. 	<ul style="list-style-type: none"> (1) Tissue-specific transcription, insertion positioning and other transgenesis-related factors are not under full control and can alter the veracity of the phenotype observed⁵⁰; (2) Potential compensatory effects during development may limit the validation of phenotype induction or suppression⁵⁰; (3) Suffer from absence of genetic models other than mouse⁵⁰.
mRNA Knockdown validation models	Synthesize RNA complementary to gene of interest. Introduce it into a cell. RNA is recognized as an exogenous genetic material, which then activate the RNAi pathway ¹⁵²⁻¹⁵⁴	<ul style="list-style-type: none"> (1) In theory, well-designed siRNA could be used to silence any gene in the body¹⁵⁵⁻¹⁵⁷; (2) siRNA technology can be used <i>in vivo</i> in established genetic models, humanized mouse models¹⁵⁸ and species other than the mouse¹⁵⁷. 	<ul style="list-style-type: none"> (1) Delivery limitations¹⁵⁸⁻¹⁶¹; (2) Time frame of the silencing process is too short, which requires repeated or continuous delivery of siRNA—with the associated risk of toxicity and the restrictions imposed by the use of viral vectors¹⁶².

Protein knockout models based on vaccination

Use of neutralizing antibodies to deplete proteins *in vivo*⁵⁰.

- (1) Enable direct target validation in adult animals avoiding activation of compensatory pathways during the embryonic development phase⁵⁰;
- (2) Applicable to circulating proteins or membrane-bound proteins⁵⁰;
- (3) Potentially applicable to all mammalian species⁵⁰;
- (4) Ability to derive the phenotype in pre-existing animal models⁵⁰.

Face the immune tolerance barrier that protects endogenous proteins from autoimmune attacks⁵⁰.

Table 01- 3 Molecular targets of FDA-approved drugs from Overington's work

Class of drug target	Species	Number of molecular targets
Targets of approved drugs	Pathogen and human	324
Human genome targets of approved drugs	Human	266
Targets of approved small-molecule drugs	Pathogen and human	248
Targets of approved small-molecule drugs	Human	207
Targets of approved oral small-molecule drugs	Pathogen and human	227
Targets of approved oral small-molecule drugs	Human	186
Targets of approved therapeutic antibodies	Human	15
Targets of approved biologicals	Pathogen and human	76

Table 01- 4 Examples of well-known drug target database

Name of target database	Description	Date of last update	Most distinguished features	Data statistics
Binding DB ⁶¹	A database of experimentally determined protein-ligand binding affinities.	2010, Jul	(1) Measured binding affinities; (2) Focus on the interactions of drug-targets with drug-like molecules.	544,641 binding data, 782 ITC data, 3,056 protein targets, and 240,203 small molecules.
DdTargets Drug Target Disease Target Database	A comprehensive database on therapeutic drug targets, and resource for biomedical and pharmaceutical researchers.	2010, Apr	(1) Contain targets reported in US patent or US/International patent applications; (2) Targets are classified according to specific drug types and disease types.	4000 drug targets.
DengueDT-DB	A database about dengue virus, containing dengue virus genes, drugs and targets and aiming to facilitate dengue research from drug discovery, vaccine development, epidemiology and comparative genomics.	2008, Jan	(1) Focus on dengue virus only; (2) Include information of dengue virus genes, dengue virus drugs targets and drugs.	13 dengue drugs targets, 80 dengue virus genes, 4 drugs information and genome sequences, and 4 attenuated vaccines.
DrugBank ⁶⁶	A bioinformatics and cheminformatics resource combining detailed drug data with comprehensive drug target information.	2008, Aug	(1) Comprehensive chemical, pharmacological and pharmaceutical data for drugs; (2) Detail sequence, structure and pathway information for targets.	1,473 FDA-approved drugs, 71 nutraceuticals, 3,243 agents in lab test. 2,500 proteins linked to FDA approved drugs.

GTD ⁶⁰ Genomic Target Database	A database consisting of putative genomic drug targets of most common human bacterial pathogens.	2009, Apr	<ul style="list-style-type: none"> (1) Focus on targets from human bacterial pathogens; (2) Selected pathogens are either drug resistant or vaccines are yet to be developed. 	58 drug targets for 4 human bacterial pathogens: <i>Aeromonas hydrophila</i> ATCC-7966, <i>Burkholderia pseudomallei</i> K96243, <i>Helicobacter pylori</i> , and <i>Vibrio cholerae</i> .
HDAPD ¹⁶³ Human Disease-Associated Protein Database	A database providing a variety of resources of disease-causing proteins, including X-ray, NMR and electron microscopy structures and pathway map.	2010, Feb	<ul style="list-style-type: none"> (1) Disease-causing proteins with X-ray, NMR and electron microscopy structures; (2) Targets information of cellular component, protein function, and biological process. 	395 disease associated proteins, 256 disease-associated proteins with X-ray/ NMR structure; 2,861 disease associated protein structures.
PDTD ¹⁶⁴ Potential Drug Target Database	A dual function database associating an informatics database to a structural database of known and potential targets.	2008, Mar	<ul style="list-style-type: none"> (1) Focus only on therapeutic targets with 3D structure; (2) Integrated with TarFisDock¹⁶⁵, a web server for identifying drug targets with docking. 	1,207 entries covering 841 known and potential drug targets with structures.
PharmGKB ⁶³ Pharmacogenomics Knowledge Base	A central repository for genetic, genomic, molecular and cellular phenotype data and clinical information of people participated in pharmacogenomics research.	2010, Jun	<ul style="list-style-type: none"> (1) Pharmacokinetic and pharmacogenomic data in cardiovascular, pulmonary, cancer, pathways, metabolic domains; (2) Provide genetic variation differentiating individuals in reaction to drugs. 	20,000 genes, 3000 diseases, 2500 drugs, 53 pathways, 470 genetic variants (SNP data) affecting drug metabolism.

STITCH ⁶⁴ Search Tool for Interactions of Chemicals	A database integrating information about interactions from metabolic pathways, crystal structures, binding experiments and drug-target relationships.	2010 Jan	(1) Create a network of interactions; (2) Incorporating BindingDB, PharmGKB and Comparative Toxicogenomics Database.	74,000 small molecules and over 2.5 million proteins in 630 organisms.
SuperTarget ⁶²	A one-stop data warehouse integrating drug-related information about medical indication areas, adverse drug effects, drug metabolism, pathways and GO terms of target proteins.	2008 Jan	(1) Information about medical indication area, drug adverse effects and metabolism, pathways and GO terms of target proteins; (2) Provide tools for 2D drug screening and sequence comparison of targets.	2500 target proteins annotated with 7300 relations to 1500 drugs.
TDR ⁶⁵ TDR Targets database	A website for finding target information and tool for prioritizing target in genome.	2009, Mar	(1) Extensive genetic and pharmacological data related to tropical disease pathogens; (2) Computational assessment of target druggability by using weight algorithm.	1,790, 862, 2,100, 1918 and 5,474 target enzyme entry for 5 tropical disease pathogens.
TTD ⁹ Therapeutic Target Database	A database providing information about known and explored therapeutic protein and nucleic acid targets, and their drugs.	2010, Mar	(1) Categorize targets into successful, clinical trial (phase I~III), and research target; (2) Identify primary target for drugs approved or in clinical trial by ignoring off-targets and targets with minor therapeutic effect.	348 successful, 292 clinical trial and 1,254 research targets, and 1,515 approved, 1,279 clinical trial and 2,332 drugs in lab test.

Chapter 2 Methods used in this thesis

In this chapter, I intend to give the audience introductions to three methodologies applied to this thesis, which organize **Chapter 2** into three sections. In **Section 2.1**, a review of strategy of pharmainformatics database development is given. In combination with three databases (TTD, IDAD and TVD) included in this work, their logical view of information construction is illustrated into detail. As the most critical job in TTD construction, the identification of primary therapeutic targets for approved drugs, drugs in clinical trial, and experimental agents is demonstrated in **Section 2.2**. In the last section of this chapter **Section 2.3**, computational approaches adopted for predicting druggable proteins are discussed. These approaches include: (1) physicochemical properties of drug targets identified by machine learning; (2) analysis on sequence similarity between drug-binding domain of a studied target and that of a successful target; (3) comparative study of structural fold of the drug-binding domains of studied and successful targets; (4) three simple system-level druggability rules.

2.1 Development of pharmainformatics databases

Database development has shown a broad spectrum of applications in not only scientific research but also our everyday life. In particular, pharmainformatics databases aiming at providing comprehensive and systematic information for pharmaceuticals-related research have been widely utilized. Despite their various applications in pharmaceutical research, the general strategy adopted for constructing these databases is similar. In **Section 2.1** of **Chapter 2**, the basic strategy used for developing knowledge-based pharmainformatics databases is demonstrated, which is extended to construct Therapeutic Targets Database (TTD), Information of Drug Activity Data (IDAD) & Target Validation Database (TVD), which are discussed later. Generally, the development of a database is a process including rational architecture design, information accumulation, optimal data storage and user-friendly data access and representation. In the following parts of **Section 2.1**, I will illustrate the strategy utilized for database construction in a stage by stage manner.

2.1.1 Rational architecture design

Before the construction of pharmainformatics database, designing a rational architecture can help us define the scope of the database, focus on relevant pharmaceutical problem, and pave the way for the information collection stage. So, in this stage, the objective and content of the database should be seriously considered. As described in **Chapter 1**, target discovery plays a very important role in drug research and development. It becomes more and more necessary to provide comprehensive and systematic information on therapeutic targets of currently available approved drugs and drugs in clinical trial, kinetic activities

of drugs and experimental agents, and how to validate a protein as the primary target of a certain drug. However, up to date, there is no such pharmainformatics database providing these kinds of information. Therefore, such kinds of knowledge-based pharmainformatics databases will provide valuable information for current pharmaceutical research. In sum, our databases (TTD, IDAD and TVD) are designed to afford therapeutic targets and drug activity related information. Following the design of the preliminary architecture of the whole database, a detailed description of the database development will be presented.

2.1.2 Information mining for pharmainformatics databases

Generally, a knowledge-based pharmainformatics database is designed to provide sufficient domain knowledge around a specific subject in pharmacology. Use TTD as an example, TTD is designed to provide some biological information for specific therapeutic targets, their relevant disease conditions, and drugs/ligands corresponding to this target, and so on. Therefore, for every entry in TTD, there are several different knowledge domains, some of which provide basic introduction to the entry itself, and some others give information derived from or relevant to this entry.

The information planned to be integrated can be selected from a comprehensive search of literatures like pharmacology textbooks and research publications. In light of the diversity of information types, the methods used for data collection vary, but one thing in common is to seek data from reliable resources. At present, no ready index or library is available and almost all the relevant information is scattered in the huge amount of biological and medical literatures. Therefore, literature information extraction is considered to be one of the most feasible ways for information mining. It is generally agreed that literatures are

typically unstructured data source, and the terms used in different sources, which may be in synonymous name, various abbreviations, or totally different expression, are difficult to be recognized by automatic language processing. An automated literature information extraction system solely relying on computational recognition, thus, cannot be invented to gather information from literature both efficiently and accurately.

In this thesis, automatic text mining methods with manual reading process was combined. Automated text retrieval programs developed in Perl were used to screen the literature that contained the key words in the local Medline abstract packages¹⁶⁶. Then, the useful subject information was picked up manually from these matched Medline abstract. If necessary, the full literature was referred to facilitate information searching. Meanwhile, in many cases, the relevant information about the same subject could also be found in the same literature. Therefore, in the first step, not only subject but also relevant information could be obtained and recorded. In the second step, detailed biological information of subject was automatically selected from some general or specific biological databases, such as SwissProt, PDB, PubChem, KEGG, and so on, by text mining program. Likewise, other information derived from the subject was also extracted from the corresponding databases in the same way. After information collection, how to store, organize and manage the data using database techniques was discussed. In the next section, the database construction is described.

2.1.3 Data organization and database structure construction

A good database system enables the user to create, store, organize, and manipulate data efficiently. By integrating databases and web sites, users and clients can open up

possibilities for data access and dynamic web content. An integrated information system of our pharmainformatics database is constructed according to some standardization strategies as follows:

1. Establishment of standardized data format and appropriate data model
2. Database structure construction
3. Development of Database Management System (DBMS)

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

2.1.3.1 The database model

Database model is an integrated collection of concepts for describing data, relationships between data, and constraint on the data. In the other word, a database model is a specific description on how a database is structured and used. Currently, there are several different basic ways of constructing databases, among which have been listed as follow:

1. The flat file model
2. The hierarchical model
3. The network model
4. The relational model
5. The object-oriented model

The flat-file model is the simplest data model, which is essentially a plain table of data. Each item in the flat file, called a record, corresponds to a single, complete data entry. A

record is made up by data elements, which is the basic building block of all data models, not just flat files. The flat-file data model is relatively simple to use; however, it is inefficient for large databases.

The hierarchical data model organizes data in a tree structure (**Figure 02-1**). It has been used in many well-known database management systems. The basic idea of hierarchical systems is to organize data into different groups, which can be divided into different subgroups. In a subgroup, there may be some sub-subgroups, and so on. That is to say, there is a hierarchy of parent and child data segments. In a hierarchical database the parent-child relationship is one-to-many. The hierarchical data model will be convenient to use and run very efficiently only if the nature of the application remains strictly hierarchical. Actually, in real world application, few database management problems remain strictly hierarchical. It is the major failing of this kind of data model.

In most cases, the relationships of data would be arbitrarily complex (**Figure 02-2**). In this model, some data are more naturally modeled with multiple parents per child. So, the network model permits the modeling of many-to-many relationships in data. This model, thus, can handle varied and complex information while remaining reasonably efficient. Even so, the biggest problem with the network data model is that databases can get excessively complicated.

The relational model was formally introduced in 1970 and has been extensively used in biological database development (**Figure 02-3**). The model is a much more versatile form of database. On the basis of this kind of data model, a novel system named relational database management system is established. A relational database allows the definition of

data structures, storage and retrieval operations and integrity constraints. In such a database the data and relations between them are organized in tables.

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

The object-oriented database (OODB) paradigm is “the combination of object-oriented programming language (OOPL) systems and persistent systems”¹⁶⁷. “The power of the OODB comes from the seamless treatment of both persistent data, as found in databases, and transient data, as found in executing programs”¹⁶⁷. The database functionality is added to object programming languages in object database management systems, which extend the semantics of the C++, Smalltalk and Java object programming languages to provide full-featured database programming capability. The combination of the application and database development with a data model and language environment is a major advantage of the object-oriented model. As a result, applications require less code, use more natural data modeling, and code bases are easier to maintain.

2.1.3.2 Construction of relational database structure

The relational model has been used in our pharmainformatics databases. It represents relevant data in the form of two-dimension tables. Each table represents relevant data

collected. The two-dimensional tables (**Figure 02-4**) for the relational database include the entry ID list table, the main information table, which contains a record for the basic information of each entry, data type table, which demonstrates the meaning represented by different number, and reference information table, which gives the general reference information following by different PubMed ID in Medline¹⁶⁶.

Figure 02-4 is a general logical view of databases (TTD, IDAD and TVD) we developed. It shows the organization of relevant data into relational tables. In these tables, certain fields may be designated as keys, by which the separated tables can be linked together to facilitate searching specific values of that field. Commonly, in relational table, the key can be divided into two types. One is primary key, which uniquely identifies each record in the table. Here it is a normal attribute that is guaranteed to be unique, such as entry ID in entry ID list table with no more than one record per entry. The other is foreign key, which is a field in a relational table that matches the primary key column of another table. The foreign key can be used to cross-reference tables. For example, in tables of our databases, there are two foreign keys: Data type ID and Reference ID. According to **Figure 02-4**, a connection between a pair of tables is established using a foreign key. The two foreign keys make three tables relevant. Generally, there are three basic types of relationships of related table: one-to-one, one-to-many, and many-to-many. In our case, these databases belong to one-to-many relationships.

2.1.3.3 Development of Database Management System

By using relational database construction software (e.g. Oracle, Microsoft SQL Server) or even the personal database systems (e.g. Access, Fox), the relational database can be

organized and managed effectively. This kind of data storage and retrieval system is called Database Management System (DBMS). An Oracle 9i DBMS is used to define, create, maintain and provide controlled access to our pharmainformatics databases and the repository. All entry data from the related tables described in previous section are brought together for user display and output using SQL queries.

2.2 Methodology for validating therapeutic targets

In the development of TTD, the most critical job is to identify primary therapeutic targets for approved drugs, drugs in clinical trial, and experimental agents. In our analysis, the primary targets and their corresponding drugs/agents were initially collected from the company websites and publications or review articles in reputable journals (e.g. *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), 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. These targets are expected to be well defined based on solid *in vitro* and *in vivo* target validation studies. However, in order to double check and have an overall understanding on the status of these targets, we have searched from literatures of reported IC_{50}/EC_{50} values against the target/targets and cell-lines and the reports of *in vivo* studies to confirm that the reported primary targets are accurate.

For those drugs/agents without explicitly reported targets in the above mentioned sources, we conducted additional literature search to first identify proteins or nucleic acid targets explicitly reported to be inhibited or activated by the drug and the interaction reported in the same or different papers to be directly responsible for the claimed therapeutic effect. We again check their IC_{50}/EC_{50} values against the targets, cell-lines and *in vivo* data. A target was confirmed if, in addition to the mentioned reports above, the corresponding

drug acts on the target with IC_{50}/EC_{50} values $< 100\text{nM}$ or with IC_{50}/EC_{50} values $< 1\ \mu\text{M}$ but acts on relevant cell-lines with $IC_{50}/EC_{50} < 1\ \mu\text{M}$.

Based on the target collection method and the results of conformation study on randomly selected targets, the collected target data is reasonably accurate. None-the-less, only ~20% of the clinical trial targets have IC_{50}/EC_{50} data, it would be better to have the relevant data for more comprehensive sets of targets. Hence, further search was conducted to collect IC_{50}/EC_{50} values against targets, cell-lines, and *in vivo* and knock-out studies for more comprehensive sets of clinical trial targets and corresponding clinical trial drugs.

For multi-target kinase inhibitors, not all drug targets with $IC_{50} < 100\text{nM}$ are necessarily primary targets. The primary targets of these kinase inhibitors were identified as follows: Step 1, all the targets of the multi-target kinase inhibitor were identified based on explicit literature report and $IC_{50} < 100\ \text{nM}$ ($< 1\ \mu\text{M}$ if cell-line $IC_{50} < 1\ \mu\text{M}$), most kinase inhibitors are multi-target and have IC_{50} values; Step 2, check literature report about each target to determine if in specific disease or disease subtype (e.g. NSCLC lung cancer), the inhibition of this target will produce a response (i.e. they can be the main target against the disease), or whether it is a bypass gene directly contributing to the resistance of the drug against the main target, then check whether the target is up-regulated or have sensitizing mutation or amplified in the specific disease cell-lines or tissues; Step 3, the primary targets are selected from those targets that are the main target or bypass genes with the right expression, mutation or amplification profile against the clinical trial disease or disease subtype (e.g. NSCLC lung cancer). Typically, the number of primary

targets of multi-target kinase inhibitors are in the range of 2-4, with a very small number of them has 5~6 primary targets.

For multi-target GPCR binders, the methods for determining their primary targets are similar to those of multi-target kinase inhibitor. The specific steps are: Step 1, identify all the targets of the multi-target GPCR binder based on explicit literature report and $IC_{50}/EC_{50} < 100 \text{ nM}$ ($< 1 \text{ }\mu\text{M}$ if cell-line $IC_{50}/EC_{50} < 1 \text{ }\mu\text{M}$); Step 2, check literature report about each target to determine in which disease or disease subtype, binding to (agonizing or antagonizing) this target will produce a response, and whether the target is up-regulated or amplified in the specific disease cell-lines or tissues. Step 3, select the primary targets from those targets reported to be effective and have the right expression and amplification profile against the clinical trial disease or disease subtype as one of the primary target. Typically, the number of primary targets of multi-target GPCR binders is in the range of 2-3, with a very small number of them has 4-5 primary targets.

2.3 Computational methods for predicting druggable proteins

In **Chapter 1 Section 1.4**, several commonly used computational methods for predicting druggable proteins have been introduced. In **Chapter 5**, I will utilize four *in silico* methods to facilitate target identification. In the following four sections, these approaches will be illustrated in detail, which include: physicochemical property of drug targets identified by machine learning (**Section 2.3.1**); sequence similarity in drug-binding domains (**Section 2.3.2**); structural fold comparison of drug-binding domains (**Section 2.3.3**); and simple system level druggability rules (**Section 2.3.4**).

2.3.1 Physicochemical properties of drug targets identified by machine learning methods

The term “machine learning” refers to algorithms and techniques that allow computers to extract information from past experience. Although it emerges as a separate research field in the early 1980s, the study of machine learning can be traced back to the 1960s¹⁶⁸. Over the past 50 years, various machine learning methods have been developed and applied in a wide spectrum of fields, such as k-nearest neighbor algorithms in text categorization, decision tree methods in pharmaceutical research, artificial neural network in stock market analysis and prediction, support vector machine in bioinformatics and cheminformatics.

Machine learning uses computational and statistical methods to build mathematical models, and makes inference from training samples¹⁶⁹. Machine learning is a branch of artificial intelligence (AI), and it is closely related to statistics and pattern recognition,

since they all study the analysis of data. However, unlike statistics and pattern recognition, machine learning is primarily concerned with algorithmic complexity of computational implementations¹⁷⁰.

In order to be learnt by computational methods, all the samples, or instances, should be represented by feature vectors, which could be categorical, binary or continuous. Machine learning could be divided into two categories: if samples are given with known classes, it is called supervised learning; otherwise, it is called unsupervised learning¹⁷¹. In supervised learning, the learning process is to optimize an objective function and predict the value of the function for any valid input object after having learnt experience from training examples. This category includes well known machine learning methods like k-nearest neighbors, support vector machines, and decision trees. On the other hand, unsupervised learning is never given the answer set, and all the answers are assumed to be latent variable. All data under investigation are allowed to speak for themselves and they are treated evenly. This category includes self organization map and clustering methods.

The machine learning method used in this work is support vector machines (SVM). We used our own developed SVM code. Websites for the freely downloadable codes of SVM and other machine learning methods are given in **Table 02-1**.

2.3.1.1 SVM Algorithm

Support vector machine (SVM) is one of newest members in supervised learning family¹⁷². It was first officially proposed by Vapnik in 1995¹⁷², and then further

explained by Dr. Burges in 1998¹⁷³. A special property of SVM is that it simultaneously minimizes the empirical classification error and maximizes the geometric margin. Over the past 20 years, SVM has been successfully applied to a wide range of real-world problems, including hand-written digit recognition¹⁷⁴, tone recognition¹⁷⁵, image classification¹⁷⁶⁻¹⁷⁹, as well as broad fields in biology, such as protein function prediction^{133,134}, protein-protein interaction prediction¹⁸⁰, protein remote homology detection^{181,182}, and classification for discriminating coronary heart disease patients¹⁸³. SVM is the primary method used in our study, and its algorithm will be discussed with more details in following sections.

2.3.1.1.1 Linear SVM

In two-class problems, SVM aims to separate examples of two classes with the maximum hyper plane (**Figure 02-5**). Mathematically, the data is composed of n examples of two classes, denoted as $\chi = \{(x_1, y_1), \dots, (x_n, y_n)\}$, where $x_i \in R^N$ is a vector in feature space and $y_i \in \{-1, +1\}$ denotes its class. A hyper plane could be drawn to separate examples of one class (positive examples) from those of the other one (negative examples). The hyper plane is represented by $w \cdot x + b = 0$, where w is slope and b is bias. Thus the objective function of SVM changes to minimize Euclidean norm $\|w\|^2$ with following limitations:

$$w \cdot x_i + b \geq +1 \text{ for } y_i = +1 \quad (\text{positive examples}) \quad (1)$$

$$w \cdot x_i + b \leq -1 \text{ for } y_i = -1 \quad (\text{negative example}) \quad (2)$$

According to which side those new instances locate, we can easily determine which class they belong to. So the decision function becomes $f_{w,b}(x) = \text{sign}(\langle w, x \rangle + b)$.

Geometrically, all the points are divided into two regions by a hyper plane H . As shown in **Figure 02-6**, there are numerous ways through which a hyper plane can separate these examples. The objective of SVM is to choose the “optimal” hyper plane. As all new examples are supposed to be located under similar distribution as training examples, the hyper plane should be chosen such that small shifts of data do not result in fluctuations in prediction result. Therefore, the hyper plane that separates examples of two classes should have the largest margin, which is expected to possess the best generalization performance. Such hyper plane is called the Optimal Separating Hyper plane (OSH)¹⁷².

Examples locating on the margins are called support vectors, whose presentation determines the location of the hyper plane. OSH could be thus represented by a linear combination of support vectors. The margin $\gamma_i(w,b)$ of a training point x_i is defined as the distance between H and x_i :

$$\gamma_i(w,b) = y_i(w \cdot x + b) \quad (3)$$

and the margin of a set of vectors $S = \{x_1, \dots, x_n\}$ is defined as the minimum distance between the hyper plane H to all the vectors in S :

$$\gamma_S(w,b) = \min_{x_i \in S} \gamma_i(w,b) = \min_{\{x|y=+1\}} \frac{w \cdot x}{\|w\|} - \max_{\{x|y=-1\}} \frac{w \cdot x}{\|w\|} \quad (4)$$

So the OSH is the solution to the optimization problem^{184,185}:

$$\text{Maximize: } \gamma_x(w, b) \quad (5)$$

Subject to:

$$\gamma_x(w, b) > 0 \quad (6)$$

$$\|w\|^2 = 1 \quad (7)$$

which is an equivalent statement of the problem

$$\text{Minimize: } \frac{1}{2} \|w\|^2 \quad (8)$$

Subject to:

$$w \cdot x_i + b \geq +1 \text{ for } y_i = +1 \quad (9)$$

$$w \cdot x_i + b \leq -1 \text{ for } y_i = -1 \quad (10)$$

This optimization problem could be efficiently solved by the Lagrange method. With the introduction of Lagrangian multipliers $\alpha_i \geq 0 (i = 1, 2, \dots, n)$, one for each of the inequality constraints, we obtain the Lagrangian:

$$L_p(w, b, \alpha) = \frac{1}{2} w^T w - \sum_{i=1}^n \alpha_i [y_i (w \cdot x_i + b) - 1] \quad (11)$$

This is a Quadratic Programming (QP) problem. We would have to minimize $L_p(w, b, \alpha)$ with respect to w , b and simultaneously require that the derivatives of $L_p(w, b, \alpha)$ with respect to the multipliers α_i vanish, $\frac{\partial}{\partial w} L_p(w, b, \alpha) = 0$ and $\frac{\partial}{\partial b} L_p(w, b, \alpha) = 0$

This leads to:

$$w = \sum_{i=1}^n \alpha_i y_i x_i \quad \text{and} \quad \sum_{i=1}^n \alpha_i y_i = 0 \quad (12)$$

By substituting these two equations into equation (11), the QP problem becomes the Wolfe dual of the optimization problem:

$$L_D(w, b, \alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \quad (13)$$

subject to constraints $\sum_{i=1}^n \alpha_i y_i = 0$ and $\alpha_i \geq 0, i = 1, 2, \dots, n$.

The corresponding bias b_0 can be calculated as:

$$b_0 = -\frac{1}{2} \left\{ \min_{\{x|y=+1\}} (w_0 \cdot x) - \max_{\{x|y=-1\}} (w_0 \cdot x) \right\} \quad (14)$$

This QP problem could be efficiently solved through several standard algorithms like Sequential Minimization Optimization¹⁸⁶ or decomposition algorithms¹⁸⁷.

Once w_0 and b_0 are determined, the hyper plane is readily drawn. The points for which $\alpha_i > 0$ are called support vectors, which lie on the margin¹⁷³.

2.3.1.1.2 Nonlinear SVM

Many real-world problems are usually too complicated to be solved with linear classifiers. With the introduction of kernel techniques, input data could be mapped to a higher-dimension space, where a new linear classifier can be used to classify these examples (Figure 02-7).

Let Φ denotes an implicit mapping function from input space to feature space F . Then all the previous equations are transformed by substituting input vector x_i and inner product (x_i, x) with $\Phi(x_i)$ and kernel $K(x_i, x)$ respectively, where

$$K(x_i, x) = \Phi(x_i) \cdot \Phi(x) \quad (15)$$

Equation (13) is then replaced by

$$L_D(w, b, \alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j K(x_i \cdot x_j) \quad (16)$$

subject to constraints $\sum_{i=1}^n \alpha_i y_i = 0$ and $\alpha_i \geq 0$, for $i = 1, 2, \dots, n$. The bias b_0 becomes

$$b_0 = -\frac{1}{2} \left\{ \min_{\{x|y=+1\}} \left[\sum_{SV} \alpha_i y_i K(x_i, x) \right] - \max_{\{x|y=-1\}} \left[\sum_{SV} \alpha_i y_i K(x_i, x) \right] \right\} \quad (17)$$

and the decision function becomes

$$f(x) = \text{sign} \left[\sum_{i=1}^n \alpha_i y_i K(x_i, x) + b_0 \right] = \text{sign} \left[\sum_{SV} \alpha_i y_i K(x_i, x) + b_0 \right] \quad (18)$$

Note that the mapping function Φ is never explicitly computed, which would significantly reduce the computation load. Another advantage is that the feature space may be infinitely dimensional, such as in the case of Gaussian kernel¹⁸⁸, where mapping function cannot be explicitly represented. A function could be used as a kernel function if and only if it satisfies Mercer's condition¹⁸⁹. Followings are well-known kernel functions:

Polynomial $k(x, z) = (\langle x, z \rangle + 1)^p$

Sigmoid $k(x, z) = \tanh(\kappa \langle x, z \rangle - \delta)$

Radial basis function (RBF) $k(x, z) = \exp(-\|x - z\|^2 / 2\sigma^2)$

In this work, RBF kernel is used due to its many advantages demonstrated in previous studies. Different SVM models could be developed by using different σ values. It is thus necessary to scan a number of σ values to find the best model, which is evaluated by their performance on classification tasks. In our work, SVM models with σ value in the range of 1~100 were developed for each classification task. **Figure 02-8** illustrates the schematic diagrams of the process of training and prediction of drug targets by SVM. Sequence-derived feature h_i , p_i , v_i ... represents such structural and physicochemical properties as hydrophobicity, polarizability, and volume.

2.3.1.1.3 Performance evaluation

The performance evaluation aims to find out whether an algorithm is able to be applied to novel data that have not been used to develop the prediction model, or measure the generalization capacity to recognize new examples from the same data domain¹⁹⁰.

In this study, several statistical measurements were explored, including sensitivity (SE), specificity (SP), positive prediction value (PPV), and overall prediction accuracy (Q).

The formulas to calculate these measurements are listed as following:

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

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

$$PPV = TP / (TP + FP)$$

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

where TP, FN, TN, and FP represent correctly predicted positive data, positive data incorrectly predicted as negative, correctly predicted negative data, and negative data incorrectly predicted as positive respectively. Another measurement named as Matthews correlation coefficient (MCC) was also used to evaluate the randomness of the prediction.

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

where MCC is within the range of -1 to 1. Negative values of MCC indicate the disagreement between prediction and measurement, while positive values of MCC indicates the agreement between prediction and measurement. A zero value means the prediction is no better than random guess.

2.3.1.2 Structural and physicochemical descriptors of proteins

A number of descriptors have been introduced to represent protein^{133,180,191-195}, the post-translational modifications and localization features¹⁹⁶⁻¹⁹⁸. The sequence-derived

structural and physicochemical descriptors include amino acid composition, dipeptide composition, sequence autocorrelation descriptors, sequence coupling descriptors, and descriptors for composition, transition and distribution of hydrophobicity, polarity, polarizability, charge, secondary structures, surface tension, and Van der Waals volume. Servers such as **PROFEAT**¹⁹⁹ (<http://jing.cz3.nus.edu.sg/cgi-bin/prof/prof.cgi>) and the **ProtParam**²⁰⁰ (<http://au.expasy.org/tools/protparam.html>) have also appeared for facilitating the computation of these descriptors, and other sequence derived features such as the cleavage sites, the nuclear export signals, and the subcellular localization can be computed from CBS Prediction (<http://www.cbs.dtu.dk/services/>).

Three groups of widely-used protein descriptors were combined into a single set of protein descriptors for predicting targets. The first group is physicochemical descriptors used for predicting druggable proteins²⁰¹ that include amino acid composition and the composition, transition and distribution of such structural and physicochemical properties as hydrophobicity, polarity, polarizability, charge, secondary structures, surface tension, and normalized Van der Waals volumes²⁰², the second is normalized Moreau-Broto autocorrelation, and the third is pseudo amino acid descriptors.

2.3.1.2.1 Amino acid composition and composition, transition and distribution

Each feature vector is constructed from the encoded representations of tabulated residue properties including amino acid composition, hydrophobicity, normalized Van der Waals volume, polarity, polarizability, charge, surface tension, secondary structure and solvent accessibility^{180,202}. For each of these properties, amino acids are divided into three groups such that those in a particular group are regarded to have the same property. For

instance, amino acids can be divided into hydrophobic (CVLIMFW), neutral (GASTPHY), and polar (RKEDQN) groups. The groupings of amino acids for each of the properties are given in **Table 02-2**. Three descriptors, composition (C), transition (T), and distribution (D), are used to describe global composition of each of the properties. C is the number of amino acids of a particular property (such as hydrophobicity) divided by the total number of amino acids in a protein sequence. T characterizes the percent frequency with which amino acids of a particular property is followed by amino acids of a different property. D measures the chain length within which the first, 25%, 50%, 75% and 100% of the amino acids of a particular property is located respectively.

Overall, there are 21 elements representing these three descriptors: 3 for C, 3 for T and 15 for D^{180,202}. The feature vector of a protein is constructed by combining the 21 elements of all of these properties and the 20 elements of amino acid composition in sequential order. In this study, totally 188 elements are used as feature vector for each protein shown in **Table 02-3**. The following is a hypothetical protein sequence:

AEAAAEAEAAAEAAAEAEAAAEAAAEAEAAAE

As shown in **Figure 02-9**, which has 16 alanines ($n_1=16$) and 14 glutamic acids ($n_2=14$).

The compositions for these two amino acids are:

$$n_1 \times 100.00 / (n_1 + n_2) = 53.33 \text{ and } n_2 \times 100.00 / (n_1 + n_2) = 46.67 \text{ respectively}$$

There are 15 transitions from A to E or from E to A in this sequence and the percent frequency of these transitions is $(15/29) \times 100.00 = 51.72$. The first, 25%, 50%, 75% and 100% of alanines are located within the first 1, 5, 12, 20, and 29 residues respectively.

The D descriptor for alanines is therefore $1/30 \times 100.00=3.33$, $5/30 \times 100.00=16.67$, $12/30 \times 100.00=40.0$, $20/30 \times 100.00=66.67$, $29/30 \times 100.00=96.67$. Likewise, the D descriptor for glutamic is 6.67, 26.67, 60.0, 76.67, and 100.0. Overall, the amino acid composition descriptors for this sequence are C=(53.33, 46.67), T=(51.72), and D=(3.33, 16.67, 40.0, 66.67, 96.67, 6.67, 26.67, 60.0, 76.67, 100.0) respectively. Descriptors for other properties can be computed by a similar procedure. Table 2-4 gives the computed descriptors of the cellular tumor antigen p53 (Swiss-Prot AC P04637). The feature vector of a protein is constructed by combining all of the descriptors in sequential order.

2.3.1.2.2 Normalized Moreau-Broto autocorrelation

Normalized Moreau-Broto autocorrelation features describe the level of correlation between two protein sequences in terms of their specific structural or physicochemical property²⁰³, which are defined based on the distribution of amino acid properties along the sequence²⁰⁴. There are eight amino acid properties used for deriving these autocorrelation descriptors. The first is hydrophobicity scale derived from the bulk hydrophobic character for the 20 types of amino acids in 60 protein structures²⁰⁵. The second is the average flexibility index derived from the statistical average of the B-factors of each type of amino acids in the available protein x-ray crystallographic structures²⁰⁶. The third is the polarizability parameter computed from the group molar refractivity values originally provided by Hansch et al²⁰⁷. The fourth is the free energy of amino acid solution in water measured by Hutchins²⁰⁷. The fifth is the residue accessible surface areas taken from average values from folded proteins²⁰⁸. The sixth is the amino acid residue volumes measured by Fisher²⁰⁹. The seventh is the steric parameters derived

from the van der Waals radii of amino acid side-chain atoms²¹⁰. The eighth is the relative mutability obtained by multiplying the number of observed mutations by the frequency of occurrence of the individual amino acids²¹¹. Each of these properties is centralized and standardized such that $P_r' = (P_r - \bar{P}) / \sigma$, where \bar{P} is the average of the property of the 20 amino acids, \bar{P} and σ are given by:

$$\bar{P} = \frac{\sum_{r=1}^{20} P_r}{20} \text{ and } \sigma = \sqrt{\frac{1}{20} \sum_{r=1}^{20} (P_r - \bar{P})^2}$$

Moreau-Broto autocorrelation is given by:

$$AC(d) = \sum_{i=1}^{N-d} P_i P_{i+d} \quad 203,212$$

which has been used for predicting transmembrane protein types²¹³ and protein secondary structural contents²¹⁴ at accuracy levels of 82%~94% and 91%~94% respectively. Here d is the lag of the autocorrelation, P_i and P_{i+d} are the amino acid property at position i and $i+d$ respectively. The normalized Moreau-Broto autocorrelation is defined as:

$$ATS(d) = AC(d) / (N - d) \text{ where } d=1, 2, 3 \dots 30.$$

2.3.1.2.3 Pseudo amino acid

Pseudo amino acid descriptor is made up of a 50-dimensional vector in which the first 20 components reflect the effect of the amino acid composition and the remaining 30 components reflect the effect of sequence order, only now, the coupling number τ_d is now

replaced by the sequence order correlation factor θ_λ ²¹⁵. The set of sequence order correlated factors is defined as follows:

$$\theta_\lambda = \frac{1}{N - \lambda} \sum_{i=1}^{L-\lambda} \Theta(R_i, R_{i+\lambda})$$

where θ_λ is the first-tier correlation factor that reflects the sequence order correlation between all of the λ -most contiguous residues along a protein chain ($\lambda=1, \dots, 30$) and N is the number of amino acid residues. $\Theta(R_i, R_j)$ is the correlation factor and is given by:

$$\Theta(R_i, R_j) = \frac{1}{3} \left\{ [H_1(R_j) - H_1(R_i)]^2 + [H_2(R_j) - H_2(R_i)]^2 + [M(R_j) - M(R_i)]^2 \right\}$$

where $H_1(R_i)$, $H_2(R_i)$ and $M(R_i)$ are the hydrophobicity, hydrophilicity, and side-chain mass of amino acid R_i , respectively. Before being substituted in the above equation, the various physicochemical properties $P(i)$ are subjected to a standard conversion.

$$P(i) = \frac{P^0(i) - \sum_{i=1}^{20} \frac{P^0(i)}{20}}{\sqrt{\frac{\sum_{i=1}^{20} \left[P^0(i) - \sum_{i=1}^{20} \frac{P^0(i)}{20} \right]^2}{20}}}$$

This sequence order correlation definition introduces more correlation factors of physicochemical effects as compared to the coupling number, and has been shown to be an improvement on the way sequence order effect information is represented. Thus, for each amino acid type, the first part of the vector is defined as:

$$X_r = \frac{f_r}{\sum_{r=1}^{20} f_r + w \sum_{d=1}^{30} \theta_d}$$

where $r = 1, 2, \dots, 20$, f_r is the normalized occurrence of amino acid type i and w is a weighting factor ($w = 0.1$), and the second part is defined as:

$$X_d = \frac{w\theta_{d-20}}{\sum_{r=1}^{20} f_r + w \sum_{d=1}^{30} \theta_d}$$

2.3.1.3 Computational implementation

In this work, a nonlinear SVM with kernel function:

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|\mathbf{x}_j - \mathbf{x}_i\|^2 / 2\sigma^2}$$

was used. SVM prediction system was developed using feature vectors of physicochemical and structural properties of 348 successful targets and it was used to screen the clinical trial and research targets for identifying potential promising targets by the procedures illustrated in **Figure 02-8**.

2.3.1.4 Sources of druggable and non-druggable proteins

Sufficiently diverse sets of druggable and non-druggable proteins are needed for training and testing a SVM prediction model. There are 1,894 successful (commercialized) and research targets in the therapeutic target database²¹⁶ with available sequence information, which together form the druggable class. Some viral and microbial targets have multiple

sequence entries, as there are significant sequence variations across strains. Based on their family distribution pattern, targets are expected to be represented by < 800 protein families including 460 covered by the known targets^{58,217}. There are 11,912 protein families in the protein family PFAM database²¹⁸ that contain no known target at present. Protein families in PFAM database are defined based on domain affiliations or sequence clustering. Therefore, without substantially reducing SVM prediction performance, putative non-druggable proteins can be tentatively derived from these non-target families, which produces a maximum possible “wrong” family representation rate of <7% even when all of the < 340 unidentified target families are misplaced²⁰¹. Representative proteins of these non-target families form the non-druggable class. Importantly, inclusion of the representative of a “wrong” family into the non-druggable class does not preclude other family members from being classified as druggable. Statistically, a substantial percentage of druggable members can be located on the druggable side of the SVM hyper-plane even if its family representative is on the non-druggable side. Therefore, in principle, a reasonably good SVM prediction model can be derived from these putative non-druggable proteins for predicting druggable ones rather than PFAM family members, as confirmed by the case studies described. The quality of the non-druggable class and performance of SVM can be further improved along with the discovery of new targets.

2.3.2 Method for analyzing sequence similarity between the drug-binding domain of a studied target and that of a successful target

Sequence similarity between the drug-binding domain of a studied target and that of a successful target is obtained by using BLAST²¹⁹ to scan the sequence of the drug-binding

domain of the studied target against those of 168 successful targets with identifiable drug-binding domain. We used BLAST program downloaded from the National Centre for Biotechnology Information (NCBI) (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>). A stricter BLAST cut-off, E-value=0.001, was used for determining the similarity domains. The drug-binding domain of the successful target with the smallest E-value that is smaller than 0.001 was selected. This E-value has been reported to give reliable predictions of the homologous relationships²²⁰ and it 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 threshold²²¹. The majority of protein pairs that share 40~50% (or higher) sequence-identity differ by <1 Å RMS deviation^{222,223}, and a larger structural deviation probably alters drug-binding properties. Therefore, the adopted E-value seems to be reasonable for selecting similarity protein domains relevant to the binding of a common set of drug-like molecules. None-the-less, low percentages of protein pairs of higher sequence-identity have been found to differ by larger RMS deviations²²³ and some protein pairs of lower sequence-identity might also have high structural similarity, which probably affects the accuracy of our analysis to some extent.

2.3.3 Comparative study of structural fold of the drug-binding domains of studied and successful targets

The rationale and procedure for comparative study of structural fold of the drug-binding domain of two proteins are similar to general strategies adopted by popular structural fold databases such as SCOP²²⁴, and are described below:

2.3.3.1 The ligand-sensing core of domain

The ligand-binding or catalytic sites are the most relevant subsets of a domain from the point of view of development of small-molecule binders, which are normally located within the so-called ligand-sensing core of the domain where the actual catalytic conversion of enzymes, or the binding event of small-molecule ligand, occurs. It has been suggested to confine structural similarity considerations to these distinct parts of a domain and grouping ligand-sensing cores, instead of whole domains, according to 3D similarities into so-called protein structure similarity clusters²²⁵.

2.3.3.2 Protein structure similarity clustering

Clustering of ligand-sensing or catalytic cores of two domains is based on visual inspection and structural superimposition and alignment tools in SYBYL (SYBYL® 6.7 Tripos Inc., St. Louis, Missouri, USA). and Insight II (Insight II® Accelrys Software Inc, San Diego, CA) following the same procedure used for generating SCOP structural folds²²⁴.

2.3.4 Simple system-level druggability rules

2.3.4.1 Druggability rules

Based on the systems characteristics of therapeutic targets described in earlier studies^{87,226,227}, systems-level druggability rules have been proposed for guiding the search of druggable proteins²²⁷, a revised version is as follows:

1) Protein preferably has less than 15 human similarity proteins outside its family. While existence of a larger number of human similarity proteins doesn't rule it out as a druggable protein, it generally increases the chance of un-wanted interferences and thus the level of difficulty for finding viable drugs. (78% of the successful targets with identifiable drug-binding domain have less than 15 human similarity proteins).

2) Protein is preferably involved in no more than three pathways in human. While association with a larger number of human pathways doesn't rule it out as a druggable protein, it generally increases the chance of un-wanted interferences with other human processes and thus the level of difficulty for finding a viable target. (87% of the successful targets with pathway information are associated with no more than 3 pathways).

3) For organ or tissue specific diseases, protein is preferably distributed in no more than five tissues in human. While distribution in a higher number of tissues doesn't rule it out as a druggable protein, it generally increases the chance of un-wanted interferences with other tissues unless the disease-relevant targets are located within blood vessels or cells lining the arteries where they have higher priority to bind drugs than targets in other tissues. The un-wanted interferences increase the level of difficulty for finding a viable target. (79% of the successful targets with tissue distribution information are distributed in no more than 5 tissues).

In this work, these rules were applied to those studied targets with sufficient information about their systems-related profiles for identifying potential promising targets.

2.3.4.2 Number of human similarity proteins of a target

Human similarity proteins of a target are those human proteins whose drug-binding domain is similar to that of the studied target by using the same BLAST method²¹⁹ as that described in Section 1 of this supplementary material.

2.3.4.3 Number of affiliated human pathways of a target

Information about the affiliated pathways of a target was obtained from KEGG database²²⁸ (see: <http://www.genome.jp/kegg/>).

2.3.4.4 Number of human tissues of a target is distributed

In estimating the number of human tissues where each target is distributed, relevant data from the Swissprot database were used. We were able to find the published literature for 92% of these data, and a random check of these publications confirmed the quality of the data. We also used the level-4 tissue-distribution data from another database, TissueDistributionDBs (http://genome.dkfz-heidelberg.de/menu/tissue_db/index.html), to derive the tissue distribution pattern of the same set of 158 successful targets. A target is assumed to be primarily distributed in a tissue if no less than 8% of the total protein contents are distributed in that tissue. Approximately 28, 24, 19, 10, 6, 6, 5, and 1% of these targets were found to be affiliated with 1 to 8 tissues, respectively, which are roughly similar to those derived from Swissprot data²²⁷, although the definition and content of these databases are somehow different. Thus, our estimated tissue distribution profiles are quite stable even though the exact percentages may differ by some degrees.

Figure 02- 1 The hierarchical data model

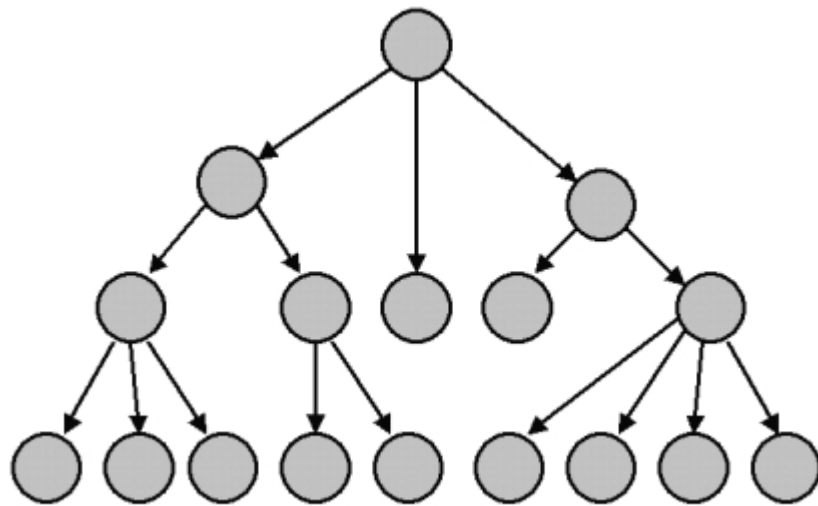


Figure 02- 2 The network data model

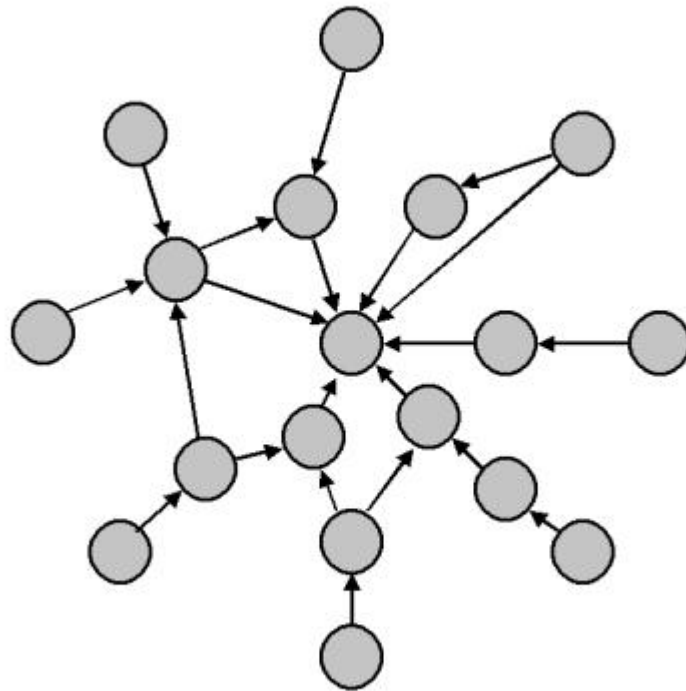


Figure 02- 3 The relational data model

	Data item 1	Data item 2	Data item 3
Record 1				
Record 2				
Record 3				
.....				

Figure 02- 4 Logical view of the database

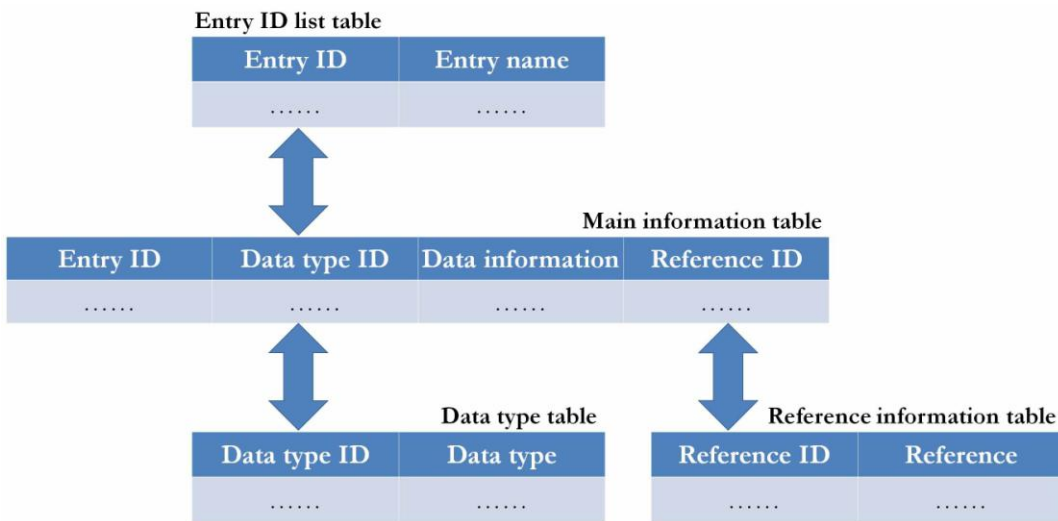


Figure 02- 5 Architecture of support vector machines

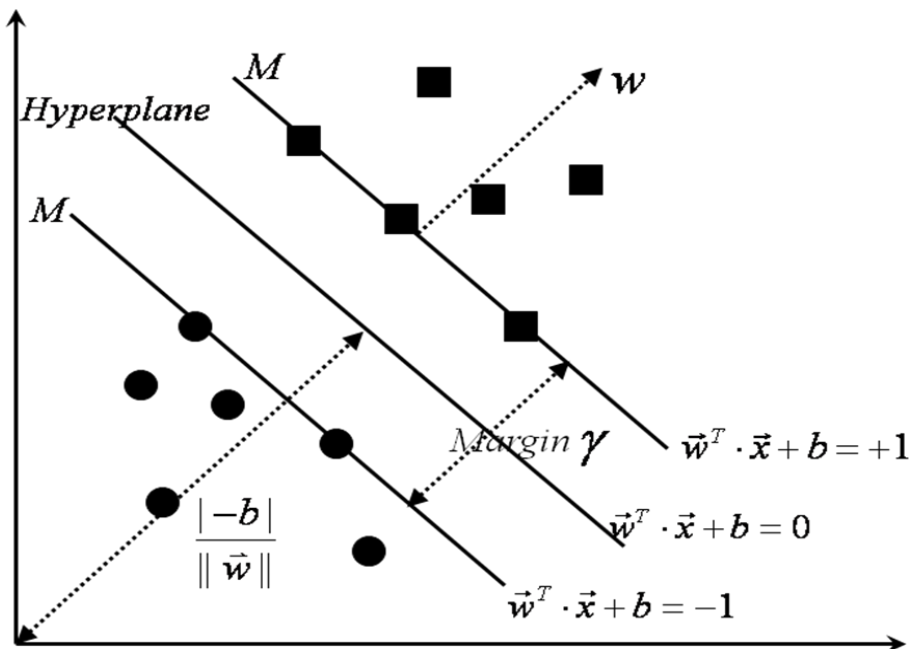


Figure 02- 6 Different hyper planes could be used to separate examples

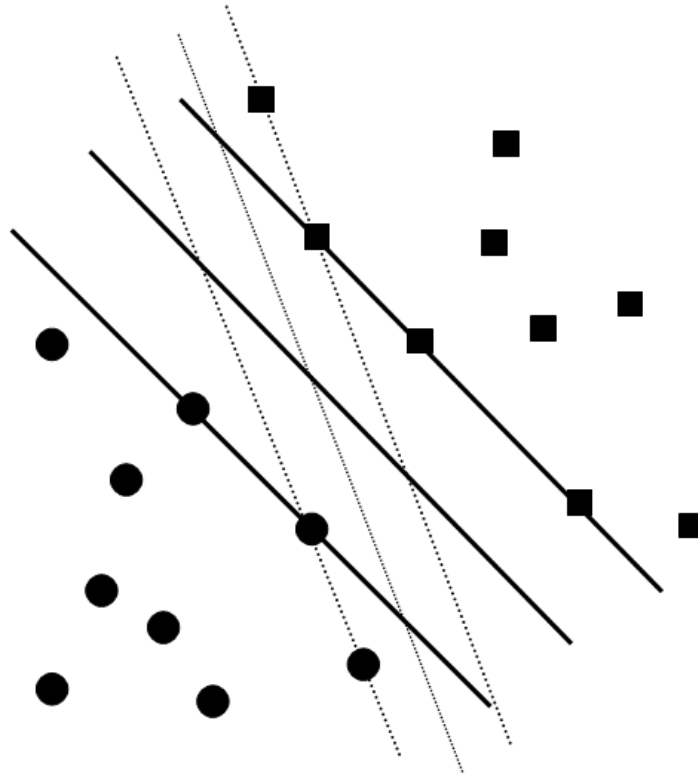


Figure 02- 7 Mapping input space to feature space

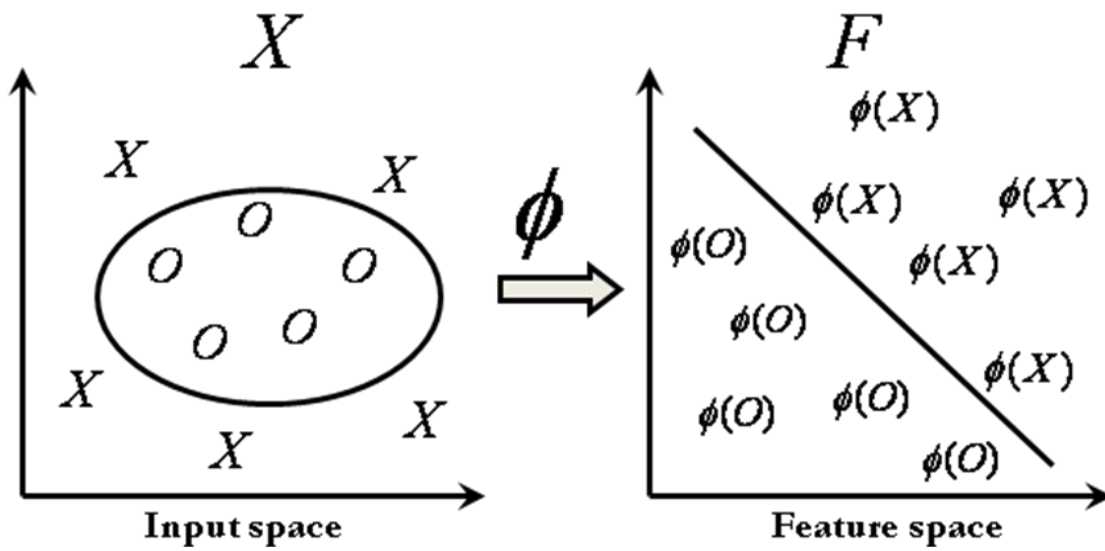


Figure 02- 8 Diagrams of the process for training and predicting targets

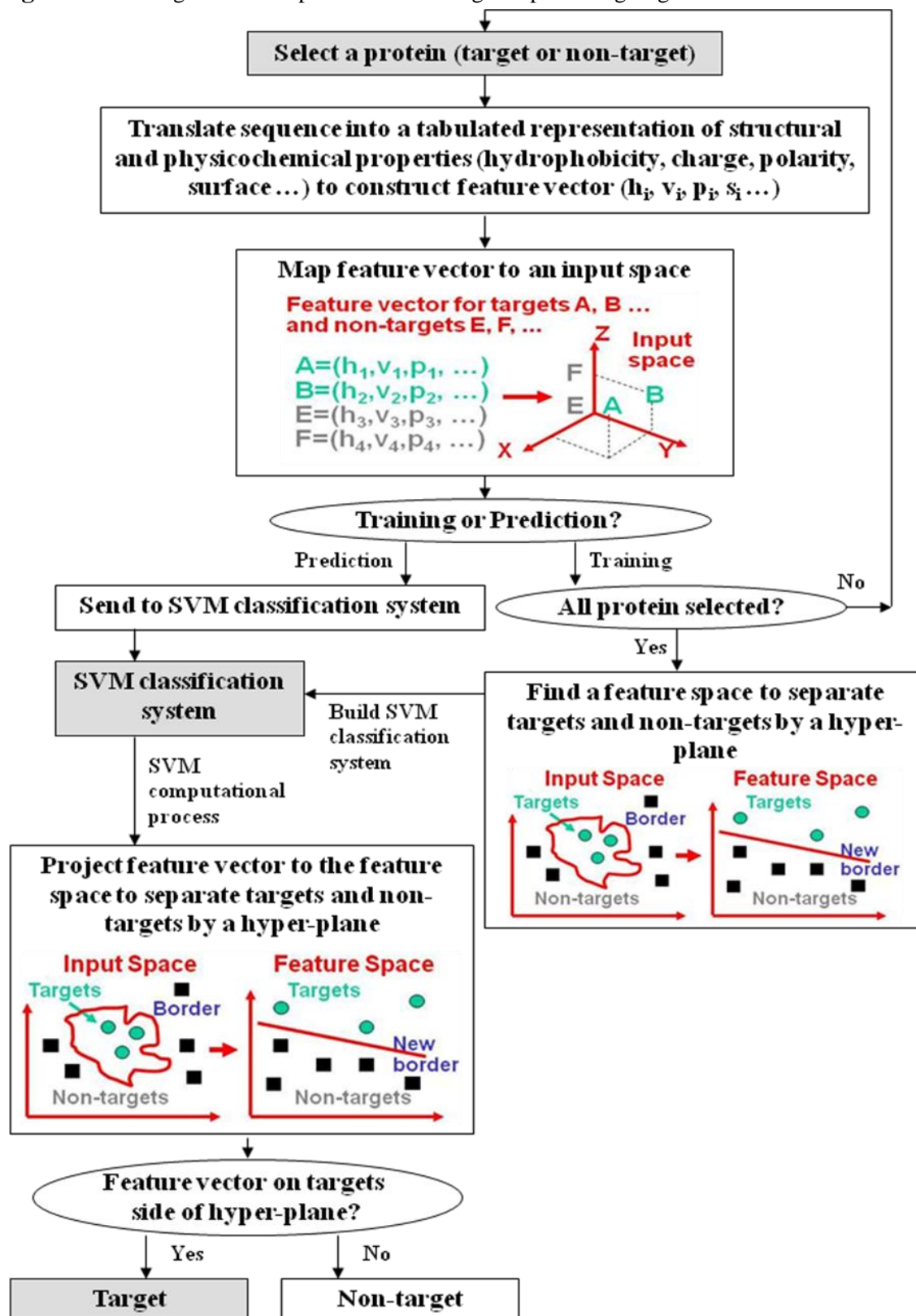


Figure 02- 9 Illustration of derivation of the feature vector*

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

* Sequence index indicates the position of an amino acid in the sequence. The index for each type of amino acids in the sequence (A or E) indicates the position of the first, second, third, ... of that type of amino acid (The position of the first, second, third, ..., A is at 1, 3, 4, ...). A/E transition indicates the position of AE or EA pairs in the sequence.

Table 02- 1 Websites that contain freely downloadable codes of machine learning methods

Machine learning	Program package	Web link for program download
Decision Tree	PrecisionTree	http://www.palisade.com.au/precisiontree/
	DecisionPro	http://www.vanguardsw.com/decisionpro/jdtree.htm
	C4.5	http://www2.cs.uregina.ca/~hamilton/courses/831/notes/ml/dtrees/c4.5/tutorial.html
	C5.0	http://www.rulequest.com/download.html
KNN	k Nearest Neighbor demo	http://www.cs.cmu.edu/~zhuxj/courseproject/knndemo/KNN.html
	PERL Module for KNN	http://aspn.activestate.com/ASPN/CodeDoc/AI-Categorize/AI/Categorize/kNN.html
	Java class for KNN	http://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/classify/old/KNN.html
Neural Network	BrainMaker	http://www.calsci.com/
	Libneural	http://pochat.online.fr/webus/tutorial/BPN_tutorial7.html
	fann	http://leenissen.dk/fann/
	NeuralWorks Predict	http://www.neuralware.com/products.jsp
	NeuroShell Predictor	http://www.mbaware.com/neurpred.html
SVM	SVM light	http://svmlight.joachims.org/
	LIBSVM	http://www.csie.ntu.edu.tw/~cjlin/libsvm/
	mySVM	http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/index.html
	SMO	http://www.datalab.uci.edu/people/xge/svm/
	BSVM	http://www.csie.ntu.edu.tw/~cjlin/bsvm/

Table 02- 2 Division of amino acids into 3 different groups by different physicochemical properties

Property		Group 1	Group 2	Group 3
Hydrophobicity	Type	Polar	Neutral	Hydrophobic
	Amino acids	RKEDQN	GASTPHY	CVLUMFW
Van der Waals volume	Value	0~2.78	2.95~4.0	4.43~8.08
	Amino acids	GASCTPD	NVEQIL	MHKFRYW
Polarity	Value	0~0.456	0.6~0.696	0.792~1.0
	Amino acids	LIFWCMVY	PATGS	HQRKNE
Polarizability	Value	0~0.108	0.128~0.186	0.219~0.409
	Amino acids	GASDT	CPNVEQIL	KMHFRYW
Charge	Type	Positive	Neutral	Negative
	Amino acids	KR	ANCQGHILMFPSTWYV	DE
Surface tension	Value	-0.20~0.16	-0.3~ -0.52	-0.98~ -2.46
	Amino acids	GQDNAHR	KTSEC	ILMFPWYV
Secondary structure	Type	Helix	Strand	Coil
	Amino acids	EALMQKRH	VIYCWFT	GNPSD
Solvent accessibility	Type	Buried	Exposed	Intermediate
	Amino acids	ALFCGIVW	RKQEND	MPSTHY

Table 02- 3 List of features for proteins

Feature Description	Number of Dimensions
Amino acid composition	20
Hydrophobicity	21
Van der Waals volume	21
Polarity	21
Polarizability	21
Charge	21
Surface tension	21
Secondary structure	21
Solvent accessibility	21
Total	188

Table 02- 4 Characteristic descriptors of cellular tumor antigen p53

Property	Elements of Descriptors									
	A	C	D	E	F	G	H	I	K	L
Amino acid composition	6.11	2.54	5.09	7.63	2.80	5.85	3.05	2.04	5.09	8.14
	M	N	P	Q	R	S	T	V	W	Y
	3.05	3.56	11.45	3.82	6.62	9.67	5.60	4.58	1.02	2.29
Hydrophobicity	31.81	44.02	24.17	26.02	16.58	19.39	0.51	33.33	58.02	81.17
	100.0	1.02	22.39	43.26	74.55	99.75	0.25	24.68	46.31	65.39
	97.96									
Van der waals volume	46.31	29.77	23.92	23.98	17.10	14.29	1.02	20.87	42.24	70.74
	100.0	0.51	24.68	51.14	72.77	98.73	0.25	34.10	59.29	81.68
	98.22									
Polarity	26.46	38.68	34.86	18.62	19.90	24.23	0.25	27.74	47.84	64.89
	97.96	1.02	21.12	39.44	74.55	99.75	0.51	34.61	58.02	81.17
	100.0									
Polarizability	32.32	43.77	23.92	28.57	13.52	17.86	1.53	22.40	46.82	76.84
	100.0	0.51	19.59	48.35	69.97	99.24	0.25	34.10	59.29	81.68

Chapter 3 Pharmainformatics databases construction

Three pharmainformatics databases have been constructed and described in detail in this thesis. They are Therapeutic Targets Database (TTD), Information of Drug Activity Data (IDAD), and Target Validation Database (TVD). TTD was first constructed in 2002 as a pioneer for providing pharmaceutical information on therapeutic target. After progress in the past 8 years on target discovery, TTD still acts as reliable knowledge base providing information on successful therapeutic target. However, the profile of drugs under clinical developing keeps changing in the past decade, and many new drugs have been approved for acting on novel targets. So it is time to update information into TTD by adding novel targets approved and identifying clinical trial targets. The identification of primary targets for approved, clinical trial, and experimental drugs partly relies on literature data mining, but sometimes target for a drug is not clearly indicated. In both situations, the validation of the target is very critical not only for double confirm the reliability of information got from those reputable journal but also for identifying reliable primary target for drugs with no target information provided. Thus, IDAD for evaluating drug potency on its target(s) and TVD for validating primary target(s) for drugs have been developed. In **Section 3.1**, therapeutic target database 2010 update has been shown. I have spent most of this chapter to demonstrate its data structure and new features. In **Section 3.2** and **Section 3.3**, IDAD and TVD have been introduced respectively. Although there are three databases, they are interrelated to each other and aiming at the same goal to provide useful information for modern target, and finally contribute to drug, discovery.

3.1 Therapeutic targets database, 2010 update

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 (like membrane) targets^{6,34}. 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)^{53,56,58,59}. Rapid advances in genomic, proteomic, structural, functional and systems studies of the known targets and other disease proteins^{79,81,229-233} enable the discovery of drugs, multi-target agents, combination therapies and new drug targets^{56,58,81,234,235}, analysis of on-target toxicity²³⁶ and pharmacogenetic responses²³⁷, and development of discovery tools²³⁸⁻²⁴¹.

To facilitate the access of therapeutic targets information, publicly accessible databases such as Drugbank⁶⁶, Potential Drug Target Database (PDTD)¹⁶⁴ and our own Therapeutic Target Database (TTD)²¹⁶ have been developed. As illustrated in **Chapter 1 Section 1.2.2**, a detail list of these therapeutic target related databases is provided (**Table 01-4**). 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 targets⁶⁶. PDTD contains active-sites as well as functional information for the potential targets with available 3D structures¹⁶⁴ in PDB. TTD provides information about the primary therapeutic targets of a comprehensive set of both approved and experimental drugs²¹⁶.

While drugs and agents typically modulate the activities of multiple proteins²⁴² and up to 14,000 drug-targeted-proteins have been published⁵⁵, the reported number of primary targets directly related to the therapeutic actions of approved drugs is limited to 324⁵⁹. 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^{56,81,234,236,237,240}. 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 targets⁵⁹ 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>.

3.1.1 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@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 comprehensively 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. 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 were 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 were 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 literatures of reported IC_{50}/EC_{50} values against the target/targets and cell-lines and the reports of *in vivo* studies to confirm that the reported primary targets are accurate. For detailed information about how primary target is identified, please refer back to **Chapter 2** methodology **Section 2.2**.

3.1.2 Ways to access therapeutic targets database

TTD data can be accessed by both whole database (**Figure 03-1**) and customized (**Figure 03-2**) keyword search, and by target sequence similarity (**Figure 03-3**) and drug Tanimoto similarity search (**Figure 03-4**). Full TTD data download is also provided (**Figure 03-5**). 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 are 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.

After input keywords and search in TTD database, the intermediate searching results will be displayed for user to choose from. For example, if we input “Dopamine receptor” into

the search box–“List search results by targets” in the home page, the intermediate search results page (**Figure 03-6**) will display *Dopamine D1 receptor*, *D(1B) dopamine receptor*, *D(2) dopamine receptor*, *D(3) dopamine receptor*, and *D(4) dopamine receptor* for users to make further selection. Another example is: if we input “influenza virus infection” into the search box–“List search results by drugs”, the intermediate results page (**Figure 03-7**) will display approved drugs *Oseltamivir*, *Zanamivir*, *Rimantadine*, phase III clinical trial drugs *Peramivir* and *CS-8958*. In the intermediate search results page, hyper-links linking to detailed target or drug information pages are provided.

Target detail information page (**Figure 03-8**) 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.

Drug detail information page (**Figure 03-9**) 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. Further drug information can be accessed via cross links to the external databases, such as PubChem, DrugBank, SuperDrug, and ChEBI

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 tool⁹⁴. Similarity drugs of an input drug structure can be searched by using molecular descriptor based Tanimoto similarity searching method^{243,244}. 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 (**Figure 03-5**). 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.

3.1.3 Target and drug similarity searching

Target similarity search (**Figure 03-3**) is based on BLAST⁹⁴ 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 (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>) is used for downloading BLAST program. The result of similarity targets searched out are ranked by E-value and BLAST score⁹⁴. E-value has been reported to give reliable predictions of the homologous relationships²²⁰ 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 threshold²²¹. The majority of protein pairs sharing ~50% (or higher) sequence-identity differ by < 1 Å RMS deviation^{222,223}. A larger structural deviation alters drug-binding properties probably.

Drug similarity search (**Figure 03-4**) is based on the Tanimoto similarity search method²⁴³. An input compound structure in MOL or SDF format is converted into a vector composed of molecular descriptor by using MODEL²⁴⁵. 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^{246,247}. Based on the results of our earlier studies²⁴⁴, a total of 98 1D and 2D descriptors are 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* then compared to drug *j* in TTD by using the Tanimoto coefficient $sim(i,j)$ ²⁴³:

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

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^{248,249}. 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^{79,81,229-233}. Apart from establishing crosslink to the emerging sources, some of the profiles extracted or derived from the relevant data⁵⁸ may be further incorporated into TTD. Target data has been used for developing target discovery methods²³⁸⁻²⁴⁰, and some of these methods may be included in TTD in addition to the BLAST tool for similarity target searching. As in the case of PDTD¹⁶⁴, some of the virtual screening methods and datasets may also be included in TTD for facilitating target oriented drug lead discovery.

3.2 Information of Drug Activity Data

The initial idea of building a drug activity database is to provide activity information for the primary targets of drugs and clinical trials compounds in TTD. With the development of this database, we feel that the scope shall not be limited only to drugs and clinical trials compounds. Compounds like natural products, promising compounds developed by the pharmaceutical companies as lead compound or preclinical candidates shall be included too. Currently, there are several similar databases that provide activity information for compounds, like BindingDB⁶¹, DrugBank⁶⁶, MDDR²⁵⁰, *et al.* (**Table 03-1**). 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. As compared to those databases, IDAD is mainly focusing on *in vitro* activity of drugs, clinical trial compounds and preclinical candidates while BindingDB collects data of all kinds of compounds binding to the targets, which are not limited to therapeutic targets. IDAD can currently be accessed at http://bidd.nus.edu.sg/group/IDAD/IDAD_Home.asp.

3.2.1 The data collection of IDAD information

Information collection was conducted by literature search on PubMed database by using keyword combinations of “therapeutic” and “target”, “drug” and “target”, “clinical trial”

and “drug”, and “clinical trial” and “target”, and by comprehensive search of such review journals as *Journal of medicinal chemistry*, *Journal of European journal of medicinal chemistry*, *Current topics in medicinal chemistry*, *Nature Reviews Drug Discovery*, *Trends of Pharmaceutical Science*, *Drug Discovery Today*, *Oncogene* and so on. In particular, these searches identified 198 recent papers reporting approved and clinical trial drugs and their targets.

3.2.2 The construction of IDAD database

IDAD is a relational database, which represents the drug-target interaction database in the form of two-dimension tables. The two-dimensional tables include IDAD ID-Drug Name pair ID table, IDAD ID-Activity ID pair main information table, Activity ID, Protein ID, Activity, Normalized Activity, Reference ID table, Protein ID–TTDID and Swiss-Prot ID information table and Reference information table. In these tables, IDAD serves as primary key; Activity ID, Protein ID, reference ID are considered as foreign keys. TTDID and Swiss-Prot ID are used to cross-link to external database like TTD and Swiss-Prot.

3.2.3 Way to accession IDAD database

Entries of IDAD are searchable by several methods. These methods include the search by compound name or ID, search by target. Case-insensitive keyword-based text search and wildcards are also supported. In a query, one can specify full name or part of the name in a text field. For instance, wild characters of '*' and '?' are allowed in the text field. In this case, '?' represents any single character, and '*' represents a string of characters of any

length. As an example, input of 'HDAC' in the field of target name enables the search of all entries containing the target name of 'HDAC' such as HDAC1, HDAC8, HDAC4, etc. In IDAD interface, all entries that satisfy search criteria are listed along with IDAD ID, target name, activity, and references. More detailed information of a compound can be obtained by clicking the corresponding TTD target ID and TTD drug ID. For a systematic comparison of compound activities, all activity values are normalized. For completeness, the relevant references are provided in the interface.

In summary, IDAD is designed to provide activity information for approved drugs, drugs in clinical trial, and important experimental agents, such as the natural product, promising compounds developed by the pharmaceutical companies as lead compound or preclinical candidates. This information will act as an informative data source to support research in pharmaceutical sciences.

3.3 Therapeutic targets validation database

In the development of TTD, the most critical job is to identify primary therapeutic targets for approved drugs, drugs in clinical trial, and experimental agents. In our analysis, the primary targets and their corresponding drugs/agents were initially collected from the company websites and publications in reputable journals, which explicitly mentioned the targets and their corresponding drugs. These drug targets were expected to be well defined based on solid *in vitro* and *in vivo* target validation studies. Because of this, we came up with the idea of constructing Target Validation Database (TVD) for collecting supporting information of the primary target(s) for approved drugs, drugs in clinical trial, and the experimental agents (such as natural products, promising compounds developed by the pharmaceutical companies as lead compound or preclinical candidates).

3.3.1 Pharmaceutical demands for target validation information

As illustrated in **Chapter 1 Section 1.1.4**, target validation is important in selecting right targets for drug discovery, which evaluates multiple profiles including the expression and relevance of targets in disease models, the potency of drugs in modulating target activity and disease model, and the correlation of these activities to the claimed therapeutic effect. Therefore, target validation data, particularly those of successful and clinical trial targets, is expected to be invaluable data which provide historical “model” for facilitating target discovery, validation and analysis. Our TVD aims at collecting *in vitro* and *in vivo* target validation data for therapeutic targets covered by TTD. These data include the potencies of drugs against their efficacy targets, and potencies of drugs against the disease relevant

cell-lines expressing these targets (potencies are measured in IC₅₀, Ki, and EC₅₀), and the effects of target knock-out or variation in target sequence, expression and activity in disease models. TVD can be accessed at <http://bidd.nus.edu.sg/group/TVDtest/TVD.asp>.

3.3.2 The data collection of TVD information

Therapeutic target validation data was collected by combinational keywords search in PubMed database by using “validation” and “target”, “drug” and “potency”, “cell line” and “potency”, “cell line” and “activity”, “knock-out” and “target”, “target” and “IC₅₀”, and by comprehensively searching literatures and research articles in reputable research journals such as *Annual Review of Pharmacology and Toxicology*, *Annual Review of Physiology*, *Nature Reviews Drug Discovery*, *Nature Reviews Cancer*, *Nature Reviews Neuroscience*, *Trends in Pharmacological Sciences*, *Pharmacology & Therapeutics*, *Drug Discovery Today*, *Clinical Pharmacology & Therapeutics*, *Current Opinion in Pharmacology*, and so on, which explicitly mentioned techniques validating targets. In particular, these searches identified 218 recent papers offering target validation data.

Currently, there are 243 successful, 233 clinical trial, and 154 research targets covering 1,006 approved, 573 clinical trial, and 311 investigative drugs. With TTD data, the majority of successful (243 out of 348) and (233 out of 292) clinical trial targets have been covered. On the other hand, only 154 out of 1254 research targets have validation information. Furthermore, TVD is an ongoing project, which aims at covering all successful and clinical trial targets, and exploring validation information for as many research targets as possible. In order to integrate target validation data into TTD, links to the relevant data in TVD will be shown in the corresponding target page in TTD.

3.3.3 Explanation on target validation data

Generally, our target validation data collected for TVD provided evidences from three different aspects: (1) the potencies of drugs against their primary efficacy targets; (2) the potencies of drugs against the disease relevant cell-lines expressing these targets; (3) the effects of drug target knock-out and variance in target sequence, expression and activity in disease models; and additional evidences of actions on target from drug-like molecules. TVD is constructed based on explaining and validating a target from those data.

Take CDK2 as an example. CDK2 is a clinical trial target inhibited by many drugs in clinical trial, such as Flavopiridol (Phase III, CLL), SCH 727965 (Phase II, NSCLC), Seliciclib (Phase II, CLL), R-roscovitine (Phase I/II, NSCLC), AT7519 (Phase I/II, NHL), R547 (Phase I, solid tumors), AT7519 (Phase I, solid tumors), Ro 31-7453 (Phase I, solid tumors), SCH 727965 (Phase I, NHL and CLL), ZK 304709 (Phase I, solid tumors), and SNS-032 (Phase I, B-lymphoid malignancy). **Table 03-2** shows potency of drugs against their primary efficacy target CDK2. Their kinetic activities are all <500 nM and majority of them are below 50 nM, which indicate CDK2 is very promising in light of its binding potency. **Table 03-3** listed potency of drugs against disease relevant cell-lines expressing CDK2. As shown, the cell-line activities are around 200~300 nM, which is a very potent value compared to cell-line activities of other drugs^{251,252}. Finally, the effects of target knock-out and variance in CDK2 sequence, expression and activity in disease models and additional evidences have been illustrated in **Table 03-4**. CDK2 *in vivo* knock-out study correlates CDK2 with the tumor development, and identifies it as tumor-causing target. Addition evidences focus on information of drug-like action on their primary target(s).

Figure 03- 1 Screenshot of home page of TTD 2010

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:

List search results by targets:

Examples: Osetamivir; 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

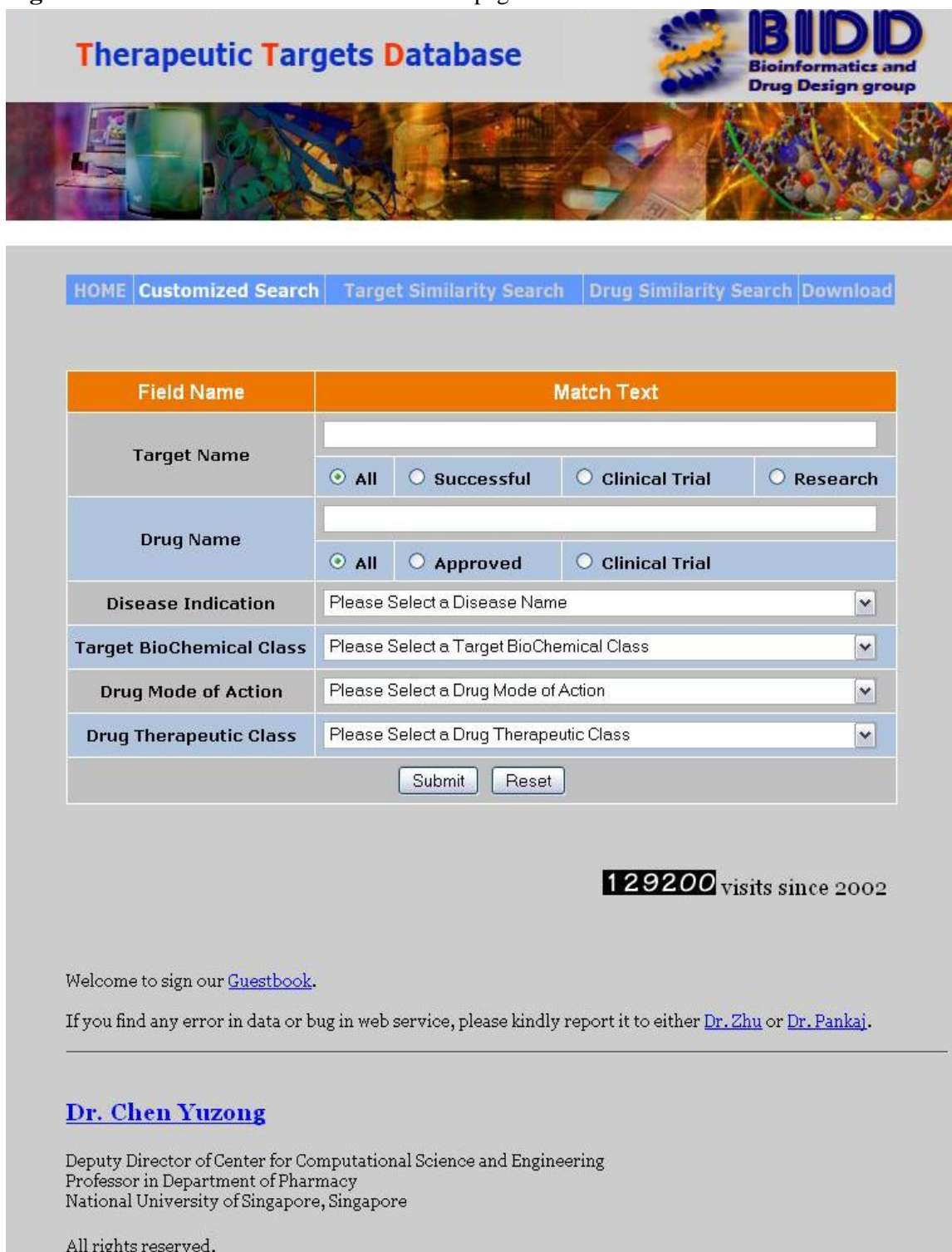
Zhu F, Han BC, Pankaj Kumar, Liu XH, Ma XH, Wei XN, Huang L, Guo YF, Han LY, Zheng CJ, Chen YZ. Update of TTD: Therapeutic Target Database. *Nucleic Acids Res.* 2009. [PubMed](#)

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

Last update by

March 19th, 2010

Figure 03- 2 Screenshot of customized search page of TTD 2010



Therapeutic Targets Database

BIDD
Bioinformatics and Drug Design group

HOME | **Customized Search** | Target Similarity Search | Drug Similarity Search | Download

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

129200 visits since 2002

Welcome to sign our [Guestbook](#).

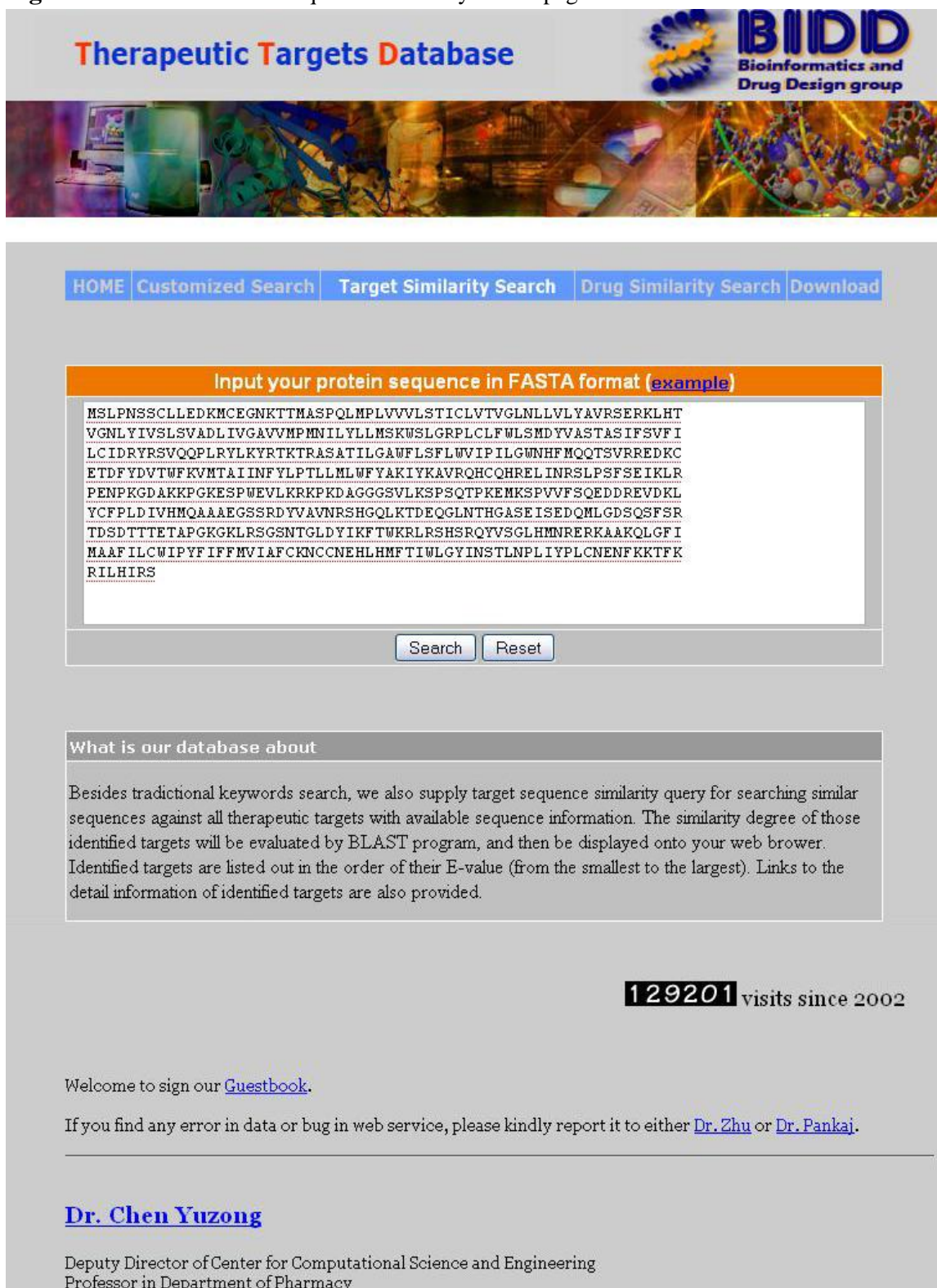
If you find any error in data or bug in web service, please kindly report it to either [Dr. Zhu](#) or [Dr. Pankaj](#).

[Dr. Chen Yuzong](#)

Deputy Director of Center for Computational Science and Engineering
 Professor in Department of Pharmacy
 National University of Singapore, Singapore

All rights reserved.

Figure 03- 3 Screenshot of sequence similarity search page of TTD 2010



Therapeutic Targets Database

BIDD
Bioinformatics and Drug Design group

HOME | Customized Search | **Target Similarity Search** | Drug Similarity Search | Download

Input your protein sequence in FASTA format (example)

```

MSLPNSSCLEDKMCENKTTMASPQLMPLVVVLSTICLVTVGLNLLVLYAVRSEKRLHT
VGNLYIVSLSVADLIVGAVVMPMNILYLLMSKWSLGRPLCLFWLSMDYVASTASIFSVFI
LCIDRYRSVQQPLRYLKYRTKTRASATILGAWFLSFLWVIPILGWNHFQQQTSVRREDKC
ETDFYDVTWFKVMTAIIINFYLPDLLMLWFYAKIYKAVROHCQHRELINRSLPSFSEIKLR
PENPKGDAKKPGKESPWEVLKRKPKDAGGGSVLKSPSQTPKEMKSPVVFSQEDDREVDKL
YCFPLDIVHMQAAAEGSSRDYVAVNRSHGQKTDQQLNTHGASEISEDQMLGDSQSFSR
TDSDTTETAPGKGLRSGSNTGLDYIKFTWKRLRSHSRQYVSGLHMNRERKAAKQLGFI
MAAFILCWIPYFIFFMVIAFCKNCCNEHLHMF TIWLG YINS TLNPLIYPLCNEFKKTFK
RILHIRS

```

Search Reset

What is our database about

Besides traditional keywords search, we also supply target sequence similarity query for searching similar sequences against all therapeutic targets with available sequence information. The similarity degree of those identified targets will be evaluated by BLAST program, and then be displayed onto your web browser. Identified targets are listed out in the order of their E-value (from the smallest to the largest). Links to the detail information of identified targets are also provided.

129201 visits since 2002

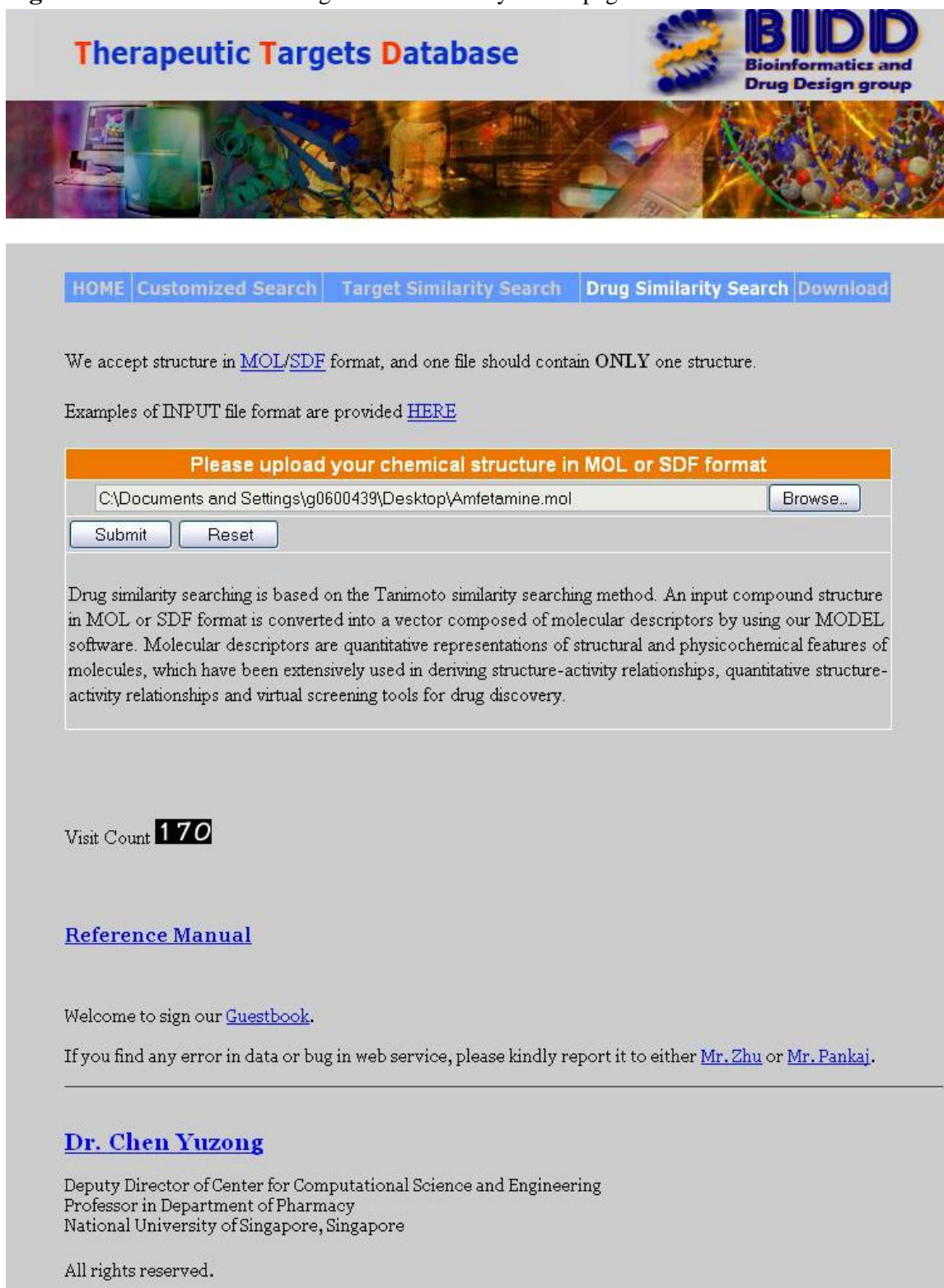
Welcome to sign our [Guestbook](#).

If you find any error in data or bug in web service, please kindly report it to either [Dr. Zhu](#) or [Dr. Pankaj](#).

Dr. Chen Yuzong

Deputy Director of Center for Computational Science and Engineering
Professor in Department of Pharmacy

Figure 03- 4 Screenshot of drug tanimot similarity search page of TTD 2010



Therapeutic Targets Database

BIDD
Bioinformatics and Drug Design group

HOME | Customized Search | Target Similarity Search | Drug Similarity Search | Download

We accept structure in [MOL/SDF](#) format, and one file should contain ONLY one structure.

Examples of INPUT file format are provided [HERE](#)

Please upload your chemical structure in MOL or SDF format

C:\Documents and Settings\g0600439\Desktop\Amfetamine.mol

Drug similarity searching is based on the Tanimoto similarity searching method. An input compound structure in MOL or SDF format is converted into a vector composed of molecular descriptors by using our MODEL software. 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 relationships and virtual screening tools for drug discovery.

Visit Count **170**

[Reference Manual](#)

Welcome to sign our [Guestbook](#).

If you find any error in data or bug in web service, please kindly report it to either [Mr. Zhu](#) or [Mr. Pankaj](#).

Dr. Chen Yuzong

Deputy Director of Center for Computational Science and Engineering
Professor in Department of Pharmacy
National University of Singapore, Singapore

All rights reserved.

Figure 03- 5 Screenshot of full database download page of TTD 2010

Therapeutic Targets Database

BIDD
Bioinformatics and Drug Design group

HOME | Customized Search | Target Similarity Search | Drug Similarity Search | **Download**

TTD Database Downloads

Download TTD targets information in raw format	Click to Save
Cross-matching ID between TTD drugs and public databases	Click to Save
Synonyms of drugs and small molecules in TTD	Click to Save

Target Sequence Downloads

Download sequence data for all targets	Click to Save
=>Download sequence data for successful targets only	Click to Save
=>Download sequence data for clinical trial targets only	Click to Save
=>Download sequence data for research targets only	Click to Save

Drug Structure Downloads


Download structure data for all drugs in MOL format	Click to Save
=>Download structure data for approved drugs only	Click to Save
=>Download structure data for clinical trial drugs only	Click to Save
=>Download structure data for experimental agents only	Click to Save
Download structure data for all drugs in SDF format	Click to Save
=>Download structure data for approved drugs only	Click to Save
=>Download structure data for clinical trial drugs only	Click to Save
=>Download structure data for experimental agents only	Click to Save
Download antisense oligonucleotide sequences in raw format	Click to Save

Last update by

March 19th, 2010

129202 visits since 2002

Figure 03- 6 Intermediate search results of “dopamine receptor” listed by targets



Therapeutic Targets Database

HOME
Customized Search
Target Similarity Search
Drug Similarity Search
Download


You are searching for: 'Dopamine receptor'

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


TTD ID	Search Result	
TTDS00015 Target Info	Target Name	D(1B) dopamine receptor
	Target type	Successful target
	Disease	Respiratory diseases
	Drugs	ZD-3638 Drug Info ; LE-300 Drug Info ...
TTDS00012 Target Info	Target Name	D(2) dopamine receptor
	Target type	Successful target
	Disease	Vomiting ; Schizophrenia ; Erectile dysfunction ...
	Drugs	Zuclopenthixol Drug Info ; Ziprasidone Drug Info ...
TTDS00013 Target Info	Target Name	D(3) dopamine receptor
	Target type	Successful target
	Disease	Schizophrenia ; Drug dependence ; Respiratory diseases ...
	Drugs	U-99194A Drug Info ; Sarizotan Drug Info ...
TTDS00011 Target Info	Target Name	Dopamine D1 receptor
	Target type	Successful target
	Disease	Parkinson's disease
	Drugs	Pergolide Drug Info ; Methylergonovine Drug Info ...
TTDC00291 Target Info	Target Name	D(4) dopamine receptor
	Target type	Clinical trial target
	Disease	Parkinson's disease ; Psychiatric illness ...
	Drugs	U-99363E Drug Info ; Sonepiprazole Drug Info ...

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

Figure 03- 7 Intermediate search results of “influenza virus infection” listed by drugs



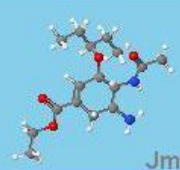


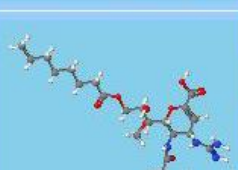
Therapeutic Targets Database



[HOME](#) | [Customized Search](#) | [Target Similarity Search](#) | [Drug Similarity Search](#) | [Download](#)

You are searching for: 'Influenza virus infection'

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

Structure	Search Result	
 Jmol	Drug Name	Oseltamivir
	Drug Status	Approved
	Disease	Influenza virus infection
	TTD Drug ID	DAP000714 Drug Info
 Jmol	Drug Name	Zanamivir
	Drug Status	Approved
	Disease	Influenza virus infection
	TTD Drug ID	DAP000715 Drug Info
 Jmol	Drug Name	Rimantadine
	Drug Status	Approved
	Disease	Influenzavirus A infection
	TTD Drug ID	DAP001087 Drug Info
 Jmol	Drug Name	Peramivir
	Drug Status	Phase III
	Disease	Influenza virus infection
	TTD Drug ID	DCL000286 Drug Info
 Jmol	Drug Name	CS-8958
	Drug Status	Phase III
	Disease	Influenza virus infection
	TTD Drug ID	DCL000297 Drug Info

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

Figure 03- 8 TTD target main information page

TTD Target ID: TTDS00174					
Target Information					
Name	Neuraminidase				
Type of target	Successful target				
Synonyms	N-acetylneuraminase glycohydrolase				
	NANase				
	STNA				
	Sialidase				
Disease	Influenza, virus not identified [1]				
Drug(s)	Oseltamivir	Drug Info	Approved	Influenza virus infection	[2][3]
	Zanamivir	Drug Info	Approved	Influenza virus infection	[2][3]
	CS-8958	Drug Info	Phase III	Influenza virus infection	[4]
	Peramivir	Drug Info	Phase III	Influenza virus infection	[4][2]
BioChemical Class	Glycosylases				
EC Number	EC 3.2.1.18				
Pathway	Lysosome				
	Other glycan degradation				
	Sphingolipid metabolism				
UniProt ID	P06818				
	P29768				
PDB Structure	1DIL ; 1DIM ; 2SIL ; 2SIM ; 3SIL .				
Function	Cleaves the terminal sialic acid (n-acetyl neuraminic acid) from carbohydrate chains in glycoproteins providing free sialic acid which can be used as carbon and energy sources. Sialidases have been suggested to be pathogenic factors in microbial infection.				
Sequence	MNPNQKIITIGSVSLTIATICFLMQIAILVTTVTLHFQYECSSPPNNQVMPCEPIIER NITEIVYLTNTTIDKEICPKLVEYRNWSPKQCKITGFAPFSKDNSIRLSAGGGIIVWTREP YVSCDPGKCYQFALGGQTTLDNKHNSNDTIHDRTPYRTLLMNLGVPPHFGTRQVCIAWSS SSCHDGKAWLHVCVGTGYDKNATASF IYDGLVDSIGSWSKNILRTQSEECVCI NGTCTVV MTDGSASERADTKILFIEEGKIVHISPLSGSAQHVEECSCYPRYPGVCVCRDNWKGSNR PVVDINVKDYSIVSSYVCSGLVGDTPRKNDRSSSYCRNPNNEKGNHGVKGNWAFDDGNDV WMGRTISEESRSGYETFKVIGGWSTPNSKLIQRQVIVDSDNRSYGSGIFSVVEGKSCINR CFYVELIRGREQETRVWTSNSIVVFCGTSPTYGTGSWPDGADINLMPI				
Related US Patent	6,509,359				
Inhibitor	CS-8958	Drug Info	[4]		
	GS4071	Drug Info	[5][6][7]		
	Oseltamivir	Drug Info	[2][3]		
	Peramivir	Drug Info	[4][2]		
	Zanamivir	Drug Info	[2][3]		
Ref 1	New millennium antivirals against pandemic and epidemic influenza: the neuraminidase inhibitors. Antivir Chem Chemother. 2002 Jul;13(4):205-17. To Reference				
Ref 2	Current and future antiviral therapy of severe seasonal and avian influenza. Antiviral Res. 2008 Apr;78(1):91-102. Epub 2008 Feb 4. To Reference				

Figure 03- 9 TTD drug main information page

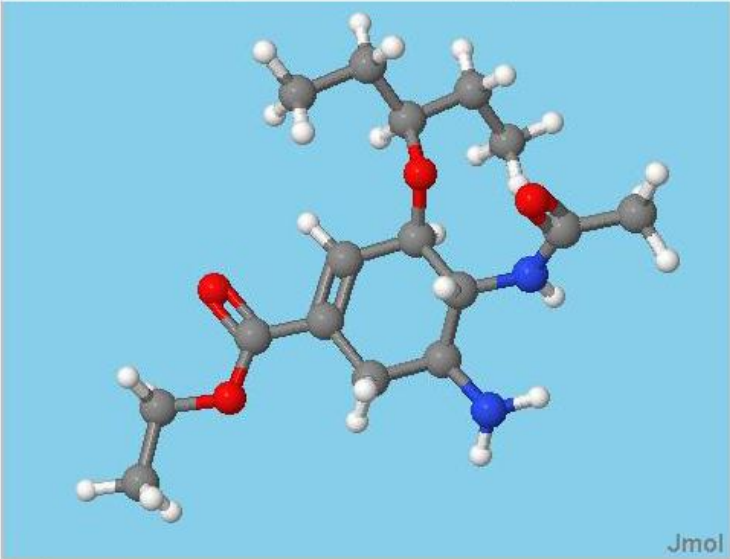
TTD Drug ID: DAP000714	
Drug Information	
Name	Oseltamivir
Synonyms	196618-13-0; 204255-11-8; AIDS-070972; AIDS070972; BSPBio_002342; C08092; C08093; D00900; DB00198; GS 4104; GS-4104; GS-4104/002; GS4104; LS-173828; Oseltamivir; Ro 64-0796/002; Ro-64-0796; Ro-64-0796/002; SPECTRUM1505822; Tamiflu; Tamiflu (TN)
Trade Name	Tamiflu
Company	Hoffmann-La Roche pharmaceutical company
Indication	Influenza virus infection Approved [1]
Structure	 <p style="text-align: right;">Jmol</p>
Therapeutic Class	Antiviral Agents
CAS Number	CAS 204255-11-8
Formular	C16H28N2O4
PubChem Compound ID	CID 65028 .
PubChem Substance ID	SID 626306 .
ChEBI	7799 ;
SuperDrug ATC ID	J05AH02 ;
SuperDrug CAS ID	196618130 ;
Target	Neuraminidase Target Info Inhibitor [2][3]
Ref 1	Natural products as sources of new drugs over the last 25 years. J Nat Prod. 2007 Mar; 70(3):461-77. Epub 2007 Feb 20. To Reference
Ref 2	Current and future antiviral therapy of severe seasonal and avian influenza. Antiviral Res. 2008 Apr; 78(1):91-102. Epub 2008 Feb 4. To Reference
Ref 3	Antiviral agents for influenza, hepatitis C and herpesvirus, enterovirus and rhinovirus infections. Med J Aust. 2001 Jul 16; 175(2):112-6. To Reference

Table 03- 1 Main drug-binding databases available online

No	Database	Address
1	BRENDA	http://www.brenda-enzymes.info/
2	DrugBank	http://www.drugbank.ca/
3	eMolecules	http://www.emolecules.com/
4	MDDR	http://www.symyx.com/products/databases/index.jsp
5	PNPDB	http://azavedolab.net/14.html
6	PubChem	http://nihroadmap.nih.gov
7	SCOWLP	http://www.scowlp.org/
8	ShikiPDB	http://azavedolab.net/14.html
9	SuperNatural	http://bioinformatics.charite.de/supernatural/
10	SuperHapten	http://bioinformatics.charite.de/superhapten/
11	WOMBAT	http://www.sunsetmolecular.com
12	ZINC	http://zinc.docking.org/

Table 03- 2 Potencies of drugs against their efficacy targets CDK2

Drug Name	Potency	Status	Disease indication
Flavopiridol	IC ₅₀ : 100 nM	Phase III	Chronic lymphocytic leukemia (CLL)
		Phase II	Acute myeloid and lymphoblastic leukemia
		Discontinued	Hepatocellular carcinoma
SCH 727965	IC ₅₀ : < 5 nM	Phase II	Advanced breast cancer, NSCLC, and acute leukemia
		Phase I	Advanced solid tumors, NHL, multiple myeloma and CLL
Seliciclib	IC ₅₀ : 220 nM	Phase II	NSCLC, lymphoid leukemia, and multiple myeloma
R-roscovitine	IC ₅₀ : 220 nM	Phase I/II completed	NSCLC
AT7519	IC ₅₀ : 47 nM	Phase I/II	Non-Hodgkin's Lymphoma
		Phase I	Solid Tumors
R547	IC ₅₀ : 1~3 nM	Phase I completed	Advanced solid tumors
SNS-032	IC ₅₀ : 48 nM	Phase I	B-lymphoid malignancies and advanced solid tumors
ZK 304709	IC ₅₀ : 4 nM	Phase I	Advanced solid tumors
AG-024322	IC ₅₀ : 1~3 nM	Discontinued	Advanced cancer

Table 03- 3 Potencies of drugs against the disease relevant cell-lines expressing CDK2

Drug Name	Effect and potency against cell line
AG-024322	Arrested multiple stages of the cell cycle and induced apoptosis in various human tumor cell lines ($IC_{50} = 30\sim 200$ nM). Displayed dose-dependent antitumor activity in mice bearing human tumor xenografts.
AT7519	Potent anti-proliferative activity in various human tumor cell lines, a lower activity in non-transformed fibroblasts and no activity in non-cycling cells; promoted tumor growth inhibition or regression in human ovarian and colon carcinoma xenogra
Flavopiridol	Induced G1–S phase and G2–M phase arrest and apoptosis at 200~300 nM concentrations in many tumor cells. Significant clinical activity in refractory CLL; clinical studies ongoing to determine efficacy of combination with anti-neoplastic agents.
R547	Induced G1–G2 arrest and apoptosis in tumor cell lines independently of RB1 or p53 status ($IC_{50} < 0.60$ mM); induced significant tumor growth reduction in human tumor xenografts and efficacious with daily oral dosing and weekly intravenous dosing. Early
R-roscovitine	Induces S–G2 arrest and apoptosis.
SCH 727965	Induced apoptosis in tumor cell lines and growth inhibition or regression in xenograft models.
SNS-032	<i>In vitro</i> ; high CDK selectivity over a panel of 12 unrelated kinases ($IC_{50} > 25$ mM); induced tumor growth reduction in a human ovarian carcinoma xenograft. Sensitized radioresistant NSCLC cells to ionizing radiation.
ZK 304709	Blocked growth of human tumor cell lines at $IC_{50} = 317$ nM, by inducing a dose-dependent G1–S arrest followed by apoptosis. Superior efficacy over standard chemotherapy in human tumor xenografts and orthotopic mouse models of human pancreatic cancer.

Table 03- 4 Effects of target knock-out in CDK2 sequence, expression and activity in disease models and additional evidences**Effects of target knock-out or variation in target sequence, expression and activity in disease models**

Target Alteration Type	Remarks	Effects on disease model
Genotype: Cdk2-/-	Loss-of-function strains	Sterility due to defective meiosis; no effect on mitotic cells
Genotype: Cdk2-/-; Cdkn1b-/-	Target validation strains	Develop tumors with similar incidence and latency to those in Cdkn1b-deficient mice, suggesting the function of p27 (encoded by Cdkn1b) is independent of CDK2

Additional Evidences

Drug Name	Drug-like Action	Remarks
Purvalanol B	Inhibition	IC ₅₀ : 6 nM
AG-024322	Inhibition	Phase I, advanced cancer: NCT00147485 sponsor: Pfizer
AT-7519	Inhibition	Phase I/II, advanced or metastatic tumours: NCT00390117; sponsor: Astex
Ro-4584820	Inhibition	Phase I, advanced solid tumours: NCT00400296; sponsor: Hoffmann-LaRoche
Roscovitine	Inhibition	Phase II, non-small cell lung cancer, nasopharyngeal cancer, haematological tumours: NCT00372073; sponsor: Cyclacel
SNS-032	Inhibition	Phase I, B-lymphoid malignancies: NCT00446342; Phase I, solid tumours: NCT00292864; sponsor: Sunesis

Chapter 4 Therapeutic targets in clinical trials

Most drugs produce their therapeutic effect by interacting and modulating the activity of selected protein targets^{6,253,254}. Based on data from TTD⁹, intensive drug exploration and target discovery efforts^{56,226,227,255} have sent 1,164 drugs to clinical trial, 690 of which direct at 292 new targets. These clinical trial drugs and targets are fruits and bear marks of past decades' application of drug discovery technologies (combinatorial chemistry²⁵⁶, HTS & virtual screening^{257,258}, ADME-Tox evaluation²⁵⁹, and fragment-based design²⁶⁰), considerations of pharmacogenetics²⁶¹, system biology and multi-target drug discovery²⁵⁴, and progresses in genomics, structural genomics and proteomics²⁶²⁻²⁶⁵.

By analyzing the therapeutic, biochemical, physicochemical, and systems features of the clinical trial targets and drugs with respect to successful targets, approved drugs and drug leads, useful information can be gained regarding general trends of past decades' drug discovery efforts^{56,59,255}, areas of focus, progress and difficulty^{201,227}, and distinguished features guiding the enhancement of specific properties in the target exploration and drug discovery^{56,59,70,240,255,266-271}. Given the key role played by target selection in clinical successes of drugs and the unique role of clinical trials in target validation in human and in evaluation of drug efficacy and safety, we systematically analyzed physicochemical, therapeutic, biochemical, and systems features of targets and drugs in clinical trials. We compared the relevant properties of clinical trial targets and drugs in different trial phases with successful targets^{56,59,201,227,240,270}, approved drugs^{70,267} and the high-throughput screening derived drug leads^{255,269,271}.

4.1 Trends in the exploration of clinical trial targets

The areas of progress in the exploration of new targets for disease treatment can be partly revealed by evaluating the distribution of clinical trial targets in different trial phases with respect to different disease classes, which are given in **Table 04-1**. Every target is assigned to the highest trial phase in which a target-directed drug has been or is being tested. The way of dividing disease classes is based on the international statistical disease classification of the World Health Organization (WHO, 1992). Neoplasm, inflammation, nervous system and sense organs disorders, diseases of circulatory system, and nutritional and metabolic diseases constitute the classes with the largest number of targets in all clinical trial phases, which include 133, 70, 57, 52 and 47 targets respectively. For phase III trial targets, the classes with the largest number of targets are neoplasm, nervous system and sense organs disorders, inflammation, diseases of circulatory system, and the nutritional and metabolic disease with 33, 21, 20, 18 and 14 targets respectively. In comparison, the disease classes with highest number of successful targets are neoplasm, infectious and parasitic diseases, nervous system and sense organs disorders, circulatory system diseases, mental disorders, and respiratory system diseases with 78, 78, 56, 54, 46, and 35 targets respectively²²⁷. Thus, in clinical trial pipelines, there is an increased pool of novel targets for inflammation and nutritional and metabolic diseases, and a steady stream of fresh drug targets for neoplasm, circulatory system diseases, and nervous system and sense organs disorders. On the other hand, a relatively decreased pool of novel targets for mental disorders, respiratory system diseases, and infectious and parasitic diseases is also observed.

The areas of progress in target exploration may be further revealed by evaluating the top protein families and biological pathways that contain high number of clinical trial targets. **Figure 04-1** and **Figure 04-2** show the top-10 PFAM protein families and top-20 KEGG pathways with a large number of phase I, II, and III clinical trial drug targets. A large number of targets are concentrated in the GPCR and kinase protein families and distributed in the upstream or upstream-linked signaling pathways, such as the neuroactive ligand-receptor, cytokine-cytokine receptor, chemokine, Jak-STAT, toll-like receptor, neurotrophin, ErbB, and VEGF signaling pathways. Evaluation of the sub-cellular distribution of the phase I, II, and III targets, in **Figure 04-3**, further demonstrates that the majority of these targets are associated with membrane or in extracellular locations.

Figure 04-4 shows the top-10 PFAM protein families that contain a large number of targets in clinical trial in comparison with the top-10 PFAM families containing a large number of successful targets. There appears to be continued progress in exploring new targets in the highly successful GPCR, kinase, trypsin, and hormone receptor families. Progresses have also been made in the exploration of new targets in new protein families such as matrixin (matrix metalloproteinase), TNFR/NGFR, and eukaryotic porin families. However, there is a substantial reduction in the number of new targets in the ion transport protein, nuclear hormone, immunoglobulin, cytochrome P450, and ABC transporter families.

Figure 04-5, **Figure 04-6** and **Figure 04-7** show the specific clinical trial and successful targets with a largest number of phase II, III and all clinical trial drugs respectively. Most

of the drugs in phase III trials target successful GPCR receptors D2, 5HT-2, 5HT-1, D3, adrenoceptor beta 2, and opioid receptor sigma 1, and successful anticancer kinase targets VEGFR2, Her2, and c-kit. But there are a fair number of drugs targeting a number of new targets such as D4, voltage-gated K channel kv1.5, VEGFR3 and endothelial nitric oxide synthase. Significantly larger number of drugs, particularly multi-target kinase inhibitors, in phase II trials target new kinase targets such as P38, VEGFR3, aurora-A and -B, Jak2, PI3K, c-Met and CDK2. Successful kinase targets such as VEGFR2, EGFR, PDGFR and VEGFR1 are also targeted by a large number of phase II drugs, and other heavily explored targets in phase II are histone deacetylase, substance-P, TNF, adrenoceptor beta2, and the adenosine A2a receptor. Overall, the new targets with the largest number of clinical trial drugs are kinases aurora-A and -B, PI3K, VEGFR3, Akt, c-met, CDK2 and P38, and the successful targets with largest number of clinical trial drugs are GPCR receptors like D2, substance P, 5HT-2, adrenoceptor beta 2 and D3, kinases VEGFR2, Her2 and EGFR, and histone deacetylase (HDAC).

The level of difficulty in target exploration is partly reflected by the time spent so far for developing target-directed drugs that enter specific stage of clinical trials. We crudely estimated the exploration time of clinical trial targets by using the number of years from the year the target was first reported in the literature to the current year. **Figure 04-8** and **Figure 04-9** shows the distribution of all clinical trial targets and clinical trial targets in individual trial phases with respect to their exploration time respectively. For comparison, the distribution of innovative successful targets approved by FDA from 1995 to 2008 with respect to target exploration time is also included in **Figure 04-8**. Here innovative successful targets refer to the successful targets that have no other subtype of the same

protein as a successful target before the first FDA approval of these targets. The target exploration time for these innovative successful targets was crudely estimated by using the number of years from the year the target was first reported in the literature to the year of FDA. The average exploration time of phase I, II, III and all clinical trial targets is 18, 16, 20, and 18 years respectively, which is compared to the average target exploration time of 20 years for the innovative successful targets. Thus it remains a very slow and difficult process for moving drugs into clinical trial and through the three trial phases, which is also reflected by the low productivity rates of innovative successful targets²²⁷.

4.2 Comparison of the characteristics of clinical trial targets with successful targets

Comparison of the characteristics of clinical trial targets with those of successful targets provide useful hints about both the common and distinguished features of clinical trial targets that can be retained, enhanced, or improved. **Figure 04-10** shows the distribution of phase I, II and III clinical trial targets and discontinued clinical trial targets in different similarity ranges with respect to successful targets. The levels of similarity to successful targets are classified into very similar, marginally similar, and un-similar based on Blast E-values in the range of ≤ 0.001 , $0.001 \sim 0.1$ and ≥ 0.1 respectively. The majority of the clinical trial targets (68%, 63%, 56%, and 50% of the phase III, II, I and discontinued clinical trial targets) are very similar in sequence to these successful targets. Non-the-less, substantial percentages of the clinical trial targets (26%, 32%, 35%, and 44% of the phase III, II, I and discontinued clinical trial targets) are un-similar in sequence to successful targets. Furthermore, target failure is not necessarily associated with the dissimilarity to successful targets, as there are comparable numbers of discontinued clinical trial targets that are very similar and un-similar to successful targets.

Figure 04-11, **Figure 04-12** and **Figure 04-13** respectively show the distribution of all clinical trial targets and successful targets with respect to the number of human similarity proteins outside the target family, number of pathways the target is associated with, and the number of tissues the target is distributed in. The distribution profiles of the clinical trial targets are comparable to those of successful targets^{201,227}. Similar to successful targets, most of the clinical trial targets have less than 15 human similarity proteins

outside their respective target family. The vast majority of the clinical trial targets, at slightly reduced percentage compared to the successful targets, are associated with no more than 3 human pathways and distributed in no more than 5 human tissues. Therefore the systems profiles of vast majority of clinical trial targets appear to be very similar to those of successful targets²⁴⁰.

Apart from similarity of sequence^{226,227} and systems profiles^{201,227,270,272-274}, the similarity or the resemblance of other features such as drug binding-site structural conformations^{275,276} and physicochemical properties^{144,227,241} are also important in protein overall druggability. **Table 04-2** shows the reported distribution of the phase I, II, and III targets that are similar or resemble the properties of successful targets in sequence, drug-binding domain structural fold, physicochemical features, and the systems profiles²⁴⁰. In **Chapter 5**, this distribution will be further illustrated. In particular, the comparison of physicochemical features of clinical trial and successful targets has been conducted by inputting the physicochemical features of a specific clinical trial target into a machine learning classification model to evaluate whether that target can be classified into the successful target class^{144,227,241}. Few percentages of the clinical trial targets, 2.4%, 8.3%, and 10% of phase I, II, and III targets, have all four profiles similar to or resemble those of successful targets, and 9.8%, 25%, and 50% of the phase I, II, and III targets have three of the four profiles similar to or resemble those of successful targets. This profile-combinatory method will be explained in detail in **Chapter 5**.

Attempts have been made to explore the individual^{144,201,226,227,241,270,274-276} and combination²⁴⁰ of the sequence similarity, drug-binding domain structural fold analysis,

the physicochemical features recognized by machine learning methods, and systems profiles with respect to successful targets for *in silico* target prediction. In particular, it has been proposed that a promising target likely has at least three of the four profiles similar to or resemble those of successful targets²⁴⁰. This proposed method recognized 7 of the 8 targets with positive phase III results, and dropped 89% of the 19 discontinued clinical trial targets and 97% of the 65 targets failed in HTS or knockout studies²⁴⁰.

4.3 The characteristics of clinical trial drugs with respect to approved drugs and drug leads

The effectiveness and trends of clinical trial targets as well as successful targets targeted by clinical trial drugs may be revealed by analyzing the potencies and physicochemical properties of their targeted drugs with respect to those of approved drugs and drug leads. Recent analyses of clinical trial drugs and patented agents with respect to approved drugs have revealed marked differences in median values of key physicochemical properties between approved oral drugs and clinical trial drugs and patented agents^{269,277}. Another analysis of recently developed drug leads and drug hits with respect to approved drugs has further shown that recently developed drug leads generally have good potencies but their median values of key physicochemical properties are substantially different from those of approved drugs primarily due to the nature of high-throughput screening hits and hit-to-lead optimization practices²⁵⁵. We extended this type of analysis to a significantly higher number of 656 clinical trial drugs and by profiling the potencies of clinical trial drugs with respect to those of approved drugs.

Figure 04-14 and **Figure 04-15** show the distribution of clinical trial and approved drugs by potency, and the distribution of phase I, II, and III drugs by potency respectively. The distribution pattern of the approved drugs is very similar to that of FDA approved drugs reported in the literature²⁷⁸. As shown in **Table 04-3**, the median potency of clinical trial drugs (32.2nM) is substantially improved against approved drugs (74.6nM), particularly for phase III drugs (19.5nM), and it very similar to that of HTS drug leads (30nM)²⁵⁵. It appears that new technologies such as HTS and virtual screening^{257,258} has enabled the

identification or design of more drug candidates with higher potencies by more extensive exploration of chemical space. Good potency seems to be one of the important factors for the advancement of some drugs into higher phases. Clinical trial drugs targeting the novel targets appear to show higher median potency (25.5nM) than clinical trial drugs targeting successful targets (39.9nM) and clinical trial targets with protein subtype as successful target (32.9nM). One possible reason for this discrepancy is that highly potent agents may be more easily identified or designed against novel targets by the ability to explore new as well as existing chemical space with reduced possibility of potential conflict with existing drugs and potential obstacles of patent protection.

Figure 04-16, Figure 04-17 and Figure 04-18 show the distribution of clinical trial drugs and approved drugs by molecular weight, the distribution patterns for drugs in different trial phases, and the distribution patterns for drugs targeting different types of targets (novel clinical trial targets, clinical trial targets with protein subtype as successful target, and successful targets). **Figure 04-19, Figure 04-20 and Figure 04-21** show distribution of clinical trial drugs and approved drugs by ALogP, the distribution pattern for drugs in specific trial phases, and the distribution pattern for drugs targeting different types of targets. As shown in **Table 04-3**, the median molecular weight, ALogP and the number of non-terminal rotatable bonds of all clinical trial drugs (403.4Da, 3.9 and 5.6), are substantially higher than the median values of 342.4Da, 2.8 and 4.6 of approved drugs. The molecular weight and flexibility level (represented by the number of non-terminal rotatable bonds) of phase I, II and III drugs show apparent descending trend, with the median values of phase III drugs closest to those of approved drugs. In contrast, AlogP of phase I, II and III drugs show no apparent trend and small variations. It seems that phase

I, II and III trials are increasingly selective towards drugs with drug-like molecular weight and flexibility level, while phase I and II trials are far less selective and phase III trials are very selective towards drugs with drug-like lipophilicity. Against HTS drug leads²⁵⁵, clinical trial drugs show slightly improved molecular weight and lipophilicity (represented by CLogP for HTS drug leads and ALogP for clinical trial drugs). Against World of Molecular BioAcTivity database compounds with potency better than 1nM²⁷⁷, clinical trial drugs show significantly improved molecular weight and flexibility level and comparable lipophilicity. Moreover, as show in **Figure 04-22**, the percentage of approved drugs obeying Lipinski's rule of five⁷⁰ is substantially higher than phase III clinical trial drugs, and the percentage of phase III drugs is substantially higher than that of phase I and II drugs.

The higher value in key physicochemical properties has been considered to be one of the important factors for the high attrition rates of clinical trial drugs^{255,269}. To some extent, preclinical and phase I trials appear to be less discriminative than phase II and III trials in selecting drugs with higher values of these two key physicochemical properties. Other factors are likely to be equally important in determining the advancement of clinical trial drugs. For instance, the key physicochemical properties of the ligands of different target protein families have been found to be substantially different²⁶⁶. Thus, there might be different ranges of good physicochemical properties for different target classes. Moreover, clinical trial drugs are more complex than approved drugs, with the computed molecular complexity level of phase I, II and III drugs showing descending trend²⁷⁷.

4.4 Perspectives

Intensive drug discovery efforts have led to a steady stream of pipeline drugs in clinical trials directed at new targets, particularly GPCRs and kinases in upstream signaling pathways for high-impact diseases that need more treatment options or more-effective drugs. Many successful targets or protein subtypes of successful targets have been heavily explored in clinical trials. Majority of clinical trial targets appear to have one or more of the sequence, structure, physicochemical and systems profiles similar to or resemble those of successful targets. In particular, targets of positive phase III results have multiple profiles similar to or resemble those of successful targets. Thus, the search of novel drugs directed at new targets with some form of similarity or resemblance to successful targets has been and will likely continue to be considered as a ‘good bet’ by the pharmaceutical industry.

There is another reason for the high number of clinical trial drugs against targets that are similar in sequence to successful targets. By taking advantage of the progress in genome sequencing and in the more extensive understanding of disease mechanisms, exploration of new targets has become increasingly subtype-specific and, for some diseases, pathogen-species-specific. It is expected that this trend will continue, and more subtype-specific and pathogen-species specific targets will be explored. Non-the-less, rapid progress in genomics, structural genomics, proteomics²⁶²⁻²⁶⁵, systems biology and multi-target drug discovery²⁵⁴ will enable the discovery of more novel targets, and the secondary targets of multi-target agents that previously cannot be explored as a primary target of single-target drugs.

Comparative analysis of multiple profiles of clinical trial targets with respect to successful targets provides useful clues to the quality of clinical trial targets and to the identification of promising targets²⁴⁰. *In silico* target identification methods have been introduced based on the analysis of the individual^{144,201,226,227,241,270,274-276} and combination²⁴⁰ of the sequence, structure, physicochemical, and systems profiles with respect to successful targets. These methods explore comparative sequence analysis^{226,276}, structural analysis^{275,276}, ligand-protein inverse docking²⁷⁹, machine learning of druggability characteristics²²⁷, system-related druggability profiles^{201,227} and combinations of these four profiles²⁴⁰ for recognizing target-like and druggable proteins. These progresses combined with increased molecular understanding of diseases²⁸⁰ and their corresponding targets²²⁷ enable the development of efficient tools for identifying innovative targets of new therapies and personalized medicine. In exploring and validating a new target, one also needs to pay attention to the capability in the identification of drug candidates with good drug-like physicochemical properties, such as lipophilicity and molecular weight, as well as other desirable features such as potency, ADME and toxicity.

In conclusion, over 1,164 drugs have entered clinical trials, 690 of which target 283 new targets. Analysis of the therapeutic, biochemical, physicochemical, and systems features of these clinical trial targets and drugs reveals areas of focuses, progresses and distinguished features. Many new targets, particularly GPCRs and kinases in upstream signaling pathways, are in advanced trial phases against cancer, inflammation, nervous and circulatory systems diseases, and nutritional and metabolic disorders. Majority of the new targets show sequence and systems profiles similar to those of successful targets, but

fewer of them show overall sequence, structure, physicochemical, and systems features resembling those of successful targets. Drugs in advanced trial phases show improved potency but increased lipophilicity and molecular weight with respect to approved drugs, and improved potency and lipophilicity but increased molecular weight (particularly for novel targets) compared to HTS leads. These suggest a need for further improvement in drug-like and target-like physicochemical features.

Table 04- 1 Number of clinical trial targets in different disease classes*

Index	Disease Class	Number of Targets in All Trial Phases, Phase III, II, and I		Number of Targets in All Trial Phases Shared by Another Disease Class																
		All Targets	Targets Exclusively for This Disease Class	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	
a	Blood and blood-forming organ disease	14, 3, 6, 5	3, 0, 2, 1		7	0	0	0	1	1	0	0	0	0	3	0	3	2	0	
b	Circulatory system diseases	52, 18, 24, 10	7, 3, 3, 1	7		2	1	2	13	0	4	4	5	2	20	2	21	9	0	
c	Digestive system diseases	7, 4, 3, 0	0, 0, 0, 0	0	2		1	1	2	0	0	2	1	1	1	0	2	4	0	
d	Genitourinary system diseases	7, 3, 4, 0	0, 0, 0, 0	0	1	1		1	2	0	2	0	2	0	1	1	0	2	0	
e	Musculoskeletal system and connective tissue diseases	12, 3, 9, 0	0, 0, 0, 0	0	2	1	1		4	0	1	0	2	1	4	0	6	0	1	
f	Nervous system and sense organ disease	57, 21, 24, 12	16, 5, 7, 4	1	13	2	2	4		0	3	4	2	3	14	7	14	9	0	
g	Respiratory system diseases	10, 3, 7, 0	0, 0, 0, 0	1	0	0	0	0	0		1	0	1	2	8	1	2	0	0	
h	Skin and subcutaneous tissue disease	17, 2, 13, 2	1, 1, 0, 0	0	4	0	2	1	3	1		0	5	2	9	1	5	2	0	
i	Endocrine disorders	20, 8, 8, 4	2, 0, 0, 0	0	4	2	0	0	4	0	0		0	1	1	0	7	14	0	
j	Immunity disorders	34, 5, 22, 7	6, 1, 3, 2	0	5	1	2	2	2	1	5	0		3	20	2	9	4	0	
k	Infectious and parasitic diseases	26, 4, 16, 6	12, 0, 9, 3	0	2	1	0	1	3	2	2	1	3		2	0	7	1	0	
l	Inflammation	70, 20, 22, 10	8, 2, 3, 1	3	20	1	1	4	14	8	9	1	20	2		4	26	6	0	
m	Mental disorders	22, 10, 7, 5	8, 2, 4, 2	0	2	0	1	0	7	1	1	0	2	0	4		4	4	0	
n	Neoplasm	133, 33, 66, 34	70, 12, 37, 21	3	21	2	0	6	14	2	5	7	9	7	26	4		6	1	
o	Nutritional and Metabolic diseases	43, 14, 24, 5	7, 4, 3, 0	2	9	4	2	0	9	0	2	14	4	1	6	4	6		0	
p	Congenital anomalies	2, 0, 1, 1	1, 0, 0, 1	0	0	0	0	1	0	0	0	0	0	0	0	0	1		0	
Total clinical trial therapeutic targets in all trial phases based on disease classes		516 (duplicate); 286 (distinct)		141	Redundancy of therapeutic targets = 145; non-redundancy of therapeutic targets = 141															

* The total number of non-redundant clinical trial targets in all trial phases is 286, 145 of which are for more than one disease classes. Because of this redundancy of targets, the sum of the number of targets in these classes is greater than 286. The number of targets shared between different disease classes is also given in the table.

Table 04- 2 Distribution of the phase III, II, and I targets that are similar or resemble the properties of successful targets in sequence (A), drug-binding domain structural fold (B), physicochemical features (C), and systems profiles (D)

Similarity in combinations of sequence (A), structural (B), physicochemical (C), and systems (D) profiles	No and Percentage of the 30 Phase III Targets in This Category	No and Percentage of the 84 Phase II Targets in This Category	No and Percentage of the 41 Phase I Targets in This Category
Similarity in all four profiles A, B, C, D	3 (10.0%)	7 (8.3%)	1 (2.4%)
Similarity in any 3 profiles of A, B, C, D	15 (50.0%)	21 (25.0%)	4 (9.8%)
Combination of A, B, C	5 (16.7%)	10 (11.9%)	1 (2.4%)
Combination of A, B, D	7 (23.3%)	11 (13.1 %)	4 (9.8%)
Combination of A, C, D	9 (30.0%)	14 (16.7%)	1 (2.4%)
Combination of B, C, D	3 (10.0%)	7 (8.3%)	1 (2.4%)
Similarity in any one profile of A, B, C, D	28 (93.3%)	51 (60.7%)	25 (61.0%)
Only A	18 (60.0%)	39 (46.4%)	17 (41.5%)
Only B	11 (36.7%)	26 (31.0%)	8 (19.5%)
Only C	13 (43.3%)	21 (25.0%)	3 (7.3%)
Only D	23 (76.7%)	31 (36.9%)	13 (31.7%)

Table 04- 3 Median potency, molecular weight, AlogP, the number of H-bond donor and H-bond acceptor, and the number of rotatable bond of approved, all clinical trial, phase , II and III drugs, and clinical trial drugs targeting novel clinical trial targets, clinical trial targets protein subtype as a successful target, and successful targets.

Type of Drugs	Median Potency (nM)	Median Molecular Weight (Da)	Average AlogP	Average Number of H-bond Donors	Average Number of H-bond Acceptors	Average Number of Non-terminal Rotatable Bonds
Approved drugs	74.608	342.386	2.787	2.190	5.847	4.554
All clinical trial drugs	32.210	403.352	3.915	2.177	6.827	5.620
Phase III clinical trial drugs	19.475	387.637	3.831	1.962	6.445	4.829
Phase II clinical trial drugs	29.587	406.791	3.993	2.195	6.802	5.725
Phase I clinical trial drugs	46.365	416.812	3.865	2.422	7.379	6.441
Clinical trial drugs targeting novel clinical trial targets	25.485	412.756	4.018	2.200	7.174	5.613
Clinical trial drugs targeting clinical trial targets with protein subtype as successful targets	32.973	412.090	3.826	2.559	7.126	6.216
Clinical trial drugs targeting successful targets	39.910	394.768	3.878	2.047	6.518	5.444
HTS drug leads ²⁵⁵	30	406	4.1 (CLogP)	NA	NA	NA
WORLD of Molecular BioAcTivity database compounds with potency better than 1nM ²⁷⁷	<1	463.6	3.82 (CLogP)	1	5	10
WORLD of Molecular BioAcTivity database compounds with potency worse than 1μM ²⁷⁷	>1000	364.4	3.00 (CLogP)	1	4	6
Inactives ²⁷⁷	NA	260.2	2.00 (CLogP)	0	3	4

Figure 04- 1 Top-10 PFAM protein families that contain high number of phase I (yellow), II (green), and III (orange) clinical trial targets along with the number of targets in each family

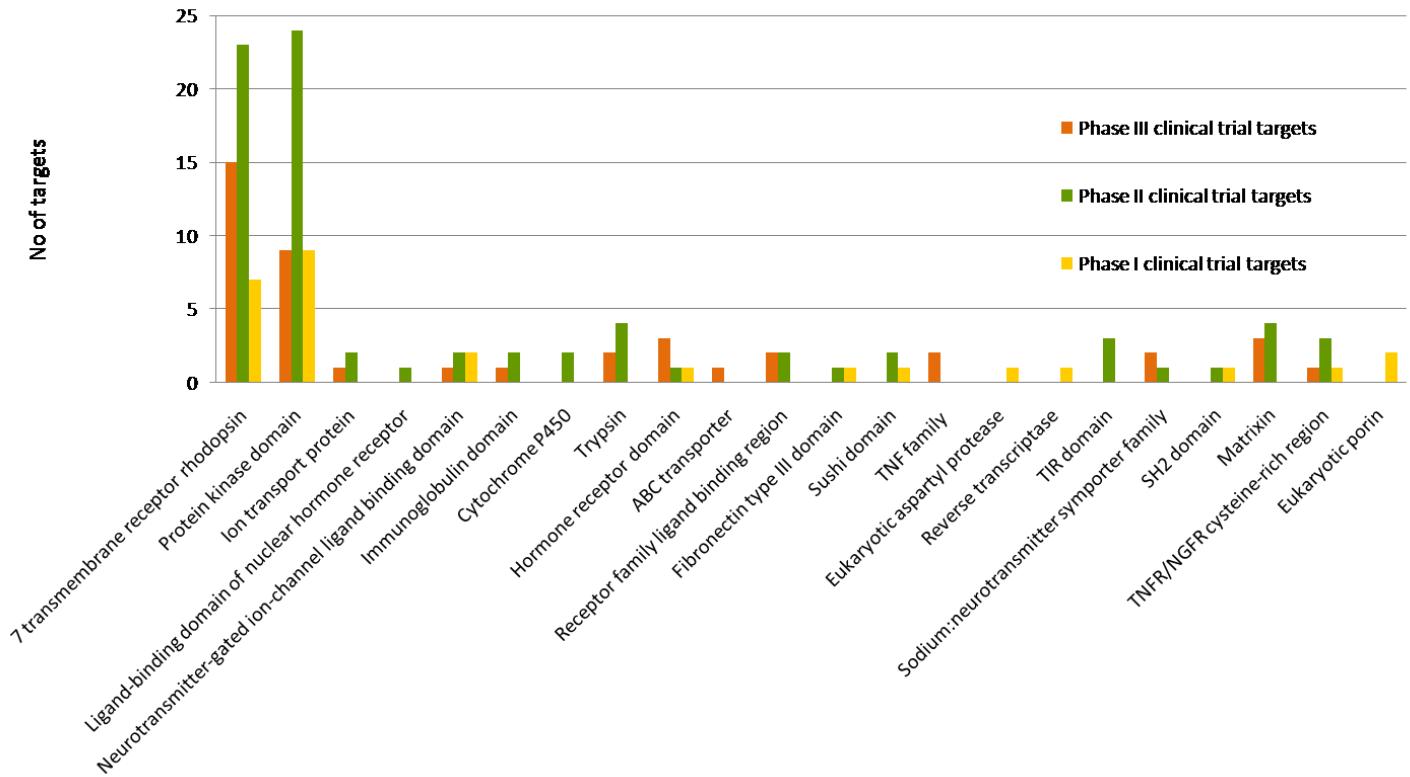


Figure 04- 2 Top-20 KEGG pathways that contain high number of phase I (yellow), II (green), and III (orange), and all clinical trial targets (brown) along with the number of targets in each pathway

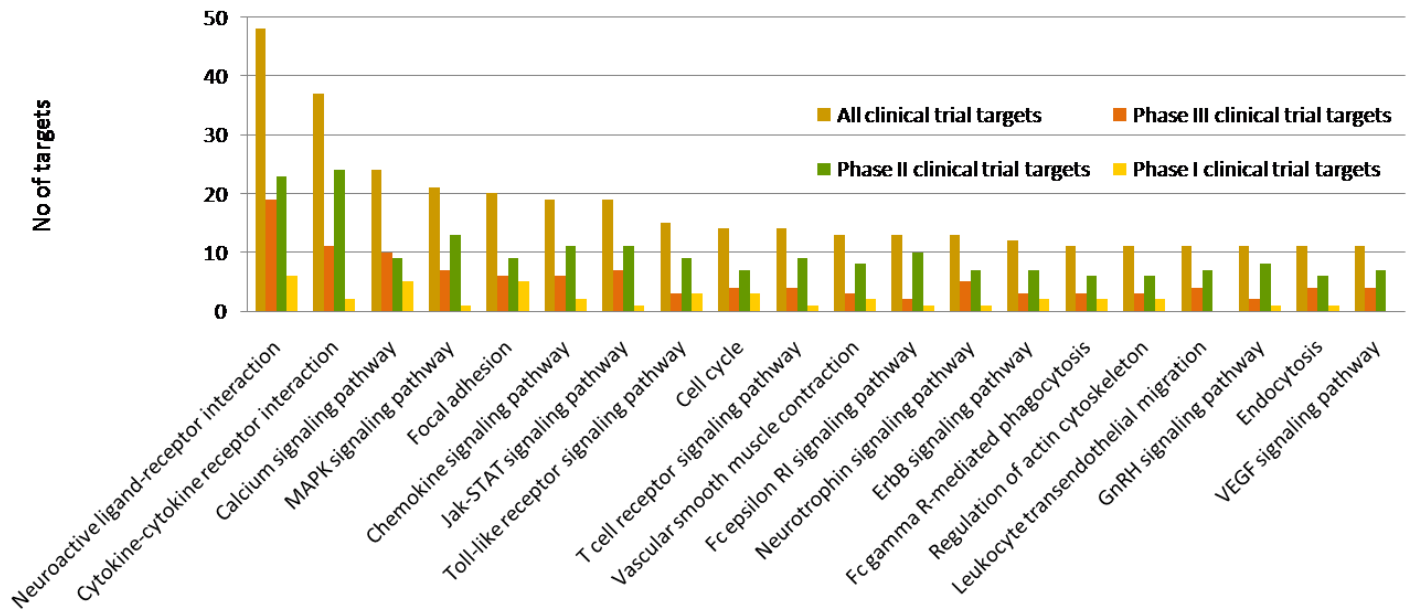


Figure 04- 3 Number of phase I (yellow), II (green), and III (orange) targets distributed in various sub-cellular locations

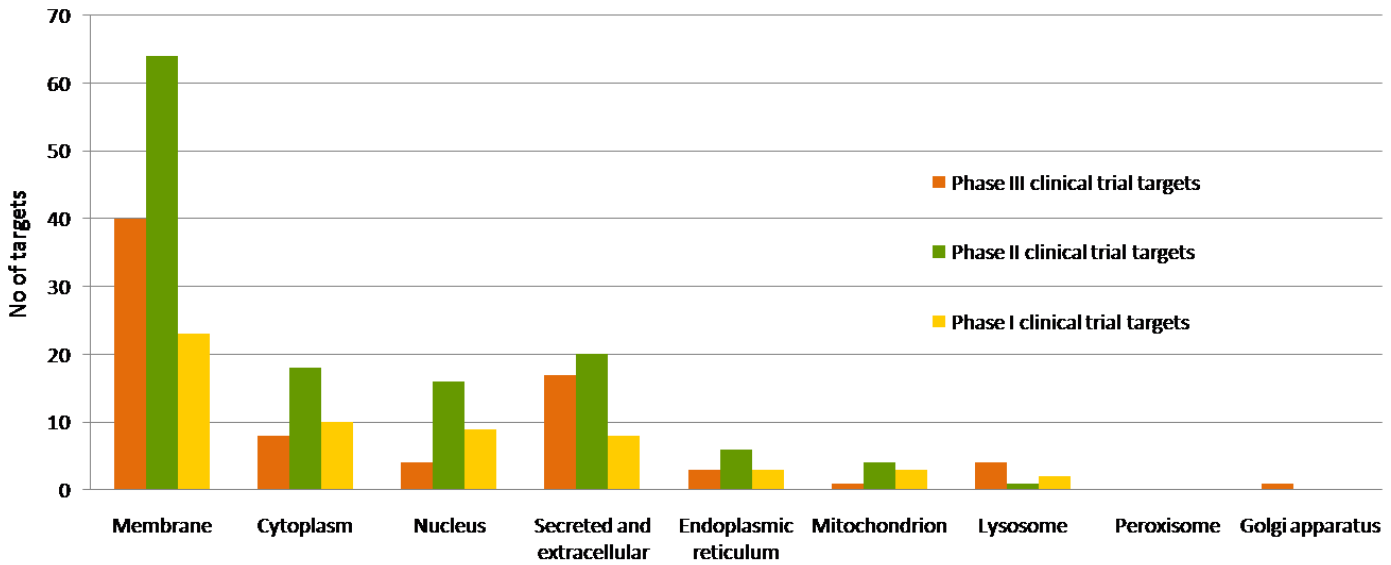


Figure 04- 4 Top-10 Pfam protein families that contain high number of clinical trial (orange) and successful (red) targets along with the number of targets in each family

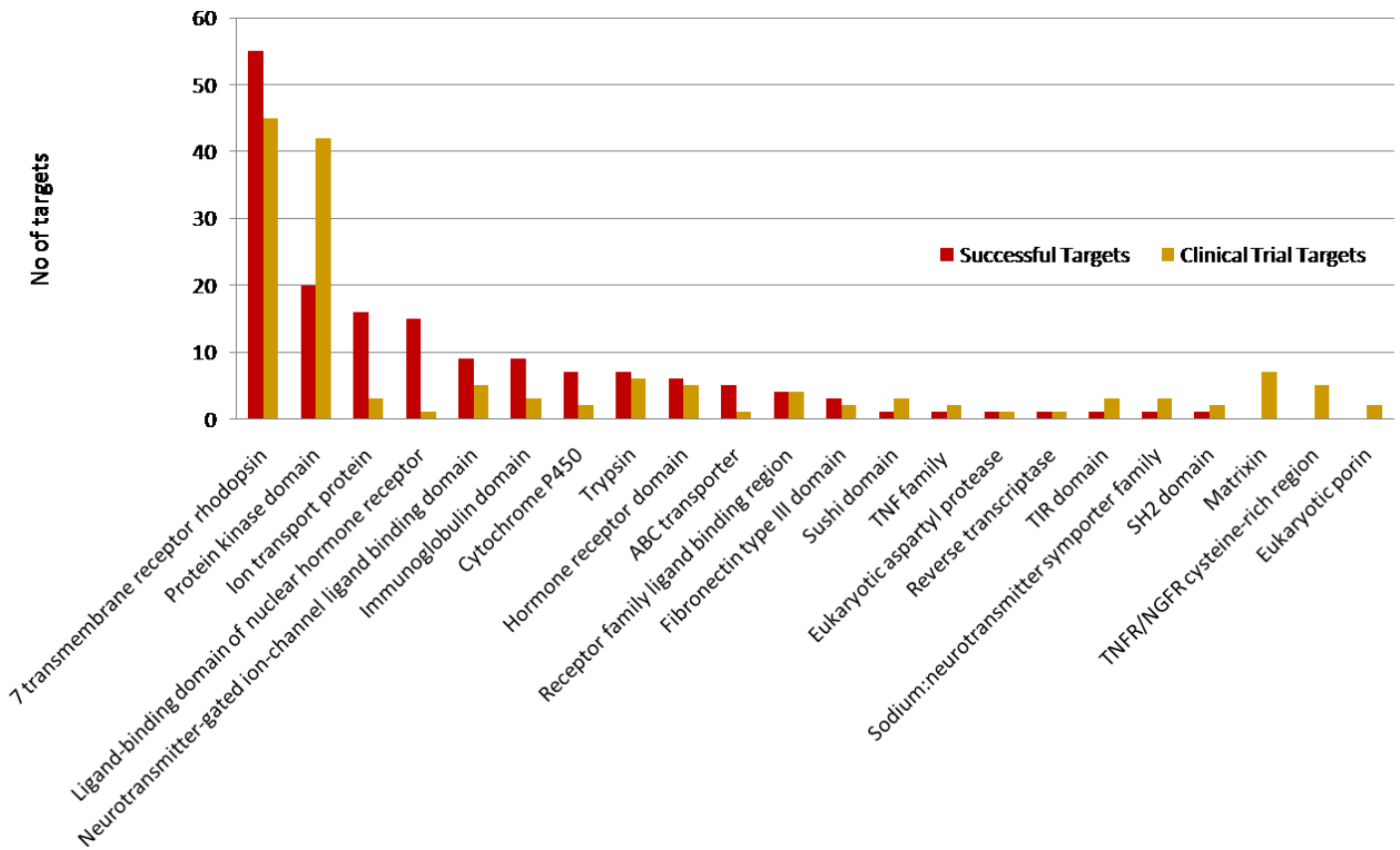


Figure 04- 5 Top-10 clinical trial (orange) and successful (red) targets targeted by phase II clinical trial drugs

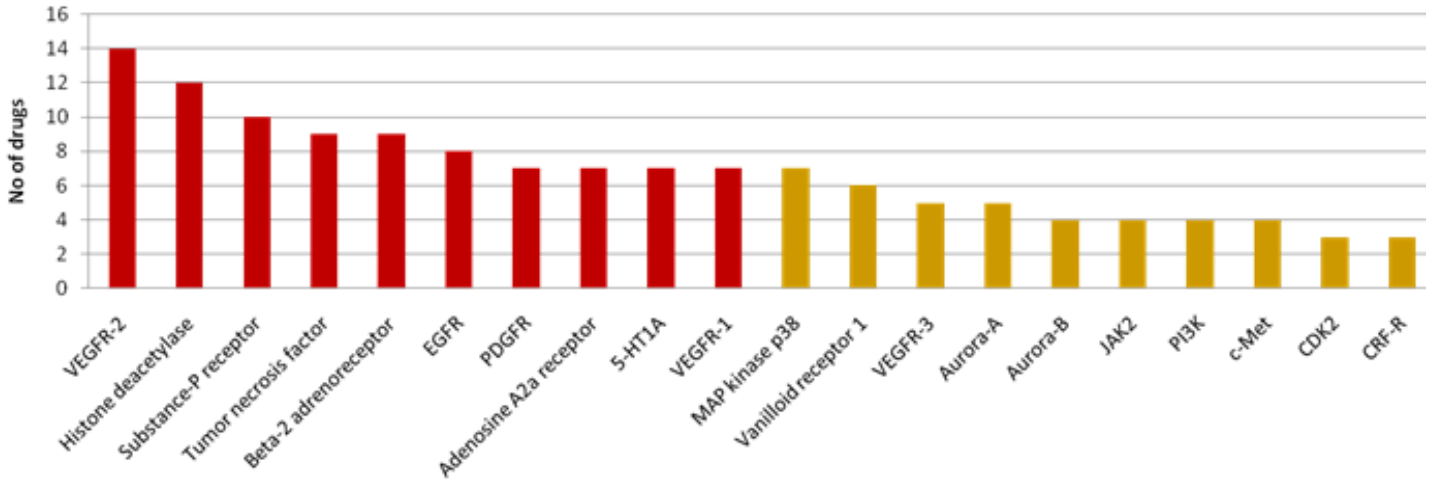


Figure 04- 6 Top-10 clinical trial (orange) and successful (red) targets targeted by phase III clinical trial drugs

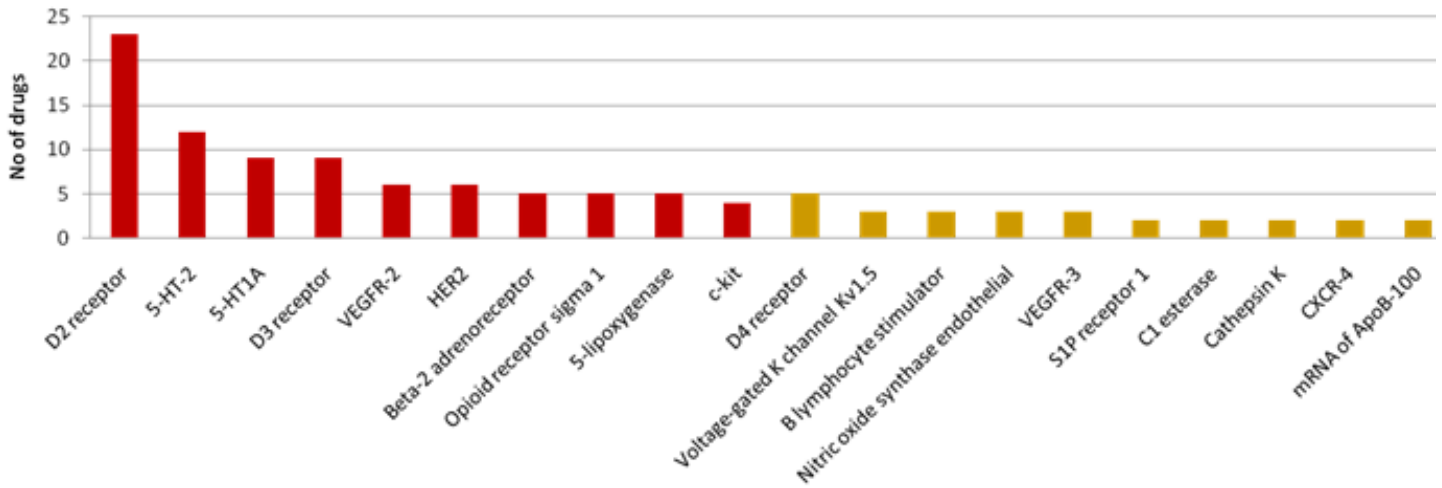


Figure 04- 7 Top-10 clinical trial (orange) and successful (red) targets targeted by all clinical trial drugs

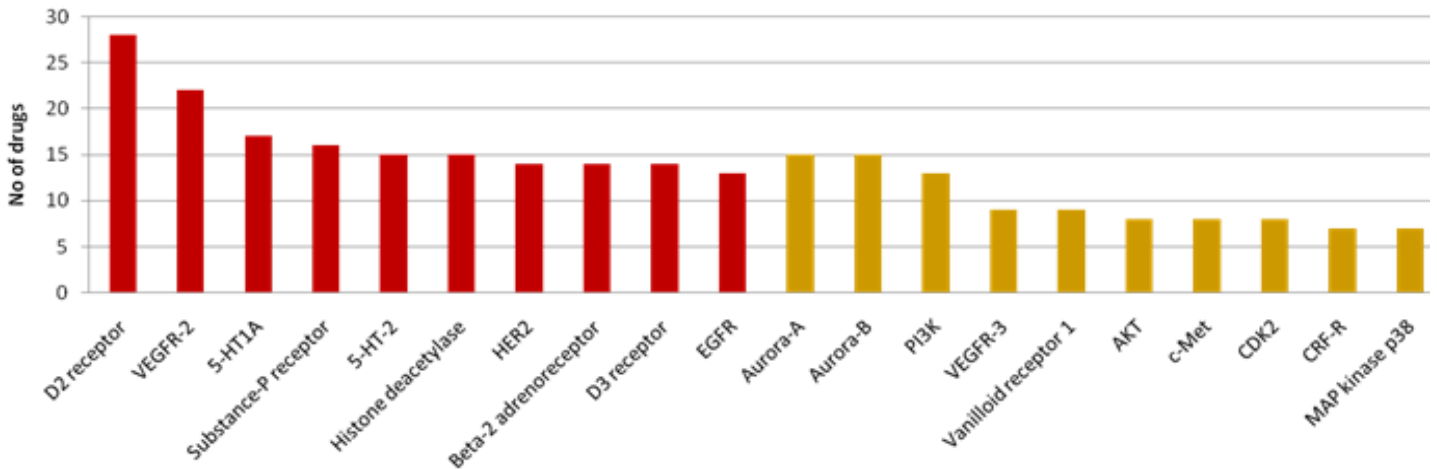


Figure 04- 8 Distribution of all clinical trial targets (orange) and the innovative successful targets (approved by FDA from 1995 to 2008) (red) by crudely estimated target exploration time

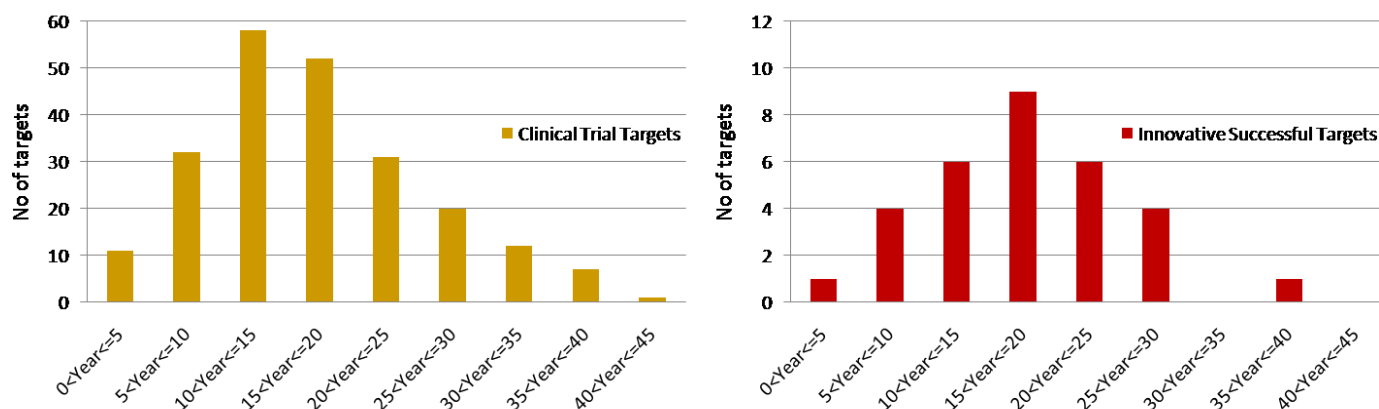


Figure 04- 9 Distribution of phase I (yellow), phase II (green), and phase III (orange) clinical trial targets by crudely estimated target exploration time

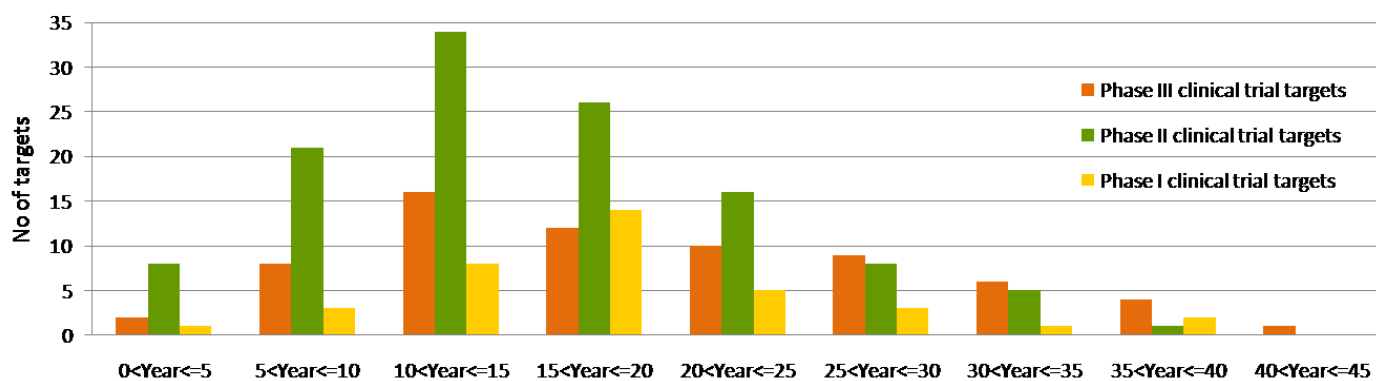
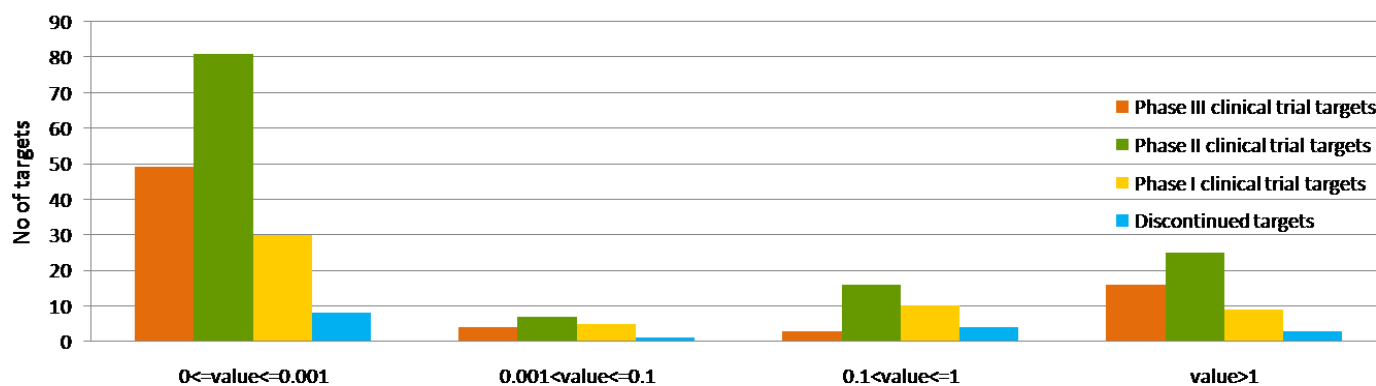


Figure 04- 10 Distribution of phase I (yellow), phase II (green), and phase III (orange) clinical trial targets and discontinued clinical trial targets (blue) by level of similarity to successful targets*



* The level of similarity to successful targets is classified into very similar, marginally similar, and un-similar with Blast E-values in the range of ≤ 0.001 , $0.001 \sim 0.1$ and ≥ 0.1 respectively.

Figure 04- 11 Distribution of all clinical trial targets and successful targets with respect to the number of human similarity proteins outside the target family

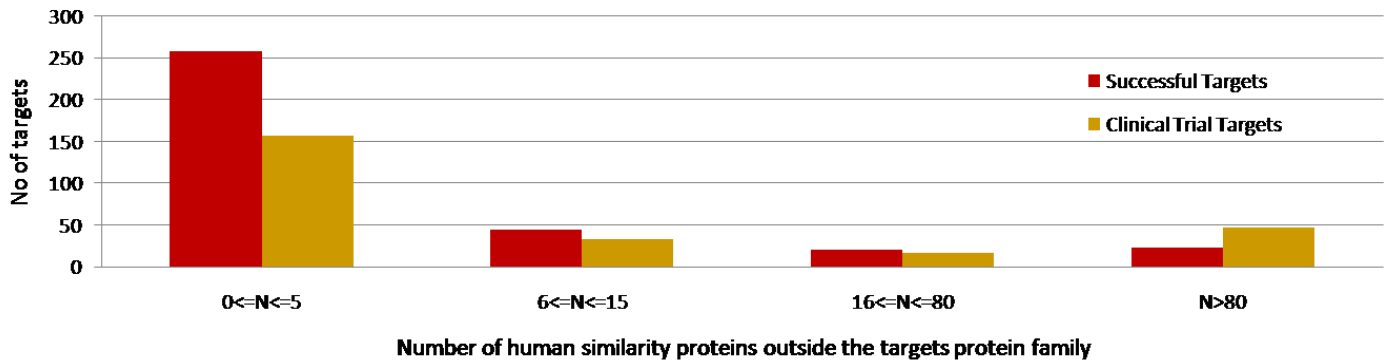


Figure 04- 12 Distribution of all clinical trial targets and successful targets with respect to the number of human pathways the target is associated with

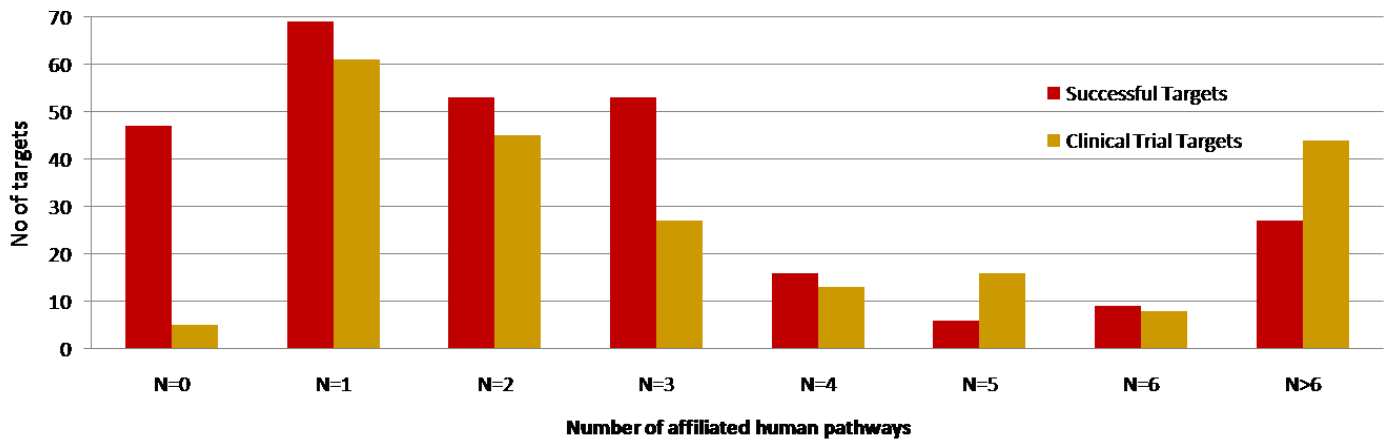


Figure 04- 13 Distribution of all clinical trial targets and successful targets with respect to the number of human tissues the target is distributed in

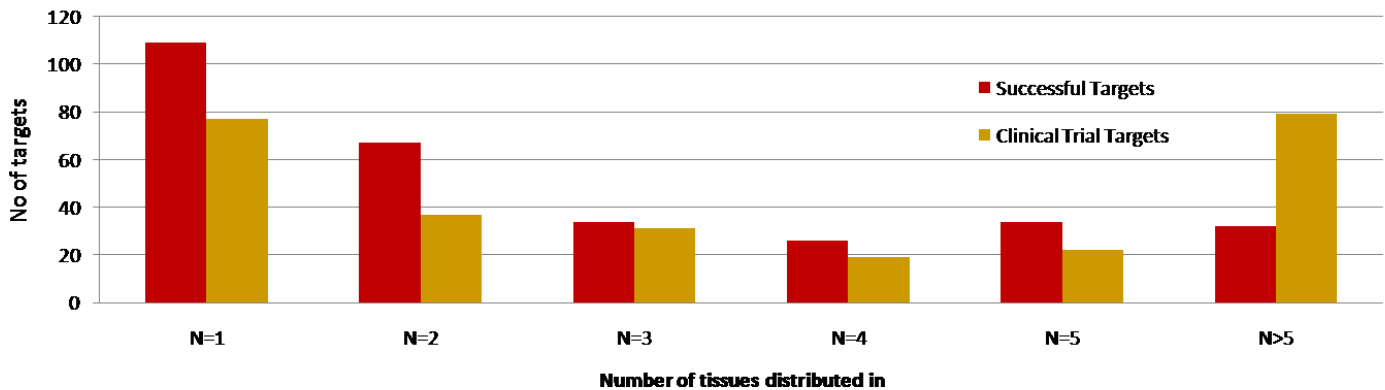


Figure 04- 14 Distribution of clinical trial drugs (orange) and approved drugs (red) by potency (IC_{50} , EC_{50} , K_i etc in units of nM)

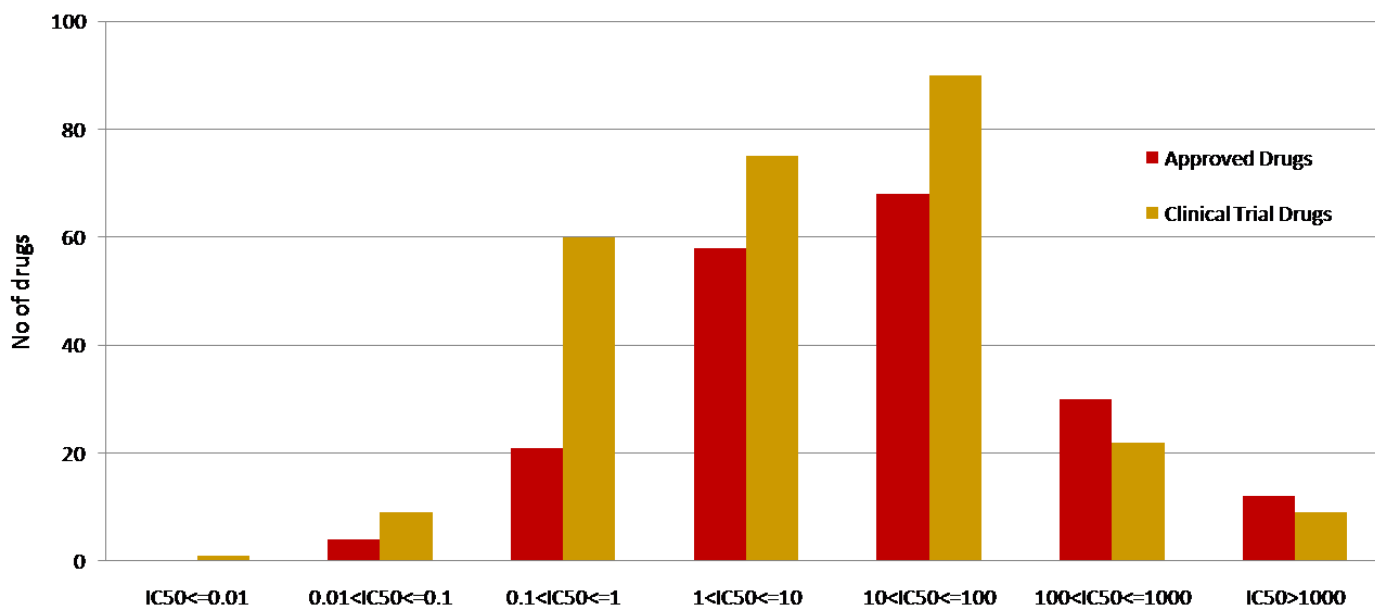


Figure 04- 15 Distribution of phase I (yellow), II (green), and III (orange) clinical trial drugs and discontinued clinical trial drugs (blue) by potency (IC_{50} , EC_{50} , K_i etc in units of nM)

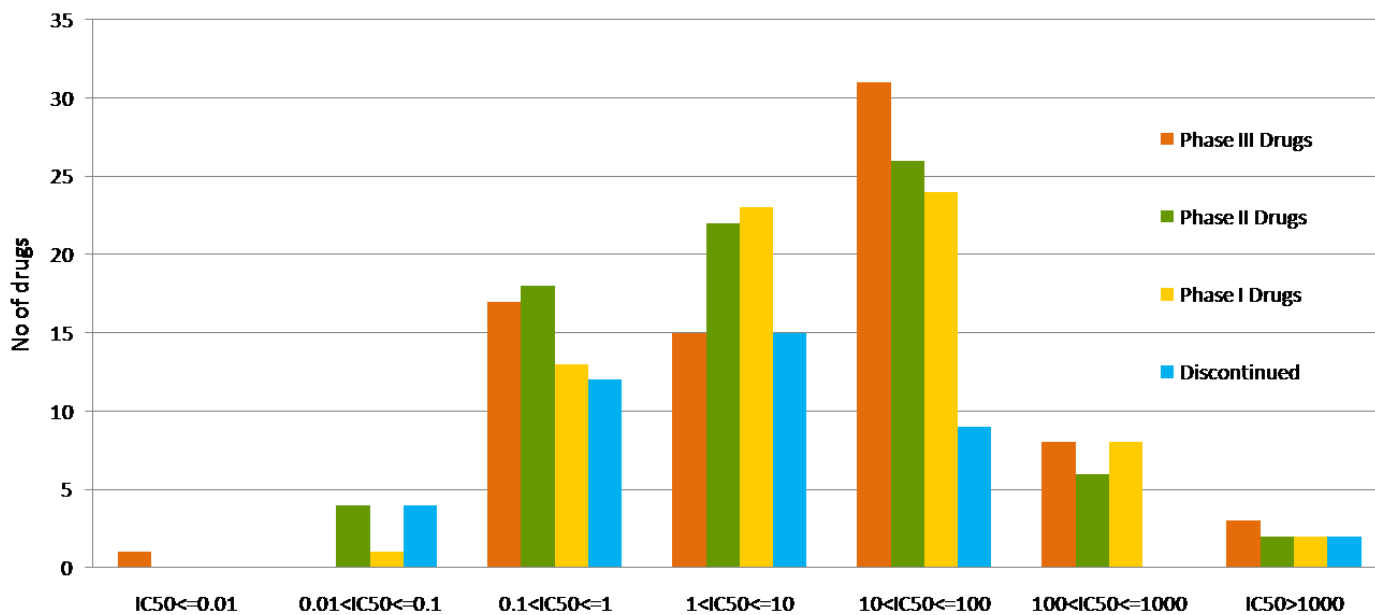


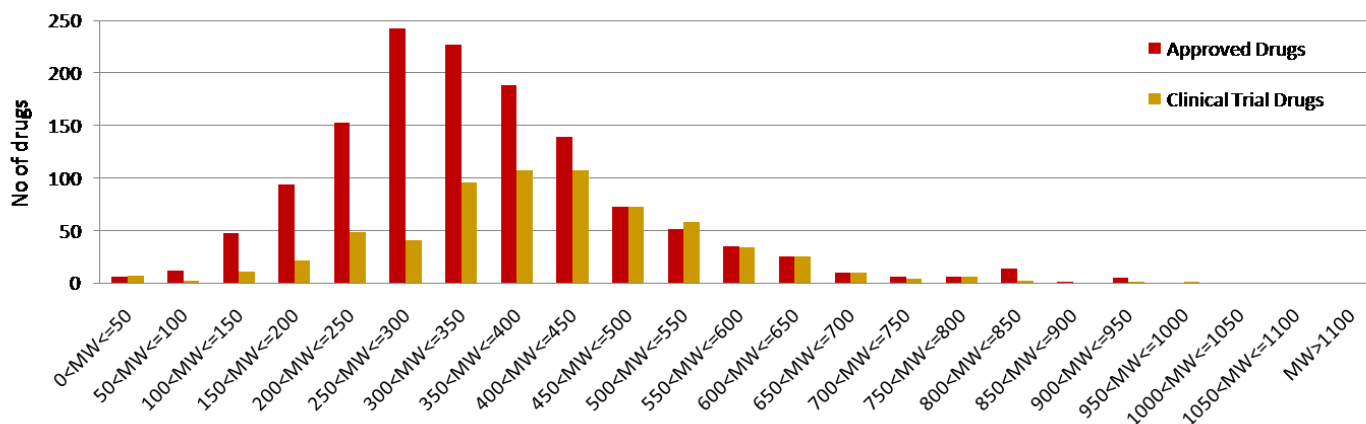
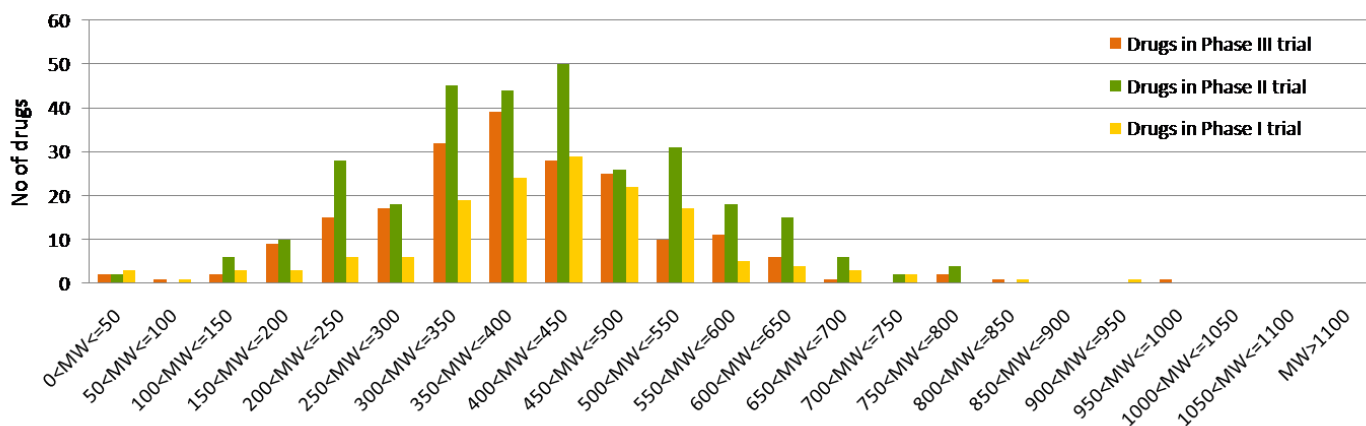
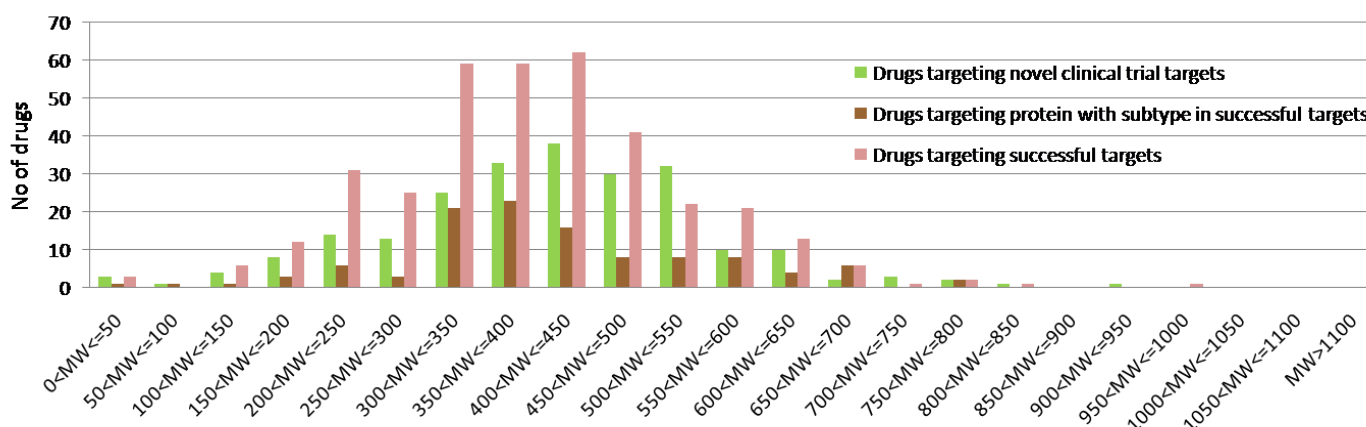
Figure 04- 16 Distribution of clinical trial drugs (orange) and approved drugs (red) by molecular weight**Figure 04- 17** Distribution of phase I (yellow), II (green), and III (orange) clinical trial drugs by molecular weight**Figure 04- 18** Distribution of clinical trial drugs targeting novel clinical trial targets (green), clinical trial targets with protein subtype as successful target (brown), and successful targets (pink) by molecular weight

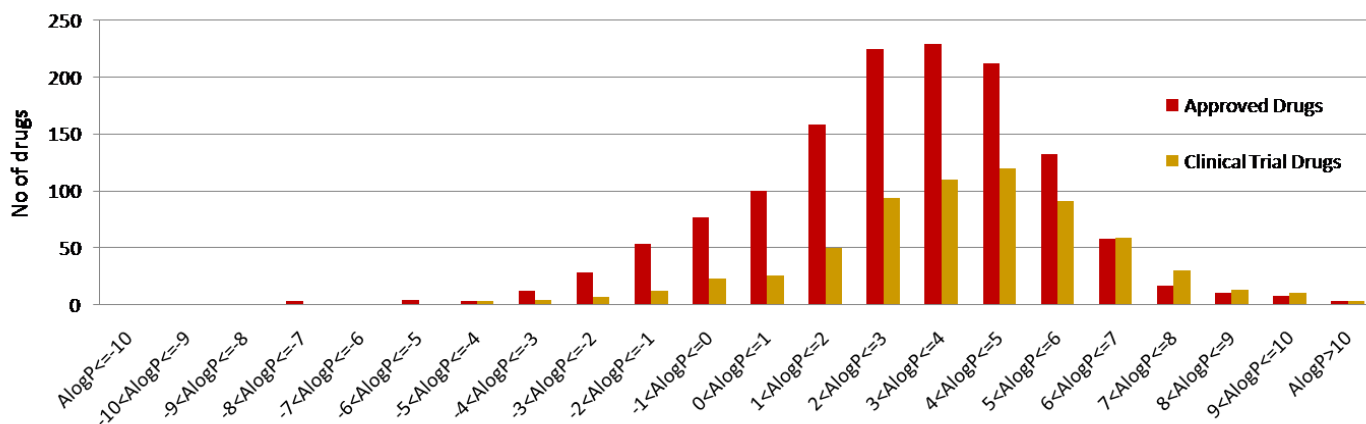
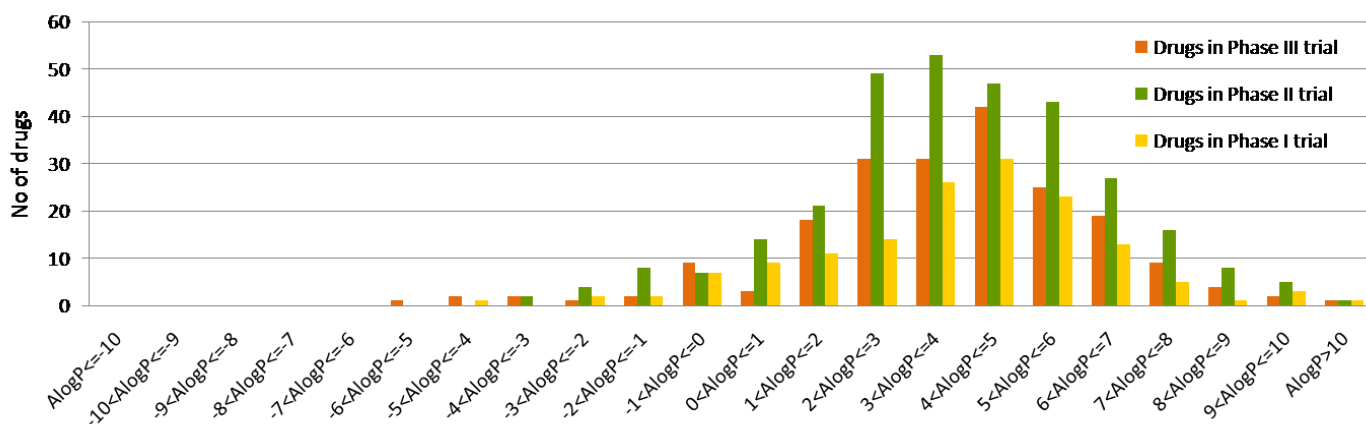
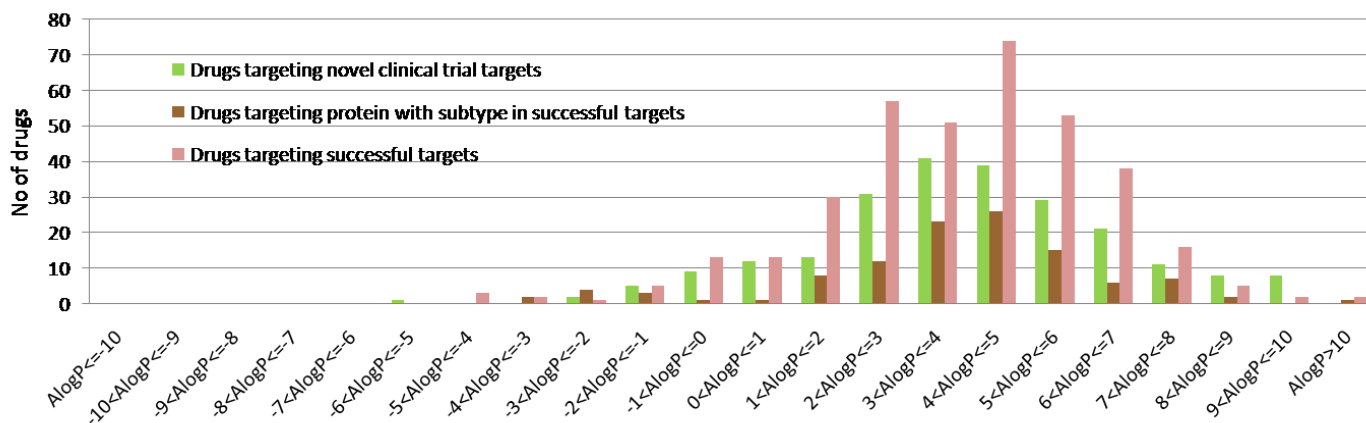
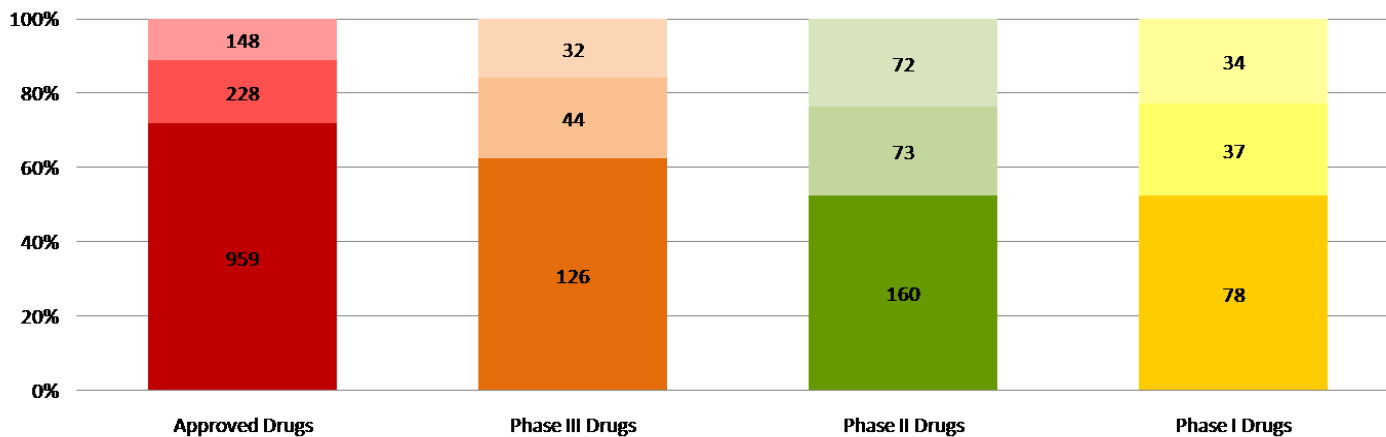
Figure 04- 19 Distribution of clinical trial drugs (orange) and approved drugs (red) by ALogP**Figure 04- 20** Distribution of phase I (yellow), II (green), and III (orange) clinical trial drugs and discontinued clinical trial drugs (blue) by ALogP**Figure 04- 21** Distribution of clinical trial drugs targeting novel clinical trial targets (green), clinical trial targets with protein subtype as successful target (brown), and successful targets (pink) by ALogP

Figure 04- 22 Percentage of phase I (yellow), II (green), III (orange) clinical trial drugs and approved drugs (red) obeying Lipinsky's rule of five (dark color), with one violation of rule of five (medium color) and the others (light color). The numbers in this figure refer to number of drugs.



Chapter 5 Identification of next generation innovative therapeutic targets: an application to clinical trial targets

The majority of clinical drugs achieve their therapeutic effects by binding and modulating activity of protein targets^{6,253,281}. Intensive efforts in searching for target^{75,76,227,282,283} have discover > 1,000 research targets (targeted by investigational agents only)²²⁷. These targets have been derived from analysis of disease relevance, functional roles, expression profiles and loss-of-function genetics of normal & disease state^{272,273,284-289}. Many targets have been targeted by target-selective lead^{227,290}. Despite heavy spending and exploration of techniques²⁹¹, fewer innovative targets have emerged²⁷². It typically takes 8~20 years to derive a marketed drug against these innovative targets²²⁷. Innovative targets refer to the targets with no other sub-type of the same protein successfully explored before.

Low productivity of innovative targets²⁷² has been attributed to problems in target selection and validation^{148,272,273}. A particular problem is inadequate physiological and clinical investigations^{272,273,292}. Drug effects are due to interactions with various sites of human physiological systems and pathways as well as its intended target, which collectively determine the success of target exploration²²⁷. Current efforts have been focused on target-selective agents minimally interacting with other human members of the target family^{6,293}. However, their possible interactions with other human proteins, pathways and tissues have not been fully considered, leading to frequent failures in subsequent development stages. Therefore, a target cannot be fully validated by considering disease relevance and target-selectivity alone^{272,273}.

Integrated target and physiology-based approaches have been proposed for target identification and validation^{272,273}. Different *in silico* approaches have been explored for target prediction based on sequence similarity^{226,227}, structural similarity and binding-site geometric and energetic features^{275,276}, target physicochemical and other characteristics detected by machine learning^{144,227,241}, and systems-profiles (similarity to human proteins, pathway and tissue distribution)^{227,270,274}. We evaluated whether target prediction can be improved by combinations of these approaches, which were tested against 155 clinical trial targets (Data are collected from CenterWatch Drugs in Clinical Trials Database 2009 <http://www.centerwatch.com/drug-information/pipeline/>), 864 non-clinical trial research targets²⁹⁴, 19 difficult targets currently discontinued in clinical trials (with clinical trial drug discontinued and no new drug entered clinical trial at the moment) (Data from CenterWatch Drugs in Clinical Trials Database 2009), and 65 non-promising targets failed in large-scale HTS campaigns²⁹⁵ or found non-viable in knockout studies²⁹⁶.

In summary, low target discovery rate has been linked to inadequate consideration of multiple factors collectively contributing to druggability. These factors include sequence, structural, physicochemical and systems profiles. Methods individually exploring each of these profiles for identifying target have been developed but have not been collectively used. In the following sections of the chapter, we evaluated the collective capability of these methods in identifying promising targets from 1,019 research targets based on the multiple profiles of up to 320 successful targets. As shown by the results, the collective consideration of multiple profiles demonstrated promising potential in identifying the innovative therapeutic targets.

5.1 Summary on materials and methods applied for drug target identification

As shown in **Chapter 2 Section 2.3**, four *in silico* approaches have been applied for the identification of drug targets, which include: physicochemical property of drug targets identified by machine learning; sequence similarity in drug-binding domains; structural fold comparison of drug-binding domains; and simple system level druggability rules.

5.1.1 Target classification based on characteristics of successful targets detected by a machine learning method

Promising targets can be separated from other proteins based on the structural and physicochemical characteristics of successful targets detected by a machine learning method. By using sequence-derived structural and physicochemical descriptors of the successful targets and those of other proteins, a machine learning algorithm attempts to separate successful targets from other proteins by searching for a projection function that maps the descriptors of successful targets and those of other proteins into separate regions in a high-dimensional feature space, and these regions are separated by easily defined borders. A research target is classified as promising if it is located in the region of successful targets, which is not necessarily similar in sequence to any successful target because the mapping to the feature space is typically nonlinear and the proteins are characterized by structural and physicochemical descriptors rather than sequence.

The machine learning method used in this work is support vector machines (SVM), which is a supervised learning methods used for classification of objects (e.g. proteins) into two classes (e.g. promising targets and other proteins) and has been applied to target prediction²²⁷. Details of SVM can be found in **Chapter 2 Methodology Section 2.3.1**. In this work, a nonlinear SVM was used with the following kernel function:

$$K(x_i, x_j) = e^{-|x_j - x_i|^2 / 2\sigma^2} \quad (1)$$

The non-linear SVM projects feature vectors into a high-dimensional feature space using the kernel function defined above. The linear SVM was then applied to produce a single hyper-plane that separates targets from non-targets. A SVM prediction system was developed by using the feature vectors of the structural and physicochemical properties of 320 successful targets and 24066 putative non-targets generated by a procedure described in our earlier study²⁴¹, which was used to screen the 1,019 research targets for identifying potential promising targets. The sequence-derived structural and physicochemical descriptors used in SVM include amino acid composition, dipeptide composition, sequence autocorrelation descriptors, sequence coupling descriptors, and the descriptors for the composition, transition and distribution of hydrophobicity, polarity, polarizability, charge, secondary structures, surface tension, and normalized Van der Waals volumes¹⁹¹.

5.1.2 Sequence similarity analysis between drug-binding domain of studied target and that of successful target

BLAST²¹⁹ was applied to determine the level of similarity between sequence of the drug-binding domain of each studied research target and the sequence of drug-binding domain

of each of the 168 successful targets with identifiable drug-binding domains. The BLAST program was downloaded from NCBI (<ftp://ftp.ncbi.nlm.nih.gov/blast/executables/>). A stricter BLAST cut-off, E-value = 0.001, was used for selecting research targets which are similar to a successful target, i.e., the E-value of the drug-binding domains is ≤ 0.001 . The detail strategy utilized by this analysis is described in **Chapter 2 Methodology Section 2.3.2**.

5.1.3 Structural comparison between drug-binding domain of studied target and that of successful target

The ligand-binding or catalytic sites are the most relevant subsets of a domain, which are normally located within the so-called ligand-sensing core where actual catalytic conversion of enzyme substrates, or the binding event of small-molecule ligands, occurs. It has been suggested that structural similarity considerations should be confined to ligand-sensing cores, instead of whole domains, according to 3D similarities with respect to so-called protein structure similarity clusters²²⁵. In this study, ligand-sensing or catalytic cores of drug-binding domain of the studied research target were clustered against those of 129 successful targets with available 3D structure based on visual inspection and structural superimposition and alignment tools in SYBYL (SYBYL® 6.7 Tripos Inc., St. Louis, Missouri, USA). and Insight II (Insight II® Accelrys Software Inc, San Diego, CA) following the same procedure used for generating SCOP structural folds²²⁴. For details information of structural comparison, please refer back to **Chapter 2 Methodology Section 2.3.3**.

5.1.4 Computation of number of human similarity proteins, number of affiliated human pathways, and number of human tissues of a target

These quantities are needed for determining whether or not a studied target obeys the simple systems-level druggability rules. Human similarity proteins of a target are those human proteins whose drug-binding domain is similar to that of the studied target by using the same BLAST method as that for analyzing sequence similarity between drug-binding domain of studied target and that of successful target²¹⁹. Information about the affiliated pathways of a target was obtained from KEGG²²⁸ (<http://www.genome.jp/kegg/>). In estimating the number of human tissues of each target is distributed, relevant data from the Swissprot database were used. We were able to find the published literature for 92% of these data, and a random check of these publications confirms the quality of the data. We have also used the level-4 tissue-distribution data from another database, TissueDistributionDBs (http://genome.dkfz-heidelberg.de/menu/tissue_db/index.html), to derive the tissue distribution pattern of the same set of 158 successful targets. A target is assumed to be primarily distributed in a tissue if no less than 8% of the total protein contents are distributed in that tissue. Approximately 28, 24, 19, 10, 6, 6, 5, and 1% of these targets were found to be affiliated with 1 to 8 tissues, respectively, which are roughly similar to those derived from Swissprot data²²⁷, although the definition and content of these databases are somehow different. Therefore, our estimated tissue distribution profiles are quite stable even though the exact percentages may differ by some degrees. Please refer back to **Chapter 2 Methodology Section 2.3.4** for detail information.

5.2 Target identification by collective analysis of sequence, structural, physicochemical, and system profiles of successful targets

Each *in silico* target prediction approach has its unique advantages and limitations. Sequence similarity to the drug-binding domain of a successful target may indicate druggability, which has been extensively explored for target identification^{226,275}. However, it cannot fully capture druggable features not reflected by homology²⁷⁵ and tend to indiscriminately select homologous proteins. Targets can be identified by structural similarity to drug-binding domain and binding-site geometric and energetic features²⁷⁵, which are less effective for covering proteins of unknown structure and for describing systems profiles.

Druggability is collectively determined by target structural and physicochemical properties, ability to conduct certain interactions and functions, and patterns of pathway, sub-cellular and tissue distributions²²⁷. Many of these individual properties can be predicted by machine learning⁶⁸ which have been explored for target prediction^{144,201,227,241}. This approach cannot fully capture such systems profiles as pathway affiliation and may disproportionately interpret certain physicochemical properties due to biases in protein descriptors or training datasets. Simple systems-level druggability rules have been derived²²⁷, which are summarized as follows: targets are similar to fewer (<15) human proteins of non-target family and associated with fewer (≤ 3) human pathways tend to bind drugs with reduced side-effects, and high efficacy drugs

may be more easily derived from targets expressed in fewer tissues (≤ 5) or located within blood vessels or cells lining the arteries where they have higher priority (P) to bind drugs than targets in other tissues. These systems-level rules are not intended for describing structural, physicochemical, and functional aspects of druggability.

These limitations may be reduced if those approaches are combined. Four *in silico* methods were developed from the relevant profiles of up to 320 successful targets in TTD database²⁹⁴. Method A measures drug-binding domain sequence similarity against those of 168 successful targets with identifiable drug-binding domains. Method B studies drug-binding domain structural similarity against those of 129 successful targets with available structures. Method C predicts druggable proteins from a machine learning model trained by 320 successful targets²⁴¹. Method D evaluates whether the systems-level druggability rules²²⁷ are satisfied. More detailed descriptions about these methods are given in **Chapter 2** Methodology.

5.3 Performance of target identification on clinical trial, non-clinical trial, difficult, and non-promising targets

The collective predictive performance of the four methods was tested against clinical trial (from CenterWatch Drugs in Clinical Trials Database) and non-clinical trial research targets²⁹⁴. Clinical trial targets that have drugs in multiple phases are only included in the highest phase category. The best overall performance was produced by the combination of at least three methods, which maximize the collective predictive capability of the methods and minimize the impact of limited structural availability. This combination identified 50% of the 30 phase III (**Table 05-1**), 25% and 10% of the 84 phase II and 41 phase I (**Table 05-2**), and 4% of the 864 non-clinical trial research targets as promising. We were unable to find a report about target success rates in different development stages. It is noted that the reported probabilities of successes in developing systemic broad-spectrum antibacterials are 67%, 50%, 25%, and 3% in phase III, phase II, phase I, and preclinical stages²⁹⁵. The percentages of the identified promising clinical trial targets are lower than but roughly follow a similar descending trend as the reported drug development rates. The overall performance of different combinations is given in **Table 05-3**. These combinations enriched Phase II and phase III target identification rate by 4.0~6.0 fold over random selection, with the combination of all four methods producing the highest enrichment.

The 15 identified promising phase III targets include 7 of the 8 targets with positive phase III results. These include 5 innovative targets without a protein-subtype as a successful target (BK-2 receptor, C1 esterase, CCK-A receptor, NK-2 receptor and

plasma kallikrein) and 2 conventional targets having a different protein-subtype as a successful target (5HT3 receptor and CXCR4). Overall, 60%, 43%, and 50% of the predicted phase III, phase II, and phase I targets are innovative, which seems to indicate substantial level of successes in exploring novel targets. Most of the identified promising clinical trial targets are from the highly successful GPCR, tyrosine kinase, serine protease and ABC transporter families for the treatment of cancers, cardiovascular diseases, neural disorders, arthritis, diabetes and obesity, which suggests that these families continue to be attractive sources for target discovery^{59,227,253}.

The 15 phase III targets dropped by the combination method (**Table 05-4**) include MMPs, kinases of CMGC, AGC and DAGK classes, farnesyltransferases, oxygenase, phospholipase and others. Only one of these, heme oxygenase, has a positive phase III result reported in 2004. It is noted that this protein is important for attenuating oxidative stress and inflammation and its inhibition may lead to some adverse effects²⁹⁷. The difficulty in exploring some of these targets has been reported²²⁷. MMP inhibitors have been explored since the early 1990s but their trials have not yielded good results due primarily to the lack of subtype selectivity, bioavailability and efficacy as well as inappropriate study design²⁹⁸. Despite successes in developing several tyrosine kinase inhibitors, kinase inhibitor discovery remains difficult particularly for non-tyrosine kinase classes partly due to broad promiscuity that causes off-target side effects²⁹⁹ and network compensatory actions³⁰⁰.

The combination method dropped 17 of the 19 difficult targets currently discontinued in clinical trials (**Table 05-5**) and 63 of the 65 non-promising targets failed in HTS

campaigns or found non-viable in knockout studies (**Table 05-6**). 12 of the 17 un-predicted difficult targets have been discontinued since 2004 without another drug entering clinical trial. In the HTS campaigns for testing 70 antibacterial targets, up to ~500,000 compounds have been screened at a concentration of 10 μM , 33 of which have yielded no hit and can thus be considered to be highly un-promising²⁹⁵. Target knockout, extensively explored for target validation, has been applied to the validation of 55 targets in *Mycobacterium tuberculosis*, 32 of which have been found to be non-viable for developing drugs²⁹⁶. The low rate in selecting these difficult and unpromising targets suggests that combinations of target prediction methods are capable of eliminating unpromising as well as selecting promising ones. As a supplementary data, **Table 05-7** shown the definitions and structures (if available) of drugs and compounds used in this chapter.

In conclusion, collective use of multiple *in silico* methods is capable of identifying high percentages of phase III targets including most of the targets of positive phase III results, and of eliminating difficult and un-promising ones. Our study suggests that comparative analysis of multiple profiles of successful targets provides useful clues to the identification of promising targets. Overall, 71 targets were predicted as promising from a pool of 1,019 targets. This number is likely constrained by the limited knowledge from the 320 known successful targets and limited structural information for large percentage of targets. Rapid progress in genomics²⁸⁴, structural genomics²⁷⁵, and proteomics²⁸⁵ is revolutionizing target discovery. In addition to high-throughput technologies⁷⁶, cellular²⁸⁸ and physiological studies^{272,273}, various *in silico* methods are being developed. These methods explore comparative sequence analysis^{226,275}, structural analysis²⁷⁵, ligand-

protein inverse docking²⁷⁹, machine learning of druggability characteristics²²⁷ and system-related druggability profiles^{201,227} for recognizing target-like and druggable proteins. These progresses combined with increased molecular understanding of diseases²⁸⁰ and their corresponding targets²²⁷ enable the development of efficient tools for identifying innovative targets of new therapies and personalized medicine.

Table 05- 1 List of phase III targets identified by combinations of at least three of the methods A, B, C and D used in this study

Target	Predicted as Promising by Combination of Methods	Number of target affiliated pathways	Number of human similarity proteins outside target family	Number of tissues target is primarily distributed	Targeted Disease Conditions	Target Exploration Status (Tested Drug)	Positive Results in Phase III Trial Reported in Company Website (Year of Report)
CCK-A receptor*	Combination of A, B, C, D	2	1	1	Irritable Bowel Syndrome	Phase III (dexloiglumide)	Favourable topline results in patients with constipation–predominant Irritable Bowel Syndrome (2007)
Coagulation factor IIa*	Combination of A, B, C, D	3	0	5	Venous Thromboembolism	Phase II/III (SR-123781A)	
NTRK1*	Combination of A, B, C, D	3	6	2	Acute Myeloid Leukemia	Phase II/III (lestaurtinib)	
5HT 3 receptor	Combination of A, C, D	1	0	2	Irritable Bowel Syndrome	Phase III (cilansetron)	Positive data for treating irritable bowel syndrome with diarrhea predominance (2004)
Heparanase*	Combination of A, C, D	2	0	2	Hepatocellular Cancer	Phase III (PI-88)	

MDR 3	Combination of A, C, D	1	0	3	Acute Myeloid Leukemia	Phase III (LY335979)	
Orexin-OX1/OX2 receptor*	Combination of A, C, D	1	0	2	Sleep Disorders	Phase III (almorexant)	
Somatostatin receptor 1	Combination of A, C, D	1	0	5	Cushing's disease, Renal Cell Carcinoma	Phase III (Pasireotide), Phase II (CAP-232)	
NK-2 receptor*	Combination of A, C, D	2	0	3	Depression	Phase III (Saregutant)	Overall statistically significant efficacy versus placebo, well tolerated (2007)
BK-2 receptor*	Combination of A, B, C	4	0	P	Hereditary Angioedema, Traumatic Brain Injuries	Phase III (icatibant), Phase II (anatibant)	Positive results for the treatment of hereditary angioedema (2006)
Thrombin receptor*	Combination of A, B, C	4	0	5	Cardiovascular Disorders	Phase III (SCH-530348)	
CXCR4	Combination of A, B, D	3	2	P	Non-Hodgkin's Lymphoma, Late-stage Solid Tumors	Phase III (AMD-3100), Phase I/II (AMD-070), Phase I (MSX-122)	Positive results for the treatment of multiple myeloma (2007)
C1 esterase*	Combination of A, B, D	1	3	P	Hereditary Angioedema	Phase III (C1-INH)	Positive results for treating hereditary angioedema, significantly decreases the

number of attacks in patients (2007)

NPYR5	Combination of A, B, D	1	0	2	Obesity	Phase III (CGP71683A)	
Plasma kallikrein*	Combination of A, B, D	1	0	5	Hereditary Angioedema	Phase III (DX-88)	Positive top-line results for treating hereditary angioedema, well tolerated (2007)

Targets marked by* are innovative targets without a protein-subtype as a successful target. Tissue distribution “P” represents cases where target is distributed in >5 tissues but the disease relevant targets are located within blood vessels or cells lining the arteries where they have higher priority to bind drugs.

Table 05- 2 List of phase II and phase I targets identified by combinations of at least three of the methods A, B, C and D used in this study

Research Target	Identified by Combination	Number of target affiliated pathways	Number of human similarity proteins outside target family	Number of tissues target is primarily distributed	Targeted Disease Conditions	Target Exploration Stage (testing drug)
C-C chemokine receptor 2*	Combination of A, B, C, D	1	0	1	Rheumatoid Arthritis, Multiple Sclerosis	Phase II (INCB3284), Phase I (CCX915)
ErbB-4	Combination of A, B, C, D	3	4	2	Breast Cancer	Phase II (CI-1033)
FGFR-3	Combination of A, B, C, D	3	0	4	Solid Tumors, Multiple Myeloma	Phase II (XL999), Phase I (CHIR-258)
Guanylate cyclase B*	Combination of A, B, C, D	3	0	1	Heart Disease	Phase I a (CD-NP), Preclinical (guanilib)
HDAC4	Combination of A, B, C, D	1	1	P	Basal Cell Carcinoma, Melanoma, Cancer	Phase II (Avugane, romidepsin, MS-275, PXD101)
Neuropeptide Y	Combination	1	0	4	Obesity	Phase II (Obinepitide)

receptor 2	of A, B, C, D					
Neuropeptide Y receptor 4	Combination of A, B, C, D	1	0	3	Schizophrenia, Schizoaffective Disorders	Phase I/II (TM30339)
Toll-like receptor 3	Combination of A, B, C, D	1	0	2	Human Papillomavirus Infections	Phase II (HspE7)
FGFR-1	Combination of A, B, C	5	0	>10	Coronary Heart Disease, Solid Tumors	Phase II (XL999), Phase II (FGF-1)
PKC-gamma	Combination of A, B, C	16	0	4	Acute Myocardial Infarction	Phase II (Midostaurin), Phase I/II (KAI-9803)
Tyrosine-protein kinase receptor HTK*	Combination of A, B, C	1	4	>10	Lung Cancer, Solid Tumors	Phase II (XL647)
Histamine H3 receptor	Combination of A, C, D	1	0	4	Attention-deficit hyperactivity disorder, Alzheimer's disease, Schizophrenia	Phase II (cipralisant), Phase I (ABT-239)
Leukotriene B4 receptor 1*	Combination of A, C, D	1	0	4	Cancer, Renal Cell Carcinoma	Phase II (LY293111), Phase I (Biomed 101)
Motilin receptor*	Combination of A, C, D	1	0	1	Irritable Bowel Syndrome, Gastrointestinal Disorders	Phase II b (mitemcinal), Phase I (KOS-2187)
NK-3 receptor*	Combination of A, C, D	2	0	1	Schizophrenia, Schizoaffective Disorders	Phase II b (osanetant), Phase II (talnetant)

Somatostatin receptor type 4	Combination of A, C, D	1	2	3	Solid Tumors	Phase II (CAP-232)
Tissue kallikrein-2*	Combination of A, C, D	1	0	2	Atopic Dermatitis	Phase II (Dermolastin)
Toll-like receptor 8	Combination of A, C, D	1	0	5	Genital Warts, Systemic Lupus Erythematosus	Phase II (resiquimod), Phase I (CPG 52364)
CDK7	Combination of A, B, D	1	1	P	B-cell malignancies	Phase I (SNS-032)
Coagulation factor IX*	Combination of A, B, D	1	5	1	Thrombosis, Venous Thromboembolism	Phase II a (REG1), Phase I completed (TTP889)
Melanocortin receptor*	Combination of A, B, D	1	0	3	Sexual (Female) and Erectile Dysfunction	Phase II b (bremelanotide)
Metabotropic glutamate receptor 2/3*	Combination of A, B, D	1	0	1	Psychosis	Phase II (LY2140023, LY354740)
Peroxisome proliferator-activated receptor delta	Combination of A, B, D	3	0	P	Obesity	Phase II (MBX-8025), Phase I (KD3010)
Serine/threonine-protein kinase Chk2	Combination of A, B, D	2	0	4	Solid Tumors	Phase I (XL844), Phase I (UCN-01)

Serine/threonine- protein kinase PLK*	Combination of A, B, D	1	1	P	Pancreatic, prostate and a number of other cancers	Phase I (HMN-214)
--	---------------------------	---	---	---	---	-------------------

Targets marked by* are innovative targets without a protein-subtype as a successful target. Tissue distribution P represents cases where target is distributed in >5 tissues but the disease relevant targets are located within blood vessels or cells lining the arteries where they have higher priority to bind drugs.

Table 05- 3 Statistics of promising targets selected from the 1,019 research targets by combinations of methods A, B, C and D, and clinical trial target enrichment factors

Method or Combination	No and Percentage of the 30 Phase III Targets Predicted by Method	No and Percentage of the 84 Phase II Targets Predicted by Method	No and Percentage of the 41 Phase I Targets Predicted by Method	No and Percentage of the 864 Non-Clinical Trial Targets Predicted as Target by Method	Target Prediction Enrichment Factor for Phase II and III Targets	Target Prediction Enrichment Factor for All Clinical Trial Targets
Combination of A, B, C, D	3 (10.0%)	7 (8.3%)	1 (2.4%)	4 (0.5%)	6.0	4.8
Any 3-combination of A, B, C, D	15 (50.0%)	21 (25.0%)	4 (9.8%)	31 (3.6%)	4.5	3.7
Combination of A, B, C	5 (16.7%)	10 (11.9%)	1 (2.4%)	8 (0.9%)	5.6	4.4
Combination of A, B, D	7 (23.3%)	11 (13.1 %)	4 (9.8%)	18 (2.1%)	4.0	3.6
Combination of A, C, D	9 (30.0%)	14 (16.7%)	1 (2.4%)	14 (1.6%)	5.4	4.2
Combination of B, C, D	3 (10.0%)	7 (8.3%)	1 (2.4%)	6 (0.7%)	5.3	4.3
Any of A, B, C, D	28 (93.3%)	51 (60.7%)	25 (61.0%)	283 (32.8%)	1.8	1.8
A	18 (60.0%)	39 (46.4%)	17 (41.5%)	125 (14.5%)	2.6	2.4
B	11 (36.7%)	26 (31.0%)	8 (19.5%)	95 (11.0%)	2.4	2.1
C	13 (43.3%)	21 (25.0%)	3 (7.3%)	75 (8.7%)	2.7	2.2
D	23 (76.7%)	31 (36.9%)	13 (31.7%)	138 (16.0%)	2.4	2.1

Targets that have drugs tested in multiple phases are only included in the highest phase category.

Table 05- 4 List of phase III targets dropped by combinations of at least three of the methods A, B, C and D used in this study

Research Target	Identified as Promising by Method or Combination	Number of target affiliated pathways	Number of human similarity proteins outside target family	Number of tissues target is primarily distributed	Targeted Disease Conditions	Target Exploration Status (Tested Drug)
AKT	Combination of A, B	25	1	P	Non-Hodgkin's Lymphoma, Multiple Myeloma, Renal Cell Carcinoma	Phase III (enzastaurin), Phase II (perifosine), Phase II (XL880), Phase I completed (RX-0201), Phase I (XL418)
CDK2	Combination of A, B	4	0	P	Lymphocytic leukemia, Lung Cancer (NSCLC), Non-Hodgkin's Lymphoma	Phase III (flavopiridol), Phase II completed (seliciclib), Phase I/II (AT7519), Phase I (SNS-032), Preclinical (capridine-beta)
Alpha-glucosidase	Combination of A, D	2	0	P	Cardiovascular Disorders	Phase III (acarbose), Phase II (celgosivir)
Squalene synthetase	Combination of C, D	2	0	4	Hyperlipidemia	Phase III (TAK-475)
Arachidonate 5-lipoxygenase-activating	Only D	1	0	1	Coronary Artery Disease, Heart Attack, Cardiovascular Disorders	Phase III (DG031), Phase I (AM803, AM103)

protein						
Heme Oxygenase #	Only D	1	0	1	Neonatal Hyperbilirubinemia, Jaundice	Phase III (stansporfin)
Farnesyl protein transferase	Only D	2	0	P	Myeloid Leukemia	Phase III (R115777)
Lipoprotein-associated phospholipase A2	Only D	1	0	P	Atherosclerosis, Cardiovascular Disorders	Phase II/III (darapladib), Phase I (659032)
MMP-12	Only D	1	0	4	Lung Cancer (NSCLC)	Phase III (AE-941)
Myophosphorylase	Only D	2	0	1	Lymphocytic Leukemia, Diabetes Mellitus	Phase III (flavopiridol), Phase IIa (PSN357)
Neutral endopeptidase	Only D	3	0	P	Hypertension, Congestive Heart Failure	Phase II/III (Ilepatril), Phase II (SLV 306)
Sphingosine kinase	Only D	3	0	4	Ovarian Cancer	Phase III (phenoxodiol)
Heat shock protein HSP 90	Only C	1	0	>10	Multiple Myeloma, Metastatic Breast Cancer, Prostate Cancer	Phase III (tanespimycin), Phase II (alvespimycin hydrochloride, IPI-504), Phase I (CNF1010, SNX-5422, STA-9090), IND filed (AT13387)
Cathepsin K	None	No-Info	0	4	Osteoporosis, Bone Metastases	Phase III (odanacatib), Phase II (relacatib), Phase I/II (MIV-701),
MMP-2 / MMP-9	None	3	0	6	Lung Cancer (NSCLC), Osteoarthritis	Phase III (Neovastat), Phase II (PG-530742)

The target marked by # has a positive phase-III result reported in 2004, but since then there has been no report about the further progress of the phase III drug.

Table 05- 5 List of difficult targets currently discontinued in clinical trials and having no new drug entering clinical trials, and the prediction results

Currently Discontinued Target	Predicted as Promising by Method or Combination	Targeted Disease Conditions	Discontinued Drug (Company)	Time of Discontinuation	Reason for Discontinuation
Gastrin/cholecystokinin B receptor	Combination of A, B, C, D	Sleep Disorders	GW150013 (GSK)	December of 2001	Not Clear
Prolactin receptor (PRLR)	Combination of A, B, D	Cancer/Tumors	Endostatin (EntreMed)	February of 2004	Not Clear
B-cell surface antigen CD40	Combination of A, B	Cancer/Tumors	Avrend (Amgen)	January of 2002	In 1998, phase I results showed few changes in circulating leukocyte subsets after a five-day course of treatment. In January 2002, Immunex announced that it was no longer developing Avrend.
C3/C5 convertase	Combination of A, B	Coronary Artery Disease	MLN2222 (Millennium)	August of 2005	Not Clear
Fungal 14-alpha demethylase	Combination of A, C	Fungal Infections, Onychomycosis	Ravuconazole (Eisai)	November of 2005	In November 2005, Eisai stated that ravuconazole had been superceded, hence development was discontinued.
Cytochrome P450 24A1	Only A	Prostate Cancer	RC-8800 (Sapphire)	August of 2006	Not Clear
Alpha-mannosidase 2	Only D	Cancer/Tumors	GD0039 (Inflazyme)	May of 2002	In May 2002, GlycoDesign discontinued the phase II

			Pharmaceuticals)		clinical trials of GD0039 for the treatment of metastatic renal cancer, due to the fact that tumor response and adverse events did not meet clinical expectations.
Acyl coenzyme A:cholesterol acyltransferase 1	Only D	Peripheral Vascular Disease	Avasimibe (Pfizer)	October of 2003	Not Clear
Carnitine O-palmitoyltransferase I	Only D	Congestive Heart Failure	Etomoxir (MediGene AG)	April of 2003	In April 2003, Medigene terminated phase II trials for etomoxir based on data suggesting an increase in side effects in treated subjects.
MMP-7	Only D	Pancreatic and Lung Cancer, Cancer/Tumors	Marimastat (Schering-Plough) BB-3644 (Vernalis)	June of 2003 Before 2006	Results of a phase I trial in cancer subjects showed that it caused musculoskeletal pain like marimastat did. At its maximum tolerated dose of 20 mg twice daily, BB-3644 does not show any advantage over marimastat. Due to these results, further trials were not initiated.
Acyl coenzyme A:cholesterol acyltransferase 2	None	Peripheral Vascular Disease	Avasimibe (Pfizer)	October of 2003	Not Clear
Calcitonin gene-related peptide 2	None	Migraine and Cluster Headaches	Olcegepant (Boehringer Ingelheim)	March of 2007	Not Clear

Cell surface glycoprotein MUC18	None	Melanoma	ABX-MA1 (Amgen)	March of 2005	Not Clear
Hexokinase	None	Prostatic Hyperplasia	Lonidamine (Threshold Pharmaceuticals)	July of 2006	In July 2006, Threshold reported negative results from both a phase II and phase III trial of lonidamine for the treatment of benign prostatic hyperplasia (BPH). Then, Threshold announced its discontinuation.
MMP-8	None	Non-small Cell Lung Cancer (NSCLC)	BMS 275291 (Celltech Group)	November of 2004	In June 2003, Celltech and Bristol Meyers Squibb announced they were discontinuing the development of BMS 275291 in its current indications due to a general lack of efficacy in phase II.
Pyruvate dehydrogenase kinase	None	Diabetes Mellitus	AZD 7545 (AstraZeneca)	November of 2002	Not Clear
Ribonucleoside-diphosphate reductase	None	Various Types of Cancer	Tezacitabine (Chiron)	March of 2004	In March 2004, Chiron announced they were discontinuing development of tezacitabine due to a lack of efficacy in phase II.
Sodium/hydrogen exchanger 1	None	Cardiac Surgery	Cariporide (Sanofi-aventis)	July of 2002	Not Clear
Sodium/hydrogen exchanger 3	None	Ovarian and Lung Cancer	Squalamine (Genaera)	January of 2007	Not Clear

Table 05- 6 List of unpromising targets failed in HTS campaigns or found non-viable in knockout studies, and the prediction results

Target Failed in HTS campaigns or found non-viable in knockout studies	Predicted as Promising by Method or Combination	Exploration Results	Target Failed in HTS campaigns or found non-viable in knockout studies	Predicted as Promising by Method or Combination	Exploration Results
DNA gyrase subunit A	Combination of A, C, D	Not viable	MabA	Combination of A, B, D	Not viable
Acyl carrier protein synthase	Combination of A, D	No hits	L tRNA synthetase	Combination of A, D	No hits
Penicillin-binding protein-2'	Combination of A, D	No hits	Ribonucleotide reductase	Combination of A, D	Not viable
V tRNA synthetase	Combination of A, D	No hits	AccD5	Combination of B, D	Not viable
Alanine racemase	Combination of C, D	Not viable	AroA	Combination of C, D	Not viable
D-Ala-D-Ala ligase	Combination of C, D	Not viable	FabH	Combination of C, D	Not viable
Thymidine monophosphate kinase	Combination of C, D	Not viable	A tRNA synthetase	Only D	No hits
AcpM	Only D	Not viable	AftA	Only D	Not viable
AroB	Only D	Not viable	AroC	Only D	Not viable
AroE	Only D	Not viable	ArgF	Only D	Not viable
AroG	Only D	Not viable	AroK	Only D	Not viable
AroQ	Only D	Not viable	Biotin ligase(BirA)	Only D	Not viable
Branched-chain amino acid aminotransferase	Only D	Not viable	C tRNA synthetase	Only D	No hits

Chorismate synthase	Only D	No hits	CoA(PanK)	Only D	Not viable
D tRNA synthetase	Only D	No hits	DNA polymerase IIIalpha	Only D	No hits
E tRNA synthetase	Only D	No hits	FtsH ATP-dependent protease	Only D	No hits
G tRNA synthetase	Only D	No hits	Galactofuraosyl transferase	Only D	Not viable
GlmU	Only D	No hits	GlnE	Only D	Not viable
H tRNA synthetase	Only D	No hits	IdeR	Only D	Not viable
K tRNA synthetase	Only D	No hits	LigA	Only D	Not viable
LS, riboflavin synthase	Only D	Not viable	MenA	Only D	Not viable
MenB	Only D	Not viable	MenC	Only D	Not viable
MenD	Only D	Not viable	MenE	Only D	Not viable
MenH	Only D	Not viable	MtrA	Only D	Not viable
MurB	Only D	No hits	N tRNA synthetase	Only D	No hits
P tRNA synthetase	Only D	No hits	Peptidyl tRNA hydrolase	Only D	No hits
Phosphopantetheine adenylyl transferase	Only D	No hits	R tRNA synthetase	Only D	No hits
Ribonuclease P	Only D	No hits	S tRNA synthetase	Only D	No hits
Signal peptidases	Only D	No hits	T tRNA synthetase	Only D	No hits

Transketolase	Only D	No hits	UDP-N-acety muramyl:L-alanine ligase(MurC)	Only D	No hits
UMP kinase inhibitor	Only D	No hits	Undecaprenyl(UDP) pyrophosphate synthase	Only D	No hits
Metallo beta-lactamase	Non	No hits	SecA subunit of preprotein translocase	Non	No hits
Serine beta-lactamase	Non	No hits			

Table 05- 7 Definitions and structures (if available) of drugs and compounds in this chapter**Drug Name** **Definition**

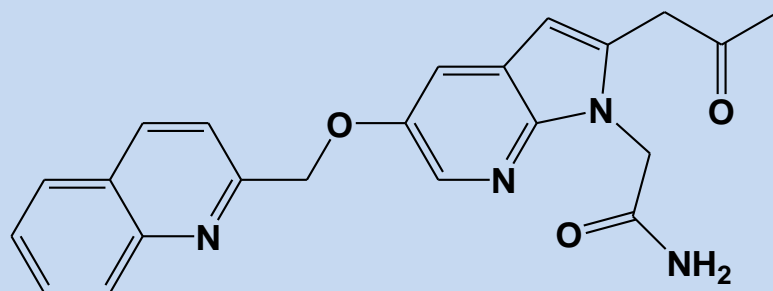
AE-941 AE-941 (Neovastat®; Aeterna Laboratories, Quebec City, Quebec, Canada) is a novel standardized water-soluble extract derived from shark cartilage that represents less than 5% of the crude cartilage. It is a multifunctional antiangiogenic product that contains several biologically active molecules being studied for its ability to prevent the growth of new blood vessels to solid tumors. However, the chemical characterization of the standardized water-soluble extract derived from shark cartilage has never been performed. United States Patent: **5,618,925**.

References: (Dupont, E, Brazeau, P, and Juneau, C (1997) Extracts of shark cartilage having an antiangiogenic activity and an effect on tumor regression; process of making thereof. *United States Patent 5,618,925*.)^{301,302}

AM103 (Amira Pharmaceuticals, Inc., San Diego, California, USA), 2-[2-(2-Oxo-propyl)-5-(quinolin-2-ylmethoxy)-pyrrolo[2,3-b]pyridin-1-yl]-acetamide, also known as 2190914, is a novel indole-based compound. United States Patent for AM103 and related compounds is **7,405,302**.

Structure:

AM103



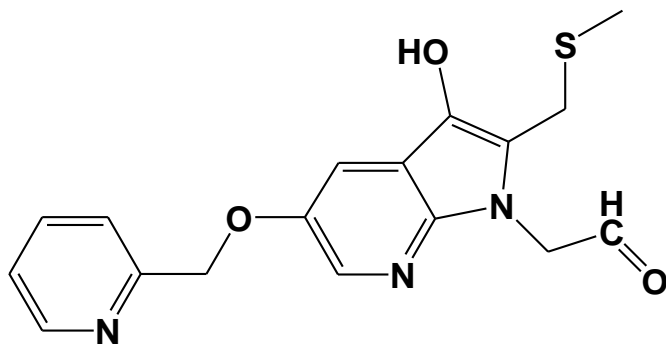
References: (Hutchinson JH, Prasit PP, Moran M, Evans JF, Zunic JE and Stock NS (2008) 5-lipoxygenase-activating protein (FLAP) inhibitors.

United States Patent 7,405,302.)³⁰³ (<http://www.amirapharm.com/pipeline.html>)

AM803 (Amira Pharmaceuticals, Inc., San Diego, California, USA), [3-Hydroxy-2-methylsulfanylmethyl-5-(pyridin-2-ylmethoxy)-pyrrolo[2,3-b]pyridin-1-yl]-acetaldehyde, is a novel inhibitor of 5-lipoxygenase-activating protein (FLAP). Information about AM803 and related compounds can be obtained from United States Patent: **7,405,302**.

Structure:

AM803



References: (Hutchinson JH, Prasit PP, Moran M, Evans JF, Zunic JE and Stock NS (2008) 5-lipoxygenase-activating protein (FLAP) inhibitors.

United States Patent 7,405,302.) (<http://www.amirapharm.com/pipeline.html>)

C1-INH (Lev Pharmaceuticals, New York, USA), also known as C1-inhibitor or C1 esterase inhibitor, has a 2-domain structure including a C-terminal serpin domain and an N-terminal tail domain.

Sequence (structure is available on *Protein Data Bank* with PDB id: 2QAY):

C1-INH

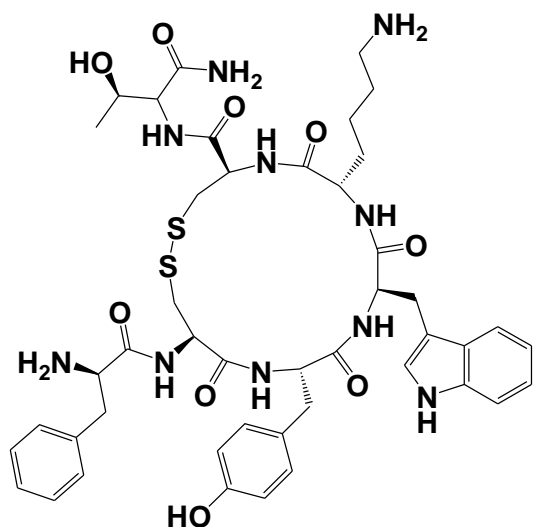
```
MASRLTLLTLLLLLAGDRASSNPATSSSSQDPESLQDRGEGKVATTVISKMLFVEPILEVSSLPTTNSTNSATKITANTTDEPTTQPTTEPTTQPTIQPTQPTTQLPTDSPTQPTTGSFCPGPVTLCSDL
ESHSTEAVLGDALVDFSLKLYHAFSAMKKVETNMAFSPFSIASLLTQVLLGAGENTKTNLESILSYPKDFTCVHQALKGFTTKGVTVSVSQIFHSPDLAIRDTFVNASRTLYSSSPRVLSNNSDANLELIN
TWVAKNTNKKISRLDLSLPSDTRLVLLNAIYLSAKWKTTDFDPKTRMEPFHFKNKNSVIKVPMMNSKKYPVAHFIDQTLKAKVGGQLQLSHNLSLVILVPQNLKHRLEDMEQALSPSVFKAI MEKLEMSK
FQPTLLTLPRIKVTTSSQDMLSIMEKLEFFDFSYDLNLCGLTEDPDLQVSAMQHQTVLELTETGVEAAAAAISVARTLLVFEVQQPFLFVLWDQQHKFPVFMGRVYDPR A
```

References:³⁰⁴⁻³⁰⁶

CAP-232 (Thallion Pharmaceuticals, Alexander-Fleming Montreal, Quebec, Canada), (1R,4S,7R,10S,13R)-4-(4-aminobutyl)-N-[(2S,3R)-1-amino-3-hydroxy-1-oxobutan-2-yl]-13-[[[(2R)-2-amino-3-phenylpropanoyl]amino]-10-[(4-hydroxyphenyl)methyl]-7-(1H-indol-3-ylmethyl)-3,6,9,12-tetraoxo-15,16-dithia-2,5,8,11-tetrazacycloheptadecane-1-carboxamide, originally named as TT-232, is a novel seven amino-acid synthetic cyclic peptide somatostatin analogue.

Structure:

CAP-232

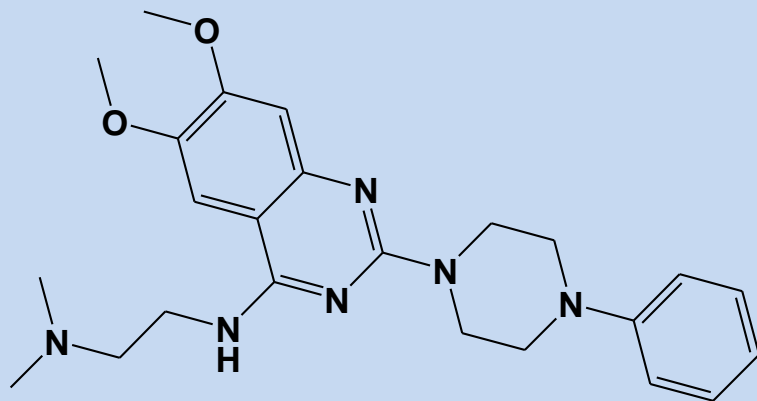


References:³⁰⁷ (<http://www.thallion.com/en/drug-development/tln-232.php>)

CPG 52364

CPG 52364 (Coley Pharmaceutical Group, Wellesley, Massachusetts, USA), N'-[6,7-Dimethoxy-2-(4-phenyl-piperazin-1-yl)-quinazolin-4-yl]-N,N-dimethyl-ethane-1,2-diamine, is a small molecule, first-in-class TLR antagonist designed to specifically inhibit TLRs 7, 8, and 9. Information about CPG 52364 and related compounds can be obtained from United States Patent: **7,410,975**.

Structure:

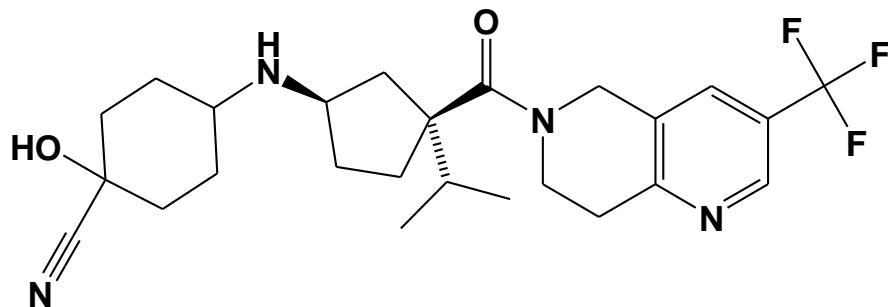


References: (Lipford GB, Forsbach A and Zepp C (2004) Small molecule toll-like receptor (TLR) antagonists. *United States Patent 7,410,975.*)
(<http://www.medicalnewstoday.com/articles/87013.php>)

INCB3284 (Incyte Corporation, Wilmington, Delaware, USA), 1-Hydroxy-4-[3-isopropyl-3-(3-trifluoromethyl-7,8-dihydro-5H-[1,6]naphthyridine-6-carbonyl)-cyclopentylamino]-cyclohexanecarbonitrile, is a small molecule CCR2 antagonist. Information about INCB3284 and related compounds can be extracted from United States Patent: **7,307,086**.

Structure:

INCB3284

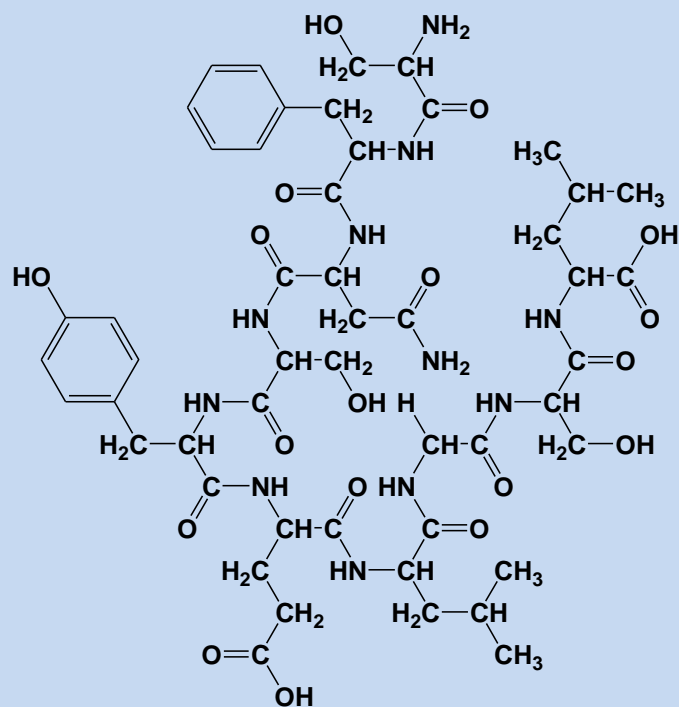


References: (Xue CB, Zheng C, Cao G, Feng H, Xia M, Anand R, Glenn J and Metcalf B (2005) 3-(4-heteroarylcylohexylamino)cyclopentanecarboxamides as modulators of chemokine receptors. *United States Patent 7,307,086.*)^{308,309}

KAI-9803 (KAI Pharmaceuticals Inc., South San Francisco, California, USA), also known as delta-V1-1, is a novel peptide derived from the first variable region of .delta.PKC conjugated via a Cys-Cys disulfide linkage to a HIV Tat-derived transporter peptide. Its sequence is SFNSYELGSL.

Structure:

KAI-9803

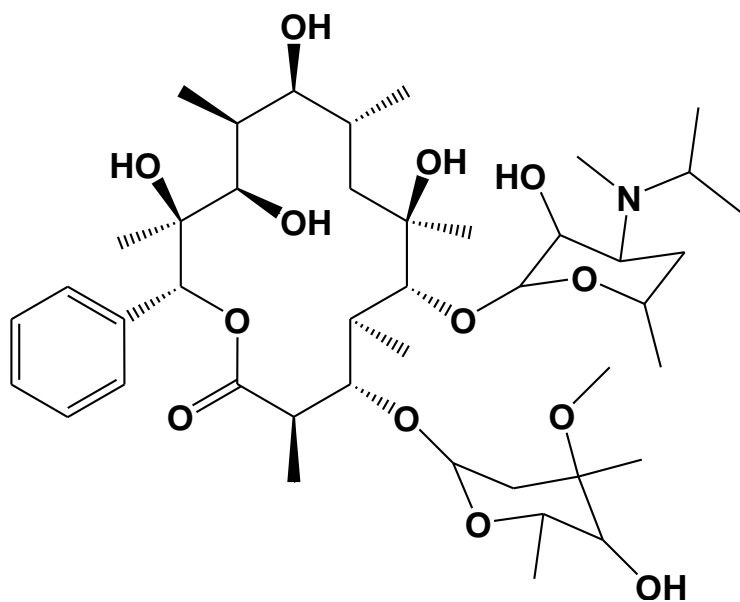


References: ³¹⁰⁻³¹² (http://www.kaipharmaceuticals.com/index.php?option=com_content&task=view&id=18&Itemid=40)

KOS-2187 (Kosan Biosciences, Hayward, California, USA), 7,10,12,13-Tetrahydroxy-6-[3-hydroxy-4-(isopropyl-methyl-amino)-6-methyl-tetrahydro-pyran-2-yloxy]-4-(5-hydroxy-4-methoxy-4,6-dimethyl-tetrahydro-pyran-2-yloxy)-3,5,7,9,11,13-hexamethyl-14-phenyl-oxacyclotetradecan-2-one, is a novel erythromycin-based second-generation motilides, a non-peptide derivative of a natural motilin agonist which demonstrates capacity in addressing the serious limitations of first-generation erythromycin analogs. Information about KOS-2187 and other related compounds can be obtained from United States Patent: **6,946,482**.

Structure:

KOS-2187



References: (Santi DV, Metcalf B, Carreras C, Liu Y, McDaniel R and Rodriguez EJ (2005) Motilide compounds. *United States Patent 6,946,482*.)³¹³

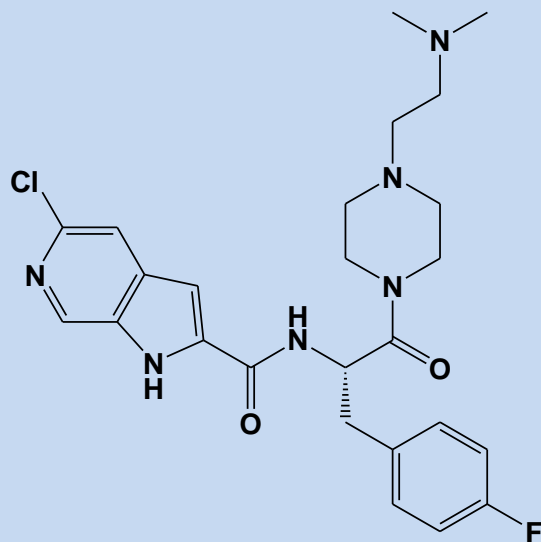
(Sandham DA (2008) Recent developments in gastrointestinal prokinetic agents. *Expert Opinion on Therapeutic Patents 18*: 501-514.)

(<http://www.kosan.com/pipeline.html>)

PSN357 (OSI Pharmaceuticals, Melville, New York, USA), 5-Chloro-1H-pyrrolo[2,3-c]pyridine-2-carboxylic acid [2-[4-(2-dimethylamino-ethyl)-piperazin-1-yl]-1-(4-fluoro-benzyl)-2-oxo-ethyl]-amide, is a glycogen phosphorylase inhibitor (GPI). Information about PSN357 and other related compounds can be obtained from United States Patent: **7,405,210**.

Structure:

PSN357

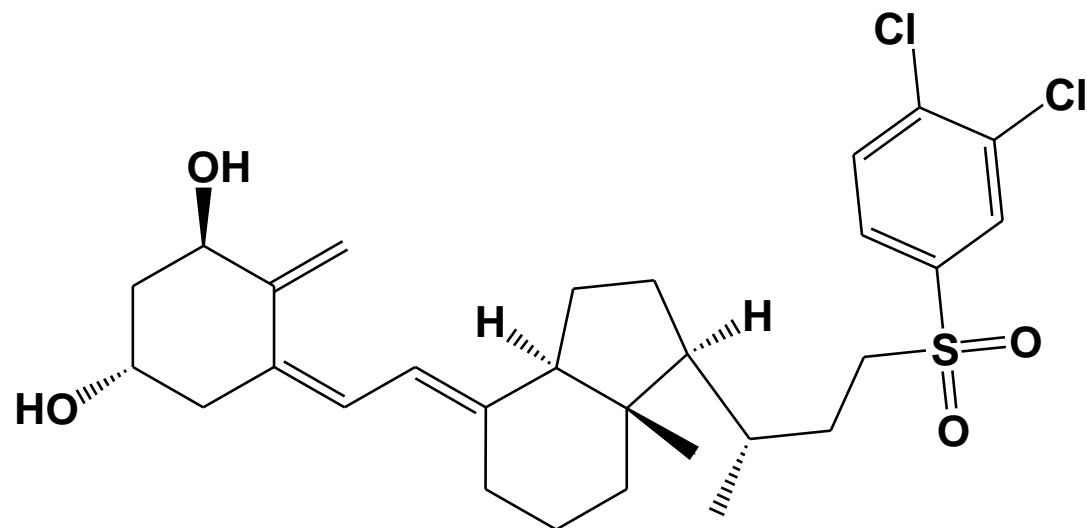


References: (Bradley SE, Krulle TM, Murray PJ, Procter MJ, Rowley RJ, Sambrook Smith CP and Thomas GH (2005) Pyrrolopyridine-2-carboxylic acid amide inhibitors of glycogen phosphorylase. United States patent **7,405,210**.)³¹⁷ (<http://www.osip.com/>)

RC-8800

RC-8800 (Speedel Pharmaceuticals, Bridgewater, New Jersey, USA), 5-(2-{1-[3-(3,4-Dichloro-benzenesulfonyl)-1-methyl-propyl]-7a-methyl-octahydro-inden-4-ylidene}-ethylidene)-4-methylene-cyclohexane-1,3-diol, is a small molecule acting on the vitamin D (calcitriol) metabolic pathway. Information about RC-8800 and other related compounds can be obtained from United States Patent: **7,166,585**.

Structure:



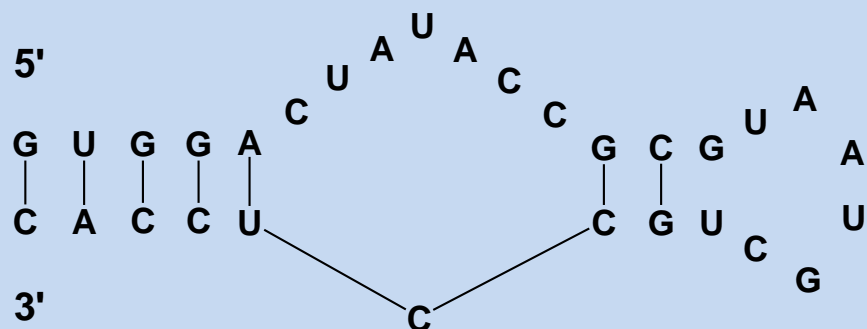
References: (Posner GH, Crawford K, Yang HW, Jeon HB, Hatcher M, Suh BC, White J and Jones G (2005) 24-Sulfur-substituted analogs of 1.alpha.,25-dihydroxy vitamin D.sub.3. *United States patent 7,166,585.*) (<http://www.helsinn.com/>)

REG1

REG1 (Regado Biosciences, Durham, North Carolina, USA) is a two-component system, consisting of an aptamer-based anticoagulant and its matched reversal agent. The REG1 anticoagulant component, RB006, is a single-stranded, nucleic acid aptamer. RB006 selectively and potently binds to and inhibits factor IXa, a protein that is critical to blood coagulation. The reversal agent, RB007, is a complementary nucleic acid that binds to and neutralizes RB006. The amount of RB007 administered allows physicians to fine tune the pharmacodynamic effect of RB006, from slight reduction in anticoagulation all the way to complete reversal. United States Patent for REG1 and related nucleic acids is **7,304,041**.

Structure:

the following is the structure of RB006, while RB007 is the complementary nucleic acid of RB006

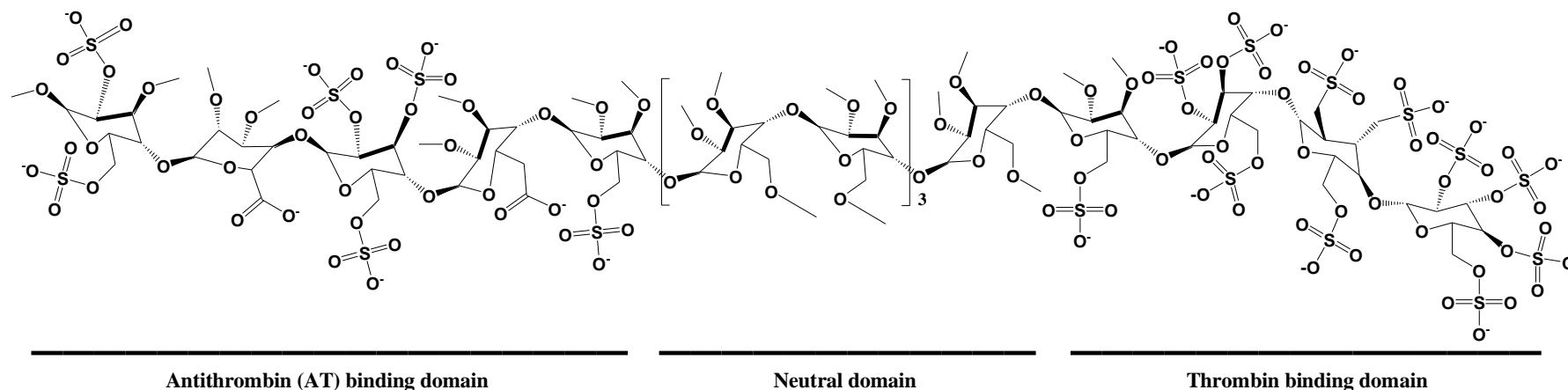


References: (Rusconi C (2005) Modulators of coagulation factors. *United States patent 7,304,041.*³¹⁸⁻³²⁰. (<http://www.regadobiosciences.com>)

SR-123781A (SanofiAventis, Paris, France) is a synthetic hexadecasaccharide comprising an antithrombin (AT) binding domain, a thrombin binding domain, and a neutral methylated hexasaccharide sequence.

Structure:

SR-123781A

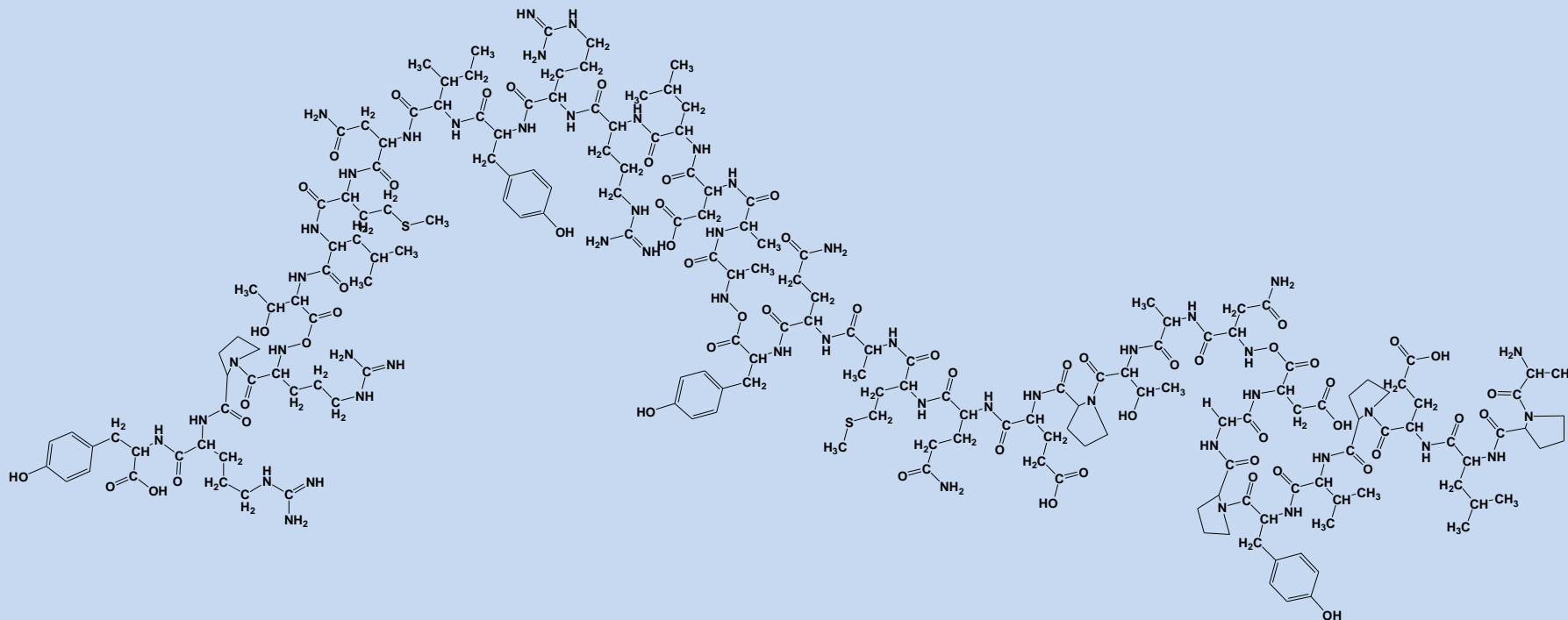


References:³²¹

TM30339 (7TM Pharma, Hørsholm, Hørsholm Municipality, Denmark) is an analogue of the natural hormone, Pancreatic Polypeptide, PP. Its sequence is APLEPVYPGDNATPEQMAQYAADLRRYINMLTRPRY.

Structure:

TM30339

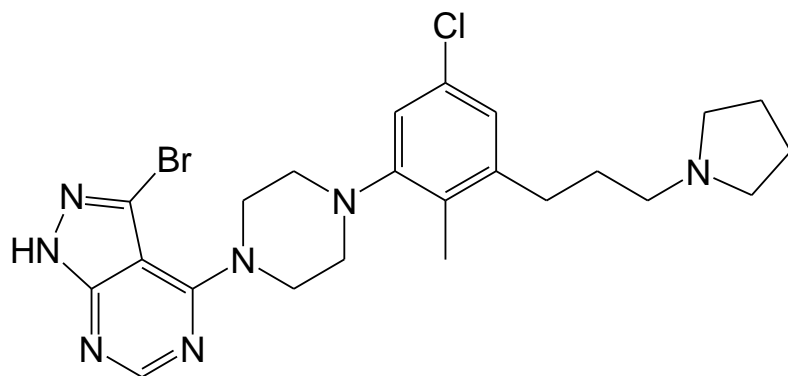


References:³²². (<http://www.7tm.com/>)

XL418

XL418 (Exelixis, San Francisco, California, USA), 3-Bromo-4-{4-[5-chloro-2-methyl-3-(3-pyrrolidin-1-yl-propyl)-phenyl]-piperazin-1-yl}-1H-pyrazolo[3,4-d]pyrimidine, is a selective, orally active small molecule mostly like a pyrazolopyrimidine.

Structure:



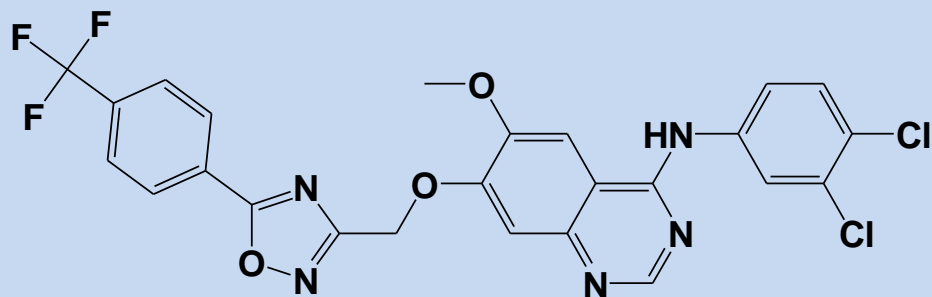
References:³²³ (Li Q (2007) Recent progress in the discovery of AKT inhibitors as anticancer agents. *Expert Opinion on Therapeutic Patents* **17**:1077-1130.) (<http://www.exelixis.com/>)

XL647 (Exelixis, San Francisco, California, USA), (3,4-Dichloro-phenyl)-{6-methoxy-7-[5-(4-trifluoromethyl-phenyl)-[1,2,4]oxadiazol-3-ylmethoxy]-quinazolin-4-yl}-amine, also named as EXEL-647 or EXEL-7647, is small molecule based on a 4-Methyl-quinazoline-6,7-diol backbone structure.

Information about XL647 and related compounds can be obtained from United States Patent: **WO2004006846**.

Structure:

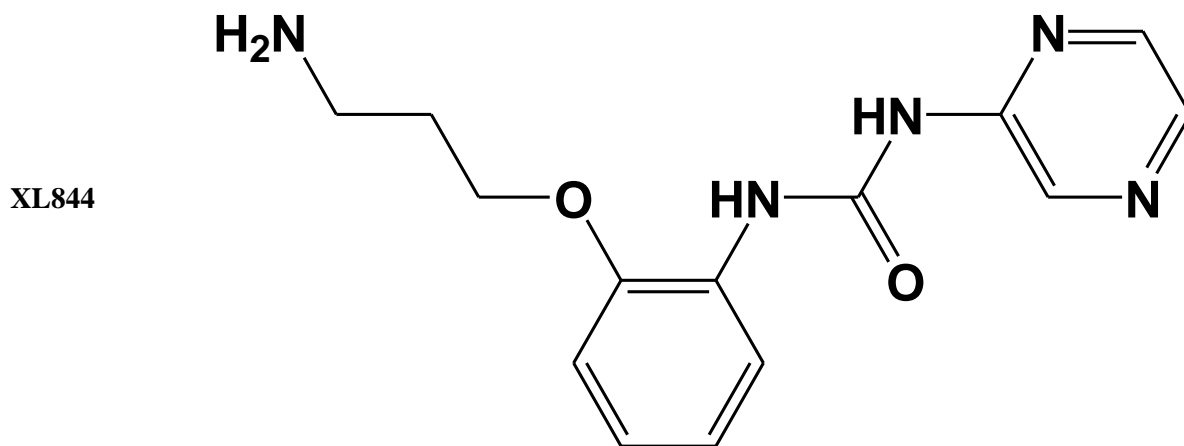
XL647



References: (Rice KD, Anand NK, Bussenius J, Costanzo S, Kennedy AR, Kim AI, Peto CJ, Tsang TH and Blazey CM (2004) Receptor-type kinase modulators and methods of use. United States patent **WO2004006846**.)^{324,325} (Paz K and Zhu Z (2007) Development of Angiogenesis Inhibitors to Vascular Endothelial Growth Factor Receptor 2 for Cancer Therapy. *Top Med Chem* 1:333-382.).

XL844 (Exelixis, San Francisco, California, USA), 1-[2-(3-Amino-propoxy)-phenyl]-3-pyrazin-2-yl-urea, also named as EXEL9844, is an aminopyrazine carboxamide. Information about XL844 and related compounds can be obtained from United States Patent: **7,202,244**.

Structure:

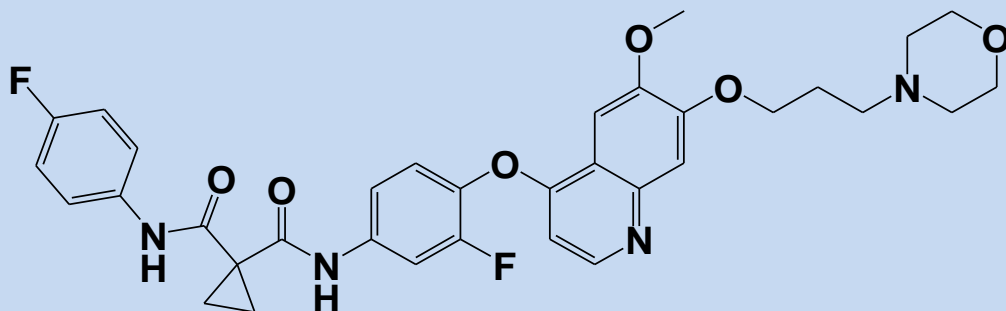


References: (Boyle RG, Imogai HJ and Cherry M (2008) Chk-1 inhibitors. *United States Patent 7,202,244*.)³²⁶⁻³²⁸

(http://www.exelixis.com/eortc/posters/EORTC08_395_XL844-002.pdf)

XL880 (Exelixis, San Francisco, California, USA), Cyclopropane-1,1-dicarboxylic acid {3-fluoro-4-[6-methoxy-7-(3-morpholin-4-yl-propoxy)-quinolin-4-yloxy]-phenyl}-amide (4-fluoro-phenyl)-amide, is also known as EXEL-2880, GSK1363089 or GSK089.

Structure:

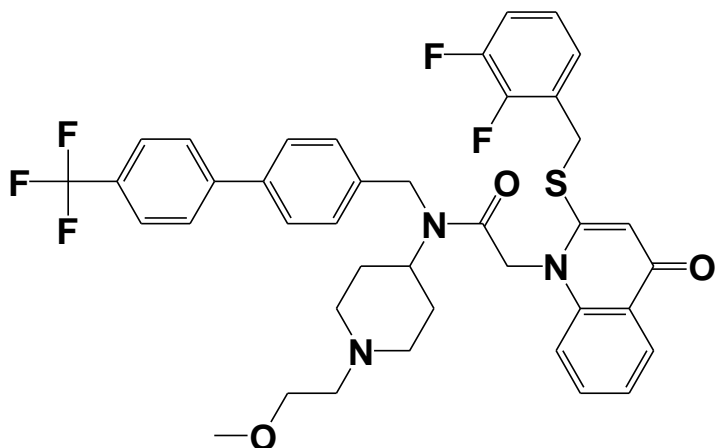


References:³²⁹ (http://www.exelixis.com/pipeline_xl880.shtml)

659032 (Human Genome Sciences, Rockville, Maryland, USA), 2-[[[(2,3-Difluorophenyl)methyl]thio]-N-[1-(2-methoxyethyl)-4-piperidinyl]-4-oxo-N-[[4'-(trifluoromethyl)[1,1'-biphenyl]-4-yl]methyl]-1(4H)-quinolineacetamide, is also named SB-659032 or Rilapladib.

Structure:

659032



References: (http://www.gsk.com/investors/product_pipeline/docs/gsk-pipeline-feb09.pdf)

AT13387

AT13387 (Astex Therapeutics Limited, Cambridge, England) is a selective small molecule of a fragment-based clinical candidate active in lung cancer and melanoma models. It is the third product derived from Astex's Pyramid™ fragment chemistry platform. Information about AT13387 and related compounds can be obtained from United States Patent: **7,385,059**.

References: (Berdini V, O'Brien MA, Carr MG, Early TR, Gill AL, Trewartha G, Woolford AJA, Woodhead AJ and Wyatt PG (2006) 3,4-disubstituted 1H-pyrazole compounds and their use as cyclin dependent kinase and glycogen synthase kinase-3 modulators. *United States patent 7,385,059*.)³³⁰.

(<http://www.astex-therapeutics.com/products/pipeline.php#AT13387>)

CCX915

CCX915 (ChemoCentryx, Mountain View, California, USA) is an orally available small molecule implicated in damaging inflammation underlying multiple sclerosis. It belongs to new class of synthetic compounds that are chemically distinct from all known inhibitors of CCR2. Information about MIV-701 and related compounds can be obtained from United States Patent: **7,435,831**.

References: (Chen W, Zhang P, Aggen JB, Dairaghi DJ, Pennell AMK, Sen S and Wright JJK (2005) Bicyclic and bridged nitrogen heterocycles.

United States patent 7,435,831.) (<http://www.chemocentryx.com/>)

(http://www.redorbit.com/news/health/311395/chemocentryx_to_initiate_clinical_studies_of_ccr2_antagonist_ccx915/index.html)

MIV-701

MIV-701 (Medivir, Huddinge, Sweden) is a highly active and selective small molecule inhibitor of cathepsin K currently under Phase I/II clinical trial, but Medivir has not disclosed any details of MIV-701 yet.

References: (Mucke HAM, Norman P, Whelan C, Yeates C (2008) Patent alert. *Current Opinion in Investigational Drugs* **9**:552-561).

(http://www.medivir.se/v3/en/RnD/projects/miv_701.cfm)

MSX-122

MSX-122 (Metastatix Inc., Atlanta, Georgia, USA) is an orally bioavailable inhibitor of CXCR4 with potential antineoplastic and antiviral activities. Currently it is under Phase I clinical trial.

References: (Zhang Y, Liang Z and Wu H (2008) MSX-122, an orally available small molecule targeting CXCR4, inhibits primary tumor growth in an orthotopic mouse model of lung cancer and improves the effect of paclitaxel. *AACR Annual Meeting 2009 Proceedings* **49**)³³¹.

(<http://clinicaltrials.gov/ct2/show/NCT00591682>)

STA-9090

STA-9090 (Synta Pharmaceuticals, Lexington, Massachusetts, USA) is a new chemical entity which contains a triazolone core structure synthesized through a 3-step cGMP process. It is a novel triazolone compound with a unique chemical structure that is distinct from 17-AAG (geldanamycin) and other ancamycin derivatives. STA-9090 is developed for treating a variety of cancers.

References:³³² (<http://www.syntapharma.com/PrdHsp90.aspx>)

TTP889

TTP889 (TransTech Pharma Inc., High Point, North Carolina, USA) is an orally available, small molecule that inhibits up to 90% of FIXa activity at therapeutic doses, using a clinical model of extended prophylaxis in hip fracture surgery. Information about TTP889 and related compounds can be obtained from United States Patent: **7,122,580**.

References: (Mjalli AMM, Andrews RC, Guo XC, Christen DP, Gohimmukkula DR, Huang G, Rothlein R, Tyagi S, Yamasu T and Behme C (2003) Aryl and heteroaryl compounds and methods to modulate coagulation. *United States patent 7,122,580*.) (Rothlein R, Shen JM, Naser N, Gohimukkula DR, Caligan TB, Andrews RC, Schmidt AM, Rose EA and Mjalli AMM (2005) TTP889, a novel orally active partial inhibitor of FIXa inhibits clotting in two a/v shunt models without prolonging bleeding times. *Blood* **106**:A1886.)³³³⁻³³⁵ (http://www.ttpharma.com/pipeline_thrombosis.html)

Chapter 6 Identification of promising therapeutic targets from influenza genomes

Influenza viruses are among the most common causes of respiratory infections in humans and are associated with high morbidity & mortality^{336,337}. Seasonal influenza affects 10% of the population annually leading to 44 deaths per 10M population³³⁸. Antigenically novel influenza strains occasionally emerge into pandemics³³⁹ that may potentially infect large populations at high death rates³⁴⁰. The widely spreading swine influenza A H1N1 may potentially develop into such a pandemics³⁴¹. The highly pathogenic avian influenza A H5N1 has caused widespread death in poultry, substantial economic loss to farmers, and reported infections of hundreds of people with a mortality rate of 60%³⁴².

Approved anti-influenza drugs^{343,344} are key lines of defense against novel pandemics as well as being used as general treatment options because the well-matched protective vaccines would not be available for at least several months³³⁷. The usefulness of these drugs may be severely reduced by the emergence of drug-resistance strains. So far, 98% of influenza A H3N2 and some percentage of influenza A H1N1 strains are adamantane-resistant³⁴⁵, and 95% and 98% of influenza A H1N1 strains circulating the World³⁴⁶ and North America³⁴⁷ are tamiflu-resistant. Tamiflu-resistant strains have also been found in influenza A H5N1 and B strains³⁴⁸. The use of alternative anti-influenza drugs may be limited by additional problems such as short supply and side effects³³⁷. Moreover, existing drugs may not be equally effective against all influenza types (e.g. tamiflu is significantly less effective against influenza B)³⁴⁹. Therefore, in addition to the tests of

several new anti-influenza agents in clinical trials^{343,344,350}, there is an urgent need for discovering new drugs for more complete coverage of potentially harmful drug-resistant strains and different influenza types and subtypes at a time of serious concern about the possible emergence of influenza pandemics³³⁷.

Intensive efforts have been directed at the development of anti-influenza agents against several successful and research targets in influenza genomes^{343,344,350,351}. Successful targets are targeted by at least one marketed drug. Research targets are targeted only by agents not yet approved. Drug development is costly, time consuming (8~20 years)²²⁷ and low in productivity particularly against novel targets²⁷². Limited resources may better be focused on anti-influenza agents against promising targets (targets that likely lead to successful drugs) so as to achieve highest possible development speed and success rates. Identification of promising targets in the influenza genomes is thus an important first step for facilitating more efficient and faster discovery of new anti-influenza agents.

Recent studies have shown that promising targets tend to show similar genetic, structural, physicochemical and system profiles as successful targets²²⁷. Comparative investigation of these profiles with respect to successful targets can thus be used for identifying promising targets, which has been validated by the identification of the targets of positive phase III results²⁴⁰. In this work, this approach was used to identify promising targets from the complete genomes of influenza A H1N1 (Swine, Mexico/InDRE4487/2009), H5N1 (Avian, Guangdong/1/96), H2N2 (Korea/426/1968), H3N2 (New York/392/2004), and H9N2 (Hong Kong/1073/99), influenza B (Lee/40), and influenza C (Ann Arbor/1/50) in the NCBI Entrez Genome database and NCBI Influenza Virus Resource³⁵².

6.1 Summary on methods applied for target identification

Each encoded protein in an influenza genome was assessed by four methods for probing its sequence, structural, physicochemical and systems profiles with respect those of successful targets. Method A measures drug-binding domain sequence similarity^{226,227} against those of 168 successful targets with identifiable drug-binding domains. Method B studies drug-binding domain structural and energetic features²⁷⁵ against those of 129 successful targets with available structures. Method C predicts druggable proteins from sequence-derived physicochemical characteristics by a machine learning model trained by 348 successful targets^{144,227,241}. Method D evaluates whether the systems-level druggability rules²²⁷ are satisfied. A protein is identified as promising if it is selected by at least 3 methods²⁴⁰, marginally promising if it is selected by 2 methods, and unpromising if it is selected by less than 2 methods. The identified promising, marginally promising, and unpromising targets were evaluated by drug discovery productivity levels against them in terms of the numbers of FDA approved³⁴³, clinical trial^{343,350} and literature-described investigative^{344,353,354} anti-influenza agents and the numbers of US patents that (<http://patft.uspto.gov/>) against each target, which was intended to determine if there is a clear trend of higher productivity levels for the identified promising targets with respect to those of the identified marginally promising and un-promising targets and to whether the identified unpromising targets show low productivity levels. Moreover, the reported structural studies were also analyzed to determine if these studies show druggable features for the identified promising and marginally promising targets and additional undesired properties for the marginally promising targets.

6.2 Target identification results from influenza genomes

Table 06-1 shows the target identification results for all encoded proteins in the genomes of the 5 subtypes of influenza A (8 proteins in the swine N1H1 and 9 proteins in the other 4 subtypes), influenza B (9 proteins) and influenza C (7 proteins). Also included in Table 1 are the lists and references of FDA approved³⁴³, clinical trial^{343,350}, and literature-described investigative^{344,353,354} agents targeting each protein and the number of US patents (<http://patft.uspto.gov/>) that target each protein for drug development. Three proteins were identified as promising targets, which are neuraminidase of influenza A and B, polymerase of influenza A, B and C, and matrix protein 2 of influenza A. Moreover, two proteins were identified as marginally promising targets by method B and D, which are haemagglutinin of influenza A and B, and hemagglutinin-esterase of influenza C.

These results are highly consistent with the current drug discovery productivity levels. For the three identified promising targets, neuraminidase and polymerase are the targets of 2 and 0 FDA approved³⁴³, 2 and 1 clinical trial^{343,350}, and 12^{344,355} and 7^{344,356} literature-described investigative drugs, and in 36 and 30 US patents for the treatment of both influenza A and B infections respectively. Matrix protein 2 is the target of 2 FDA approved³⁴³ and 10 literature-described investigative drugs^{344,357}, and in 9 US patents for the treatment of influenza A infection. In contrast, there is no approved or clinical trial drug for the two identified marginally promising targets. One target, hemagglutinin, is targeted by 5 literature-described investigative agents^{353,358} and in 3 US patents. We found no literature-described investigative agent or US patent for the second target,

hemagglutinin-esterase, probably because this influenza C protein has not been sufficiently focused in drug development as flu induced by influenza C is rare compared to those of influenza A and B³⁵⁹. For the six identified non-promising targets, we found no literature-described agent or US patent for all but two of them. One protein, nonstructural protein 1, is a target of one literature-described investigative agent³⁵⁴. Another protein, nucleocapsid protein, is the target of drug development in two US patents (US Patent 6,242,478 and 6,316,190).

It is noted that a hemagglutinin inhibitor, arbidol, has been clinically used in Russia³⁶⁰ and tested in humans in China³⁶¹. This agent has been reported to be effective in inhibiting the replication of all influenza A and B in *in vitro*, *in vivo* and clinical trials in Russia^{360,361}. But drug application to FDA has not been approved partly because an independent test appears to show that arbidol is effective only at a concentration approaching 50% cytotoxic dosage with an estimated selectivity index (SI) of 4, which is substantially lower than the minimum safety SI value of 10 (<http://arbidol.org/sidwell/>). We found no report about clinical trial of arbidol in US and Europe. Therefore, hemagglutinin is conservatively regarded as a research target with no drug approved by FDA or in clinical trial.

6.3 Discussion on target identification results

For a protein to be identified as a promising target, apart from its disease roles and systems profiles, its drug-binding site must have certain structural and physicochemical features that accommodate the selective binding of drug-like molecules and the subsequent modulation of the activities of the protein²²⁷, and some of these druggable features may be similarly and partially reflected from the sequence, structural, and physicochemical profiles of other validated targets²⁴⁰. For the identified promising targets, some of the druggable features have already been revealed from the reported structural investigations of these proteins. For the identified marginally promising targets, some druggable and undesirable features at their drug-binding sites have also been reported.

Neuraminidase of influenza A and B was identified as a promising target by all four methods. This protein plays critical roles in viral life cycles by facilitating virion progeny release and general mobility of the virus in the respiratory³⁶². Its active site consists of a number of distinct adjoining pockets that are lined by eight highly conserved amino-acid residues making direct contact with different inhibitors, and the architecture of the active site is further stabilized by ten amino-acid residues invariant in all influenza strains within the vicinity³⁶³. Neuraminidases in different influenza types or subtypes may show structurally distinct features near the active sites. For instance, in some subtypes of influenza A, there is an additional cavity adjacent to the active sites that closes upon ligand binding, and this additional cavity may be explored for developing new anti-influenza drugs³⁶⁴. Structural elucidation of drug-resistant mutants has found an altered hydrophobic pocket in the active site that allows effective inhibition of alternative

drugs³⁶⁵. Molecular dynamics simulation has suggested flexible dynamic features at multiple sites that allow the access and binding of drugs³⁶⁶. These features make neuraminidase a highly attractive target for drug development³⁶².

Polymerase of all influenza types was identified as a promising target by three methods (A, B, and D). This protein is responsible for the replication and transcription of the eight separate segments of the viral RNA genome in the nuclei of infected cells, and all of its three subunits PB1, PB2 and PA are involved in protein activities³⁶⁷. PB1 carries a polymerase active site³⁶⁷, PB2 includes a capped-RNA recognition domain³⁶⁷, and PA contains an endonuclease active site³⁶⁸ and is involved in the assembly of the functional complex³⁶⁷. Several residues at the ligand-sensing or catalytic cores of PB1³⁶⁷, PB2^{369,370} and PA^{368,371} are highly conserved in all influenza viruses. Polymerases of different influenza types are high in sequence identify (50%~63%) and similarity (E-value < 10⁻¹⁰⁰). Therefore, the 3D structures of the influenza B and C proteins, which are unavailable, can be generated from those of influenza A protein by using homology modeling. The C-terminal RNA-binding domain of PB2 has a unique phi-shaped structure with a highly basic groove along the loop³⁶⁹. The cap-binding domain of PB2 has a previously unknown fold and the ligand binding mode is similar to but distinct from other cap binding proteins³⁷⁰. The endonuclease domain of PA has a structural core closely resembling resolvases and type II restriction endonucleases^{368,371}. The carboxy-terminal domain of PA forms a novel fold with a deep and highly hydrophobic groove into which the amino-terminal residues of PB1 can fit³⁶⁷. Based on these observed structural features, the ligand-sensing and catalytic sites have been suggested as possible sites for drug development³⁶⁷⁻³⁷⁰.

Matrix protein 2 of all influenza types (named M2, BM2, and CM2 channel for influenza A, B, and C respectively) was identified as a promising target for influenza A H1N1, H5N1, and H2N2 by four methods, for influenza A H3N2 and H9N2 by three methods (A, C and D), but as a non-promising target for influenza B and C. This protein is a pH-activated proton channel mediating acidification of the interior of viral particles entrapped in endosomes³⁷². Structural study of M2 has shown druggable features. In the closed state, four tightly packed transmembrane helices define a narrow channel, in which a 'tryptophan gate' is locked by intermolecular interactions with aspartic acid. A carboxy-terminal, an amphipathic helix oriented nearly perpendicularly to the transmembrane helix, forms an inward-facing base³⁷³. Lowering the pH destabilizes the transmembrane helical packing and unlocks the gate. There are four equivalent drug-binding sites near the gate on the lipid-facing side of the channel and drug binding stabilizes the closed conformation of the pore³⁷³. Drug-resistance mutations may counter the effect of drug binding by either increasing the hydrophilicity of the pore or weakening helix-helix packing, thus facilitating channel opening^{372,373}. BM2 contains a substantially higher number of polar residues in the pore than M2, which is a probable reason for the insensitivity of BM2 to the known M2 inhibitors such as amantadine and which has led to a suggestion to seek more polar compounds as effective inhibitors³⁷⁴. The highly polar nature of the pore of BM2 was likely recognized as non-druggable by the sequence, structure, and physicochemical methods, leading to the identification of BM2 as a non-promising target.

Hemagglutinin of influenza A and B was identified as a marginally promising target by method B and D. It is a surface glycoprotein responsible for receptor binding and the

fusion of virus and cell membranes³⁵³. Some inhibitors bind in a hydrophobic pocket formed at an interface between hemagglutinin monomers to stabilize the neutral pH structure through inter-subunit and intra-subunit interactions that presumably inhibit the conformational rearrangements required for membrane fusion³⁵³. The drug-binding site has good structural features but appears to be highly hydrophobic. Thus, it has been suggested that the development of effective inhibitors with sufficient potency requires an improvement of hydrophobic interactions and the creation of additional polar interactions towards the membrane distal region of the site³⁵³. The highly hydrophobic nature at the binding-site was likely recognized as non-druggable by the sequence and physicochemical methods, leading to the identification of this protein as a non-promising target.

Hemagglutinin-esterase of influenza C was also identified as a marginally promising target by method B and D. This influenza C protein plays the dual roles of both hemagglutinin and neuraminidase³⁷⁵. Alignment of the amino-acid sequences of hemagglutinin-esterase and hemagglutinin based on their 3D structures has shown that, in spite of the low 12% sequence identity, both the overall structure and the detailed folds of individual segments of the two proteins are quite similar³⁷⁵. Specifically, the receptor-binding domain is structurally similar to the sialic acid-binding domain of haemagglutinin. The esterase domain is structurally similar to the esterase from *Streptomyces scabies* and a brain acetylhydrolase. The receptor domain is inserted into a surface loop of the esterase domain and the esterase domain is inserted into a surface loop of the stem. The stem domain is similar to that of haemagglutinin, except that the triple-stranded alpha-helical bundle diverges at both of its ends, and its amino terminus is

partially exposed. Moreover, ligand binding has been found to be a dynamic process that involves conformational rearrangement at the esterase active site³⁷⁶. The high structural similarity to haemagglutinin and the involvement of conformational rearrangement for ligand-binding may be possible reasons for the identification of this protein as a marginally promising target.

In conclusion, recent emergence of swine and avian influenza A H1N1 and H5N1 outbreaks and various drug-resistant influenza strains underscore the urgent need for developing new anti-influenza drugs. Drug development is costly and time-consuming. Limited resources may better be focused on the development of drugs against promising targets. Recent studies have shown that promising targets show genetic, structural, physicochemical and systems profiles resembling those of successful targets, which can be explored for identifying promising targets. We used this approach to identify promising targets from the genomes of influenza A (H1N1, H5N1, H2N2, H3N2, H9N2), B and C. The identified promising targets are neuraminidase of influenza A and B, polymerase of influenza A, B and C, and matrix protein 2 of influenza A. The identified marginally promising targets are haemagglutinin of influenza A and B, and hemagglutinin-esterase of influenza C. These are consistent with reported druggable structural features for the promising and marginally promising targets and undesired properties for the unpromising targets. The promising targets show fair drug discovery productivity levels (1~4 FDA approved or clinical trial drug, 7~13 literature-described agents, and 9~36 patents for each target) in comparison to the modest levels for the marginally promising targets (no FDA approved or clinical trial drug, 5 literature-described agents and 3 patents for one target) and low levels for the unpromising proteins

(only 1 investigative agent for one protein and 2 patents for another protein). These results are highly consistent with the current drug discovery productivity levels against these proteins. Literature reported drug-binding site structural studies of the identified targets also exhibit druggable features or undesirable properties consistent with the identification of these proteins as promising or marginally promising targets. These suggest that the integrated target analysis method is useful for facilitating the identification of promising targets from influenza genomes. This method may be used for the identification of promising targets from the genomes of other viruses or viral strains for facilitating target discovery and subsequent drug development for the treatment of viral infections that are also in urgent need for new drugs.

Table 06- 1 Target identification results for all encoded proteins in the genomes of the 5 subtypes of influenza A, B and C*

Protein	Identification Status for Different Influenza (Method that Selects the Protein as Promising Target)							Drugs Approved [applicable viral type] References	Drugs in Clinical Trial [applicable viral type] References	Literature-described Investigative Agents [applicable viral type] References	Number of US Patents [applicable viral type]
	A H1N1 Swine Mexico	A H5N1 Avian	A H2N2 Korea	A H3N2 New York	A H9N2 Hong Kong	B Lee	C Ann Arbor				
Neuraminidase (NA)	Promising (A,B,C,D)	Promising (A,B,C,D)	Promising (A,B,C,D)	Promising (A,B,C,D)	Promising (A,B,C,D)	Promising (A,B,C,D)	Not encoded in genome	Oseltamivir (Tamiflu ®) Zanamivir (Relenza ®) [A, B] ³⁴³	CS-8958 (Phase III) Peramivir (prepare Phase III) [A, B] ^{343,350}	A-315675; BCX-1812; BCX-1827; BCX-1898; BCX-1923; DANA; FANA; Cyclopentane amide derivatives 1-4; A-192558; GS4071. [A, B] ^{344,355,377}	36 [A, B]
Polymerase	Promising (A,B,D)	Promising (A,B,D)	Promising (A,B,D)	Promising (A,B,D)	Promising (A,B,D)	Promising (A,B,D)	Promising (A,B,D)	None	T-705 (Phase II) [A, B, C] ^{343,378}	ANX-201; T-1105; Flutimide; FdG; T-1106; Pyrimidinyl acylthiourea; 2,4-dioxo-4-phenylbutanoic acid; Thiadiazolo[2,3-a]pyrimidine. [A, B, C] ^{344,356,379,380}	30 [A, B, C]
Matrix protein 2 (M2)	Promising (A,B,C,D)	Promising (A,B,C,D)	Promising (A,B,C,D)	Promising (A,C,D)	Promising (A,C,D)	Non-Promising (D)	Non-Promising (D)	Amantadine (Symmetrel ®) Rimantadine (Flumadine ®) [A] ³⁴³	None	2-(1-adamantyl)-2-methyl-pyrrolidine; Rimantadine isomer 1 & 2; 2-(1-adamantyl) piperidine; 2-(1-adamantyl) pyrrolidine; 3-(2-adamantyl) pyrrolidine; 2-(2-adamantyl) piperidine; Spiro[piperidine-2,2'-adamantane]; Spiro[cyclopropane-1,2'-adamantan]-2-amine; Spiro[pyrrolidine-2,2'-adamantane]. [A] ^{344,357,381}	9 [A]

Hemagglutinin (HA)	Marginally Promising (B,D)	Marginally Promising (B,D)	Marginally Promising (B,D)	Marginally Promising (B,D)	Marginally Promising (B,D)	Marginally Promising (B,D)	Not encoded in genome	None	None	Arbidol (Clinically used in Russia, tested in China but not in clinical trial in US & western Europe up to 2009), TBHQ; BMY-27709; Stachyflin; CL 385319. [A, B] ^{353,358,382-384}	3 [A, B]
Hemagglutinin-esterase (HE)	Not encoded in genome	Not encoded in genome	Not encoded in genome	Not encoded in genome	Not encoded in genome	Not encoded in genome	Marginally Promising (B,D)	None	None	None	0
Nonstructural protein 1 (NS1)	Non-Promising (D)	Non-Promising (D)	Non-Promising (D)	Non-Promising (D)	Non-Promising (D)	Non-Promising (D)	Non-Promising (D)	None	None	CPSF30 [A] ^{354,385}	0
Matrix protein 1 (M1)	Non-Promising (D)	Non-Promising (D)	Non-Promising (D)	Non-Promising (D)	Non-Promising (D)	Non-Promising (D)	Non-Promising (D)	None	None	None	0
NB glycoprotein protein	Not encoded in genome	Not encoded in genome	Not encoded in genome	Not encoded in genome	Not encoded in genome	Non-Promising (D)	Not encoded in genome	None	None	None	0
Nucleocapsid protein	Non-Promising (D)	Non-Promising (D)	Non-Promising (D)	Non-Promising (D)	Non-Promising (D)	Non-Promising (D)	Non-Promising (D)	None	None	None	0
Nonstructural protein 2 (NS2)	Non-Promising (D)	Non-Promising (D)	Non-Promising (D)	Non-Promising (D)	Non-Promising (D)	Non-Promising (D)	Non-Promising (D)	None	None	None	0

PB1-F2 protein	Not encoded in genome	Non-Promising (D)	Non-Promising (D)	Non-Promising (D)	Non-Promising (D)	Not encoded in genome	Not encoded in genome	None	None	None	0
----------------	-----------------------	-------------------	-------------------	-------------------	-------------------	-----------------------	-----------------------	------	------	------	---

* Influenza A viruses used here include: H1N1 (Swine, Mexico/InDRE4487/2009), H5N1 (Avian, Guangdong/1/96), H2N2 (Korea/426/1968), H3N2 (New York/392/2004), H9N2 (Hong Kong/1073/99), influenza B virus is influenza B (Lee/40) and influenza C virus is influenza C (Ann Arbor/1/50)

Chapter 7 Concluding remarks

This last chapter summarizes the major findings and contributions of this study (**Section 7.1**). Limitation of present study and suggestion on possible areas for further studies are discussed in **Section 7.2**.

7.1 Major findings and contributions

Although the study described in this thesis has a main focus on the identification of drug targets, it is composed of three major findings. In **Section 7.1.1**, the merits of TTD update in facilitating target discovery have been illustrated, then, followed by a discussion on the performance of the collective method on target identification in **Section 7.1.2**. In **Section 7.1.3**, the finding on accelerating the development of influenza therapeutics is proposed. The contributions for each finding are also discussed in the following sections.

7.1.1 Merits of TTD in facilitating target discovery

TTD was a pioneer for providing pharmaceutical information on therapeutic target. After progress in the past 8 years on target discovery, TTD still acts as reliable knowledge base for established therapeutic target. 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 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. 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.

Since 2002, TTD has been the primary resource providing comprehensive information on drug targets. Its popularity has been reflected by its visit count (near 130,000 times) and its citation records (79 times) in Aug 9th 2010. In particular, after the update of TTD 2010 in Jan 2010, we have accumulated near 24,000 visit counts and 4 times of citation, which further prove its important role in pharmaceutical research.

7.1.2 Merits of collective decision made by four *in silico* systems in target identification from clinical trial targets

In this work, four systems were constructed to facilitate the identification of drug targets, which include: (1) statistical classification system established by machine learning; (2) homology identification system built by BLAST; (3) drug-binding domain 3-D structural

comparison set up by structure fold analysis; and (4) the simple system-level druggability rules summarized from established therapeutic target. Statistical analyses have proven the reliability and robustness of this collective method.

The collective predictive performance of the four systems was tested against clinical trial and non-clinical trial targets. The best overall performance was produced by combining at least three systems. This combination identified 50% of the phase III, 25% of the phase II, 10% of the phase I, and 4% of the non-clinical trial targets as promising. Comparing to reported drug successful rate, our results of the identified promising clinical trial targets are lower than but roughly follow a similar descending trend as the report. The collective methods enriched identification rate of phase II & III targets by 4.0~6.0 fold over random selection, with the combination of all four methods producing the highest enrichment. On the other hand, the 15 identified promising phase III targets include 7 of the 8 targets with positive phase III results reported. The only one exception has a positive phase III result reported in 2004 and no further news has been released since then. Moreover, a possible adverse effects caused by its inhibition has been reported. On the other hand, the difficulty in exploring some of the un-promising targets has been reported like MMP and kinases of non-tyrosine kinase classes.

Collective use of multiple *in silico* methods is capable of identifying high percentages of phase III targets including most of the targets of positive III results, and of eliminating difficult and un-promising ones. Our study suggests that comparative analysis of multiple profiles of successful targets provides useful clues to the identification of promising drug targets.

7.1.3 Merits of collective decision made by four *in silico* systems in target identification from influenza genome

The collective prediction method of the four systems was used to identify promising drug targets from genomes of influenza A H1N1 (Swine, Mexico/InDRE4487/2009), H5N1 (Avian, Guangdong/1/96), H2N2 (Korea/426/1968), H3N2 (New York/392/2004), and H9N2 (Hong Kong/1073/99), influenza B (Lee/40), and influenza C (Ann Arbor/1/50) in the NCBI Entrez Genome database and NCBI Influenza Virus Resource. 3 proteins were identified as promising, which are neuraminidase of influenza A and B, polymerase of influenza A, B and C, and matrix protein 2 of influenza A. Additionally, 2 proteins were identified as marginally promising by two prediction systems, which are haemagglutinin of influenza A and B, and hemagglutinin-esterase of influenza C.

Further study reveals that the promising targets show fair drug discovery productivity levels in comparison to the modest levels for the marginally promising targets and low levels for the unpromising proteins. These results are highly consistent with the current drug discovery productivity levels against these proteins.

7.2 Limitations and suggestions for future studies

As a robust classification system, machine learning (especially support vector machine) plays a major role in the collective methods used for identifying targets. The performance of machine learning methods critically depends on the diversity of protein targets in the training dataset and the appropriate representation of these proteins. The dataset used in this work are not expected to fully represent all proteins possessing and not possessing a

specific property. This is particularly true for proteins not possessing a specific property given the vast protein space of several millions of proteins in current protein databases. Hence, inadequate representation of proteins from vast protein space to a certain extent will affect the performance of the models developed.

In the construction of the *in silico* machine learning system, only 320 successful protein targets have been used for capturing common features embedded in these targets. Based on previous analysis 320 only consist of a very small portion of the whole protein space, which makes identification of drug target biased to the non-target data by using diversity-dependent machine learning methods. Besides the including of more target data, there are many *in silico* ways to increase the diversity, which generally include direct variation on protein sequences and indirect variation on protein vectors. Therefore, it is necessary for us to try these methods for increasing data diversity, which may increase the prediction performance.

According to **Chapter 2 Section 2.3.1.2**, only three groups of descriptors were combined for generating protein vector, and they are (1) amino acid composition and composition, transition and distribution; (2) normalized Moreau-Broto autocorrelation; and (3) pseudo amino acid composition. However, there are several other groups of descriptors which are popular in representing protein, like: (4) Geary autocorrelation; (5) Moran autocorrelation; (6) sequence order, and so on. The reason why we choose the first three groups is because they have demonstrated the best performance in predicting protein functional families³⁸⁶, but this may not be true for therapeutic targets. Moreover, the physicochemical properties used in this study are hydrophobicity, polarity, polarizability, charge, secondary structures,

surface tension, and normalized Van der Waals volumes, but we have not sought whether there is other property which can be good complementary to them. On the other hand, as shown in **Table 02-2**, all physicochemical properties are classified into only three groups. Take the property hydrophobicity as an example, amino acids RKEDQN are grouped in polar, GASTPHY in neutral, and CVLUMFW in hydrophobic, but the difference among amino acid inside each group has been ignored. Thus, different combination of descriptor groups, increasing number of physicochemical properties, and more detail classification of amino acids by properties will help to validate the prediction system, and may further enhance our model.

One of the most distinguished features of current TTD is the availability of clinical trial information for both drugs and targets. However, the clinical trial status keeps changing for our modern pharmacology is a dynamically moving process, the clinical trial status in TTD will not change only when manually update is applied onto the original TTD data. This kind of manual update is expected to be delayed for even several months after the original information is released, which makes TTD data not up-to-date. A solution to this may be to integrate automatic information system which helps TTD to search latest data update from reliable sources, like reputable journals, famous pharmaceutical companies and so on, then latest information can be automatically updated.

As multi-functional tools in pharmaceutical research, TTD is ambitious for integrating more functions. For instance, we can probably extend TTD by add pharmacophore model to predict antagonists/agonists for certain target. Similarly, Docking and QSAR model can also be added as extensions of TTD functions.

Bibliography

- 1 Terstappen, G. C. & Reggiani, A. In silico research in drug discovery. *Trends Pharmacol Sci* **22**, 23-26 (2001).
- 2 Ashburn, T. T. & Thor, K. B. Drug repositioning: Identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery* **3**, 673-683 (2004).
- 3 Paul, S. M. *et al.* How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov* **9**, 203-214 (2010).
- 4 Westhouse, R. A. Safety assessment considerations and strategies for targeted small molecule cancer therapeutics in drug discovery. *Toxicol Pathol* **38**, 165-168 (2010).
- 5 Sollano, J. A., Kirsch, J. M., Bala, M. V., Chambers, M. G. & Harpole, L. H. The economics of drug discovery and the ultimate valuation of pharmacotherapies in the marketplace. *Clin Pharmacol Ther* **84**, 263-266 (2008).
- 6 Ohlstein, E. H., Ruffolo, R. R., Jr. & Elliott, J. D. Drug discovery in the next millennium. *Annu Rev Pharmacol Toxicol* **40**, 177-191 (2000).
- 7 Cardon, L. R. & Watkins, H. Waiting for the working draft from the human genome project. A huge achievement, but not of immediate medical use. *Bmj* **320**, 1223-1224 (2000).
- 8 Hopkins, A. L. & Groom, C. R. The druggable genome. *Nat Rev Drug Discov* **1**, 727-730 (2002).
- 9 Zhu, F. *et al.* Update of TTD: Therapeutic Target Database. *Nucleic Acids Res* **38**, D787-791 (2010).
- 10 Russ, A. P. & Lampel, S. The druggable genome: an update. *Drug Discov Today* **10**, 1607-1610 (2005).
- 11 Qiu, S., Adema, C. M. & Lane, T. A computational study of off-target effects of RNA interference. *Nucleic Acids Res* **33**, 1834-1847 (2005).
- 12 Jackson, A. L. *et al.* Expression profiling reveals off-target gene regulation by RNAi. *Nat Biotechnol* **21**, 635-637 (2003).
- 13 Segal, E., Friedman, N., Kaminski, N., Regev, A. & Koller, D. From signatures to models: understanding cancer using microarrays. *Nat Genet* **37 Suppl**, S38-45 (2005).

- 14 Boon, K. *et al.* An anatomy of normal and malignant gene expression. *Proc Natl Acad Sci U S A* **99**, 11287-11292 (2002).
- 15 Lash, A. E. *et al.* SAGEmap: a public gene expression resource. *Genome Res* **10**, 1051-1060 (2000).
- 16 Loging, W. T. *et al.* Identifying potential tumor markers and antigens by database mining and rapid expression screening. *Genome Res* **10**, 1393-1402 (2000).
- 17 Brazhnik, P., de la Fuente, A. & Mendes, P. Gene networks: how to put the function in genomics. *Trends Biotechnol* **20**, 467-472 (2002).
- 18 Parkinson, H. *et al.* ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res* **37**, D868-872 (2009).
- 19 Sese, J. *et al.* BodyMap incorporated PCR-based expression profiling data and a gene ranking system. *Nucleic Acids Res* **29**, 156-158 (2001).
- 20 Ikeo, K., Ishi-i, J., Tamura, T., Gojobori, T. & Tateno, Y. CIBEX: center for information biology gene expression database. *C R Biol* **326**, 1079-1082 (2003).
- 21 Kaminuma, E. *et al.* DDBJ launches a new archive database with analytical tools for next-generation sequence data. *Nucleic Acids Res* **38**, D33-38 (2010).
- 22 Aach, J., Rindone, W. & Church, G. M. Systematic management and analysis of yeast gene expression data. *Genome Res* **10**, 431-445 (2000).
- 23 Barrett, T. *et al.* NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res* **37**, D885-890 (2009).
- 24 Smith, C. M. *et al.* The mouse Gene Expression Database (GXD): 2007 update. *Nucleic Acids Res* **35**, D618-623 (2007).
- 25 Haverty, P. M. *et al.* HugeIndex: a database with visualization tools for high-density oligonucleotide array data from normal human tissues. *Nucleic Acids Res* **30**, 214-217 (2002).
- 26 Diehn, M. *et al.* SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res* **31**, 219-223 (2003).
- 27 Imoto, S., Tamada, Y., Savoie, C. J. & Miyano, S. Analysis of gene networks for drug target discovery and validation. *Methods Mol Biol* **360**, 33-56 (2007).

- 28 Shmulevich, I., Dougherty, E. R., Kim, S. & Zhang, W. Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics* **18**, 261-274 (2002).
- 29 Tamada, Y., Imoto, S., Tashiro, K., Kuhara, S. & Miyano, S. Identifying drug active pathways from gene networks estimated by gene expression data. *Genome Inform* **16**, 182-191 (2005).
- 30 Magin, T. M. Lessons from keratin transgenic and knockout mice. *Subcell Biochem* **31**, 141-172 (1998).
- 31 Hrabe de Angelis, M. H. *et al.* Genome-wide, large-scale production of mutant mice by ENU mutagenesis. *Nat Genet* **25**, 444-447 (2000).
- 32 Balling, R. ENU mutagenesis: analyzing gene function in mice. *Annu Rev Genomics Hum Genet* **2**, 463-492 (2001).
- 33 Sioud, M. Therapeutic siRNAs. *Trends Pharmacol Sci* **25**, 22-28 (2004).
- 34 Zambrowicz, B. P. & Sands, A. T. Knockouts model the 100 best-selling drugs--will they model the next 100? *Nat Rev Drug Discov* **2**, 38-51 (2003).
- 35 Abu-Elheiga, L., Matzuk, M. M., Abo-Hashema, K. A. & Wakil, S. J. Continuous fatty acid oxidation and reduced fat storage in mice lacking acetyl-CoA carboxylase 2. *Science* **291**, 2613-2616 (2001).
- 36 Resing, K. A. & Ahn, N. G. Proteomics strategies for protein identification. *FEBS Lett* **579**, 885-889 (2005).
- 37 Lee, W. *et al.* The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**, 473-477 (2010).
- 38 Thorgeirsson, T. E. *et al.* A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* **452**, 638-642 (2008).
- 39 Chen, Y. *et al.* Variations in DNA elucidate molecular networks that cause disease. *Nature* **452**, 429-435 (2008).
- 40 Fenwick, A. Waterborne infectious diseases--could they be consigned to history? *Science* **313**, 1077-1081 (2006).

- 41 Nemecek, J. C., Wuthrich, M. & Klein, B. S. Global control of dimorphism and virulence in fungi. *Science* **312**, 583-588 (2006).
- 42 Pyne, S. Air pollution. Small particles add up to big disease risk. *Science* **295**, 1994 (2002).
- 43 Thaker, N. G. *et al.* Identification of survival genes in human glioblastoma cells by small interfering RNA screening. *Mol Pharmacol* **76**, 1246-1255 (2009).
- 44 Dragunow, M. The adult human brain in preclinical drug development. *Nat Rev Drug Discov* **7**, 659-666 (2008).
- 45 Bult, C. J., Kadin, J. A., Richardson, J. E., Blake, J. A. & Eppig, J. T. The Mouse Genome Database: enhancements and updates. *Nucleic Acids Res* **38**, D586-592 (2010).
- 46 McCaffrey, A. P. *et al.* RNA interference in adult mice. *Nature* **418**, 38-39 (2002).
- 47 Naftzger, C. *et al.* Immune response to a differentiation antigen induced by altered antigen: a study of tumor rejection and autoimmunity. *Proc Natl Acad Sci U S A* **93**, 14809-14814 (1996).
- 48 Abina, M. A. *et al.* Thrombopoietin (TPO) knockout phenotype induced by cross-reactive antibodies against TPO following injection of mice with recombinant adenovirus encoding human TPO. *J Immunol* **160**, 4481-4489 (1998).
- 49 Horuk, R. Chemokine receptor antagonists: overcoming developmental hurdles. *Nat Rev Drug Discov* **8**, 23-33 (2009).
- 50 Vidalin, O., Muslmani, M., Estienne, C., Echchakir, H. & Abina, A. M. In vivo target validation using gene invalidation, RNA interference and protein functional knockout models: it is the time to combine. *Curr Opin Pharmacol* **9**, 669-676 (2009).
- 51 Drews, J. Genomic sciences and the medicine of tomorrow. *Nat Biotechnol* **14**, 1516-1518 (1996).
- 52 Drews, J. & Ryser, S. Classic drug targets. *Nat Biotechnol* **15**, 1318-1319 (1997).
- 53 Golden, J. B. Prioritizing the human genome: knowledge management for drug discovery. *Curr Opin Drug Discov Devel* **6**, 310-316 (2003).
- 54 Golden, J. Towards a tractable genome: knowledge management in drug discovery. *Curr. Drug Discov.*, 17-20 (2003).

- 55 Wishart, D. S. *et al.* DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* **34**, D668-672 (2006).
- 56 Imming, P., Sinning, C. & Meyer, A. Drugs, their targets and the nature and number of drug targets. *Nat Rev Drug Discov* **5**, 821-834 (2006).
- 57 Zheng, C., Han, L., Yap, C. W., Xie, B. & Chen, Y. Progress and problems in the exploration of therapeutic targets. *Drug Discov Today* **11**, 412-420 (2006).
- 58 Zheng, C. J. *et al.* Therapeutic targets: progress of their exploration and investigation of their characteristics. *Pharmacol Rev* **58**, 259-279 (2006).
- 59 Overington, J. P., Al-Lazikani, B. & Hopkins, A. L. How many drug targets are there? *Nat Rev Drug Discov* **5**, 993-996 (2006).
- 60 Barh, D., Kumar, A. & Misra, A. N. Genomic Target Database (GTD): a database of potential targets in human pathogenic bacteria. *Bioinformatics* **4**, 50-51 (2009).
- 61 Liu, T., Lin, Y., Wen, X., Jorissen, R. N. & Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res* **35**, D198-201 (2007).
- 62 Gunther, S. *et al.* SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res* **36**, D919-922 (2008).
- 63 Klein, T. E. *et al.* Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. *Pharmacogenomics J* **1**, 167-170 (2001).
- 64 Kuhn, M. *et al.* STITCH 2: an interaction network database for small molecules and proteins. *Nucleic Acids Res* **38**, D552-556 (2010).
- 65 Agüero, F. *et al.* Genomic-scale prioritization of drug targets: the TDR Targets database. *Nat Rev Drug Discov* **7**, 900-907 (2008).
- 66 Wishart, D. S. *et al.* DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* **36**, D901-906 (2008).
- 67 Plewczynski, D. & Rychlewski, L. Meta-basic estimates the size of druggable human genome. *J Mol Model* **15**, 695-699 (2009).

- 68 Han, L. *et al.* Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity. *Proteomics* **6**, 4023-4037 (2006).
- 69 Sakharkar, M. K. & Sakharkar, K. R. Targetability of human disease genes. *Curr Drug Discov Technol* **4**, 48-58 (2007).
- 70 Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* **46**, 3-26 (2001).
- 71 Bailey, D., Zanders, E. & Dean, P. The end of the beginning for genomic medicine. *Nat Biotechnol* **19**, 207-209 (2001).
- 72 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
- 73 Claverie, J. M. Gene number. What if there are only 30,000 human genes? *Science* **291**, 1255-1257 (2001).
- 74 Drews, J. Drug discovery: a historical perspective. *Science* **287**, 1960-1964 (2000).
- 75 Walke, D. W. *et al.* In vivo drug target discovery: identifying the best targets from the genome. *Curr Opin Biotechnol* **12**, 626-631 (2001).
- 76 Ilag, L. L., Ng, J. H., Beste, G. & Henning, S. W. Emerging high-throughput drug target validation technologies. *Drug Discov Today* **7**, S136-142 (2002).
- 77 Lindsay, M. A. Finding new drug targets in the 21st century. *Drug Discov Today* **10**, 1683-1687 (2005).
- 78 Sams-Dodd, F. Target-based drug discovery: is something wrong? *Drug Discov Today* **10**, 139-147 (2005).
- 79 Kramer, R. & Cohen, D. Functional genomics to new drug targets. *Nat Rev Drug Discov* **3**, 965-972 (2004).
- 80 Ryan, T. E. & Patterson, S. D. Proteomics: drug target discovery on an industrial scale. *Trends Biotechnol* **20**, S45-51 (2002).
- 81 Lindsay, M. A. Target discovery. *Nat Rev Drug Discov* **2**, 831-838 (2003).

- 82 Nicolette, C. A. & Miller, G. A. The identification of clinically relevant markers and therapeutic targets. *Drug Discov Today* **8**, 31-38 (2003).
- 83 Jackson, P. D. & Harrington, J. J. High-throughput target discovery using cell-based genetics. *Drug Discov Today* **10**, 53-60 (2005).
- 84 Austen, M. & Dohrmann, C. Phenotype-first screening for the identification of novel drug targets. *Drug Discov Today* **10**, 275-282 (2005).
- 85 Wang, S., Sim, T. B., Kim, Y. S. & Chang, Y. T. Tools for target identification and validation. *Curr Opin Chem Biol* **8**, 371-377 (2004).
- 86 Hajduk, P. J., Huth, J. R. & Tse, C. Predicting protein druggability. *Drug Discov Today* **10**, 1675-1682 (2005).
- 87 Hardy, L. W. & Peet, N. P. The multiple orthogonal tools approach to define molecular causation in the validation of druggable targets. *Drug Discov Today* **9**, 117-126 (2004).
- 88 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990).
- 89 Fetrow, J. S. & Skolnick, J. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J Mol Biol* **281**, 949-968 (1998).
- 90 Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**, 443-453 (1970).
- 91 Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J Mol Biol* **147**, 195-197 (1981).
- 92 Reichert, T. A., Cohen, D. N. & Wong, A. K. An application of information theory to genetic mutations and the matching of polypeptide sequences. *J Theor Biol* **42**, 245-261 (1973).
- 93 Lipman, D. J. & Pearson, W. R. Rapid and sensitive protein similarity searches. *Science* **227**, 1435-1441 (1985).
- 94 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).

- 95 Shah, I. & Hunter, L. Predicting enzyme function from sequence: a systematic appraisal. *Proc Int Conf Intell Syst Mol Biol* **5**, 276-283 (1997).
- 96 Wilson, C. A., Kreychman, J. & Gerstein, M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J Mol Biol* **297**, 233-249 (2000).
- 97 Todd, A. E., Orengo, C. A. & Thornton, J. M. Plasticity of enzyme active sites. *Trends Biochem Sci* **27**, 419-426 (2002).
- 98 Devos, D. & Valencia, A. Intrinsic errors in genome annotation. *Trends Genet* **17**, 429-431 (2001).
- 99 Rost, B. Enzyme function less conserved than anticipated. *J Mol Biol* **318**, 595-608 (2002).
- 100 Katoh, M. & Katoh, M. Identification and characterization of human SNAIL3 (SNAI3) gene in silico. *Int J Mol Med* **11**, 383-388 (2003).
- 101 Enright, A. J. & Ouzounis, C. A. GeneRAGE: a robust algorithm for sequence clustering and domain detection. *Bioinformatics* **16**, 451-457 (2000).
- 102 Benner, S. A., Chamberlin, S. G., Liberles, D. A., Govindarajan, S. & Knecht, L. Functional inferences from reconstructed evolutionary biology involving rectified databases--an evolutionarily grounded approach to functional genomics. *Res Microbiol* **151**, 97-106 (2000).
- 103 Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* **96**, 4285-4288 (1999).
- 104 Lubec, G., Afjehi-Sadat, L., Yang, J. W. & John, J. P. Searching for hypothetical proteins: theory and practice based upon original data and literature. *Prog Neurobiol* **77**, 90-127 (2005).
- 105 Rost, B. & Valencia, A. Pitfalls of protein sequence analysis. *Curr Opin Biotechnol* **7**, 457-461 (1996).
- 106 Hulo, N. *et al.* The PROSITE database. *Nucleic Acids Res* **34**, D227-230 (2006).
- 107 Attwood, T. K., Beck, M. E., Bleasby, A. J. & Parry-Smith, D. J. PRINTS--a database of protein motif fingerprints. *Nucleic Acids Res* **22**, 3590-3596 (1994).

- 108 Westhead, D. R., Parish, J. H. & Twyman, R. M. Instant Notes in Bioinformatics. *BIOS Scientific Publishers Limited, Oxford*. (2002).
- 109 Bru, C. *et al.* The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res* **33**, D212-215 (2005).
- 110 Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Res* **38**, D211-222 (2010).
- 111 Apweiler, R. *et al.* The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res* **29**, 37-40 (2001).
- 112 Eddy, S. R. A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Comput Biol* **4**, e1000069 (2008).
- 113 Karplus, K., Barrett, C. & Hughey, R. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14**, 846-856 (1998).
- 114 Whisstock, J. C. & Lesk, A. M. Prediction of protein function from protein sequence and structure. *Q Rev Biophys* **36**, 307-340 (2003).
- 115 Henikoff, S. *et al.* Gene families: the taxonomy of protein paralogs and chimeras. *Science* **278**, 609-614 (1997).
- 116 Tramontano, A. Of men and machines. *Nat Struct Biol* **10**, 87-90 (2003).
- 117 Chothia, C. & Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *Embo J* **5**, 823-826 (1986).
- 118 Holm, L. & Sander, C. Protein structure comparison by alignment of distance matrices. *J Mol Biol* **233**, 123-138 (1993).
- 119 Orengo, C. A. *et al.* CATH--a hierarchic classification of protein domain structures. *Structure* **5**, 1093-1108 (1997).
- 120 Greene, L. H. *et al.* The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res* **35**, D291-297 (2007).
- 121 Lo Conte, L., Brenner, S. E., Hubbard, T. J., Chothia, C. & Murzin, A. G. SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res* **30**, 264-267 (2002).

- 122 Andreeva, A. *et al.* Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res* **36**, D419-425 (2008).
- 123 Rost, B., Liu, J., Nair, R., Wrzeszczynski, K. O. & Ofran, Y. Automatic prediction of protein function. *Cell Mol Life Sci* **60**, 2637-2650 (2003).
- 124 Skolnick, J. & Fetrow, J. S. From genes to protein structure and function: novel applications of computational approaches in the genomic era. *Trends Biotechnol* **18**, 34-39 (2000).
- 125 Bernstein, F. C. *et al.* The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol* **112**, 535-542 (1977).
- 126 Consortium., U. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res* **38**, D142-148 (2010).
- 127 Wallace, A. C., Borkakoti, N. & Thornton, J. M. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci* **6**, 2308-2323 (1997).
- 128 Ivanciuc, O. *et al.* Using property based sequence motifs and 3D modeling to determine structure and functional regions of proteins. *Curr Med Chem* **11**, 583-593 (2004).
- 129 Stark, A. & Russell, R. B. Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures. *Nucleic Acids Res* **31**, 3341-3344 (2003).
- 130 Aravind, L. Guilt by association: contextual information in genome analysis. *Genome Res* **10**, 1074-1077 (2000).
- 131 Lin, H. H., Han, L. Y., Cai, C. Z., Ji, Z. L. & Chen, Y. Z. Prediction of transporter family from protein sequence by support vector machine approach. *Proteins* **62**, 218-231 (2006).
- 132 Ben-Hur, A. & Noble, W. S. Kernel methods for predicting protein-protein interactions. *Bioinformatics* **21 Suppl 1**, i38-46 (2005).
- 133 Karchin, R., Karplus, K. & Haussler, D. Classifying G-protein coupled receptors with support vector machines. *Bioinformatics* **18**, 147-159 (2002).
- 134 Yabuki, Y., Muramatsu, T., Hirokawa, T., Mukai, H. & Suwa, M. GRIFFIN: a system for predicting GPCR-G-protein coupling selectivity using a support vector machine and a hidden Markov model. *Nucleic Acids Res* **33**, W148-153 (2005).

- 135 Bhasin, M. & Raghava, G. P. Classification of nuclear receptors based on amino acid composition and dipeptide composition. *J Biol Chem* **279**, 23262-23266 (2004).
- 136 Cai, Y. D., Zhou, G. P. & Chou, K. C. Support vector machines for predicting membrane protein types by using functional domain composition. *Biophys J* **84**, 3257-3263 (2003).
- 137 Wang, M., Yang, J., Liu, G. P., Xu, Z. J. & Chou, K. C. Weighted-support vector machines for predicting membrane protein types based on pseudo-amino acid composition. *Protein Eng Des Sel* **17**, 509-516 (2004).
- 138 Lin, H. H. *et al.* Prediction of the functional class of lipid binding proteins from sequence-derived properties irrespective of sequence similarity. *J Lipid Res* **47**, 824-831 (2006).
- 139 Cai, C. Z., Han, L. Y., Ji, Z. L. & Chen, Y. Z. Enzyme family classification by support vector machines. *Proteins* **55**, 66-76 (2004).
- 140 Zhang, Z., Kochhar, S. & Grigorov, M. G. Descriptor-based protein remote homology identification. *Protein Sci* **14**, 431-444 (2005).
- 141 Han, L. Y. *et al.* Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. *Nucleic Acids Res* **32**, 6437-6444 (2004).
- 142 Cui, J. *et al.* Prediction of functional class of novel bacterial proteins without the use of sequence similarity by a statistical learning method. *J Mol Microbiol Biotechnol* **9**, 86-100 (2005).
- 143 Hou, Y., Hsu, W., Lee, M. L. & Bystroff, C. Remote homolog detection using local sequence-structure correlations. *Proteins* **57**, 518-530 (2004).
- 144 Xu, H. *et al.* Learning the drug target-likeness of a protein. *Proteomics* **7**, 4255-4263 (2007).
- 145 Bao, L. & Sun, Z. Identifying genes related to drug anticancer mechanisms using support vector machine. *FEBS Lett* **521**, 109-114 (2002).
- 146 Bhasin, M. & Raghava, G. P. GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. *Nucleic Acids Res* **32**, W383-389 (2004).
- 147 Bhardwaj, N., Langlois, R. E., Zhao, G. & Lu, H. Kernel-based machine learning protocol for predicting DNA-binding proteins. *Nucleic Acids Res* **33**, 6486-6493 (2005).
- 148 Smith, C. Drug target validation: Hitting the target. *Nature* **422**, 341, 343, 345 passim (2003).

- 149 Ostertag, E. M., Madison, B. B. & Kano, H. Mutagenesis in rodents using the L1 retrotransposon. *Genome Biol* **8 Suppl 1**, S16 (2007).
- 150 Scrable, H. Say when: reversible control of gene expression in the mouse by lac. *Semin Cell Dev Biol* **13**, 109-119 (2002).
- 151 Lewandoski, M. Conditional control of gene expression in the mouse. *Nat Rev Genet* **2**, 743-755 (2001).
- 152 Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806-811 (1998).
- 153 Ameres, S. L., Martinez, J. & Schroeder, R. Molecular basis for target RNA recognition and cleavage by human RISC. *Cell* **130**, 101-112 (2007).
- 154 Voorhoeve, P. M. & Agami, R. Knockdown stands up. *Trends Biotechnol* **21**, 2-4 (2003).
- 155 Kleinman, M. E. *et al.* Sequence- and target-independent angiogenesis suppression by siRNA via TLR3. *Nature* **452**, 591-597 (2008).
- 156 DiFiglia, M. *et al.* Therapeutic silencing of mutant huntingtin with siRNA attenuates striatal and cortical neuropathology and behavioral deficits. *Proc Natl Acad Sci U S A* **104**, 17204-17209 (2007).
- 157 Zimmermann, T. S. *et al.* RNAi-mediated gene silencing in non-human primates. *Nature* **441**, 111-114 (2006).
- 158 Kumar, P. *et al.* T cell-specific siRNA delivery suppresses HIV-1 infection in humanized mice. *Cell* **134**, 577-586 (2008).
- 159 Hogrefe, R. I. *et al.* Chemically modified short interfering hybrids (siHYBRIDS): nanoimmunoliposome delivery in vitro and in vivo for RNAi of HER-2. *Nucleosides Nucleotides Nucleic Acids* **25**, 889-907 (2006).
- 160 Peer, D., Zhu, P., Carman, C. V., Lieberman, J. & Shimaoka, M. Selective gene silencing in activated leukocytes by targeting siRNAs to the integrin lymphocyte function-associated antigen-1. *Proc Natl Acad Sci U S A* **104**, 4095-4100 (2007).
- 161 Kumar, P. *et al.* Transvascular delivery of small interfering RNA to the central nervous system. *Nature* **448**, 39-43 (2007).

- 162 Frank-Kamenetsky, M. *et al.* Therapeutic RNAi targeting PCSK9 acutely lowers plasma cholesterol in rodents and LDL cholesterol in nonhuman primates. *Proc Natl Acad Sci U S A* **105**, 11915-11920 (2008).
- 163 Lin, Y. R., Wei, H. Y., Tsai, T. L. & Lin, T. H. HDAPD: a web tool for searching the disease-associated protein structures. *BMC Bioinformatics* **11**, 88 (2010).
- 164 Gao, Z. *et al.* PDTD: a web-accessible protein database for drug target identification. *BMC Bioinformatics* **9**, 104 (2008).
- 165 Li, H. *et al.* TarFisDock: a web server for identifying drug targets with docking approach. *Nucleic Acids Res* **34**, W219-224 (2006).
- 166 Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic acids research* **38**, D5-16 (2010).
- 167 Rao, J. Reasoning about probabilistic parallel programs. *ACM Transactions on Programming Languages and Systems* **16**, 798-842 (1994).
- 168 Briscoe, G. & Caelli, T. *A compendium of machine learning* Vol. 1 (Ablex, 1996).
- 169 Alpaydm, E. *Introduction to Machine learning* (The MIT Press, 2004).
- 170 Dietterich, T. G. in *Nature Encyclopedia of Cognitive Science* (Macmillan, 2003).
- 171 Kotsiantis, S. B. Supervised Machine Learning: A Review of Classification Techniques. *Informatica* **31**, 249-268 (2007).
- 172 Vapnik, V. N. *The Nature of Statistical Learning Theory*. (Springer-Verlag New York Inc, 1995).
- 173 BURGESS, C. J. C. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery* **2**, 121-167 (1988).
- 174 Vapnik, V. *The nature of statistical learning theory*. (Springer-Verlag New York, Inc., 1995).
- 175 Nuttakorn Thubthong & Boonserm Kijirikul. Support vector machines for Thai phoneme recognition. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **9** (2001).
- 176 Ben-Yacoub, S., Abdeljaoued & Y. & Mayoraz, E. Fusion Face and Speech Data for Person Identity Verification. *IEEE Transactions on Neural Networks* **10**, 1065-1074 (1999).

- 177 Karlsen, R. E., Gorsich, D. J. & Gerhart, G. R. Target classification via support vector machines. *Optical Engineering* **39**, 704-711 (2000).
- 178 Papageorgiou, C. & Poggio, T. A trainable system for object detection. *International Journal of Computer Vision* **38**, 15-33 (2000).
- 179 Huang, C., Davis, L. S. & Townshend, J. R. G. An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing* **23**, 725-749 (2002).
- 180 Bock, J. R. & Gough, D. A. Predicting protein--protein interactions from primary structure. *Bioinformatics* **17**, 455-460 (2001).
- 181 Busuttill, S., Abela, J. & Pace, G. J. Support vector machines with profile-based kernels for remote protein homology detection. *Genome Inform* **15**, 191-200 (2004).
- 182 Webb-Robertson, B. J., Oehmen, C. & Matzke, M. SVM-BALSA: remote homology detection based on Bayesian sequence alignment. *Comput Biol Chem* **29**, 440-443 (2005).
- 183 Hongzong, S. *et al.* Support vector machines classification for discriminating coronary heart disease patients from non-coronary heart disease. *West Indian Med J* **56**, 451-457 (2007).
- 184 Vapnik, V. *The nature of statistical learning theory*. (Springer, 1995).
- 185 Cristianini, N. & Shawe-Taylor, J. *An introduction to Support Vector Machines : and other kernel-based learning methods*. (Cambridge University Press, 2000).
- 186 Platt, J. C. Sequential Minimal Optimization: A fast algorithm for training support vector machines. *Microsoft Research. Technical Report MSR-TR-98-14* (1998).
- 187 Osuna, E., Freund, R. and Girosi, F. An improved training algorithm for support vector machines. *Neural Networks for Signal Processing VII-Proceedings of the 1997 IEEE Workshop*, 276-285 (1997).
- 188 Aizerman, M. A., Braverman, E. M. & er, L. I. R. Theoretical foundations of the potential function method in pattern recognition and learning. *Automation and Remote Control* **25**, 821--837 (1964).
- 189 Courant, R. & Hilbert, D. *Methods of Mathematical Physics*. (John Wiley & Sons, 1989).
- 190 Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. & Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **16**, 412-424 (2000).

- 191 Cai, C. Z., Han, L. Y., Ji, Z. L., Chen, X. & Chen, Y. Z. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic acids research* **31**, 3692-3697 (2003).
- 192 Schneider, G. & Wrede, P. The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site. *Biophys J* **66**, 335-344 (1994).
- 193 Chou, K. C. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem Biophys Res Commun* **278**, 477-483 (2000).
- 194 Chou, K. C. & Cai, Y. D. Prediction of protein subcellular locations by GO-FunD-PseAA predictor. *Biochem Biophys Res Commun* **320**, 1236-1239 (2004).
- 195 Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D. and Bairoch, A. *The Proteomics Protocols Handbook*. J. M. (ed.), 571-607 (2005).
- 196 Jensen, L. J. *et al.* Prediction of human protein function from post-translational modifications and localization features. *J Mol Biol* **319**, 1257-1265 (2002).
- 197 Carr, A. M. *et al.* Analysis of a histone H2A variant from fission yeast: evidence for a role in chromosome stability. *Mol Gen Genet* **245**, 628-635 (1994).
- 198 de Lichtenberg, U., Jensen, T. S., Jensen, L. J. & Brunak, S. Protein feature based identification of cell cycle regulated proteins in yeast. *J Mol Biol* **329**, 663-674 (2003).
- 199 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.* **In Press** (2006).
- 200 Gasteiger, E. *et al.* in *The Proteomics Protocols Handbook* (ed M. Walker John) 571-607 (Humana Press 2005).
- 201 Zheng, C., Han, L., Yap, C. W., Xie, B. & Chen, Y. Progress and problems in the exploration of therapeutic targets. *Drug discovery today* **11**, 412-420 (2006).
- 202 Cai, C. Z., Han, L. Y., Ji, Z. L., Chen, X. & Chen, Y. Z. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res* **31**, 3692-3697 (2003).
- 203 Broto, P., Moreau, G. & Vandicke, C. Molecular structures: perception, autocorrelation descriptor and SAR studies. *Eur. J. Med. Chem.* **19**, 71-78 (1984).

- 204 Kawashima, S. & Kanehisa, M. AAindex: amino acid index database. *Nucleic Acids Res* **28**, 374 (2000).
- 205 Cid, H., Bunster, M., Canales, M. & Gazitua, F. Hydrophobicity and structural classes in proteins. *Protein Eng* **5**, 373-375 (1992).
- 206 Bhaskaran, R. & Ponnuswammy, P. K. Positional flexibilities of amino acid residues in globular proteins. *Int. J. Pept. and Protein Res.* **32**, 242-255 (1988).
- 207 Charton, M. & Charton, B. I. The structural dependence of amino acid hydrophobicity parameters. *J Theor Biol* **99**, 629-644 (1982).
- 208 Chothia, C. The nature of the accessible and buried surfaces in proteins. *J Mol Biol* **105**, 1-12 (1976).
- 209 Bigelow, C. C. On the average hydrophobicity of proteins and the relation between it and protein structure. *J Theor Biol* **16**, 187-211 (1967).
- 210 Charton, M. Protein folding and the genetic code: an alternative quantitative model. *J Theor Biol* **91**, 115-123 (1981).
- 211 Dayhoff, H. & Calderone, H. Composition of Proteins. *Atlas of Protein Sequence and Structure* **5**, 363-373 (1978).
- 212 Moreau, G. & Broto, P. Autocorrelation of molecular structures, application to SAR studies. *Nour. J. Chim.* **4**, 757-764 (1980).
- 213 Feng, Z. P. & Zhang, C. T. Prediction of membrane protein types based on the hydrophobic index of amino acids. *J Protein Chem* **19**, 269-275 (2000).
- 214 Lin, Z. & Pan, X. M. Accurate prediction of protein secondary structural content. *J Protein Chem* **20**, 217-220 (2001).
- 215 Chou, K. C. Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins: Structure Function and Genetics* **43**, 246-255 (2001).
- 216 Chen, X., Ji, Z. L. & Chen, Y. Z. TTD: Therapeutic Target Database. *Nucleic Acids Res* **30**, 412-415 (2002).
- 217 Chantry, D. G protein-coupled receptors: from ligand identification to drug targets. 14-16 October 2002, San Diego, CA, USA. *Expert Opin Emerg Drugs* **8**, 273-276 (2003).

- 218 Finn, R. D. *et al.* Pfam: clans, web tools and services. *Nucleic Acids Res* **34**, D247-251 (2006).
- 219 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25**, 3389-3402 (1997).
- 220 George, R. A. & Heringa, J. Protein domain identification and improved sequence similarity searching using PSI-BLAST. *Proteins* **48**, 672-681 (2002).
- 221 Gerstein, M. Measurement of the effectiveness of transitive sequence comparison, through a third 'intermediate' sequence. *Bioinformatics* **14**, 707-714 (1998).
- 222 Koehl, P. & Levitt, M. Sequence variations within protein families are linearly related to structural variations. *J Mol Biol* **323**, 551-562 (2002).
- 223 Wood, T. C. & Pearson, W. R. Evolution of protein sequences and structures. *J Mol Biol* **291**, 977-995 (1999).
- 224 Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology* **247**, 536-540 (1995).
- 225 Koch, M. A. & Waldmann, H. Protein structure similarity clustering and natural product structure as guiding principles in drug discovery. *Drug discovery today* **10**, 471-483 (2005).
- 226 Hopkins, A. L. & Groom, C. R. The druggable genome. *Nature reviews* **1**, 727-730 (2002).
- 227 Zheng, C. J. *et al.* Therapeutic targets: progress of their exploration and investigation of their characteristics. *Pharmacol Rev* **58**, 259-279 (2006).
- 228 Kanehisa, M. *et al.* From genomics to chemical genomics: new developments in KEGG. *Nucleic acids research* **34**, D354-357 (2006).
- 229 Edwards, A. Large-scale structural biology of the human proteome. *Annu Rev Biochem* **78**, 541-568 (2009).
- 230 Lundstrom, K. Structural genomics: the ultimate approach for rational drug design. *Mol Biotechnol* **34**, 205-212 (2006).
- 231 Dey, R., Khan, S. & Saha, B. A novel functional approach toward identifying definitive drug targets. *Curr Med Chem* **14**, 2380-2392 (2007).

- 232 Hopkins, A. L. Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol* **4**, 682-690 (2008).
- 233 Giallourakis, C., Henson, C., Reich, M., Xie, X. & Mootha, V. K. Disease gene discovery through integrative genomics. *Annu Rev Genomics Hum Genet* **6**, 381-406 (2005).
- 234 Zimmermann, G. R., Lehar, J. & Keith, C. T. Multi-target therapeutics: when the whole is greater than the sum of the parts. *Drug Discov Today* **12**, 34-42 (2007).
- 235 Jia, J. *et al.* Mechanisms of drug combinations: interaction and network perspectives. *Nat Rev Drug Discov* **8**, 111-128 (2009).
- 236 Liebler, D. C. & Guengerich, F. P. Elucidating mechanisms of drug-induced toxicity. *Nat Rev Drug Discov* **4**, 410-420 (2005).
- 237 Eichelbaum, M., Ingelman-Sundberg, M. & Evans, W. E. Pharmacogenomics and individualized drug therapy. *Annu Rev Med* **57**, 119-137 (2006).
- 238 Barcellos, G. B. *et al.* Molecular modeling as a tool for drug discovery. *Curr Drug Targets* **9**, 1084-1091 (2008).
- 239 Lee, G. M. & Craik, C. S. Trapping moving targets with small molecules. *Science* **324**, 213-215 (2009).
- 240 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* **330**, 304-315 (2009).
- 241 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* **12**, 304-313 (2007).
- 242 Yildirim, M. A., Goh, K. I., Cusick, M. E., Barabasi, A. L. & Vidal, M. Drug-target network. *Nat Biotechnol* **25**, 1119-1126 (2007).
- 243 Willett, P. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci* **38**, 983-996 (1998).
- 244 Ma, X. H. *et al.* Evaluation of virtual screening performance of support vector machines trained by sparsely distributed active compounds. *J Chem Inf Model* **48**, 1227-1237 (2008).

- 245 Li, Z. R. *et al.* MODEL-molecular descriptor lab: a web-based server for computing structural and physicochemical features of compounds. *Biotechnol Bioeng* **97**, 389-396 (2007).
- 246 Yap, C. W., Li, H., Ji, Z. L. & Chen, Y. Z. Regression methods for developing QSAR and QSPR models to predict compounds of specific pharmacodynamic, pharmacokinetic and toxicological properties. *Mini Rev Med Chem* **7**, 1097-1107 (2007).
- 247 Li, H. *et al.* Machine learning approaches for predicting compounds that interact with therapeutic and ADMET related proteins. *J Pharm Sci* **96**, 2838-2860 (2007).
- 248 Bostrom, J., Hogner, A. & Schmitt, S. Do structurally similar ligands bind in a similar fashion? *J. Med. Chem* **49**, 6716-6725 (2006).
- 249 Huang, N., Shoichet, B. K. & Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem* **49**, 6789-6801 (2006).
- 250 Ragno, R. *et al.* Class II-selective histone deacetylase inhibitors. Part 2: alignment-independent GRIND 3-D QSAR, homology and docking studies. *Eur J Med Chem* **43**, 621-632 (2008).
- 251 Lapenna, S. & Giordano, A. Cell cycle kinases as therapeutic targets for cancer. *Nat Rev Drug Discov* **8**, 547-566 (2009).
- 252 Gainetdinov, R. R. & Caron, M. G. Monoamine transporters: from genes to behavior. *Annu Rev Pharmacol Toxicol* **43**, 261-284 (2003).
- 253 Zambrowicz, B. P. & Sands, A. T. Knockouts model the 100 best-selling drugs--will they model the next 100? *Nat Rev Drug Discov* **2**, 38-51 (2003).
- 254 Keith, C. T., Borisy, A. A. & Stockwell, B. R. Multicomponent therapeutics for networked systems. *Nature reviews* **4**, 71-78 (2005).
- 255 Keseru, G. M. & Makara, G. M. The influence of lead discovery strategies on the properties of drug candidates. *Nature reviews* **8**, 203-212 (2009).
- 256 Huwe, C. M. Synthetic library design. *Drug Discov Today* **11**, 763-767 (2006).
- 257 Bajorath, J. Integration of virtual and high-throughput screening. *Nature reviews* **1**, 882-894 (2002).
- 258 Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **432**, 862-865 (2004).

- 259 MacCoss, M. & Baillie, T. A. Organic chemistry in drug discovery. *Science* **303**, 1810-1813 (2004).
- 260 Hajduk, P. J. & Greer, J. A decade of fragment-based drug design: strategic advances and lessons learned. *Nature reviews* **6**, 211-219 (2007).
- 261 Lindpaintner, K. The impact of pharmacogenetics and pharmacogenomics on drug discovery. *Nature reviews* **1**, 463-469 (2002).
- 262 Debouck, C. & Metcalf, B. The impact of genomics on drug discovery. *Annu Rev Pharmacol Toxicol* **40**, 193-207 (2000).
- 263 Dollery, C. T. Beyond genomics. *Clin Pharmacol Ther* **82**, 366-370 (2007).
- 264 Schmid, M. B. Seeing is believing: the impact of structural genomics on antimicrobial drug discovery. *Nat Rev Microbiol* **2**, 739-746 (2004).
- 265 Dove, A. Proteomics: translating genomics into products? *Nat Biotechnol* **17**, 233-236 (1999).
- 266 Black, D. Has the NHS failed? *Health Bull (Edinb)* **41**, 56-60 (1983).
- 267 Oprea, T. I., Davis, A. M., Teague, S. J. & Leeson, P. D. Is there a difference between leads and drugs? A historical perspective. *J Chem Inf Comput Sci* **41**, 1308-1315 (2001).
- 268 Morphy, R. The influence of target family and functional activity on the physicochemical properties of pre-clinical compounds. *J Med Chem* **49**, 2969-2978 (2006).
- 269 Leeson, P. D. & Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nature reviews* **6**, 881-890 (2007).
- 270 Yao, L. & Rzhetsky, A. Quantitative systems-level determinants of human genes targeted by successful drugs. *Genome Res* **18**, 206-213 (2008).
- 271 Lipinski, C. A. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies* **1**, 337-341 (2004).
- 272 Lindsay, M. A. Finding new drug targets in the 21st century. *Drug discovery today* **10**, 1683-1687 (2005).
- 273 Sams-Dodd, F. Target-based drug discovery: is something wrong? *Drug discovery today* **10**, 139-147 (2005).

- 274 Sakharkar, M. K., Li, P., Zhong, Z. & Sakharkar, K. R. Quantitative analysis on the characteristics of targets with FDA approved drugs. *Int J Biol Sci* **4**, 15-22 (2008).
- 275 Hajduk, P. J., Huth, J. R. & Fesik, S. W. Druggability indices for protein targets derived from NMR-based screening data. *J Med Chem* **48**, 2518-2525 (2005).
- 276 Hajduk, P. J., Huth, J. R. & Tse, C. Predicting protein druggability. *Drug discovery today* **10**, 1675-1682 (2005).
- 277 Oprea, T. I. *et al.* Lead-like, drug-like or "Pub-like": how different are they? *J Comput Aided Mol Des* **21**, 113-119 (2007).
- 278 Overington, J. P., Al-Lazikani, B. & Hopkins, A. L. How many drug targets are there? *Nat Rev Drug Discov* **5**, 993-996 (2006).
- 279 Chen, Y. Z. & Zhi, D. G. Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule. *Proteins* **43**, 217-226 (2001).
- 280 Macdonald, I. A. Obesity: are we any closer to identifying causes and effective treatments? *Trends Pharmacol Sci* **21**, 334-336 (2000).
- 281 Agnati, L. F., Fuxe, K. & Ferre, S. How receptor mosaics decode transmitter signals. Possible relevance of cooperativity. *Trends Biochem Sci* **30**, 188-193 (2005).
- 282 Chiesi, M., Huppertz, C. & Hofbauer, K. G. Pharmacotherapy of obesity: targets and perspectives. *Trends Pharmacol Sci* **22**, 247-254 (2001).
- 283 Matter, A. Tumor angiogenesis as a therapeutic target. *Drug Discov Today* **6**, 1005-1024 (2001).
- 284 Kramer, R. & Cohen, D. Functional genomics to new drug targets. *Nature reviews* **3**, 965-972 (2004).
- 285 Ryan, T. E. & Patterson, S. D. Proteomics: drug target discovery on an industrial scale. *Trends in biotechnology* **20**, S45-51 (2002).
- 286 Lindsay, M. A. Target discovery. *Nature reviews* **2**, 831-838 (2003).
- 287 Nicolette, C. A. & Miller, G. A. The identification of clinically relevant markers and therapeutic targets. *Drug discovery today* **8**, 31-38 (2003).
- 288 Jackson, P. D. & Harrington, J. J. High-throughput target discovery using cell-based genetics. *Drug discovery today* **10**, 53-60 (2005).

- 289 Austen, M. & Dohrmann, C. Phenotype-first screening for the identification of novel drug targets. *Drug discovery today* **10**, 275-282 (2005).
- 290 Simmons, D. L. What makes a good anti-inflammatory drug target? *Drug discovery today* **11**, 210-219 (2006).
- 291 Booth, B. & Zimmel, R. Prospects for productivity. *Nature reviews* **3**, 451-456 (2004).
- 292 Rosenberg, L. Physician-scientists--endangered and essential. *Science* **283**, 331-332 (1999).
- 293 Drews, J. Strategic choices facing the pharmaceutical industry: a case for innovation. *Drug Discov. Today*. **2**, 72-78 (1997).
- 294 Chen, X., Ji, Z. L. & Chen, Y. Z. TTD: Therapeutic Target Database. *Nucleic Acids Res* **30**, 412-415 (2002).
- 295 Payne, D. J., Gwynn, M. N., Holmes, D. J. & Pompliano, D. L. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nature reviews* **6**, 29-40 (2007).
- 296 Mdluli, K. & Spigelman, M. Novel targets for tuberculosis drug discovery. *Curr Opin Pharmacol* **6**, 459-467 (2006).
- 297 Angermayr, B. *et al.* Heme oxygenase attenuates oxidative stress and inflammation, and increases VEGF expression in portal hypertensive rats. *J Hepatol* **44**, 1033-1039 (2006).
- 298 Ramnath, N. & Creaven, P. J. Matrix metalloproteinase inhibitors. *Curr Oncol Rep* **6**, 96-102 (2004).
- 299 Fedorov, O. *et al.* A systematic interaction map of validated kinase inhibitors with Ser/Thr kinases. *Proc Natl Acad Sci U S A* **104**, 20523-20528 (2007).
- 300 Sergina, N. V. *et al.* Escape from HER-family tyrosine kinase inhibitor therapy by the kinase-inactive HER3. *Nature* **445**, 437-441 (2007).
- 301 Gingras, D., Batist, G. & Beliveau, R. AE-941 (Neovastat): a novel multifunctional antiangiogenic compound. *Expert Rev Anticancer Ther* **1**, 341-347 (2001).
- 302 Dupont, E. *et al.* Antiangiogenic and antimetastatic properties of Neovastat (AE-941), an orally active extract derived from cartilage tissue. *Clin Exp Metastasis* **19**, 145-153 (2002).
- 303 Evans, J. F., Ferguson, A. D., Mosley, R. T. & Hutchinson, J. H. What's all the FLAP about?: 5-lipoxygenase-activating protein inhibitors for inflammatory diseases. *Trends Pharmacol Sci* **29**, 72-78 (2008).

- 304 Prematta, M. J., Prematta, T. & Craig, T. J. Treatment of hereditary angioedema with plasma-derived C1 inhibitor. *Ther Clin Risk Manag* **4**, 975-982 (2008).
- 305 Bork, K., Barnstedt, S. E., Koch, P. & Traupe, H. Hereditary angioedema with normal C1-inhibitor activity in women. *Lancet* **356**, 213-217 (2000).
- 306 Beinrohr, L. *et al.* C1 inhibitor serpin domain structure reveals the likely mechanism of heparin potentiation and conformational disease. *J Biol Chem* **282**, 21100-21109 (2007).
- 307 Keri, G. *et al.* A tumor-selective somatostatin analog (TT-232) with strong in vitro and in vivo antitumor activity. *Proc Natl Acad Sci U S A* **93**, 12513-12518 (1996).
- 308 Johnson, Z., Schwarz, M., Power, C. A., Wells, T. N. & Proudfoot, A. E. Multi-faceted strategies to combat disease by interference with the chemokine system. *Trends Immunol* **26**, 268-274 (2005).
- 309 Blakeney, J. S., Reid, R. C., Le, G. T. & Fairlie, D. P. Nonpeptidic ligands for peptide-activated G protein-coupled receptors. *Chem Rev* **107**, 2960-3041 (2007).
- 310 Chen, L. *et al.* Opposing cardioprotective actions and parallel hypertrophic effects of delta PKC and epsilon PKC. *Proc Natl Acad Sci U S A* **98**, 11114-11119 (2001).
- 311 Bates, E. *et al.* Intracoronary KAI-9803 as an adjunct to primary percutaneous coronary intervention for acute ST-segment elevation myocardial infarction. *Circulation* **117**, 886-896 (2008).
- 312 Metzler, B., Xu, Q. & Mayr, M. Letter by Metzler et al regarding article, "Intracoronary KAI-9803 as an adjunct to primary coronary intervention for acute ST-segment elevation myocardial infarction". *Circulation* **118**, e80 (2008).
- 313 Herbert, M. K. & Holzer, P. Standardized concept for the treatment of gastrointestinal dysmotility in critically ill patients--current status and future options. *Clin Nutr* **27**, 25-41 (2008).
- 314 Wang, N. PPAR-delta in Vascular Pathophysiology. *PPAR Res* **2008**, 164163 (2008).
- 315 Higgins, P. J. *et al.* A soluble chimeric complement inhibitory protein that possesses both decay-accelerating and factor I cofactor activities. *J Immunol* **158**, 2872-2881 (1997).
- 316 Ricklin, D. & Lambris, J. D. Complement-targeted therapeutics. *Nat Biotechnol* **25**, 1265-1275 (2007).

- 317 Chen, J. *et al.* Pentacyclic triterpenes. Part 3: Synthesis and biological evaluation of oleanolic acid derivatives as novel inhibitors of glycogen phosphorylase. *Bioorg Med Chem Lett* **16**, 2915-2919 (2006).
- 318 Chan, M. Y. *et al.* Phase 1b randomized study of antidote-controlled modulation of factor IXa activity in patients with stable coronary artery disease. *Circulation* **117**, 2865-2874 (2008).
- 319 Chan, M. Y. *et al.* A randomized, repeat-dose, pharmacodynamic and safety study of an antidote-controlled factor IXa inhibitor. *J Thromb Haemost* **6**, 789-796 (2008).
- 320 Dyke, C. K. *et al.* First-in-human experience of an antidote-controlled anticoagulant using RNA aptamer technology: a phase 1a pharmacodynamic evaluation of a drug-antidote pair for the controlled regulation of factor IXa activity. *Circulation* **114**, 2490-2497 (2006).
- 321 Dementiev, A., Petitou, M., Herbert, J. M. & Gettins, P. G. The ternary complex of antithrombin-anhydrothrombin-heparin reveals the basis of inhibitor specificity. *Nat Struct Mol Biol* **11**, 863-867 (2004).
- 322 Kamiji, M. M. & Inui, A. Neuropeptide y receptor selective ligands in the treatment of obesity. *Endocr Rev* **28**, 664-684 (2007).
- 323 Burley, S. K. Cancer and kinases: reports from the front line. *Genome Biol* **7**, 314 (2006).
- 324 Trowe, T. *et al.* EXEL-7647 inhibits mutant forms of ErbB2 associated with lapatinib resistance and neoplastic transformation. *Clin Cancer Res* **14**, 2465-2475 (2008).
- 325 Gendreau, S. B. *et al.* Inhibition of the T790M gatekeeper mutant of the epidermal growth factor receptor by EXEL-7647. *Clin Cancer Res* **13**, 3713-3723 (2007).
- 326 Antoni, L., Sodha, N., Collins, I. & Garrett, M. D. CHK2 kinase: cancer susceptibility and cancer therapy - two sides of the same coin? *Nat Rev Cancer* **7**, 925-936 (2007).
- 327 Matthews, D. J. *et al.* Pharmacological abrogation of S-phase checkpoint enhances the anti-tumor activity of gemcitabine in vivo. *Cell Cycle* **6**, 104-110 (2007).
- 328 Bucher, N. & Britten, C. D. G2 checkpoint abrogation and checkpoint kinase-1 targeting in the treatment of cancer. *Br J Cancer* **98**, 523-528 (2008).
- 329 Bayes, M. Gateways to clinical trials. *Methods Find Exp Clin Pharmacol* **29**, 153-173 (2007).

- 330 Hajduk, P. J. & Greer, J. A decade of fragment-based drug design: strategic advances and lessons learned. *Nat Rev Drug Discov* **6**, 211-219 (2007).
- 331 Wong, D. & Korz, W. Translating an Antagonist of Chemokine Receptor CXCR4: from bench to bedside. *Clin Cancer Res* **14**, 7975-7980 (2008).
- 332 Lin, T. Y. *et al.* The novel HSP90 inhibitor STA-9090 exhibits activity against Kit-dependent and -independent malignant mast cell tumors. *Exp Hematol* **36**, 1266-1277 (2008).
- 333 Eriksson, B. I. *et al.* Partial factor IXa inhibition with TTP889 for prevention of venous thromboembolism: an exploratory study. *J Thromb Haemost* **6**, 457-463 (2008).
- 334 Howard, E. L., Becker, K. C., Rusconi, C. P. & Becker, R. C. Factor IXa inhibitors as novel anticoagulants. *Arterioscler Thromb Vasc Biol* **27**, 722-727 (2007).
- 335 Tomillero, A. & Moral, M. A. Gateways to clinical trials. *Methods Find Exp Clin Pharmacol* **30**, 383-408 (2008).
- 336 Memoli, M. J., Morens, D. M. & Taubenberger, J. K. Pandemic and seasonal influenza: therapeutic challenges. *Drug Discov Today* **13**, 590-595 (2008).
- 337 Layne, S. P., Monto, A. S. & Taubenberger, J. K. Pandemic influenza: an inconvenient mutation. *Science* **323**, 1560-1561 (2009).
- 338 Doshi, P. Trends in recorded influenza mortality: United States, 1900-2004. *Am J Public Health* **98**, 939-945 (2008).
- 339 Cox, N. J. & Subbarao, K. Global epidemiology of influenza: past and present. *Annu Rev Med* **51**, 407-421 (2000).
- 340 Simonsen, L. The global impact of influenza on morbidity and mortality. *Vaccine* **17 Suppl 1**, S3-10 (1999).
- 341 Smith, G. J. *et al.* Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459**, 1122-1125 (2009).
- 342 Gambotto, A., Barratt-Boyes, S. M., de Jong, M. D., Neumann, G. & Kawaoka, Y. Human infection with highly pathogenic H5N1 influenza virus. *Lancet* **371**, 1464-1475 (2008).
- 343 Beigel, J. & Bray, M. Current and future antiviral therapy of severe seasonal and avian influenza. *Antiviral Res* **78**, 91-102 (2008).

- 344 De Clercq, E. Antiviral agents active against influenza A viruses. *Nat Rev Drug Discov* **5**, 1015-1025 (2006).
- 345 Weinstock, D. M. & Zuccotti, G. The evolution of influenza resistance and treatment. *Jama* **301**, 1066-1069 (2009).
- 346 WHO, R. Influenza A(H1N1) virus resistance to oseltamivir - 2008/2009 influenza season, northern hemisphere. *Official web site of World Health Organization* (2009).
- 347 Dharan, N. J. *et al.* Infections with oseltamivir-resistant influenza A(H1N1) virus in the United States. *Jama* **301**, 1034-1041 (2009).
- 348 Poland, G. A., Jacobson, R. M. & Ovsyannikova, I. G. Influenza virus resistance to antiviral agents: a plea for rational use. *Clin Infect Dis* **48**, 1254-1256 (2009).
- 349 Kawai, N. *et al.* Comparison of the effectiveness of Zanamivir and Oseltamivir against influenza A/H1N1, A/H3N2, and B. *Clin Infect Dis* **48**, 996-997 (2009).
- 350 Hayden, F. Developing new antiviral agents for influenza treatment: what does the future hold? *Clin Infect Dis* **48 Suppl 1**, S3-13 (2009).
- 351 De Clercq, E. & Neyts, J. Avian influenza A (H5N1) infection: targets and strategies for chemotherapeutic intervention. *Trends Pharmacol Sci* **28**, 280-285 (2007).
- 352 Bao, Y. *et al.* The influenza virus resource at the National Center for Biotechnology Information. *J Virol* **82**, 596-601 (2008).
- 353 Russell, R. J. *et al.* Structure of influenza hemagglutinin in complex with an inhibitor of membrane fusion. *Proc Natl Acad Sci U S A* **105**, 17736-17741 (2008).
- 354 Twu, K. Y., Noah, D. L., Rao, P., Kuo, R. L. & Krug, R. M. The CPSF30 binding site on the NS1A protein of influenza A virus is a potential antiviral target. *J Virol* **80**, 3957-3965 (2006).
- 355 Chand, P. *et al.* Comparison of the anti-influenza virus activity of cyclopentane derivatives with oseltamivir and zanamivir in vivo. *Bioorg Med Chem* **13**, 4071-4077 (2005).
- 356 Julander, J. G., Shafer, K., Smee, D. F., Morrey, J. D. & Furuta, Y. Activity of T-705 in a hamster model of yellow fever virus infection in comparison with that of a chemically related compound, T-1106. *Antimicrob Agents Chemother* **53**, 202-209 (2009).

- 357 Stamatiou, G. *et al.* Heterocyclic rimantadine analogues with antiviral activity. *Bioorg Med Chem* **11**, 5485-5492 (2003).
- 358 Minagawa, K. *et al.* Novel stachyflin derivatives from *Stachybotrys* sp. RF-7260. Fermentation, isolation, structure elucidation and biological activities. *J Antibiot (Tokyo)* **55**, 239-248 (2002).
- 359 Gouarin, S. *et al.* Study of influenza C virus infection in France. *J Med Virol* **80**, 1441-1446 (2008).
- 360 Leneva, I. A., Russell, R. J., Boriskin, Y. S. & Hay, A. J. Characteristics of arbidol-resistant mutants of influenza virus: implications for the mechanism of anti-influenza action of arbidol. *Antiviral Res* **81**, 132-140 (2009).
- 361 Liu, M. Y. *et al.* Pharmacokinetic properties and bioequivalence of two formulations of arbidol: an open-label, single-dose, randomized-sequence, two-period crossover study in healthy chinese male volunteers. *Clin Ther* **31**, 784-792 (2009).
- 362 von Itzstein, M. The war against influenza: discovery and development of sialidase inhibitors. *Nat Rev Drug Discov* **6**, 967-974 (2007).
- 363 Colman, P. M., Varghese, J. N. & Laver, W. G. Structure of the catalytic and antigenic sites in influenza virus neuraminidase. *Nature* **303**, 41-44 (1983).
- 364 Russell, R. J. *et al.* The structure of H5N1 avian influenza neuraminidase suggests new opportunities for drug design. *Nature* **443**, 45-49 (2006).
- 365 Collins, P. J. *et al.* Crystal structures of oseltamivir-resistant influenza virus neuraminidase mutants. *Nature* **453**, 1258-1261 (2008).
- 366 Amaro, R. E., Cheng, X., Ivanov, I., Xu, D. & McCammon, J. A. Characterizing loop dynamics and ligand recognition in human- and avian-type influenza neuraminidases via generalized born molecular dynamics and end-point free energy calculations. *J Am Chem Soc* **131**, 4702-4709 (2009).
- 367 Obayashi, E. *et al.* The structural basis for an essential subunit interaction in influenza virus RNA polymerase. *Nature* **454**, 1127-1131 (2008).
- 368 Dias, A. *et al.* The cap-snatching endonuclease of influenza virus polymerase resides in the PA subunit. *Nature* **458**, 914-918 (2009).

- 369 Kuzuhara, T. *et al.* Structural basis of the influenza A virus RNA polymerase PB2 RNA-binding domain containing the pathogenicity-determinant lysine 627 residue. *J Biol Chem* **284**, 6855-6860 (2009).
- 370 Guilligay, D. *et al.* The structural basis for cap binding by influenza virus polymerase subunit PB2. *Nat Struct Mol Biol* **15**, 500-506 (2008).
- 371 Yuan, P. *et al.* Crystal structure of an avian influenza polymerase PA(N) reveals an endonuclease active site. *Nature* **458**, 909-913 (2009).
- 372 Stouffer, A. L. *et al.* Structural basis for the function and inhibition of an influenza virus proton channel. *Nature* **451**, 596-599 (2008).
- 373 Schnell, J. R. & Chou, J. J. Structure and mechanism of the M2 proton channel of influenza A virus. *Nature* **451**, 591-595 (2008).
- 374 Ma, C. *et al.* Identification of the pore-lining residues of the BM2 ion channel protein of influenza B virus. *J Biol Chem* **283**, 15921-15931 (2008).
- 375 Rosenthal, P. B. *et al.* Structure of the haemagglutinin-esterase-fusion glycoprotein of influenza C virus. *Nature* **396**, 92-96 (1998).
- 376 Mayr, J. *et al.* Influenza C virus and bovine coronavirus esterase reveal a similar catalytic mechanism: new insights for drug discovery. *Glycoconj J* **25**, 393-399 (2008).
- 377 Chand, P. *et al.* Syntheses and neuraminidase inhibitory activity of multisubstituted cyclopentane amide derivatives. *J Med Chem* **47**, 1919-1929 (2004).
- 378 Furuta, Y. *et al.* T-705 (favipiravir) and related compounds: Novel broad-spectrum inhibitors of RNA viral infections. *Antiviral Res* **82**, 95-102 (2009).
- 379 Tomassini, J. *et al.* Inhibition of cap (m7GpppXm)-dependent endonuclease of influenza virus by 4-substituted 2,4-dioxobutanoic acid compounds. *Antimicrob Agents Chemother* **38**, 2827-2837 (1994).
- 380 Nakazawa, M. *et al.* PA subunit of RNA polymerase as a promising target for anti-influenza virus agents. *Antiviral Res* **78**, 194-201 (2008).
- 381 Zoidis, G. *et al.* Are the 2-isomers of the drug rimantadine active anti-influenza A agents? *Antivir Chem Chemother* **14**, 153-164 (2003).

- 382 Plotch, S. J. *et al.* Inhibition of influenza A virus replication by compounds interfering with the fusogenic function of the viral hemagglutinin. *J Virol* **73**, 140-151 (1999).
- 383 Deshpande, M. S. *et al.* An approach to the identification of potent inhibitors of influenza virus fusion using parallel synthesis methodology. *Bioorg Med Chem Lett* **11**, 2393-2396 (2001).
- 384 Deng, H. Y. *et al.* Efficacy of arbidol on lethal hantaan virus infections in suckling mice and in vitro. *Acta Pharmacol Sin* **30**, 1015-1024 (2009).
- 385 Das, K. *et al.* Structural basis for suppression of a host antiviral response by influenza A virus. *Proc Natl Acad Sci U S A* **105**, 13093-13098 (2008).
- 386 Ong, S. A., Lin, H. H., Chen, Y. Z., Li, Z. R. & Cao, Z. Efficacy of different protein descriptors in predicting protein functional families. *BMC Bioinformatics* **8**, 300 (2007).