國立交通大學

生物資訊及系統生物研究所

博士論文

蛋白質-配體結合模式預測與其結合區域定性研究

A study for predicting protein-ligand binding modes and characterizing protein-ligand binding sites in structure-based drug design

研 究 生：陳彥甫

指導教授：楊進木　教授

中華民國九十九年九月

蛋白質-配體結合模式預測與其結合區域定性研究

A study for predicting protein-ligand binding modes and characterizing protein-ligand binding sites in structure-based drug design

研究生：陳彥甫　　Student：Yen-Fu Chen

指導教授：楊進木　　Advisor：Jinn-Moon Yang

國立交通大學

生物資訊及系統生物研究所

博士論文

A Thesis Submitted to Institute of Bioinformatics and Systems Biology
National Chiao Tung University in partial Fulfillment of the Requirements
for the Degree of Ph.D. in
Bioinformatics and Systems Biology

September 2010

Hsinchu, Taiwan, Republic of China

中華民國九十九年九月

蛋白質-配體結合模式預測與其結合區域定性研究

研究生：陳彥甫　　　　　　　　　　　　　　　　　　　指導教授：楊進木博士

國立交通大學 生物資訊及系統生物研究所 博士班

摘　　要

隨著蛋白質結晶結構的快速增加，以結構為基礎之藥物設計與虛擬藥物篩選(virtual screening)在先導藥物開發過程日漸重要。目前一系列的分子對接(protein-ligand docking)以及虛擬藥物篩選方法已經被應用到先導藥物發展中，並且已獲得數個成功的藥物開發案例。即使如此，目前由巨量的虛擬藥物篩選資料中找出真正具有活性的先導藥物仍然是一個困難的挑戰。其問題肇因於目前對蛋白質-配體之間的結合機制了解仍然有所不足，使得已發展的蛋白質-配體結合計分方程式不夠周全。

針對上述議題，我們已提出了以藥物孔洞為基礎的計分方程式(pharmacophore-based scoring function)與虛擬藥物篩選共通計分方法應用準則(consensus scoring criteria)之研究。其中，共通計分方法是透過結合數個計分方程式的共同處，相較於單一計分方法可以有更好的虛擬藥物篩選準確性。然而虛擬藥物篩選的計分方程式通常無法辨識蛋白質-配體間的關鍵結合特性[例如: 藥效基團熱點(pharmacophore hotspot)]，而這些關鍵特性卻通常是觸發或抑制目標蛋白質對其調控的生物反應必要條件。雖然應用藥物孔洞方法與相關計分方程式可以找出關鍵結合特性，但是這些方法需要一系列已知的活性配體，這些資料必須由實驗取得，使應用性受到限制。因此，對於虛擬藥物篩選過程發展更好的篩選後分析(post-screening analysis)與關鍵特性之發現方法，將對於藥物發展具有重要價值。

在本研究中，我們已經發展出 site-moiety map (簡稱 SiMMap)方法，並且將其延伸應用到辨識與定性垂直同源蛋白質(ortholog)的共通結合環境 (orthologous SiMMap)研究之中。SiMMap 透過統計對目標蛋白質與一群對其預測或共結晶之配體所產生的交互作用，推測位於目標蛋白質結合區域內之錨點(anchor)，並用以描述分布在結合區域中的配體官能基偏好(moiety preference)以及物化特性集合。每一個錨點具有三個基本構成要件：1)由具一致交互作用之殘基構成的結合袋點(binding pocket)；2)複數個虛擬配體構成的官能基組成；3)結合袋點與官能基之交互作用關係(包含靜電力、氫鍵及凡德瓦力交互作用)。實驗證據已顯示錨點通常是蛋白質-配體結合區域中的熱點。同時 site-moiety map 也可提供將官能基團(靜電力、氫鍵及凡德瓦力特性之官能基)之組合最佳化的建議，有助於設計潛在先導藥物。實驗結果也證實當小分子化合物與 site-moiety map 描述的錨點特性高度相符時，通常有高度潛力成為目標蛋白質的抑制劑或促進劑。SiMMap 已提供全球服務，網址為 http://simfam.life.nctu.edu.tw/。我們相信我們對於藥物孔洞為基之計分方程式、虛擬藥物篩選之共通計分方法應用準則的成果、以及 site-moiety map 之研究，將對藥物發現與了解蛋白質-配體機制有所幫助。

A study for predicting protein-ligand binding modes and characterizing protein-ligand binding sites in structure-based drug design

Student : Yen-Fu Chen Adviser : Dr. Jinn-Moon Yang

Institute of Bioinformatics and Systems Biology

National Chiao Tung University

ABSTRACT

As the number of protein structures increases rapidly, structure-based drug design and virtual screening approaches are becoming important and helpful in lead discovery. A number of docking and virtual screening (VS) methods have been utilized to identify lead compounds, and some success stories have been reported. However, identifying lead compounds by exploiting thousands of docked protein-compound complexes is still a challenging task. The major weakness of virtual screenings is likely due to incomplete understandings of ligand binding mechanisms and the subsequently imprecise scoring algorithms.

To address these issues, we have proposed a pharmcophore-based scoring function approach and a consensus strategies among different scoring methods in VS. The consensus scores would improve the performance and, on average, the performance of the combined method performs better than the average of the individual scoring functions. Nevertheless, the approaches generally cannot identify the key features (e.g., pharmacophore spots) that are essential to trigger or block the biological responses of the target protein. Although pharmacophore techniques have been applied to derive the key features, these methods require a set of known active ligands that were acquired experimentally. Therefore, the more powerful techniques for post-screening analysis to identify the key features through docked compounds and to characterize the binding site provide a great potential value for drug design.

Recently, we have developed the site-moiety map (SiMMap) method and extended to characterize the consensus binding environments (*i.e.*, anchors) of orthologous targets (orthSiMMap). SiMMap statistically derived anchors from the interaction profiles between query target protein and its docked (or co-crystallized) compounds, and then described the relationship between the moiety preferences and physico-chemical properties of the binding site. Each anchor includes three basic elements: a binding pocket with conserved interacting residues, the moiety composition of query compounds, and pocket-moiety interaction type (electrostatic, hydrogen-bonding, or van der Waals). Experimental results showed that an anchor is often a hot spot and the site-moiety map can be helpful to assemble potential leads by optimal steric, hydrogen-bonding, and electronic moieties. When a compound highly agrees with anchors of site-moiety map, this compound often activates or inhibits the target protein. The SiMMap web server is available at http://simfam.life.nctu.edu.tw/. We believe that our evolutionary approach with pharmacophore-based scoring functions, consensus scoring criteria for virtual screening, and the method of site-moiety map are useful for drug discovery and understanding biological mechanisms.

Acknowledgement

來到新竹交大十二個年頭，從一個懵懂的大學生一路念到了博士學位。回首這些生命中的精彩日子，每一個畫面都充滿了無數的回憶。在這段學習的黃金時代，難免遇到瓶頸與困難，感謝許多貴人的幫助讓我得以安然度過，並且學到如何面對壓力與思考問題的能力。我認為這些能力將是未來面對下一階段的人生路程之重要基礎與助力。

第一個貴人是我的指導教授楊進木博士。在這段學習的過程中，我的指導教授總是嚴格且不厭其煩的指出我的缺點與問題，並提出可能的方向讓我依循。雖然克服困難的過程中，難免跌跌撞撞，但是這些解決問題的經驗、創新思考與邏輯歸納能力將是我未來人生中的寶藏。

接著我要特別感謝的是實驗室同組的凱程、伸融與御哲，由於有了他們的配合與日以繼夜的努力，我們才能在最短的時間之內突破困難的關卡，完成這本論文中的許多重點工作。除此之外，指導教授、凱程與我之間的討論與合作過程，也是突破這些困難的另外一個重要關鍵。許多的深入探討與解決困難的經驗，都是在這個時候得到啟發與進展，並且讓我獲益良多。

口試期間承蒙所內黃鎮剛教授、趙瑞益教授與梁美智教授，及所外王雯靜教授、徐祖安教授與謝興邦教授百忙之中不吝賜教並且提供許多寶貴建言，使得本文得以更加完善。同時我也從口試委員以及指導教授身上看到頂尖學術研究人才所應具備的條件以及視野，這將是我未來需要繼續努力的目標。

在本文的理論發展過程中，我的指導教授提供許多與國內外頂尖實驗室合作的機會，使得我有機會可以與跨領域的研究人才合作與交流，這樣的交流過程中經常會激發出許多新的想法與題材。同時在本中有引用到與這些頂尖實驗室(美國 Fordham 大學 D. Frank Hsu 教授、清大王雯靜教授、國衛院徐祖安教授與交大楊昀良教授)在合作過程中所完成與提供寶貴實驗資料，特此感謝。

在這段辛苦而充實的研究生活中，實驗室的同學們也不時的給予我許多生活上的幫助或研究討論的空間，像是已經畢業的再威、右儒與振寧，或是現在的宇書、章維、怡馨、一原、志達...等許多同學，在這裡也要感謝他們。除此之外在論文的撰寫過程中，感謝已經畢業的陳俊辰博士給予我許多的撰寫方面的建議，使我事半功倍。

最後我必須要特別感謝我的父母與家人。在許多困難的時候，他們總是默默的支持著我，無怨無悔的付出。每當我覺得筋疲力盡的時候，我知道總是有一盞溫暖的燈火等著我回家，隨時幫我加滿重新面對挑戰的能量。因此我將此論文獻給我的父母，希望他們可以以我為榮。

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Virtual screening (VS) of molecular compound libraries has emerged as a powerful and inexpensive method for the discovery of novel lead compounds for drug development [1-10]. Given the structure of a target protein active site and a potential small ligand database, VS predicts the binding mode and the binding affinity for each ligand and ranks a series of candidate ligands. There are four main reasons for the rapid acceptance and success of VS: 1) The availability of the growing number of protein crystal structures; 2) The advent of structural proteomics technologies; 3) The enrichment and speed of VS [2,11]; and 4) The contribution of VS to the reduction in the cost of drug discovery.

Each VS computational method involves two basic critical elements: efficient molecular docking and a reliable scoring method. Scoring methods for VS should effectively discriminate between correct binding states and non-native docked conformations during the molecular docking phase and distinguish a small number of active compounds from hundreds of thousands of non-active compounds during the post-docking analysis. There are three general classes of scoring functions that calculate the binding free energy, including knowledge-based [12], physics-based [13], and empirical-based [14] scoring functions.

However, the performance of these scoring functions is often inconsistent across different systems from a database search. The inaccuracy of the scoring methods, *i.e.*, inadequately predicting the true binding affinity of a ligand for a receptor, is probably the major weakness for VS. It has been reported that fusion among different scoring methods in VS would improve the performance and, on average, the performance of the combined method performs better than the average of the individual scoring functions. More recently, the same phenomena has been previously reported in information retrieval (IR) and in molecular similarity measurement. Charifson *et al.* (1999)[15], presented a study in which they used an intersection-based consensus approach to combine scoring functions. The evidences showed an enrichment in the ability to discriminate between active and inactive enzyme inhibitions for three different enzymes (p38 MAP kinase, inosine monophosphate dehydrogenase, and HIV protease) using two different docking methods (DOCK[16] and GAMBLER) and thirteen scoring functions. Then, Bissantz *et al*. (2000)[17] , Stahl and Rarey (2001)[18], and Verdonk *et al.* (2004)[19] *et.al*., also reported their works for consensus scores (CS) improving VS. Wang and Wang (2004)[20] presented an idealized computer experiment to explore how consensus scoring works based on the assumption that the error of a scoring function is a random number in a normal distribution. They also studied the relationship between the hit-rates and the number of scoring functions and the performance of several ranking strategies (the rank-by-score, the rank-by-rank, and the rank-by-vote strategy) for consensus scorings.

These reported results seem to depend on the method of combination (by rank, by score, by intersection, by MIN, by MAX, and by voting) and the number and nature of individual scoring functions involved in the combination. While researchers focus to realize the benefit of method combination and consensus scorings, the major issues of how and when these individual scoring functions should be combined remain a challenging problem not only for researchers but also perhaps more importantly, for practitioners in virtual screening.

In addition, some of these VS methods are capable of identifying so-called "pharmacological preference" that is often the important interactions or binding-site hot spots typically evolved from known active ligands and the target protein [21-22]. These preferences might improve screening accuracy and guide the design and selection of lead compounds for subsequent investigation and refinement during lead discovery and lead optimization processes. However, identifying lead compounds by exploiting thousands of docked protein-compound complexes is still a challenging task, too. The major weakness of virtual screenings is likely due to incomplete understandings of ligand binding mechanisms and the subsequently imprecise scoring algorithms [2,6,9].

Most of docking programs[16,23-24] use energy-based scoring methods which are often biased toward both the selection of high molecular weight compounds and charged polar compounds. These approaches[25-26] generally cannot identify the key features (*e.g.,* pharmacophore spots) that are essential to the biological responses of the target protein. Although pharmacophore techniques[27] have been applied to derive the key features, these methods are restricted by a set of known active ligands that were acquired experimentally. Therefore, the more powerful techniques for post-screening analysis to identify the key features through docked compounds and to understand the binding mechanisms provide a great potential value for drug design.

1.2 Thesis overview

For addressing above issues, some studies have been reported (Fig. 1.1). Three of our related studies were briefly described in Chapter 2. The study of the pharmacophore-based scoring function proposed a target-specific scoring function by utilizing the protein-ligand interactions and physic-chemical properties of known actives to improve the accuracy and precision for the ranking of VS data (Fig. 1.1a). The studies of consensus scoring and cluster analysis addressed the issues of improving enrichment for the post-screening analysis stage (Fig. 1.1b). Furthermore, we also applied these methods on the inhibitor discoveries of the

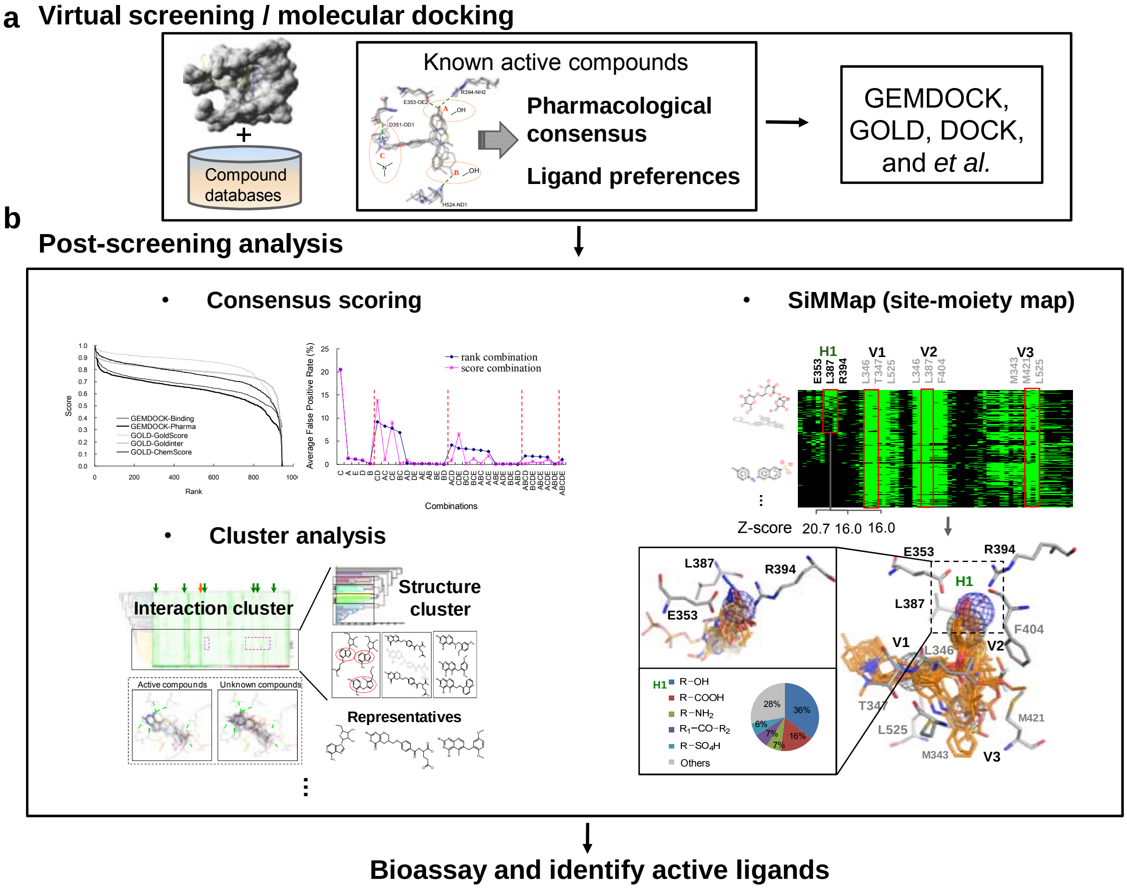


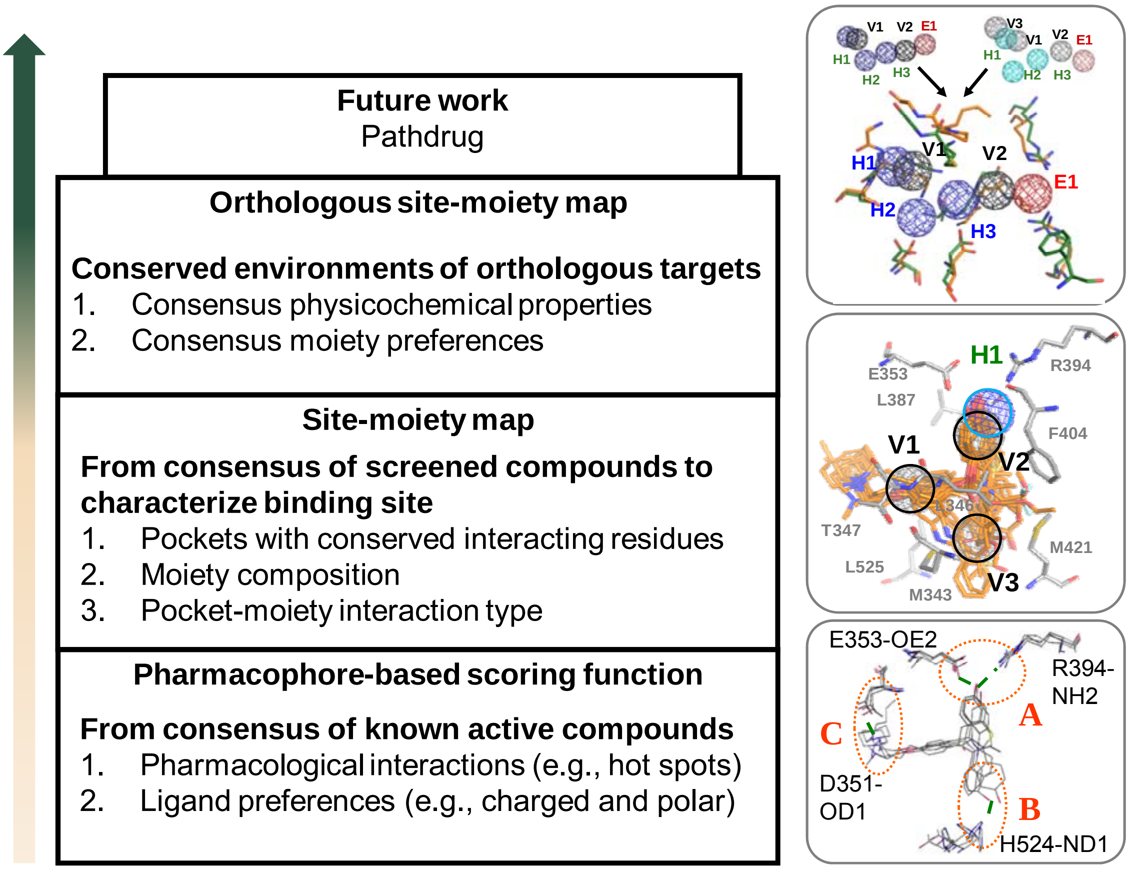
Figure 1.1. Overview of structure-based drug design and related works. The major steps of structure-based drug design include (a) virtual screening and (b) post-screening analysis and following bioassay.

dengue virus E protein and the influenza virus neuraminidase. Although some of novel inhibitors were discovered in these researches, we still found the drawbacks of these previous studies. Firstly, the pharmacophore-based scoring function is limited by the consensus of known active compounds. Second, the consensus scoring criteria and cluster analysis are helpful for improving the enrichment of VS, but these methods does not use the protein-ligand interaction data and ligand structures produced in the VS process for investigating the key environment of the protein-ligand binding site.

To address these issues, we developed the SiMMap approach to infer the key features by a site-moiety map describing the relationship between the moiety preferences and the physico-chemical properties of the binding site in Chapter 3 (Fig. 1.1b). The further application and validation of SiMMap was presented in the Chapter 4. According to our knowledge, SiMMap is the first public server that identifies the site-moiety map from a query protein structure and its docked (or co-crystallized) compounds. The server characterizes a binding site by pocket-moiety interaction preferences (anchors) including binding pockets with conserved interacting residues, moiety preferences, and interaction type.

In Chapter 4, we further extended SiMMap to orthologous SiMMap. We derived the orthologous site-moiety maps (orthologous SiMMap) from identifying consensus binding environments of orthologous proteins; orthologous SiMMap represents the conserved binding environment or "hot spots" among orthologous targets in an aim to investigate the protein-ligand interface family and apply for discovering potential leads across multiple species. Finally, Chapter 5 described some conclusions and future perspectives.

The research framework of this thesis is shown as Figure 1.2. The concept of the research of pharmacophore-based scoring function is that utilizing the consensus of known active compounds identifies the key feature of binding site. However, such approach needs the known active compounds and prefers the compounds similar with the known set. To address these limitations, we extract the consensus of screening compounds to characterize the binding site and further validate on the inhibitor discovery of orthologous shikimate kinases.



Future work
Pathdrug

Orthologous site-moiety map

Conserved environments of orthologous targets
1. Consensus physicochemical properties
2. Consensus moiety preferences

Site-moiety map

From consensus of screened compounds to characterize binding site
1. Pockets with conserved interacting residues
2. Moiety composition
3. Pocket-moiety interaction type

Pharmacophore-based scoring function

From consensus of known active compounds
1. Pharmacological interactions (e.g., hot spots)
2. Ligand preferences (e.g., charged and polar)

Figure 1.2. The research framework for predicting protein-ligand binding modes and characterizing protein-ligand binding sites in structure-based drug design.

Chapter 2

Related works

Virtual screening (VS) of molecular compound libraries has emerged as a powerful and inexpensive method for the discovery of novel lead compounds for drug development [2-3] (Fig. 2.1). The VS computational method involves two basic critical elements: efficient molecular docking and a reliable scoring method. Scoring methods for VS should effectively discriminate between correct binding states and non-native docked conformations during the molecular docking phase and distinguish a small number of active compounds from hundreds of thousands of non-active compounds during the post-docking analysis. The scoring functions that calculate the binding free energy mainly include knowledge-based[12], physics-based[13], and empirical-based [14] scoring functions.

In addition, some of these VS methods are capable of identifying so-called "pharmacological preference" that is often the important interactions or binding-site hot spots typically evolved from known active ligands and the target protein[21-22] (Fig. 2.1b). These preferences might improve screening accuracy and guide the design and selection of lead compounds for subsequent investigation and refinement during lead discovery and lead optimization processes. However, the pharmacological preferences for each protein target and corresponded ligands are limited by the demand of pre-studied bioassays or structure data.

Currently, the screening quality of docking methods using energy-based scoring functions alone is often influenced by the molecular weight and the structure of the ligand being screened (*e.g.*, the numbers of charged and polar atoms) (Fig. 2.2). These methods are often biased toward both the selection of high molecular weight compounds (due to the contribution of the compound size [28-29]) and charged polar compounds (due to the pair-atom potentials of the electrostatic energy and hydrogen-bonding energy).

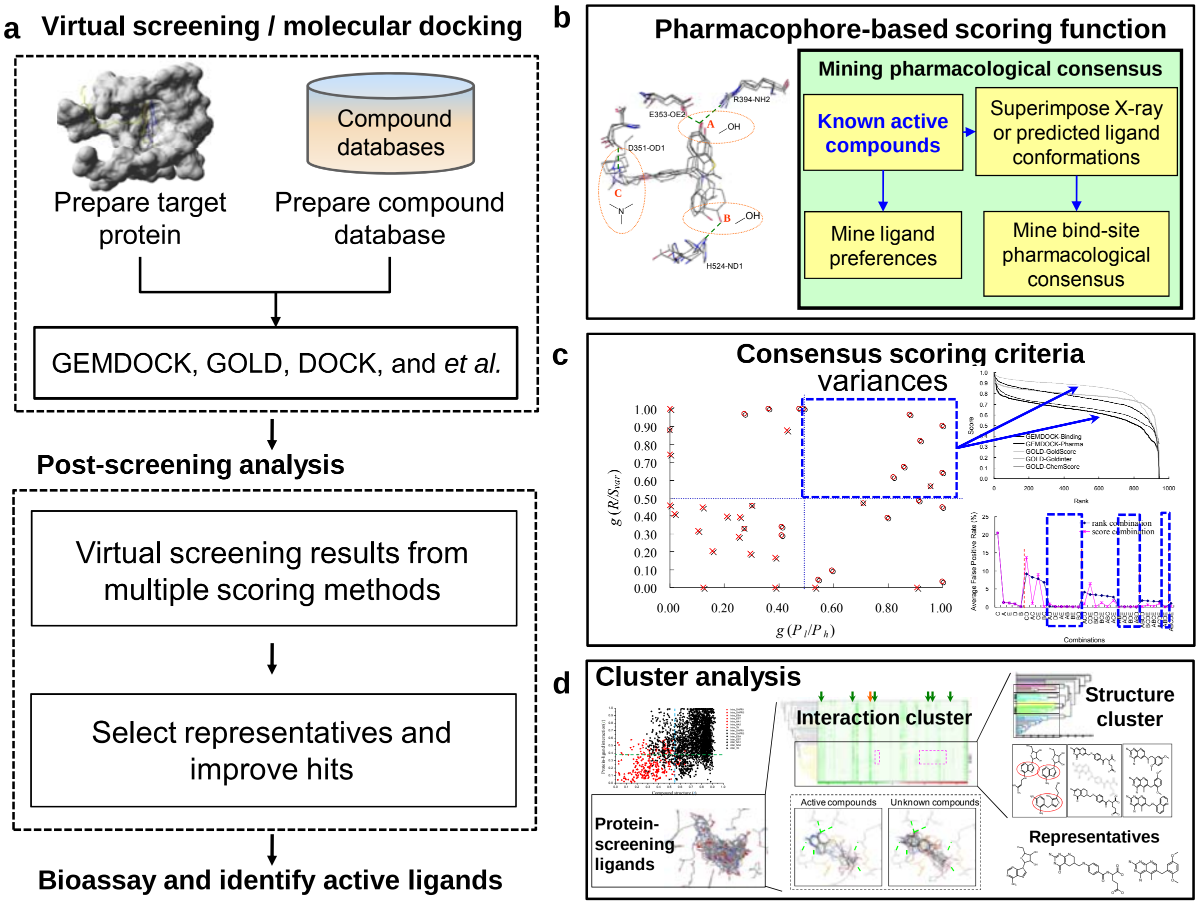

a Virtual screening / molecular docking
Compound databases
Prepare target protein
Prepare compound database
GEMDOCK, GOLD, DOCK, and et al.
Post-screening analysis
Virtual screening results from multiple scoring methods
Select representatives and improve hits
Bioassay and identify active ligands
b
Pharmacophore-based scoring function
Mining pharmacological consensus
Known active compounds
Superimpose X-ray or predicted ligand conformations
Mine ligand preferences
Mine bind-site pharmacological consensus
R394-NH2
E353-OE2
D351-OD1
H524-ND1
A
B
C
OH
c
Consensus scoring criteria
variances
g (R/S_var)
g (P_l/P_h)
Score
Rank
GEMDOCK-Binding
GEMDOCK-Pharma
GOLD-GoldScore
GOLD-Goldinter
GOLD-ChemScore
Average False Positive Rate (%)
rank combination
score combination
Combinations
d
Cluster analysis
Interaction cluster
Structure cluster
Protein-screening ligands
Active compounds
Unknown compounds
Representatives

Figure 2.1. Main procedure of structure-based virtual screening. (a) The major steps of structure-based virtual screening, including virtual screening, post-screening analysis, and bioassay. (b) Pharmacophore-based scoring function for virtual screening step. Post-screening analysis step is usually utilized for improving including (c) consensus scoring and (d) cluster analysis.

In the meanwhile, the performance of these scoring functions is often inconsistent across different systems from a database search [17-18]. The inaccuracy of the scoring methods, i.e., inadequately predicting the true binding affinity of a ligand for a receptor, is probably the major weakness for VS. Furthermore, the application of VS[2,30], to the drug discovery process invariably produces a large number of potential lead candidates. These prospective ligands need to be filtered in order to reduce their number for more precise and labor-intensive studies. Hence, it is urgent that the utilizations of post-analysis to minimize the number of false positives in the selection list and to propagate the true hits to the top of the list. (Fig. 2.1a, 2.1c and 2.1d)

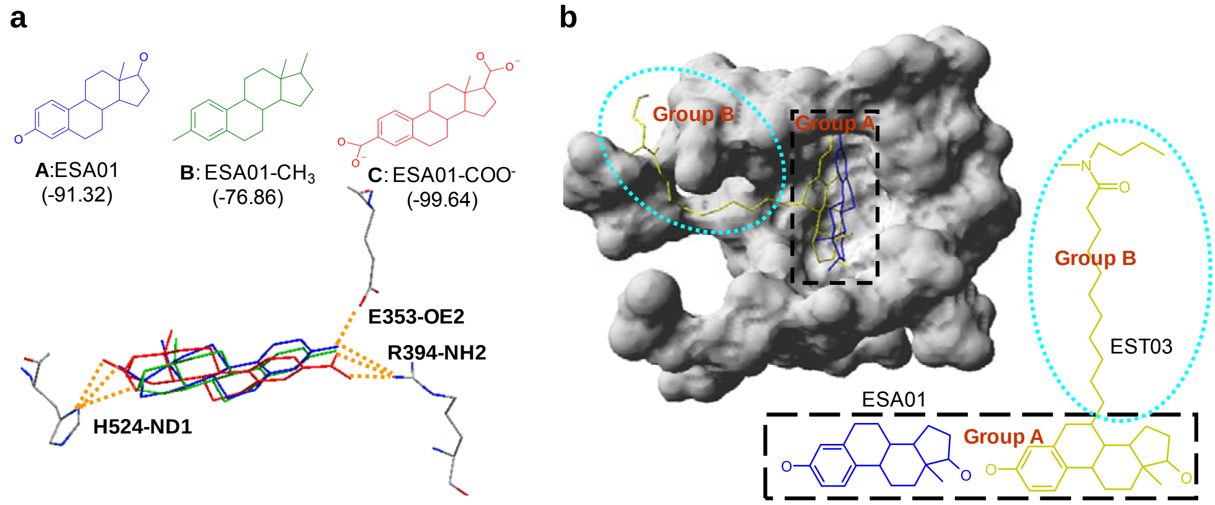


Figure 2.2. The influences of ligand structures and molecular weight on docking energy. (a) The fraction of polar atoms in ESA01-C is the smallest among these 3 ligands, whereas that of ESA01-COO is the largest. The docked positions are similar, but the docking energies differ: -91.32 for ESA01, -76.86 for ESA01-CH_3, and -99.64 for ESA01-COO. (b) ESA01 (blue) and EST03 (yellow) have a common group A, and EST03 has an additional substructure group B. The docked conformations (into reference protein 3ert) are similar, and the docking energies are -82.82 for ESA01 and -127.27 for EST03.

It has been reported that fusion among different scoring methods in VS would improve the performance and, on average, the performance of the combined method performs better than the average of the individual scoring functions.[15,18-20,31] These reported results are significant and potentially robust in that the performance results of these consensus scoring (CS) methods seem to be independent of the target receptor and the docking algorithm. The reported results seem to depend on the method of combination (by rank, by score, by intersection, by MIN, by MAX, and by voting) and the number and nature of individual scoring functions involved in the combination. While researchers have come to realize the advantage and benefit of method combination and consensus scorings, the major issues of how and when these individual scoring functions should be combined remain a challenging problem not only for researchers but also perhaps more importantly, for practitioners in virtual screening.

Another frequently used technique for post-screening analysis is cluster analysis. Clustering methods based on compound structural similarity or interacting profiles can group VS data, reduce complexity of observation, and improve the performance of the scoring function[32-34]. Through the cluster analysis, the enormous data produced by VS process is able to easily visualize and efficiently handle. However, most of researchers only consider the descriptors of protein-ligand interactions or compound structures individually. The combination of protein-ligand interactions and compound topology could provide more detail and pure classifications for following biological assay and refinement. Therefore, some of related studies are briefly introduced as following (Fig. 2.1b and 2.2d).

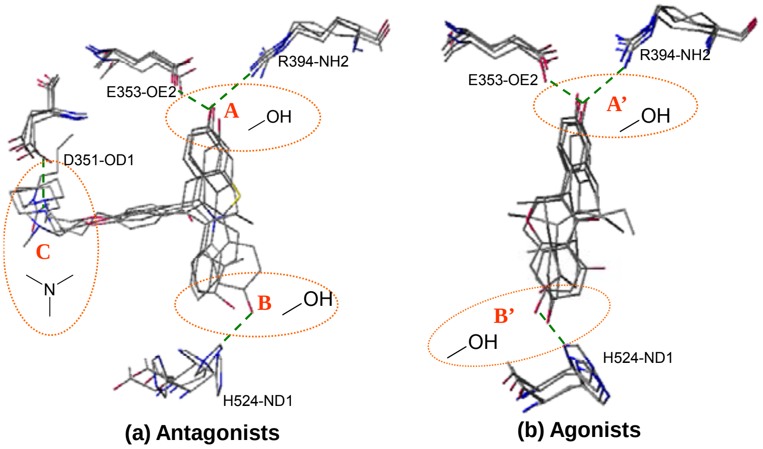


Figure 2.3. The binding-site pharmacological consensuses are identified by overlapping the docked conformations of (a) 10 known ER antagonists and (b) 10 known ER agonists against the reference proteins 3ert and 1gwr, respectively. (a) Four pharmacological interactions were identified and circled as A (phenolic hydroxyl group), B (phenolic hydroxyl group), and C (piperidine nitrogen). (b) Three pharmacological interactions were identified and circled as A (phenolic hydroxyl group) and B (phenolic hydroxyl group). The dashed lines indicate the hydrogen bonds formed between the ligand and the target protein. These pharmacological interactions are consistent with those evolved from X-ray structures.

2.1 Pharmacophore-based scoring functions

The screening quality of docking methods using energy-based scoring functions alone is often influenced by the molecular weight and the structure of the ligand being screened (*e.g.*, the numbers of charged and polar atoms). These methods are often biased toward both the selection of high molecular weight compounds (due to the contribution of the compound size [28-29]) and charged polar compounds (due to the pair-atom potentials of the electrostatic energy and hydrogen-bonding energy).

A pharmacophore-based evolutionary approach for virtual screening was developed to address these issues. This tool, termed the Generic Evolutionary Method for molecular DOCKing (GEMDOCK), combines an evolutionary approach[23,35-37] with a new pharmacophore-based scoring function. The former integrates discrete and continuous global search strategies with local search strategies to expedite convergence. The latter, integrating an empirical-based energy function and pharmacological preferences (binding-site pharmacological interactions and ligand preferences shown as Fig. 2.3), simultaneously serves as the scoring function for both molecular docking and post-docking analyses to improve screening accuracy (Fig. 2.4). We apply pharmacological-interaction preferences to select the ligands that form pharmacological interactions with target proteins, and use the ligand preferences to eliminate the ligands that violate the electrostatic or hydrophilic constraints. We assessed the accuracy of our approach using human estrogen receptor (ER) and a ligand database from the comparative studies of Bissantz et al.[17] Using GEMDOCK, the average goodness-of-hit (GH) score was 0.83 and the average false positive rate was 0.13% for ER antagonists, and the average GH score was 0.48 and the average false positive rate was 0.75% for ER agonists. The performance of GEMDOCK was superior to competing methods such as GOLD and DOCK. We found that our pharmacophore-based scoring function indeed is able to reduce the number of false positives; moreover, the resulting pharmacological interactions at the binding site as well as ligand preferences are important for assigning confidence to the results of virtual screening experiments. These results suggest that GEMDOCK constitutes a robust tool for virtual database screening.

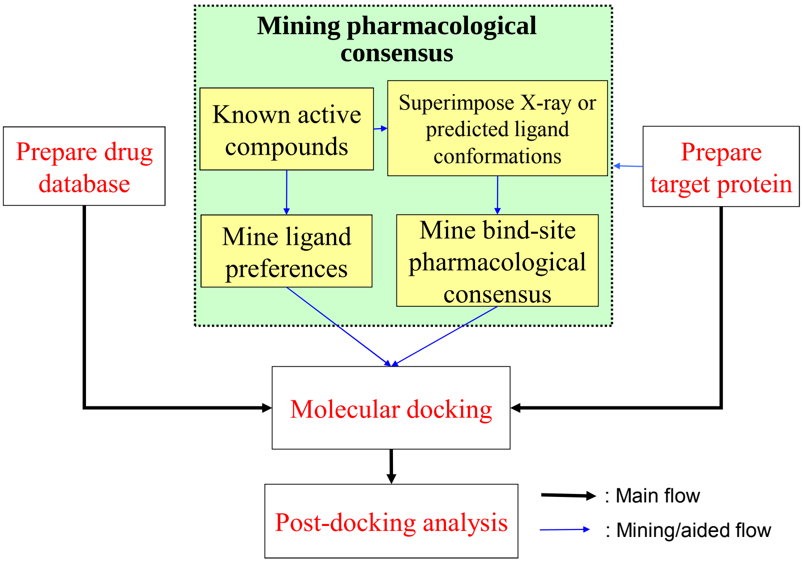


Figure 2.4. The main steps of GEMDOCK for virtual database screening, including the target protein and compound database preparation, flexible docking, and post-docking analysis. GEMDOCK mines a pharmacological consensus from the target protein and known active ligands when available.

2.2 Consensus scoring criteria

The performance of these scoring functions is often inconsistent across different systems from a database search [18,31]. The inaccuracy of the scoring methods, i.e., inadequately predicting the true binding affinity of a ligand for a receptor, is probably the major weakness for VS. It has been demonstrated that combining multiple scoring functions (consensus scoring) improves enrichment of true positives. Previous efforts at consensus scoring have largely focused on empirical results, but they are yet to provide theoretical analysis that gives insight into real features of combinations and data fusion for VS.

We explore consensus scoring (CS) criteria and provide a consensus scoring procedure for improving the enrichment in VS using data fusion and exploring diversity on scoring characteristics between individual scoring functions (Fig. 2.5). In particular, we demonstrate that combining multiple scoring functions improves enrichment of true positives only if (a) each of the individual scoring functions has relatively high performance, and (b) the scoring characteristics of each of the individual scoring functions are quite different (Fig. 2.6). These two prediction variables are also indicative criteria for the performance between rank combination and score combination. Moreover our second criterion (b) using the rank/score characteristics as the scoring diversity is independent of the performance of the individual scoring function. It is therefore very useful in practical settings in the VS process when the performance of an individual scoring function (such as in criterion (a)) is not known or cannot be evaluated at the juncture. We have developed a novel CS system, available online http://gemdock.life.nctu.edu.tw/dock/download.php, which was tested for five scoring systems with two evolutionary docking algorithms on four targets, thymidine kinase (TK), human dihydrofolate reductase (DHFR), and estrogen receptors (ER) of antagonists and agonists (Fig. 2.7). Our procedure is computationally efficient, able to adapt to different situations, and scalable to a large number of compounds and to a greater number of combinations. Results of the experiment show a fairly significant improvement on the goodness-of-hit (GH) scores, false positive (FP) rate, and enrichment factors over average individual performance. This approach has practical utility for cases where the basic tools are known or believed to be generally applicable, but where specific training sets are absent.

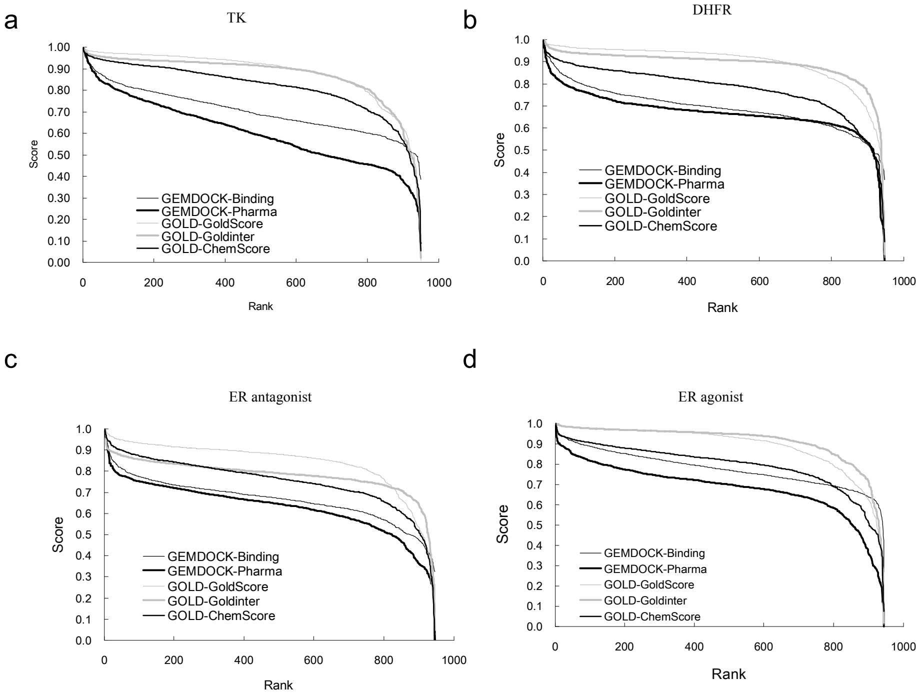


Figure 2.5. Rank/score curves of five methods for four virtual screening targets: (a) TK, (b) DHFR, (c) ER-antagonist receptor, and (d) ER-agonist receptor.

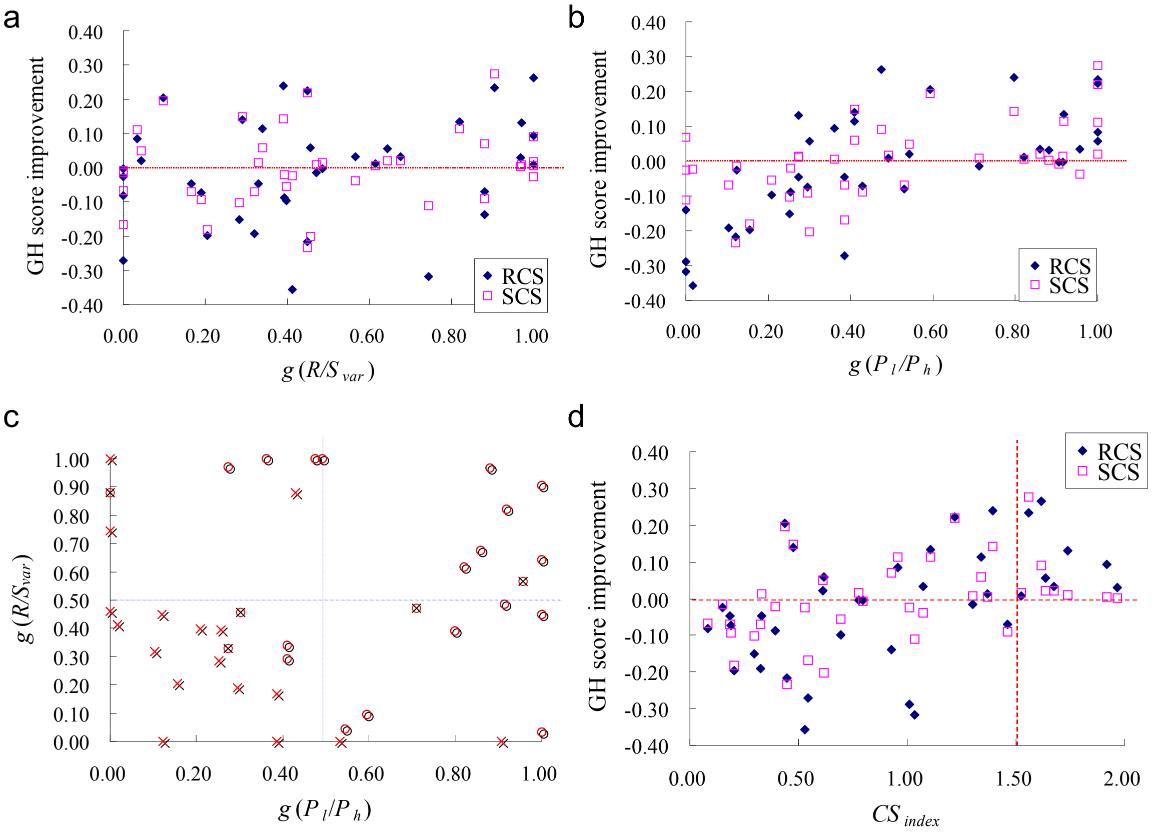


Figure 2.6. The relationships between the GH-score improvement with (a) normalized value of variance of rank/score graph and (b) normalized value of P_l/ P_h of 40 pairing combinations of five methods for four virtual screening targets. (c) The GH-score improvements with normalized variances of rank/score graphs (R/S_{var}) and normalized relative performance measurement (P_l/ P_h) of 40 RCS and SCS pairing combinations of five methods for four virtual screening targets. (d) The positive and negative GH-score improvements are denoted with circle and cross, respectively.

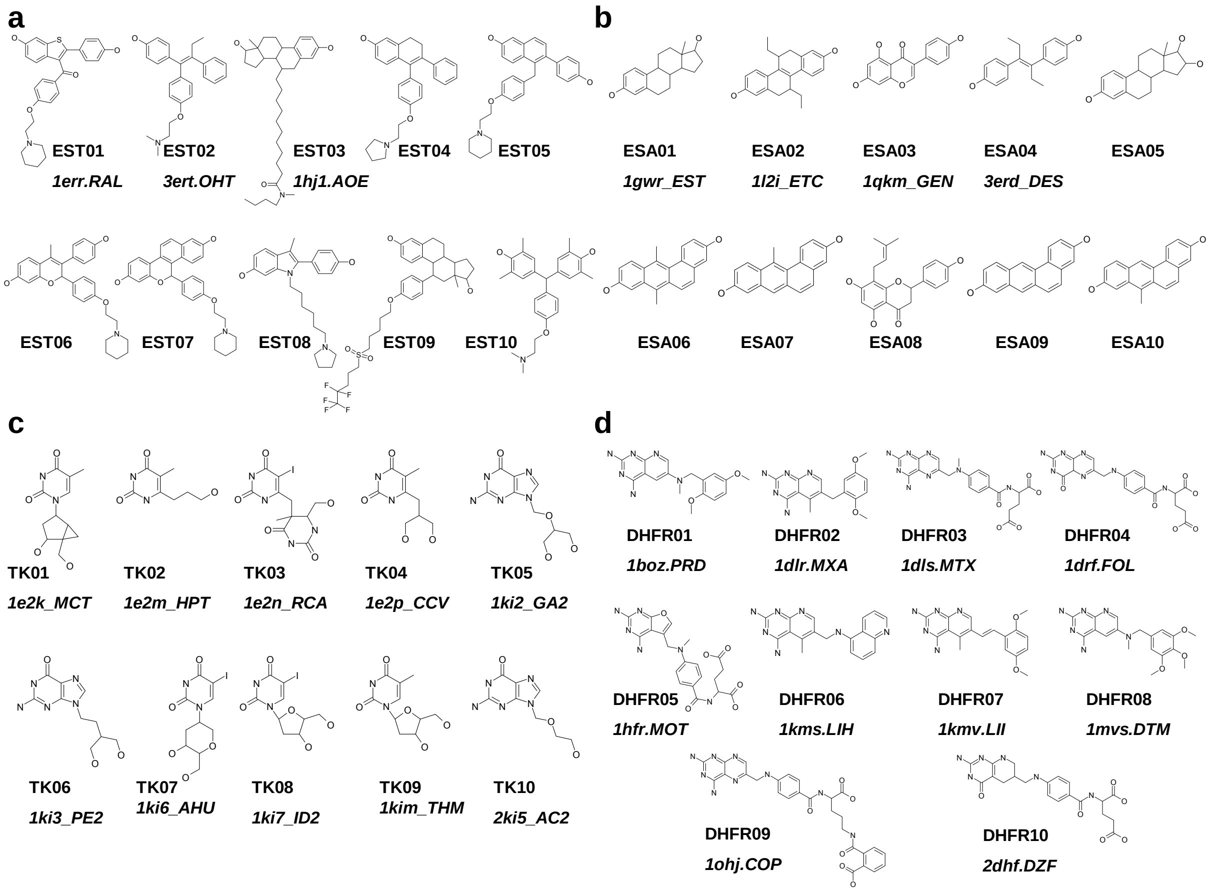


Figure 2.7. The known active ligands of four VS targets, estrogen receptors (ER) of antagonists (a) and agonists (b), (c) thymidine kinase (TK), and (d) human dihydrofolate reductase (DHFR). The ligand data set from the comparative studies of Bissantz *et al.* [17]was used to evaluate the screening accuracy of different CS on TK, DHFR, ER, and ERA. For each target protein, the ligand database included 10 known active compounds and 990 random compounds.

2.3 Combinative clustering analysis

The increasing numbers of 3D compounds and protein complexes stored in databases contribute greatly to current advances in biotechnology, being employed in all kinds of pharmaceutical and industrial applications. However, screening and retrieving appropriate candidates as well as handling false positives presents a challenge for all post-screening analysis methods employed in retrieving therapeutic and industrial targets.

Using the combinative clustering method (Fig. 2.8), virtually screened compounds were clustered based on their protein-ligand interactions then structure clustering employing physical-chemical features was done to retrieve the final compounds. Based on the protein-

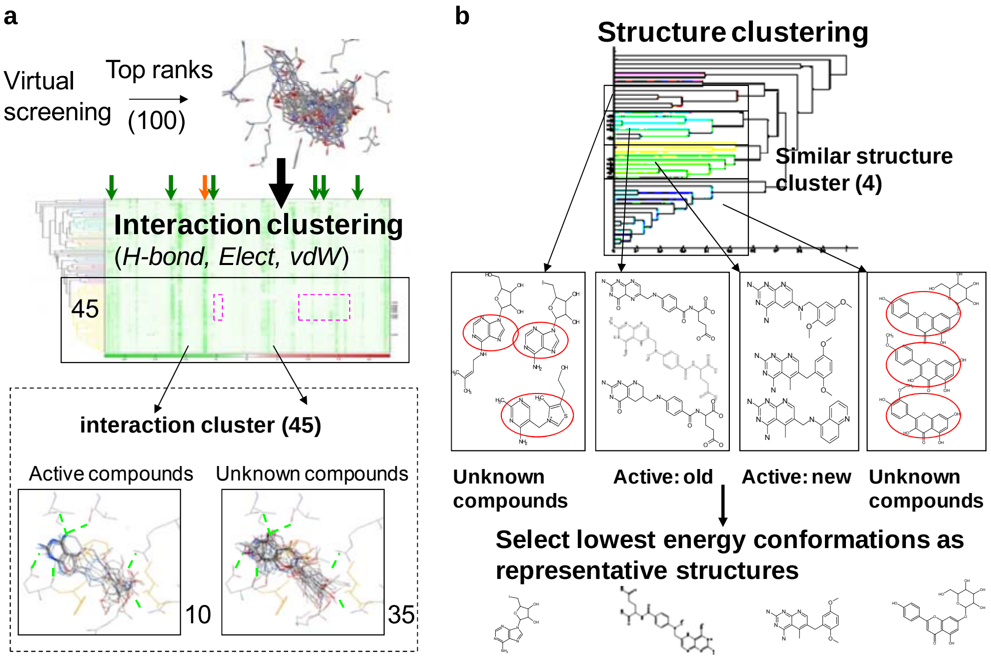


Figure 2.8. Overall process of the two-stage combinative cluster analysis. (a) First stage clustering using protein-ligand interactions generated via GEMDOCK. (b) Second stage clustering of first stage results done using physical-chemical features.

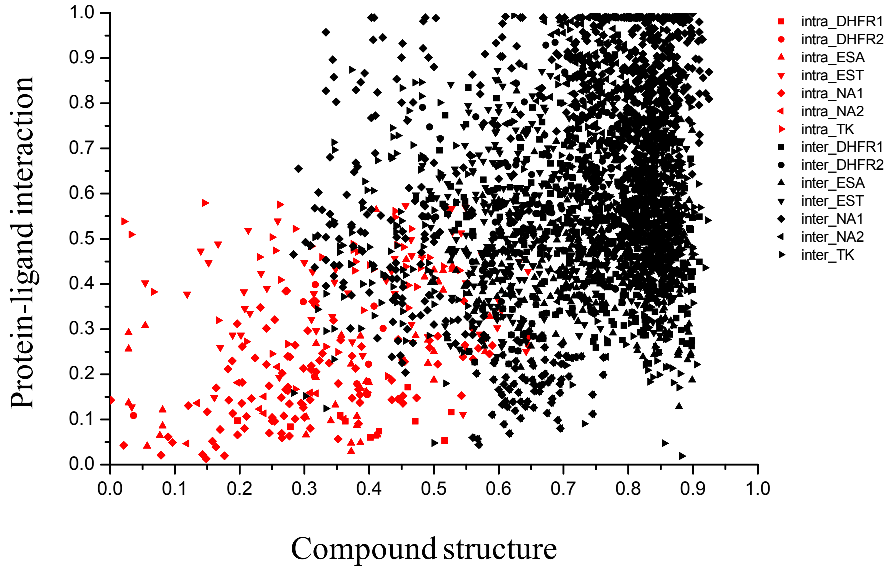


Figure 2.9. Designing a reference threshold of P-L interaction and atom-pair descriptors. The complementation between atom-pair descriptor and the protein-ligand interaction descriptor is also show in this figure. The distance threshold of atom-pair descriptor was 0.55 (tanimoto coefficient). The threshold of distance of protein-ligand interaction descriptor was 0.39 (correlation coefficient).

ligand interaction profile (first stage), docked compounds can be clustered into groups with distinct binding interactions. Structure clustering (second stage) grouped similar compounds obtained from the first stage into similar structures clusters; the lowest energy compound from each cluster being selected as a final candidate. By representing interactions at the atomic-level and including measures of interactions strength (Fig. 2.9), better descriptions of protein-ligand interactions and a more specific analysis of virtual screening was achieved. The two-stage clustering approach enhanced our post-screening analysis by revealing accurate performances in clustering, mining and visualizing compound candidates, thus, improving virtual screening enrichment.

2.4 Summary

As the number of protein structures increases rapidly, structure-based drug design and virtual screening approaches are becoming important and helpful in lead discovery[1-2,6]. A number of docking and virtual screening methods [16,23-24,35] have been utilized to indentify lead compounds, and some success stories have been reported [1-2,4-5,7-8,10]. However, identifying lead compounds by exploiting thousands of docked protein-compound complexes is still a challenging task. The major weakness of virtual screenings is likely due to incomplete understandings of ligand binding mechanisms and the subsequently imprecise scoring algorithms . In the related works, several studies were proposed for improving the accuracy and precision in the VS processes. First, the scoring function of GEMDOCK evolves the pharmacological preferences from a number of known active ligands to take advantage of the similarity of a putative ligand to those that are known to bind to a protein's active site, thereby guiding the docking of the putative ligand. In the post-screening analysis process, the consensus scoring strategy using data fusion and exploring diversity on scoring characteristics between individual scoring functions for improving VS is proposed. When the huge amount of VS data needs to be interpreted, the combinative cluster analysis is applied for effectively mining the representatives and easily visualizing the VS data. Although we have been successfully applied these methods on the VS studies of two important virus targets, dengue virus and influenza virus, some shortcomings are needed to be addressed.

Chapter 3

Site-moiety map for recognizing interaction preferences between protein pockets and compound moieties

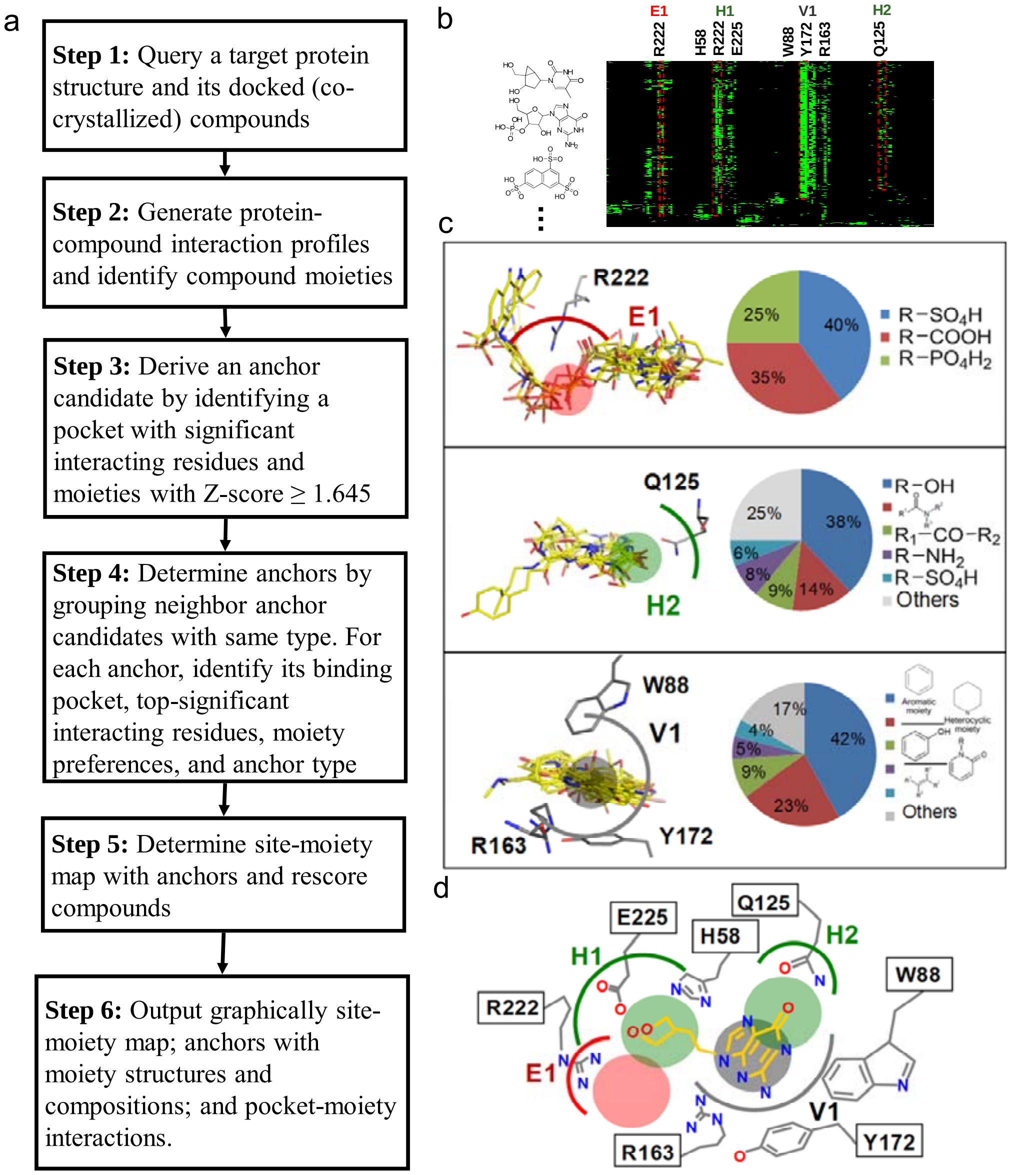
3.1 Introduction

Most of docking programs[16,23-24] use energy-based scoring methods which are often biased toward both the selection of high molecular weight compounds and charged polar compounds in the top ranks. Meanwhile, these approaches generally cannot identify the key features (*e.g.,* pharmacophore spots) that are essential to trigger or block the biological responses of the target protein. Although pharmacophore techniques[27] have been applied to derive the key features, these methods require a set of known active ligands that were acquired experimentally. Therefore, the more powerful techniques for post-screening analysis to identify the key features through docked compounds and to understand the binding mechanisms provide a great potential value for drug design.

To address these issues, we presented the SiMMap method to infer the key features by a site-moiety map describing the relationship between the moiety preferences and the physico-chemical properties of the binding site. This method also provides the web server for public access. According to our knowledge, SiMMap is the first public server that identifies the site-moiety map from a query protein structure and its docked (or co-crystallized) compounds. The server provides pocket-moiety interaction preferences (anchors) including binding pockets with conserved interacting residues; moiety preferences; and interaction type. We verified the site-moiety map on three targets, thymidine kinase, and estrogen receptors of antagonists and agonists. Experimental results show that an anchor is often a hot spot and the site-moiety map is useful to identify active compounds for these targets. We believe that the site-moiety map is able to provide biological insights and is useful for drug discovery and lead optimization.

3.2 Method

Figure 3.1 presents an overview of the SiMMap server for identifying the site-moiety map with anchors, describing moiety preferences and physico-chemical properties of the binding site, from a query protein structure and docked compounds. The server first uses checkmol (http://merian.pch.univie.ac.at/~nhaider/cheminf/cmmm) to recognize the compound moieties



Step 1: Query a target protein structure and its docked (co-crystallized) compounds

Step 2: Generate protein-compound interaction profiles and identify compound moieties

Step 3: Derive an anchor candidate by identifying a pocket with significant interacting residues and moieties with Z-score ≥ 1.645

Step 4: Determine anchors by grouping neighbor anchor candidates with same type. For each anchor, identify its binding pocket, top-significant interacting residues, moiety preferences, and anchor type

Step 5: Determine site-moiety map with anchors and rescore compounds

Step 6: Output graphically site-moiety map; anchors with moiety structures and compositions; and pocket-moiety interactions.

Figure 3.1. Overview of the SiMMap server for the site-moiety map using herpes simplex virus type-1 thymidine kinase (TK) and 1000 docked compounds as the query. (a) Main procedure; (b) The merged protein-compound interaction profile; (c) The pocket-moiety interaction preferences of the anchors: H (hydrogen-bonding). Each anchor consists of a binding pocket with conserved interacting residues, the moiety composition and anchor type; The site-moiety map has one hydrogen-bonding (H) and three van der Waals (V) anchors for ER. Each anchor contains the moiety structures and composition, anchor type, and key residues in the binding pocket.

and utilizes GEMDOCK[35] to generate a merged protein-compound interaction profile (Fig. 3.1b), including electrostatic (E), hydrogen-bonding (H) and van der Waals (V) interactions. According to this profile, we infer anchor candidates by identifying the pockets with significant interacting residues and moieties with Z-score ≥ 1.645. The neighbor anchor candidates, which are the same interaction type and the distances between their centers are less than 3.5Å, are grouped into one anchor. These anchors form the site-moiety map describing interaction preferences between compound moieties and the binding site of the query (Figs. 3.1c and 3.1d). Finally, this server provides graphic visualization for the site-moiety map; anchors with moiety structures and compositions; pocket-moiety interactions; and the relationship between anchors and moieties of query compounds.

3.2.1 Definitions of site-moiety map, anchor and pocket

The anchor (pocket-moiety interaction preference) is the core of a site-moiety map. An anchor possesses three essential elements: (1) a binding pocket with conserved interacting residues and specific physico-chemical properties; (2) moiety preferences of the pocket; (3) pocket-moiety interaction type (E, H, or V). An anchor can be considered as "key features" for representing the conserved binding environment element or a "hot spot" which involves biological functions. In addition, we regard a binding pocket, which consists of several residues significantly interacting to compound moieties, as a part of the binding site. The binding pocket often possesses specific physico-chemical properties and geometric shape to bind preferred moieties. The site-moiety map, which can help to assemble potential leads by optimal steric, hydrogen-bonding, and electronic moieties, is useful for drug discovery and understanding biological mechanisms.

3.2.2 Constructing site-moiety map

The SiMMap server performs six main steps for a query (Fig. 3.1a). Here, we used TK as an example for describing these steps.

Generating protein-compound interaction profiles and identification of compound moieties

First, users input a protein structure and its docked compounds. The server used checkmol to identify moieties of docked compounds and GEMDOCK to generate E, H and V interaction profiles. For each profile, the matrix size is $N \times K$ where N and K are the numbers of compounds and interacting residues of query protein, respectively. An interaction profile matrix P(*I*) with type *I* (E, H, or V) is represented as

$$P(I)=\begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1K} \\ p_{2,1} & p_{2,2} & \cdots & p_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N,1} & p_{N,2} & \cdots & p_{NK} \end{bmatrix}$$

where $p_{i,j}$ is a binary value for the compound *i* interacting to the residue *j* (Fig. 4.2B). For H and E profiles, $p_{i,j}$ is set to 1 (green) if an atom pair between the compound *i* and the residue *j* forms hydrogen-bonding or electrostatic interactions, respectively; conversely, the interaction is set to 0 (black). For van der Waals (vdW) interaction, an interaction is set to 1 when the energy is less than -4 (kcal/mol).

SiMMap identified consensus interactions between residues and compound moieties with similar physical-chemical properties through the profiles. For each interacting residue (a column of the matrix P(*I*)) (Fig. 3.1b), we used Z-score value to measure the interacting conservation between this residue and moieties. The standard deviation (σ) and mean (μ) were derived by random shuffling 1,000 times in a profile. The Z-score of the residue *j* is defined as $Z_j=\dfrac{f_j-\mu}{\sigma}$, where f_j is the interaction frequency and given as $f_j=\sum_{i=1}^{N}\dfrac{p_{ij}}{N}$.

We treated protein-compound interactions as a binomial distribution, and then consensus interactions with statistical significance could be identified by their normal approximation. Statistically, a binomial distribution is approximated by a normal distribution when either $p\leq 0.5$ and $np>5$ or $p>0.5$ and $n(1-p)>5$, where *n* is the number of trials and *p* is the probability of success. Here, *n* is the number of selected compounds and *p* is the probability of forming an interaction between a protein and a compound, that is, $p_{i,j}=1$. Typically, the *p* values ranged between 0.01 and 0.03 in this study. While the binomial distribution is a normal approximation, at least 500 compounds should be selected for constructing an interaction profile matrix.

Deriving anchor by identifying a pocket with significant interacting residues and moieties

Spatially neighbor interacting residues and moieties with statistically significant Z-score ≥ 1.645 were referred as an anchor candidate. Neighbor anchor candidates, which are spatially overlapped and the same anchor type, were clustered as an anchor and the anchor center is the weighted geometric center of their interacting compound moieties. Here, two anchors were merged if the distance of two anchor centers is less than 3.5 Å. In each anchor, top three residues with the highest Z-score values were regarded as key residues forming a binding pocket. For each anchor, we identified its moieties of docked compounds according to the moiety library derived from checkmol, and calculated the moiety composition (Fig. 3.1c). These anchors form the site-moiety map (Fig. 3.1d) of the query.

Outputting graphically site-moiety map and identifying active compounds

SiMMap can be applied to identify active compounds for structure-based virtual screening. One of weaknesses of virtual screening is likely incomplete understanding of the chemistry involved in ligand binding and the subsequently imprecise scoring algorithms. When a compound highly agrees with the anchors of the site-moiety map, this compound often activates or inhibits the target. The SiMMap server scores a compound by combining predicted binding energy of GEMDOCK and the anchor score between the map and the compound. The SiMMap score, $S(i)$, for a compound i is defined as

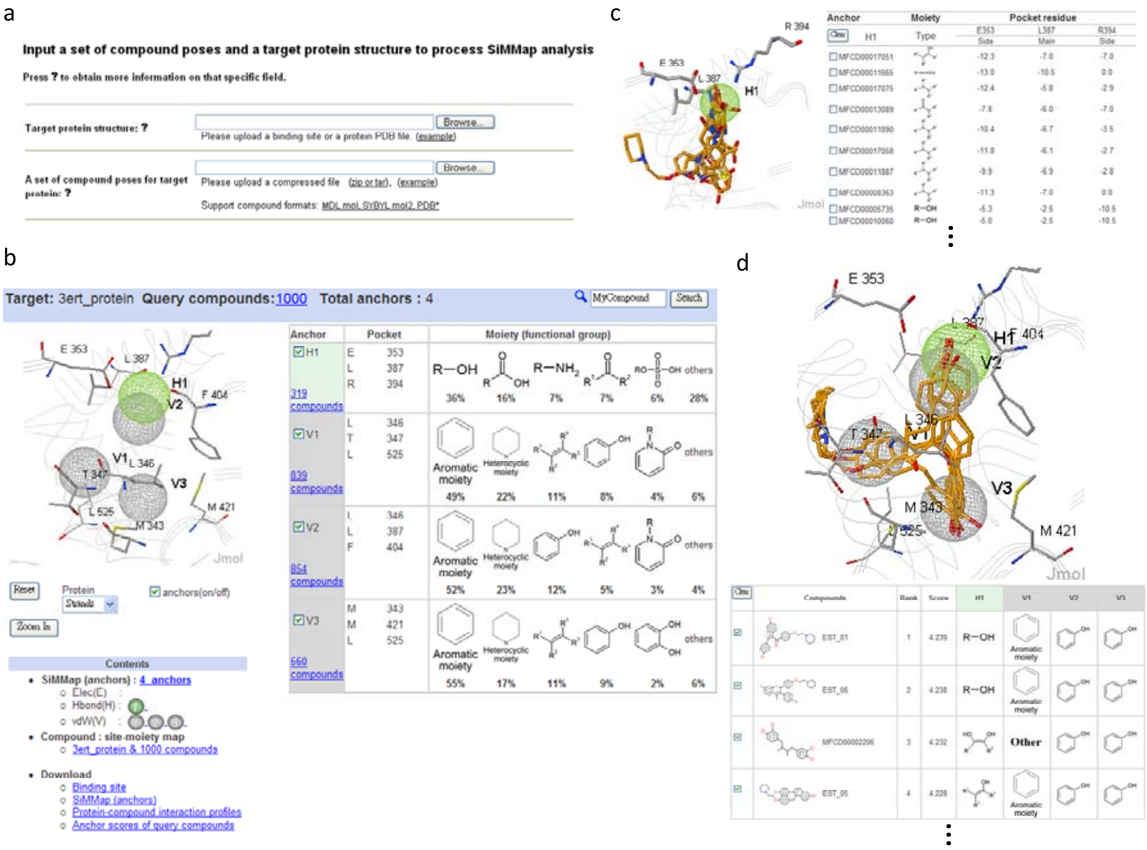
$$S(i) = \sum_{a=1}^{n} AS_a(i) + (-0.001)\frac{E(i)}{M^{0.5}} \tag{1}$$

where $AS_a(i)$ is the anchor score of compound i in the anchor a, n is the number of anchors, $E(i)$ is the docked energy of compound i, and M is the atom number of compound i. The anchor score is set to 1 when the compound i agrees the moiety preference of the anchor a. Here, the anchor score and the term $M^{0.5}$ are useful to reduce the deleterious effects of selecting high molecular weight compounds[26]. Based on SiMMap scores, we can obtain new ranks of query compounds.

3.3 Results

3.3.1 Web service

SiMMap is an easy-to-use web server (Fig. 3.2). Users input a protein structure without ligands in PDB format and its docked or co-crystallized compounds in MDL mol, SYBYL mol2, or PDB format (Fig. 3.2a). These docked compounds should be generated by any external docking methods (*e.g.*, DOCK, FlexX, GOLD and GEMDOCK) before users uploaded these compounds. Typically, the SiMMap server yields a site-moiety map within 5 minutes if the number of query compounds is less than 100. This server provides the graphic



Input a set of compound poses and a target protein structure to process SiMMap analysis

Press **?** to obtain more information on that specific field.

Target protein structure: ? [ ] Browse...
Please upload a binding site or a protein PDB file. (example)

A set of compound poses for target protein: ? [ ] Browse...
Please upload a compressed file (zip or tar). (example)
Support compound formats: MDL mol, SYBYL mol2, PDB*

Anchor	Moiety	Pocket residue		
H1	Type	E353 Side	L387 Main	R394 Side
□ MFCD00017051		-12.3	-7.0	-7.0
□ MFCD00011665		-13.0	-10.5	0.0
□ MFCD00017075		-12.4	-5.8	-2.9
□ MFCD00013089		-7.8	-6.0	-7.0
□ MFCD00011890		-10.4	-6.7	-3.5
□ MFCD00017058		-11.8	-6.1	-2.7
□ MFCD00011887		-9.9	-6.9	-2.8
□ MFCD00008363		-11.3	-7.0	0.0
□ MFCD00005735	R—OH	-5.3	-2.5	-10.5
□ MFCD00010060	R—OH	-5.0	-2.5	-10.5

⋮

Target: 3ert_protein **Query compounds:**1000 **Total anchors : 4**

Anchor	Pocket		Moiety (functional group)					
☑ H1 319 compounds	E L R	353 387 394	R—OH 36%	16%	R—NH$_2$ 7%	7%	6%	others 28%
☑ V1 839 compounds	L T L	346 347 525	Aromatic moiety 49%	Heterocyclic moiety 22%	11%	8%	4%	others 6%
☑ V2 854 compounds	L L F	346 387 404	Aromatic moiety 52%	Heterocyclic moiety 23%	12%	5%	3%	others 4%
☑ V3 560 compounds	M M L	343 421 525	Aromatic moiety 55%	Heterocyclic moiety 17%	11%	9%	2%	others 6%

Contents

- SiMMap (anchors) : 4 anchors
  - Elec(E) :
  - Hbond(H) : 1
  - vdW(V) : 1 2 3
- Compound : site-moiety map
  - 3ert_protein & 1000 compounds
- Download
  - Binding site
  - SiMMap (anchors)
  - Protein-compound interaction profiles
  - Anchor scores of query compounds

	Compounds	Rank	Score	H1	V1	V2	V3
☑	EST_01	1	4.239	R—OH	Aromatic moiety		
☑	EST_06	2	4.238	R—OH	Aromatic moiety		
☑	MFCD00002206	3	4.232		Other		
☑	EST_05	4	4.228		Aromatic moiety		

⋮

Figure 3.2. The SiMMap server analysis results using estrogen receptor (ER) and 1000 docked compounds as the query. (a) The user interface for uploading target protein structure and docked compounds. (b) The site-moiety map has one hydrogen-bonding and three van der Waals anchors for ER. Each anchor contains the moiety structures and composition, anchor type, and key residues in the binding pocket. (c) The details of moiety structures and residue-moiety interactions in the H1 anchor. (d) The SiMMap scores, ranks and the relationships between anchors and moieties of query compounds.

visualization of the site-moiety map and anchors elements, including a binding pocket with interacting residues, moiety compositions and structures, numbers of involved compounds, and anchor types (Fig. 3.2b). For each anchor, this server shows docked conformations of compounds and the detailed atomic interactions between pocket residues and moieties (Fig. 3.2c). In addition, SiMMap shows the new rank and compound moiety structures fitting the anchors for each query compound (Fig. 3.2d). SiMMap uses two open source tools for graphic visualization: Jmol (http://www.jmol.org/) for displaying three-dimensional protein and compound structures with anchors and OASA (http://bkchem.zirael.org/oasa_en.html) for visualizing compound structures. The server allows users to download the anchor coordinates in the PDB format; interaction profiles; new ranks and anchor scores of query compounds.

3.3.2 Thymidine kinase and estrogen receptor

The SiMMap server inferred the site-moiety map of TK. This map consisted of four anchors (*i.e.*, E1, H1, H2, and V1 (Fig. 3.1d) and the moiety composition and conserved interacting residues of each anchor (Fig. 3.1c). For example, the E1 anchor possesses a binding pocket with residue R222, and three moiety types (*i.e.*, sulfuric acid monoester (40%), carboxylic group (35%) and phosphoric acid monoester (25%)) derived from 57 compounds. Meanwhile, the E1 includes the phosphate moiety of ATP and its residue R222 playing a major role to interact with the substrate [38-39]. The preferred moiety types of an anchor are suitable groups interacting to conserved residues of the binding pocket. The moiety preference is able to guide the suggestion of functional group substitutions for lead structures.

We used estrogen receptor (ER), a therapeutic target for osteoporosis and breast cancer[40], as the example. Based on 1000 docked compounds and ER, the SiMMap server identifies four anchors (H1, V1, V2, and V3) and provides moiety preferences and compositions in these anchors (Fig. 3.2b and 3.3). The H1 anchor comprises three residues (E353, L387, and R394) and five main moiety types: hydroxyl group (36%), carboxylic acid (16%), amine (7%), ketone (7%), and sulfuric acid monoester (6%) summarized from 319 compounds. Furthermore, three residues (L346, T347, and L525) and 839 compounds are involved in the V1 anchor, preferring five moiety types (*i.e.*, aromatic ring (49%), heterocyclic group (22%), alkenes (11%), phenol (8%), and oxohetarene (4%)). The anchor V2 is a hydrophobic pocket containing L346, F404, and L387, and the former two re sidues are highly conserved[41]. These hydrophobic residues interact with aromatic ring (52%), heterocyclic group (23%), phenol (12%), alkenes (5%), and oxohetarene (3%). Finally, aromatic rings (55%), heterocyclic groups (17%), alkenes (11%), and phenols (9%) summarized from 560 compounds often form vdW contacts with the long side chains of M343, M421, and L525 in the anchor V3. The ring groups of antagonists are often stabilized by the side chains of M343, L346, T347, L387, M421, and L525. In this case, most selective estrogen receptor modulators of ER (*e.g.*, EST_01 (raloxifene), EST_06 (LY-326315,) and EST_05 (EM-343)) agree with these four anchors (Fig. 3.2d and 3.3c). Anchors

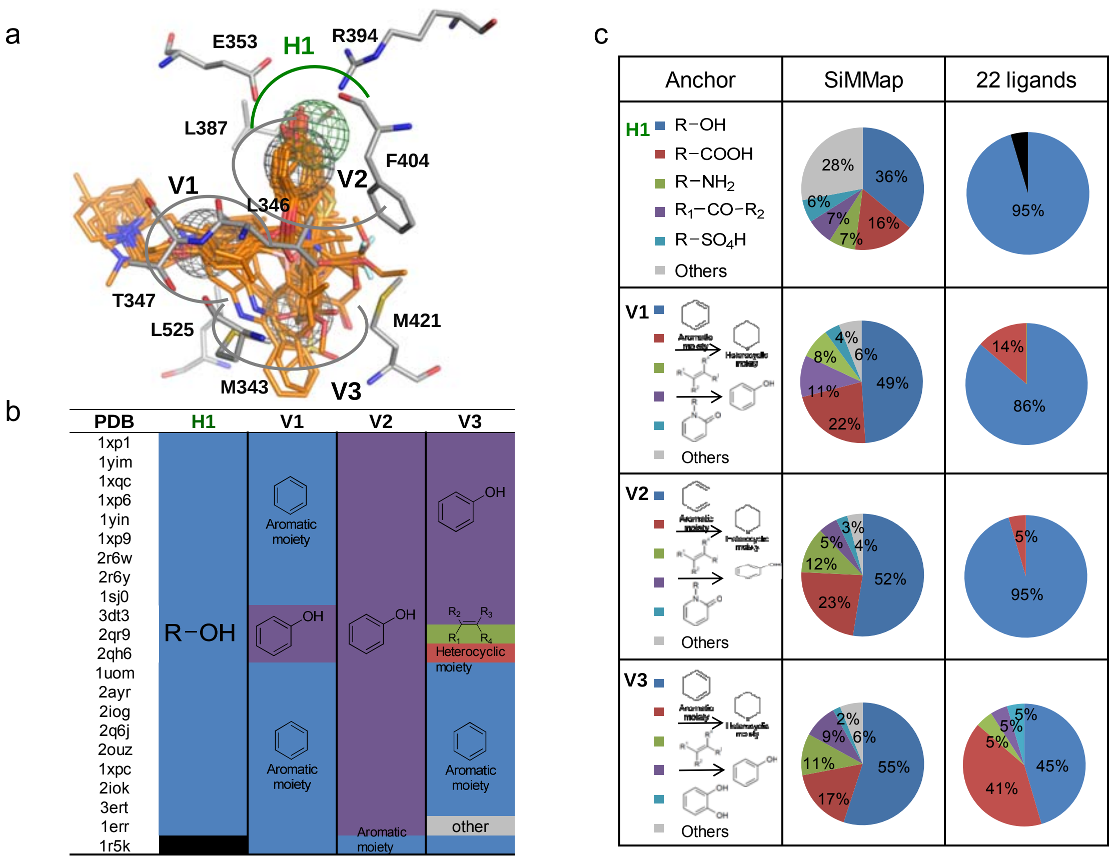
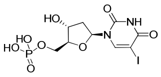
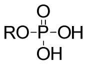
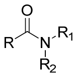
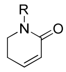
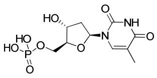
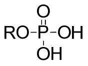
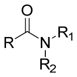
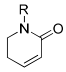
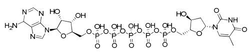
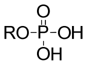
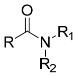
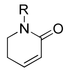
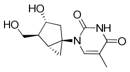
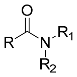
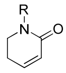


Figure 3.3. The relationships between the site-moiety map and 22 co-crystallized ligands of ER. (a) The mapping between four inferred anchors (binding pocket with conserved interacting residues) and these 22 ligands in the active site. (b) The moieties of these 22 ligands in each anchor. Black cell presents that the moiety of the compound does not agree with the anchor H1. (c) The moiety compositions of 1000 docked compounds (SiMMap) and these 22 ligands.

Table 3.1. The relationship between the anchors and moieties of 15 co-crystallized ligands for TK

PDB code	Compound structure	Anchor			
		E1	H1	H2	V1
3vtk			R-OH		
1vtk			R-OH		
1p7c			R-OH		
1of1			R-OH		

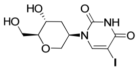
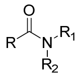
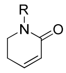
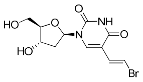
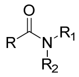
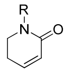
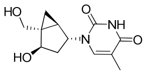
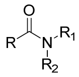
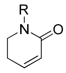
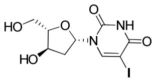
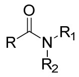
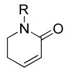
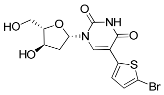
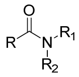
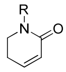
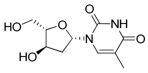
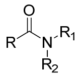
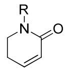
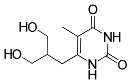
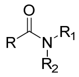
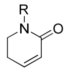
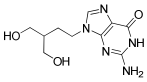
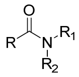
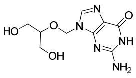
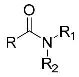
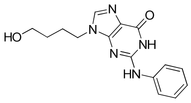
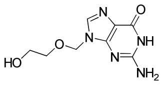
1ki6		R-OH		
1ki8		R-OH		
1e2k		R-OH		
1ki7		R-OH		
1ki4		R-OH		
1kim		R-OH		
1e2p		R-OH		
1ki3		R-OH		Aromatic moiety
1ki2		R-OH		Aromatic moiety
1qhi		R-OH	R-NH$_2$	Aromatic moiety
2ki5		R-OH	R-NH$_2$	Aromatic moiety

Table 3.2. Comparing SiMMap with other methods on thymidine kinase and estrogen receptor by false-positive rates

True positive (%)	Thymidine kinase (TK)				Estrogen receptor (ER)			
	SiMMap	DOCK[a]	FlexX[a]	GOLD[a]	SiMMap	DOCK[a]	FlexX[a]	GOLD[a]
80	6.3[b]	23.4	8.8	8.3	1.1	13.3	57.8	5.3
90	6.8	25.5	13.3	9.1	1.1	17.4	70.9	8.3
100	6.8	27	19.4	9.3	7.5	18.9	NA	23.4

[a] Summarized from Bissantz *et al.*[17]

identified by the SiMMap server often contain key pockets and moieties. To initially validate the anchors for biological mechanisms (*e.g.*, ligand binding and catalysis mechanisms), we selected 15 TK and 22 ER co-crystallized ligands (Table 3.1 and Fig. 3.3). The corresponding moieties of these co-crystallized ligands were highly matched the anchors derived from 1000 docked compounds (10 known active ligands and 990 randomly selected compounds described in Data sets). The site-directed mutagenesis shows that the conserved interacting residues of the anchors are often essential for ligand binding and catalysis mechanisms. For ER target, 22 ER co-crystallized ligands contain three consistent moieties that are hydroxyl group and aromatic rings (Fig. 3.3b). The hydroxyl group forms hydrogen bonds with R394 and E353 in H1, and the aromatic ring yields vdW contacts with L346, L387, and F404 in V2. The other consistent aromatic ring forms vdW contacts with L346, T347, and L525 in V1. These results show that an anchor is often a hot spot and involved in biological functions.

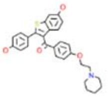
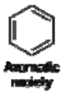
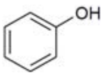
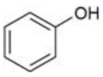
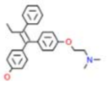
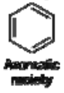
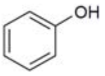
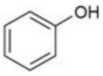
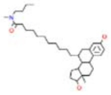
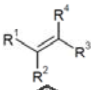
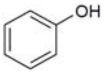
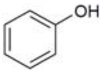
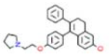
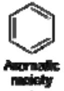
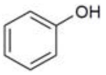
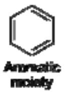
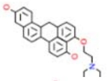
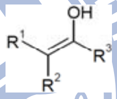
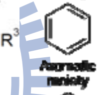
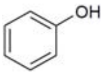
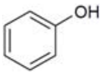
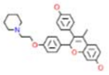
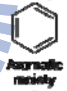
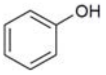
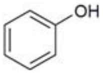
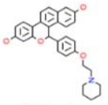
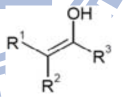
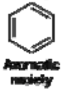
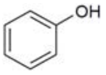
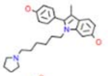
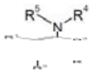
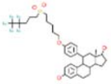
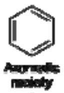
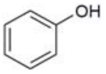
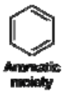
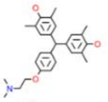
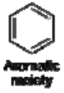
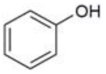
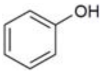
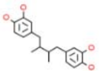
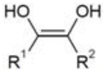
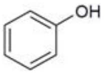
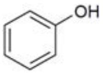
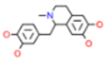
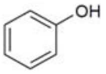
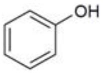
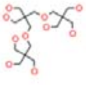
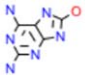
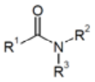
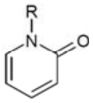
To provide initial validation of the SiMMap server for virtual screening, we selected TK, ER, and ERA with 1000 compounds as test sets. First, we compared the accuracies of SiMMap with those of GEMDOCK on these three targets based on true positive rates. SiMMap, combining anchor scores and docking energies (Equation 1), outperforms GEMDOCK on these cases. We then compared SiMMap with other three programs (DOCK, FlexX, and GOLD) on TK and ER sets. All approaches were tested using the same proteins and compound sets (Table 3.2). When the positive rate was 90%, the false positive rates were 6.8% (SiMMap), 25.5% (DOCK), 13.3% (FlexX), and 9.1% (GOLD) for TK and were 1.1% (SiMMap), 17.4% (DOCK), 70.9% (FlexX), and 8.3% (GOLD) for ER.

The compound, which agrees with anchors of the site-moiety map, is often able to activate or inhibit the target protein (Tables 3.1 and 3.3). In addition, the anchor score (i.e. $AS(i)$ defined in Equation 1) of SiMMap can be used to reduce the ill-effect of the energy-based scoring methods which are often biased toward both the selection of high molecular weight compounds and charged polar compounds[25-26]. For example, according to the SiMMap scores (Equation 1), the top ranks of ER, MFCD0002206 (masoprocol) and MFCD00012748 were identified as the analogs of the active compounds (Table 3.3). The anchor score of SiMMap was helpful to reduce the highly polar compounds (*e.g.*, MFCD00004690 and MFCD00013089 in ER) whose anchor scores are low. The anchor score of SiMMap can easily combine with other energy-based scoring functions.

3.4 Summary

The utility and feasibility of SiMMap method is demonstrated for statistically inferring the site-moiety map describing the relationship between the moiety preferences and physico-chemical properties of the binding site. The validation results show that the site-moiety map is useful to reflect biological functions and identify active compounds from thousands of compounds. In addition, the site-moiety map can guide to assemble potential leads by optimal steric, hydrogen-bonding, and electronic moieties. We believe that the SiMMap serve is able to provide the biological insights of protein-ligand binding models, enrich the screening accuracy, and guide the processes of lead optimization.

Table 3.3. The mapping between the anchors and active and typical compounds for ER

Compound	Structure	GEMDOCK rank	SiMMap rank	SiMMap score	H1	V1	V2	V3
EST_01		2	1	4.239	R—OH			
EST_02		32	19	4.216	R—OH			
EST_03		28	16	4.217	R—OH			
EST_04		8	5	4.226	R—OH			
EST_05		6	4	4.228				
EST_06		3	2	4.238	R—OH			
EST_07		21	13	4.218				Other
EST_08		10	7	4.225	R—OH	Other		Other
EST_09		30	20	4.216	R—OH			
EST_10		246	84	4.193	R—OH			
MFCD00002206		4	3	4.232		Other		
MFCD00012748		17	11	4.221	R—OH	Other		
MFCD00004690		5	154	3.23	R—OH		Other	Other
MFCD00013089		25	617	2.218				

Chapter 4

The application of site-moiety map for characterizing protein-ligand binding sites and discovering adaptive inhibitors for orthologous protein targets

4.1 Introduction

The expanding number of protein structures and advances in bioinformatics tools have offered an exciting opportunity for structure-based virtual screening methods in drug discovery[42]. Although there are some successful agents in the antibiotic development, few agents act at novel molecular sites to target multiple antibiotic–resistant pathogenic bacteria[43-44]. However, the screening tools are often designed for one-target paradigm and the scoring methods are highly target-dependent and energy-based, and cannot persuasively identify true leads, which results in a lower success rate [2-3,45]. To discover adaptive inhibitors of multiple targets is an emergent task for drug design[46-48].

SiMMap modeling provides an *in silico* post-screening analysis to establish a target binding site-moiety map that comprises three crucial elements (conserved interacting residues, the moiety preference, and pocket-moiety interaction type [electrostatic (E), hydrogen-bonding (H), or van der Waals (V)])[49]. For structural-based virtual screening [1-2,9,45], this method offers a more efficient solution for drug discovery, particularly in the absence of a pharmacophore model describing the structure-activity relationship extracted from experiments.

We introduce the concept of orthSiMMap to represent the conserved binding environment elements or "hot spots" among orthologous targets. Identification of these consensus features in the orthSiMMap can then be used for identifying novel binding partners of orthologous targets. We then developed a method to extract the conserved features of the ligand-binding environments of orthologous targets, thus establishing the orthSiMMap using structure-based virtual screening. We focused on shikimate kinase (EC 2.7.1.71), the fifth enzyme in the shikimate pathway that is present in bacteria, fungi, and plants, but not animals. This expression pattern makes shikimate kinase an attractive target for the development of new antimicrobial agents, herbicides, and antiparasitic agents[50]. In using these models, six potent inhibitors with low IC_{50} values (<8.0 μM) were identified. Site-directed mutagenesis studies revealed that critical conserved interacting residues contribute to specific pocket-moiety interaction anchors (S15, D33, F48, R57, R116, and R132). These results illustrate a robust orthSiMMap-based approach to identify selective SK inhibitors and shed insight to a new induced-fit mechanism by an inhibitor. The works of the biological assay and the crystal structures in this chapter are done by Dr. Wen-Ching Wang of National Tsing Hua University[51].

4.2 orthSiMMap methods

The main steps of the orthSiMMap method for producing SiMMaps and an orthSiMMap from orthologous targets are described as follows (Fig. 4.1):

(1) Virtual screening of orthologous targets. We used in-house GEMDOCK program[26,35] to screen Maybridge (65,947 compounds) and NCI (236,962 compounds) databases for both HpSK and MtSK (apo/closed forms). The top-ranked 2% (6,000 compounds) of each target were selected from the screening results for the subsequent protein-compound profiling.

(2) Profiling analysis of target-compound interactions. The target-compound interactions of clustered top ranked ~3,000 compounds by discarding similar compounds were assessed to derive the anchor with the interacting type, including hydrogen-bonding, electrostatic, and van der Waals interactions. For H and E profiles, an interaction is set to 1 (green) if an atom pair forms hydrogen-bonding or electrostatic interactions; conversely, the interaction is set to 0 (black). For van der Waals interaction, an interaction is set to 1 when the energy is less than -4 (kcal/mol) (Fig. 4.1b).

(3) Identification of anchors (Fig. 4.1c). We identified consensus interactions between residues and compound moieties through the profiles. For an interacting residue, we used Z-score to measure the interacting conservation between the residue and moieties. The interaction is treated as a binomial distribution that is approximated to a normal distribution when either $p \leq 0.5$ and $np > 5$ or $p > 0.5$ and $n(1-p) > 5$, where n is the number of selected compounds and p is the probability of forming an interaction. Theoretically, at least 500 compounds should be selected for constructing a target-compound interaction profile. Spatially neighbor interacting residues and moieties with statistically significant Z-score ≥ 1.645 were referred as an anchor. A set of anchors derived from the target-compound interacting profile can be used

to establish a site-moiety map for each orthologous target.

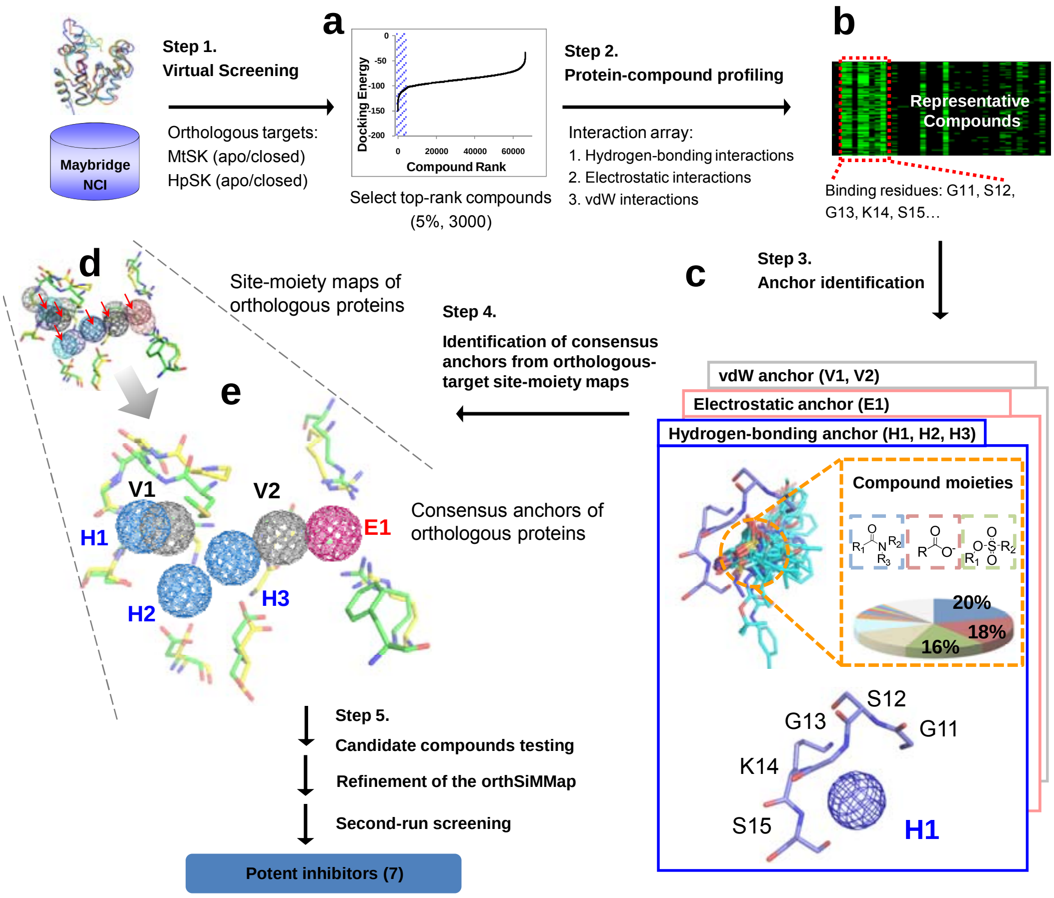

a
Step 1.
Virtual Screening
Maybridge
NCI
Orthologous targets:
MtSK (apo/closed)
HpSK (apo/closed)
Docking Energy
0
-50
-100
-150
-200
0
20000
40000
60000
Compound Rank
Select top-rank compounds
(5%, 3000)
Step 2.
Protein-compound profiling
Interaction array:
1. Hydrogen-bonding interactions
2. Electrostatic interactions
3. vdW interactions
b
Representative Compounds
Binding residues: G11, S12, G13, K14, S15…
Step 3.
Anchor identification
c
vdW anchor (V1, V2)
Electrostatic anchor (E1)
Hydrogen-bonding anchor (H1, H2, H3)
Compound moieties
20%
18%
16%
S12
G13
G11
K14
S15
H1
Step 4.
Identification of consensus anchors from orthologous-target site-moiety maps
d
Site-moiety maps of orthologous proteins
e
V1
V2
H1
E1
H2
H3
Consensus anchors of orthologous proteins
Step 5.
Candidate compounds testing
Refinement of the orthSiMMap
Second-run screening
Potent inhibitors (7)

Figure 4.1. Framework of the orthSiMMap method. In Step 1, GEMDOCK was used to generate docked poses for HpSK and MtSK by screening compound libraries (Maybridge and NCI). For each target (HpSK or MtSK), the protein-compound interacting profile was derived from fusing the top ranked 5% (~3000) compounds. In Step 3, conserved interactions of the target protein and chemical moieties of ligands are identified to deduce the anchors of HpSK and MtSK. The orthSiMMap is constructed based on the conserved features between orthologous target site-moiety maps, which will be used to select candidate compounds for the enzymatic assay. Finally, the model is refined based on the bioassay of candidate compounds.

(4) Establishment of the orthSiMMap of the orthologous targets (Fig. 4.1d and 4.1e). The superimposed SiMMaps (anchors) of orthologous proteins (HpSK and MtSK) revealed an over-

lapping region of matched anchors which form the orthSiMMap (Fig. 4.2c and 4.2d). For the compound *x*, the orthSiMMap score is defined as $CAS(x)=\sum_{i=1}^{a} w_i AS_i(x) - 0.001E(x)$, where w_i is the conservation of the anchor *i* on orthologous targets; $AS_i(x)$ is the anchor score of the compound *x* in the anchor *i*; *a* is the number of anchors; $E(x)$ is the docked energy of the compound *x*. The orthSiMMap rank of the compound *x* was obtained by sorting $CAS(x)$ into the descending order.

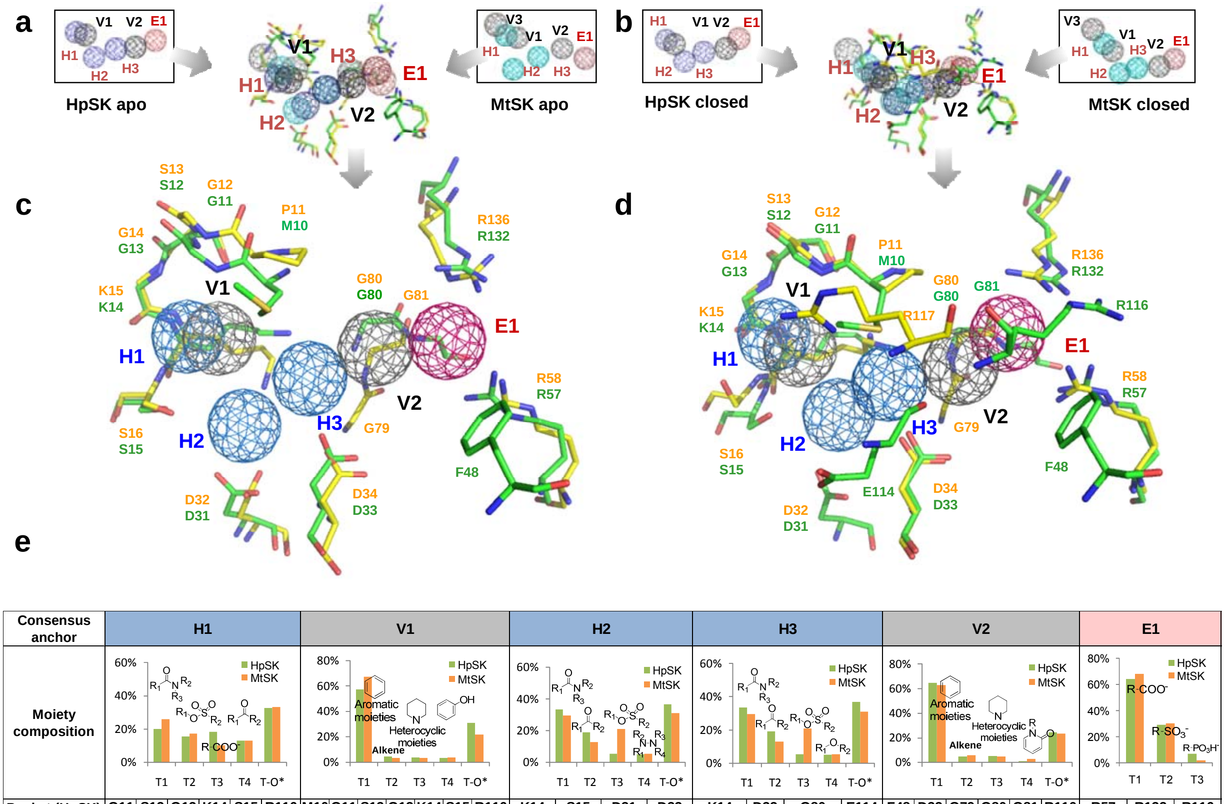
(5) Inhibition assay. We selected top-ranked compounds using rank-based consensus scoring (RCS) for inhibitory assay. For a compound *x*, we calculated its RCS by combining the ranks of *m* (apo/closed) forms of *n* orthologous targets as follows: $S_R(x)=\sum_{i=1}^{n}\sum_{k=1}^{m} R_{C_{ki}}(x)/2mn$, where $R_{C_{ki}}(x)$ is the orthSiMMap rank of *x* on the *k* (apo/closed) form of the target *i*. Here, *m* and *n* are 2. We yielded the RCS rank of the compound *x* by sorting $S_R(x)$ into the ascending order.

(6) Refinement of orthSiMMaps. Active and inactive compounds from the enzyme inhibition assay were used to evaluate and refine the orthSiMMaps.

4.3 Results

4.3.1 orthSiMMap method

We have previously reported "SiMMap" to construct the site-moiety map of a target protein from a set of screening compounds. This map consists of several anchors, which is useful in providing biological insights and guiding the process in drug discovery including hit search and lead optimization[49]. Here, an orthSiMMap method is established to derive a core site-moiety map, referred as orthSiMMap, from orthologous site-moiety maps by structure-based virtual screening on these targets (Fig. 4.2 and Fig. 4.1). The consensus anchors of an orthSiMMap derived from multiple orthologous targets can be considered as "key features" that represent the conserved binding environment involved in biological functions. An orthSiMMap is defined by a set of consensus anchors between orthologous proteins and compound moieties. The following criteria are considered: (1) screening targets are orthologous proteins; (2) the binding sites of orthologous targets share conserved physical-chemical features; (3) the site-moiety maps of orthologous targets often share comparable anchors with respect to their sites and crucial protein-ligand interactions.



Consensus anchor	H1						V1							H2				H3				V2						E1		
Pocket (HpSK)	G11	S12	G13	K14	S15	R116	M10	G11	S12	G13	K14	S15	R116	K14	S15	D31	D33	K14	D33	G80	E114	F48	D33	G79	G80	G81	R116	R57	R132	R116
Pocket (MtSK)	G12	S13	G14	K15	S16	R117	P11	G12	S13	G14	K15	S16	R117	K15	S16	D32	D34	K15	D34	G80	T115	F49	D34	G79	G80	G81	R117	R58	R136	R117
Motif / Domain	A-motif					LID	A-motif						LID	A-motif		SB		A-motif	SB	B-motif	LID	SB		B-motif			LID	SB		LID
HpSK apo	●	●	●	●	●		●	●	●	●	●	●			●	●	●	●	●	●		●	●		●	●		●	●	
MtSK apo	●	●	●	●			●	●	●	●	●			●	●	●		●	●	●			●	●	●			●	●	
HpSK closed	●	●	●	●	●		●	●	●	●		●		●	●	●	●	●	●	●	●	●			●	●	●	●	●	●
MtSK closed	●	●	●		●	●		●		●	●	●	●	●	●		●	●	●	●			●	●	●			●	●	

Figure 4.2. Shikimate kinase orthSiMMaps. (a) Superimposed apo-form anchors of HpSK and MtSK. (b) Superimposed closed-form anchors of HpSK and MtSK. (c) The apo-form orthSiMMap model and (d) closed-form orthSiMMap include six consensus anchors derived from consensus anchors (a) and (b), respectively. Each consensus anchor shares conserved residues between HpSK and MtSK and the same type of binding environment. (e) Features of the six consensus anchors of the apo-form orthSiMMap. Each of T groups (T1–T4) represents a given chemical moiety and T-O* indicates other chemical groups. H1, V1, and H2 are situated at the ATP-binding site, while H3, V2, and E1 are at the shikimate-binding site. Each consensus anchor includes conserved interacting residues (●) and the major chemical moieties of the compound candidates.

For each orthologous target, we used top ranked 2% (~3000) compounds obtained by screening compound libraries to analyze target-compound interaction profiles in order to establish the site-moiety map (SiMMap) comprised of several anchors (Fig. 4.2a and 4.2b). Each anchor represents a local binding environment with specific physico-chemical property or pharmacophore spot, which is derived by identifying statistically significant interacting residues and compound moieties (Fig. 4.3). The orthSiMMap (Fig. 4.2c and 4.2d) that consists of the matched anchors, referred as "hot spots", of orthologous proteins is generated by extracting the consensus anchors of orthologous SiMMaps. We were able to derive the orthSiMMaps of apo-form and closed-form HpSK and MtSK according to the above steps.

To validate the orthSiMMap method, we collected a dataset of 37 orthologous target pairs (Table 4.1) with biological function annotations summarized from UniProt[52]. Experimental results show that the consensus anchors of an orthSiMMap often reveal the binding pocket with conserved interacting residues involving biological functions.

4.3.2 Orthologous SiMMaps and orthSiMMaps of SKs

Figure 4.2 shows the orthSiMMaps of SKs. We first generated the apo-form SiMMaps of HpSK (6 anchors) and MtSK (7 anchors), respectively, allowing us to derive the orthSiMMap with 6 consensus anchors (Fig. 4.2a and 4.2c). In parallel, the closed-form orthSiMMap with 6 consensus anchors was also derived (Fig. 4.2b and 4.2d). It is noted that the apo-form and closed-form orthSiMMaps share six comparable anchors (hot spots): E1, H1–H3, V1, and V2 (Fig. 4.2c and 4.2d). H1, H2, and V1 sit at the ATP site, while H3, V2, and E1 are situated at the shikimate site (Fig. 4.3). The protein-ligand relationship was analyzed for each hot spot; a set of chemically related entities that contribute to intermolecular interactions were then identified (Fig. 4.2e). Our results support the notion that a hot spot shares a unique chemical-physical binding environment, which may be used to guide combinatorial library design for further compound development and lead optimization. The compounds moieties, anchors, SiMMaps and orthSiMMaps are available at http://simmap.life.nctu.edu.tw/orthsimmap/.

Of the six consensus anchors (Fig. 4.2e), E1 is a negatively charged site that interacts with R57 (R58 in MtSK), R116 (R117 in MtSK), and R132 in HpSK (R136 in MtSK); these arginines are highly conserved in SKs and are critical for binding to shikimate[53]. The chemical entities on E1 consisted of carboxyl, sulfonate, and phosphate groups. H1 is enclosed with a tight turn (Walker A motif) that binds the β-phosphate of ATP[53]. The identified moieties were carboxylic amide, sulfonate ester, carboxyl acid, and ketone. H2 is situated between H1 and H3 and possesses a hydrogen bonding environment from Walker A motif (K14 and S15 in HpSK; K15 and S16 in MtSK) and a DT/SD motif (D31 and D33 in HpSK; D32 and D34 in MtSK).

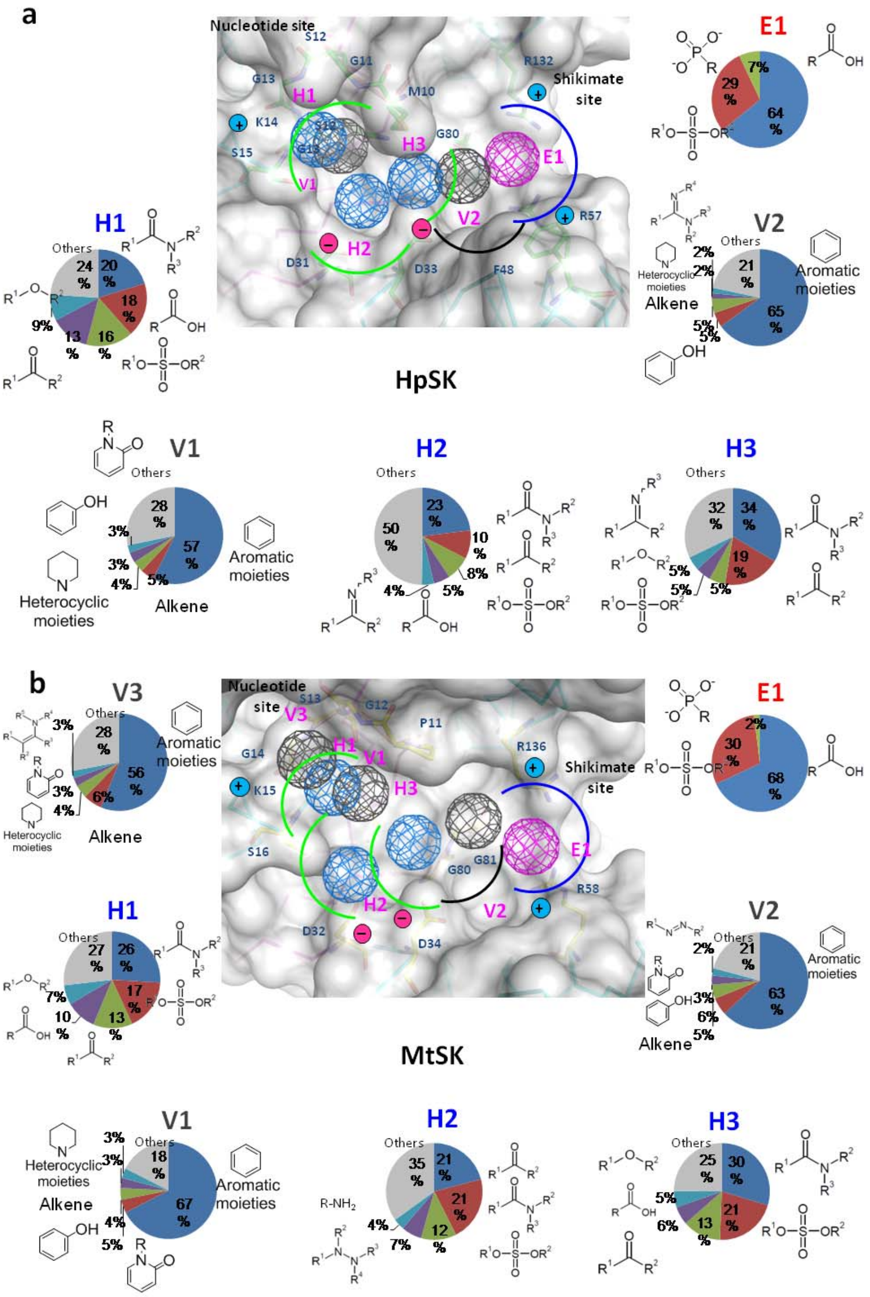


Figure 4.3. The site-moiety maps of (a) HpSK and (b) MtSK. Each anchor represents one of three binding environments (electrostatic: blue; hydrogen-bonding: green; van der Waals: black). The distribution of identified chemical moieties for each anchor is shown as a pie chart. For HpSK, H1, V1, and H2 are situated at the nucleotide site, while H3, V2, and E1 are at the shikimate site. For MtSK, H1, V1, and H2 are at nucleotide site, while H3, V2, and E1 are at the shikimate site.

Table 4.1. Summary of 37 pairs of orthologous targets

ID	Description	Gene Name	Species		PDB code		Bound ligand
			A	B	A	B	
1	3-phosphoshikimate 1-carboxyvinyl transferase	aroA	Escherichia coli	Agrobacterium sp.	1x8t	2pqc	RC1
2	Adenosine deaminase	Ada	Mus musculus	Bos taurus	1a4m	1krm	PRH
3	Androgen receptor	AR	Homo sapiens	Rattus norvegicus	1t5z	1i37	DHT
4	Arginase-1	ARG1	Homo sapiens	Rattus norvegicus	2pll	1d3v	ABH
5	Aspartate aminotransferase	aspC	Thermus thermophilus	Escherichia coli	1bkg	1aia	PMP
6	ATP-dependent hsl protease ATP-binding subunit hslU	hslU	Escherichia coli	Haemophilus influenzae	1do0	1g3i	ATP
7	Bifunctional protein glmU	glmU	Haemophilus influenzae	Escherichia coli	2v0i	1fwy	UD1
8	cAMP-dependent protein kinase catalytic subunit α	Prkaca	Mus musculus	Bos taurus	1atp	1q24	ATP
9	Cytochrome b	MT-CYB	Gallus gallus	Bos taurus	3l71	1sqb	AZO
10	Dihydrofolate reductase	DHFR	Homo sapiens	Mus musculus	3gyf	3k47	D09
11	Dihydrofolate reductase	folA	Escherichia coli	Mycobacterium tuberculosis	1ddr	1df7	MTX
12	DNA mismatch repair protein mutS	mutS	Escherichia coli	Thermus aquaticus	1e3m	1fw6	ADP
13	Elongation factor Tu-A	tufA	Thermus thermophilus	Escherichia coli	1ha3	1d8t	GDP
14	Fatty acid-binding protein, adipocyte	Fabp4	Mus musculus	Homo sapiens	1lie	2hnx	PLM
15	Fructose-1,6-bisphosphatase 1	FBP1	Sus scrofa	Homo sapiens	1eyj	1fta	AMP
16	Glucose-6-phosphate isomerase	Gpi	Mus musculus	Oryctolagus cuniculus	2cxr	1dqr	6PG
17	Hemagglutinin-neuraminidase	HN	Newcastle disease virus	Newcastle disease virus	1e8v	1usr	DAN
18	HTH-type transcriptional regulator qacR	qacR	Staphylococcus aureus	Staphylococcus aureus	1jt6	3br1	DEQ
19	Inositol-1-monophosphatase	suhB	Methanocaldococcus jannaschii	Archaeoglobus fulgidus	1g0h	1lbx	IPD
20	Methionine aminopeptidase	map	Escherichia coli	Mycobacterium tuberculosis	1xnz	3iu7	FCD
21	Neuraminidase	NA	Influenza A virus (H11N9)	Influenza A virus (H1N1)	1nnc	3b7e	ZMR
22	NH(3)-dependent NAD(+) synthetase	nadE	Bacillus subtilis	Bacillus anthracis	1ih8	2pz8	APC
23	Orotidine 5'-phosphate decarboxylase	pyrF	Methanobacterium	Pyrococcus horikoshii	1lol	2czf	XMP

			thermoautotrophicum				
24	Peptide deformylase	def	Escherichia coli	Leptospira interrogans	1g2a	1szz	BB2
25	Protein farnesyltransferase subunit beta	Fntb	Rattus norvegicus	Homo sapiens	1d8d	1tn6	FII
26	Protein recA	recA	Mycobacterium smegmatis	Mycobacterium tuberculosis	1ubg	1mo6	DTP
27	Purine nucleoside phosphorylase	PNP	Bos taurus	Homo sapiens	1a9s	1rct	NOS
28	Pyridoxal kinase	PDXK	Homo sapiens	Ovis aries	2yxu	1lhr	ATP
29	Ribulose bisphosphate carboxylase large chain	rbcL	Nicotiana tabacum	Spinacia oleracea	1rlc	1ir1	CAP
30	Shikimate kinase	aroK	Helicobacter pylori	Mycobacterium tuberculosis	hpsa	1zyu	S3P
31	Thymidine kinase	TK	Human herpesvirus 1	Equine herpesvirus 4	1e2j	1p6x	THM
32	Thymidylate synthase	thyA	Escherichia coli	Lactobacillus casei	1aiq	1lca	CB3
33	Tyrosine-protein kinase ABL1	Abl1	Mus musculus	Homo sapiens	1opk	1opl	P16
34	UDP-glucose 4-epimerase	galE	Escherichia coli	Homo sapiens	1lrj	1hzj	UD1
35	UDP-N-acetylglucosamine 1-carboxyvinyltransferase	murA	Enterobacter cloacae	Escherichia coli	1ryw	3iss	EPU
36	Vitamin D3 receptor	VDR	Homo sapiens	Rattus norvegicus	1db1	1rk3	VDX
37	Xylose isomerase	xylA	Streptomyces rubiginosus	Arthrobacter sp.	1xig	1xlc	XYL

Amide, ketone, sulfonate ester, and azine-contained compounds fit into this site. H3 is situated above the central sheet in which two conserved residues (D33, and G80 in HpSK; D34, G80 in MtSK) contribute to H3. Amide, sulfonate ester, and ester groups were frequently identified.

V1, which is adjacent to H1, bears a vdW-binding environment and also contains residues from Walker A motif. V2, in proximity to H3, is situated at the border between shikimate and the nucleotide binding regions. V1 and V2, allowing the interactions with large chemical groups, prefer aromatic groups (over 60% on average). Analysis of the closed-form SiMMaps revealed that E114 and R116 (T115 and R117 in MtSK) located in LID are conserved interacting residues.

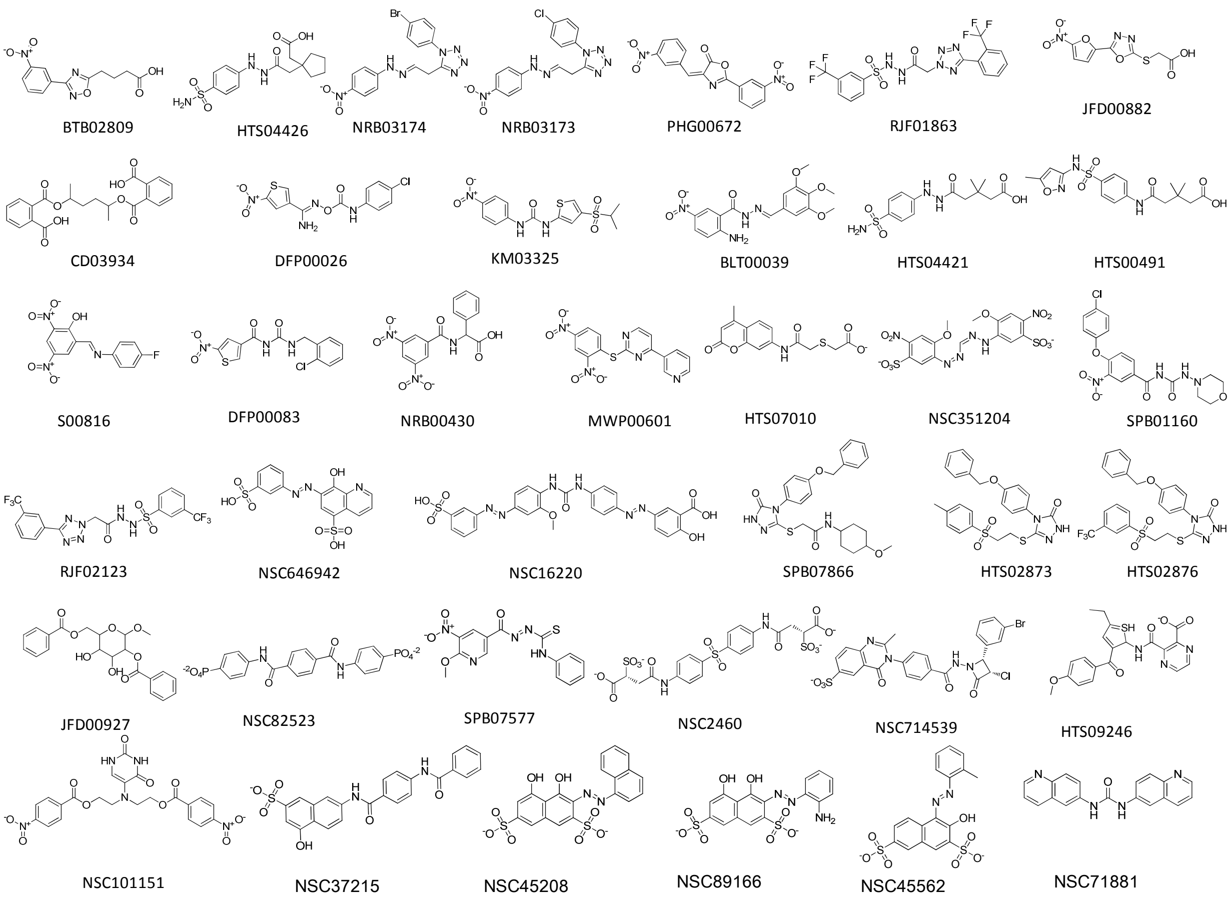


Figure 4.4. Structures of the 38 inactive compounds from the NCI and Maybridge databases.

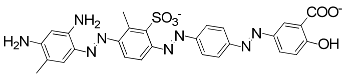
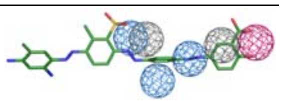
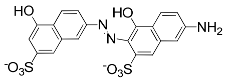
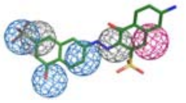
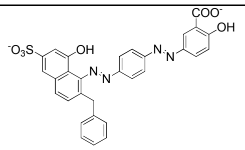
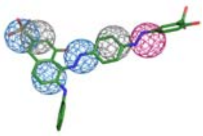
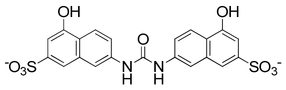
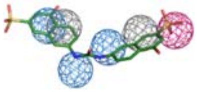
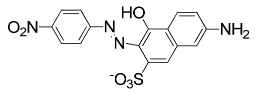
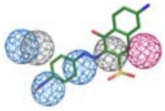
4.3.3 Inhibitors and inhibition assay

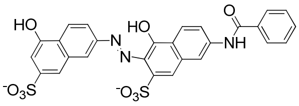
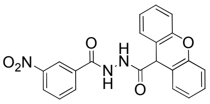
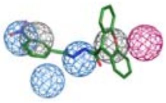
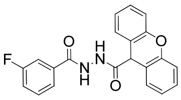
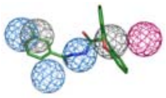
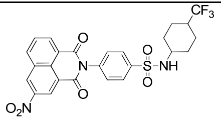
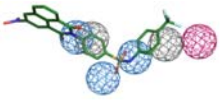
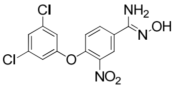
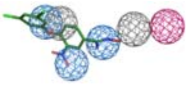
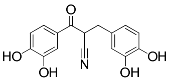
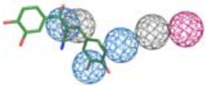
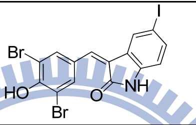
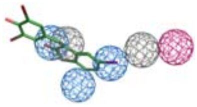
Following the SiMMap analysis, compounds were rescored using the rank-based consensus scoring (RCS[54]), which combines energy-based and anchor-based scoring functions. Since a compound simultaneously docked into apo and closed-form binding sites of orthSiMMaps was considered as a potentially useful hit, we selected common top-ranked compounds from the closed-form and apo-form orthSiMMap analysis for subsequent bioassay. After RCS ranks in both the Maybridge and NCI databases, 48 compounds that were available (either requested or purchased) were then subjected to MtSK and HpSK inhibitory assays (Fig. 4.4). Among those, 10 compounds had an IC_{50} value ≦ 100 μM for both HpSK and MtSK (Table 4.2), in which six (NSC45611, NSC162535, NSC45612, NSC45174, NSC45547, and NSC45609) demonstrated IC_{50} values of ∨∥10 μM (Table 4.3). In parallel, 65 existing kinase inhibitors were tested to evaluate their inhibitory effects against shikimate kinase. Of the two compounds

(AG538 and GW5074) that showed inhibitory effects, AG538 had a low IC_{50} value.

Enzymatic kinetic analysis showed that NSC45611, NSC162535, NSC45612, NSC45174, and AG538 were competitive inhibitors of ATP, in agreement with the docked poses (Table 4.2 and Table 4.3). Of these, NSC45611, NSC162535 and NSC45612 also competed with shikimate. Notably, NSC45611, NSC162535, NSC45612 and NSC45174 had low IC_{50} and αK_i values, showing potent inhibition. Figure 4.5 shows that three (NSC45611, NSC162535, and NSC45612) had lower values of IC_{50} ($\leq$ 10 μM) and fit well over five hot spots (H1, V1, H3, V2, and E1). For those with $IC_{50} \geq 20$ μM, these compounds lack of negatively charged groups to form electrostatic interactions with arginines (R57 and R136 in HpSK) on E1. On the other hand, kinase inhibitors AG538 and GW5074 did not occupy the shikimate site. Moieties with 1−3 rings were present at V1 and V2, yielding a number of vdW contacts. The binding groups of active inhibitors matched well with the identified moieties found from the consensus anchors. For example, the sulfonate groups of NSC162535, NSC45611, and NSC45612 were found to occupy H1. The moieties of NSC162535 (SO_3^- group), NSC45611 (CO_2^- group), and NSC45612 (CO_2^- group) occupied E1.

Table 4.2. Summary of 12 inhibitors with inhibition assay, compound structures, docked poses, and consensus anchors

Compound ID	SK species	IC_{50} (μM)^a	Compound structure	Docked pose ATP site SKM site H1 V1 H2 H3 V2 E1
NSC45611	Hp	4.8		
	Mt	1.5		
NSC162535	Hp	4.9		
	Mt	1.6		
NSC45612	Hp	6.1		
	Mt	2.8		
NSC45174	Hp	7.8		
	Mt	2.8		
NSC45547	Hp	7.8		
	Mt	3.4		

NSC45609	Hp	7.0		
	Mt	2.0		
RH00037	Hp	23.8		
	Mt	<100		
RH00016	Hp	40.2		
	Mt	<100		
GK01385	Hp	<100		
	Mt	<100		
SPB01099	Hp	<100		
	Mt	<100		
AG538	Hp	2.3		
	Mt	0.4		
GW5074	Hp	31.4		
	Mt	29.6		

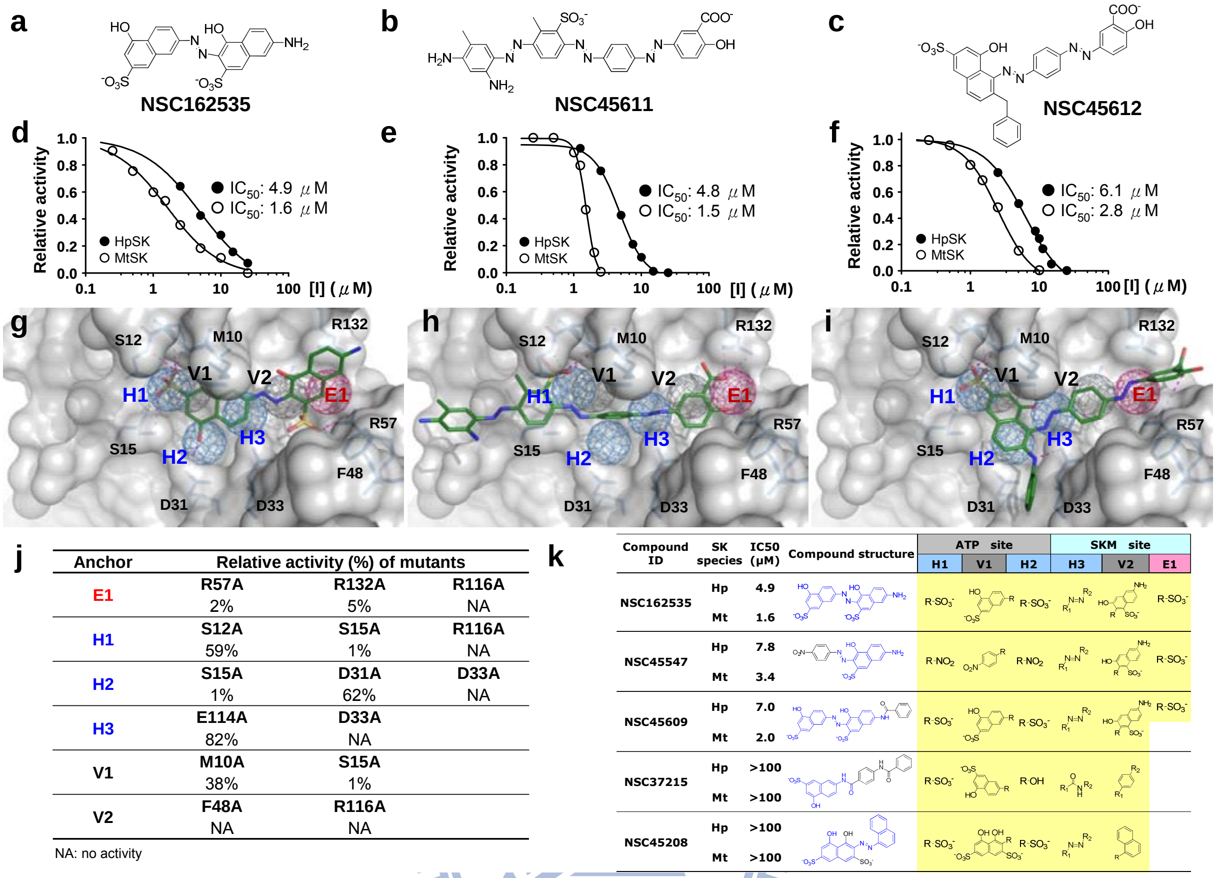
[a] The inhibition assay is done by Dr. Wen-Ching Wang of National Tsing Hua University[51].

Table 4.3. Properties of potent inhibitors for HpSK and MtSK [a]

Compound ID	SK species	Inhibition mode[b]		αKi, ATP (μM)	αKi, SKM (μM)
		ATP	SKM		
NSC45611	Hp	■	■	1.1	1.7
	Mt	■	■	0.3	0.7
NSC162535	Hp	■	■	1.9	1.8
	Mt	■	■	0.2	0.6
NSC45612	Hp	■	■	2.0	2.4
	Mt	■	■	0.7	1.0
NSC45174	Hp	■	□	1.7	12.8
	Mt	■	□	0.4	2.7
AG538	Hp	■	□	3.1	5.4
	Mt	■	□	0.04	0.4

[a] The inhibition assay is done by Dr. Wen-Ching Wang of National Tsing Hua University[51].

[b] ■: Competitive inhibition; □: Non-competitive inhibition



Anchor	Relative activity (%) of mutants		
E1	R57A 2%	R132A 5%	R116A NA
H1	S12A 59%	S15A 1%	R116A NA
H2	S15A 1%	D31A 62%	D33A NA
H3	E114A 82%	D33A NA	
V1	M10A 38%	S15A 1%	
V2	F48A NA	R116A NA	

NA: no activity

Compound ID	SK species	IC50 (μM)
NSC162535	Hp	4.9
	Mt	1.6
NSC45547	Hp	7.8
	Mt	3.4
NSC45609	Hp	7.0
	Mt	2.0
NSC37215	Hp	>100
	Mt	>100
NSC45208	Hp	>100
	Mt	>100

Figure 4.5. Characterization of shikimate kinase inhibitors by enzyme assay, orthSiMMaps, site-mutagenesis studies and analogues. (a–c) Structures of three inhibitors, NSC162535, NSC45611, and NSC45612. (d–f) The inhibitions of these compounds were analyzed by enzyme IC$_{50}$ test on HpSK (filled) and MtSK. (g–i) The relationship between anchors and the docked mode of each inhibitor for HpSK. These compounds consistently include two negative charge moieties (SO_3^- or CO_2^-) that form hydrogen bonds with conserved interactions residues of anchors E1 and H1. (j) Comparison of relative activities of HpSK mutants. The conserved interacting residues for each anchor were mutated, respectively. R57, R132, R116, and F48 located in the shikimate site are critical for the enzymatic functions. (k) The potency of NSC162535 analogues. The substitution moieties of analogues are indicated in black. Those that lack the E1 moiety greatly lost the inhibitory effects (IC$_{50}$ >100 μM). The inhibition assay is done by Dr. Wen-Ching Wang of National Tsing Hua University[51].

4.3.4 Site-directed mutagenesis

A consensus anchor of orthologous targets, identified from the conserved binding pockets shared with conserved interacting residues and specific physico-chemical property, usually engages with specific functions in the enzymatic catalysis. We sought to investigate the roles of identified consensus anchor residues of the orthSiMMaps in catalysis. The site-directed mutagenesis study is done by Dr. Wen-Ching Wang of National Tsing Hua University[51]. We first investigated mutants of E1 residues (R57, R116, and R132) that contact with shikimate[55]. Enzymatic analysis revealed that these arginines had extremely low activity (Fig. 4.5j), suggesting the importance of these residues in catalysis. Indeed, R117 of MtSK that corresponds to R116 of HpSK has thus been suggested as a primary candidate to stabilize the transition-state intermediate[56].

For the H3 (D33) and V2 (F48) residues, D33A completely lost the enzymatic activity while F48A exhibited hardly any detectable activity (1%). D33 and F48 are in direct contact with shikimate. More importantly, it should be noted that D33 forms a hydrogen bond to the 3-OH group of shikimate, which may increase the nucleophilicity of the O atom or accept the proton from the 3-OH group of shikimate, facilitating the catalysis. E114A, a LID residue whose side chain faces the solvent, retained 82% relative activity. On the other hand, the F48 side chain contacts with those from several residues nearby (V44, E53, F56, R57 and P117), which may form a stable platform to interact with the ligand for subsequent catalytic reaction.

We then evaluated residues from H1, H2, and V1 located at the nucleotide site. H1 residues are primarily from the Walker A motif (P loop; residues 11−16, GSGKSS) surrounding the phosphate groups of the nucleotides. Of the three mutants (S12A, S15A, and S16A), S12A and S16A remained >50% of the relative activity, while S15A had extremely low activity (1%). The S15 side chain resides nearby the β-phosphate of ADP. Furthermore, the adjacent lysine (K14) corresponding to K15 of MtSK has been identified as a critical catalytic residue in MtSK since its side chain points toward the γ-phosphate[56]. The other H2 mutant D31A retained 62% of the relative activity (62%), possibly due to its remote location to the phosphate group. For V1 that is just next to H1, several H1 residues are also shared by V1. Enzymatic analysis showed that M10A remained 38% relative activity. These results suggest that the conserved interacting residues from E1 (R57, R116 and R132), H1 (S15 and R116), H2 (S15 and D33), H3 (D33), V1 (S15) and V2 (F48) contribute significantly to catalytic power and substrate binding.

4.3.5 Analogues assay and orthSiMMap

To validate the moiety preferences of consensus anchors, we identified four analogues (NSC45547, NSC45609, NSC37215, and NSC45208) of NSC162535 for inhibitory assays (Fig. 4.5k). NSC45547 and NSC45609 that occupy E1 (SO_3^- group) and H1 (SO_3^- and NO_2 groups) retained good IC_{50} values (7.8 and 7.0 µM for HpSK; 3.4 and 2.0 µM for MtSK). Conversely, NSC37215 and NSC45208, that cannot anchor at E1, lost the inhibitory.

To evaluate the significance of pocket-moiety interaction preferences of consensus anchors in the orthSiMMaps, we performed clustering analysis on 27 inhibitory assay compounds. These compounds can be roughly clustered into three groups (Fig. 4.6). The potent inhibitors of group I (NSC162535, NSC45609, NSC45547, NSC45174, NSC45611, and NSC45612) match more than 5 consensus anchors (Fig. 4.5g-i, and Table 4.2). For Group II compounds (RH00037, RH00016, GK01385, and SPB01099), each compound matches four of six anchors; Group III are kinase inhibitors (AG538 and GW5074) and these compounds share anchors of ATP site. For inactive compounds, there are fewer matched consensus anchors in the HpSK/MtSK (usually 4), particularly E1 is the least seen. While the inhibitors of group I and II agreed with anchors of ATP site and shikimate site, the kinetic assay showed competitive inhibitions for ATP and shikimate acid (Table 4.3). The kinase inhibitors of group III occupied the anchors of ATP site, and only showed the competitive inhibitions for ATP. Generally, the pocket environment of ATP is conserved for kinase family, and the inhibitors of group III also have the broadband inhibition for multiple kinases, such as the inhibition of AG538 observed on insulin-like growth factor-1 receptor (IGF-1R)[57], IR, EGFR[58], and Src kinases[59].

While there are the same number of consensus anchors (E1, H1, H2, H3, V1 and V2), the spatial arrangement of these anchors were closer in the closed form (Fig. 4.2c and 4.2d). Residues (D31 and D33) that contribute to H2 of the apo form were in closer proximity in the closed conformation, resulting in a reduced volume at this site. Likewise, the corresponding site at V2 surrounded by F48, G80, and G81 in HpSK had less space in the closed form, hindering the accommodation of large moieties carrying one or two rings at this site. The above evidences demonstrate that induced LID conformation of shikimate kinases was sensitive in the structure-based drug discovery strategy.

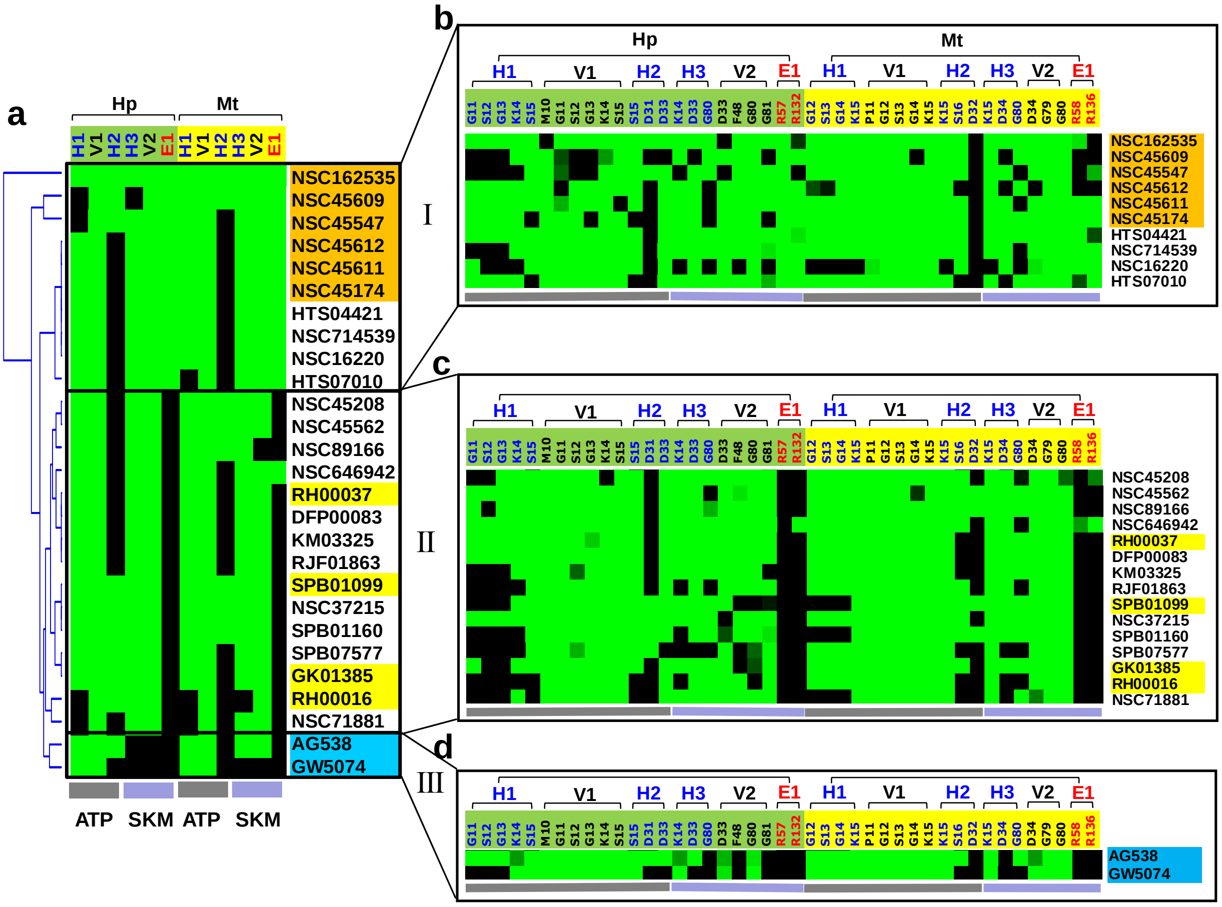


Figure 4.6. Interaction profiles between selected anchor residues and 27 tested compounds. (a) The anchor profile of tested compounds on shikimate kinases. (b) Group I: the NCI compounds (orange). (c) Group II: the Maybridge compounds (yellow). (d) Group III: kinase inhibitors (cyan). The NCI compounds consistently occupy anchors E1 and V2 locating in both ATP and shikimate sites. Except for NSC45174, the NCI compounds are competitive inhibitors with both ATP and shikimate. For the Maybridge compounds, none form electrostatic interactions with R57 and R132 on the consensus anchor E1. The two kinase compounds are located at the ATP site, in good agreement with the kinetic results showing that they exhibited competitive inhibition with ATP and noncompetitive inhibition with shikimate.

4.3.6 Structural mechanism of the inhibitor binding for shikimate kinases

Superposition of various structures (apo HpSK, HpSK·shikimate·PO_4, HpSK·S3P·ADP, and E114A·162535) reveals a significant conformational change in the LID-containing segment after β4 of the CORE domain (residues 101 to 138; α5, LID and α6) (Fig. 4.7). Furthermore, the SB region (residues 32−60) shows a small rotation among the different unliganded/liganded states, in accord with MtSK structures[53].

Of the three conserved arginines (R57, R116, and R132), it is noted the Cα atom of R57 superimposes relatively well, while that of R132 has a small shift among various structures (Fig. 4.7a-4.7d). A shorter Cα-atom (R57-R132) distance (~0.6 Å) is noted for HpSK·S3P·ADP and HpSK·shikimate·PO_4 as compared to the apo-form HpSK. On the other hand, there is a significant drift for R116 due to the distinct conformations of the LID loop (Fig. 4.7a-4.7d). Our results suggest that these arginines contribute to the movement of the lid region and the shikimate-binding domain upon ligand binding. R116, when visible, makes a significant shift to form hydrogen-bonding interactions with various ligands in the binding pocket: (i) shikimate in HpSK·shikimate·PO_4; (ii) β-phosphate of ADP in HpSK·S3P·ADP; and (iii) NSC162535 in E114A·162535. In the MtSK·shikimate·AMPPCP structure, a direct contact is also observed between R117 (corresponding to R116 in HpSK) and γ-phosphate of AMPPCP, an ATP analogue, which supports its catalytic role in the γ-phosphoryl transfer[56].

To evaluate whether NSC162535 will come in contact with R116 in other forms, we have docked NSC162535 into the binding pockets of various HpSK structures (Fig. 4.7e-4.7i). In the apo form, HpSK with a flexible LID presents a wide-opening pocket, allowing entry of promising substrates (Figs. 4.7a, 4.7e and 4.7j). No close contacts are found between R116 and the docked NSC162535 in the binding pockets of the apo and HpSK·shikimate·PO_4 forms (Fig. 4.7e and 4.7f). In the HpSK·S3P·ADP state, NSC162535 is docked into a site where the Nη1 and Nη2 of the guanidino group in R116 make no significant contacts. NSC162535, on the other hand, is docked into a comparable site in the E114A·162535 form, where it contacts directly with the Nη1 and Nη2 atoms of R116 just like that of the crystal structure. Thus, it is likely that R116 plays a crucial role during the course of a conformational cycle in conducting a catalytic event (Fig. 4.7i and 4.7j). Upon diffusion into the binding pocket, NSC162535 that carries two SO_3^- and a -N=N- groups may bind to the active site, interact with R57 and R132,

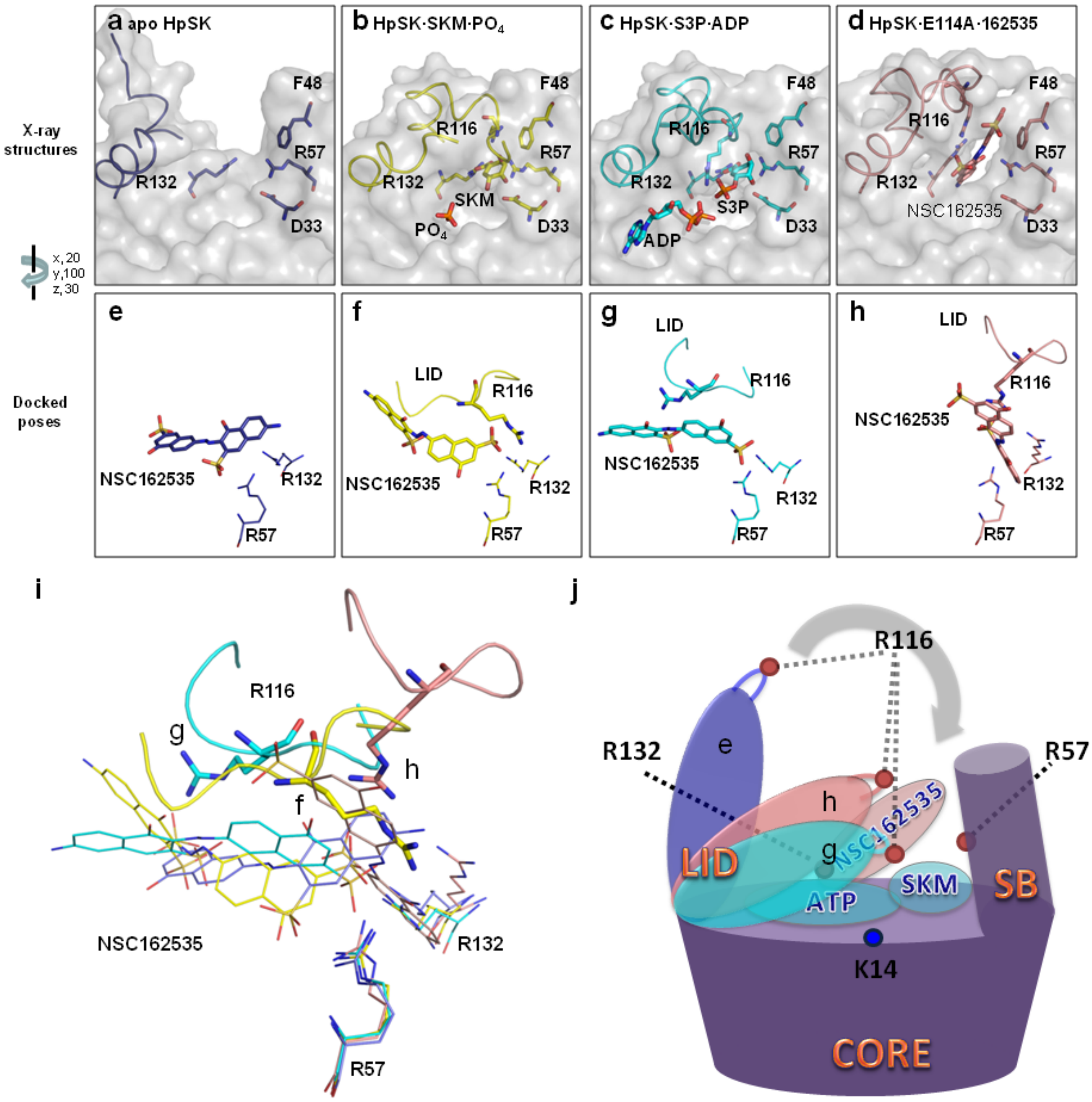


Figure 4.7. Probing the affinity pockets in HpSK. (a-d) The binding pockets of HpSK (a), HpSK·shikimate·SO$_4$ (b), HpSK·S3P·ADP (c), and E114A·162535 (d) structures. The bound ligands, D33, F48, R57, R116, and R132 are drawn as sticks. The LID segments (residues 109–123) are drawn as the ribbon structures. (e–h) The docked NSC162535 models in the binding pockets of HpSK (e), HpSK·shikimate·SO$_4$ (f), HpSK·S3P·ADP (g), and E114A·162535 (h) structures. Superposition of three residues (R57, R116, and R132), docked and bound NSC162535 among HpSK (blue), HpSK·SKM·SO$_4$ (yellow), HpSK·S3P·ADP (cyan), and E114A·162535 (orange) structures. (i) Superimposed docked structures (e–h). The conformation of LID segment (residues 113–119, ribbon) having R116 (thick stick) demonstrates the greatest conformational changes induced by bound ligands. (j) Schematic

diagram of induced-fit conformational changes upon binding to ligands. The view of LID regions corresponding to apo HpSK, HpSK·S3P·ADP, and E114A·162535 is colored as blue, cyan, and orange, respectively. The crystallized structures studies are done by Dr. Wen-Ching Wang of National Tsing Hua University.

and then trigger a conformational change cycle. As a result, R116 along with R57 and R132 will trap the inhibitor, yielding an optimized anchor (E1). These results also suggest an unusually elastic LID region, which allows accommodating various ligands, as demonstrated here. In this section, the crystallized structures studies are done by Dr. Wen-Ching Wang of National Tsing Hua University.

4.3.7 Performance of the orthSiMMap method

We then evaluated the accuracy of the orthSiMMap approach. The orthSiMMap score (solid lines) significantly outperformed those (dashed lines) of energy-based scoring methods, which are often used in docking tools, on apo-form HpSK and MtSK (Fig. 4.8a). The average enrichments of 3.73 (HpSK), 1.59 (MtSK), and 2.74 (fusion of HpSK and MtSK) were obtained using energy-based scoring methods, as compared to 11.18 (HpSK), 35.51 (MtSK), and 93.69 (fusion of HpSK and MtSK) using the orthSiMMap scoring method. Additionally, the orthSiMMap scores exhibited a higher accuracy than that of the SiMMap score from a single target (HpSK or MtSK).

The orthSiMMap is able to reduce the deleterious effects of screening ligand structures that are rich in charged or polar atoms. Generally, energy-based scoring functions favor the selection of high-molecular-weight compounds yielding high vdW potentials, as well as polar compounds that produce hydrogen-bonding and/or electrostatic potentials[54]. The average molecular weights of the top ranked 100 compounds of the orthSiMMap and the energy-based scoring methods were 459.9 and 532.6, respectively; the average numbers of polar atoms were 11.3 (orthSiMMap method) and 14.1 (energy-based method) (Fig. 4.8b,c). The ranks of those 10 active compounds were much higher in the orthSiMMap scoring analysis than in the energy-based analysis. It should be noted that NSC162535 was ranked as 1 and 1821 using the apo-form orthSiMMap and energy-based scoring methods, respectively (Table 4.4).

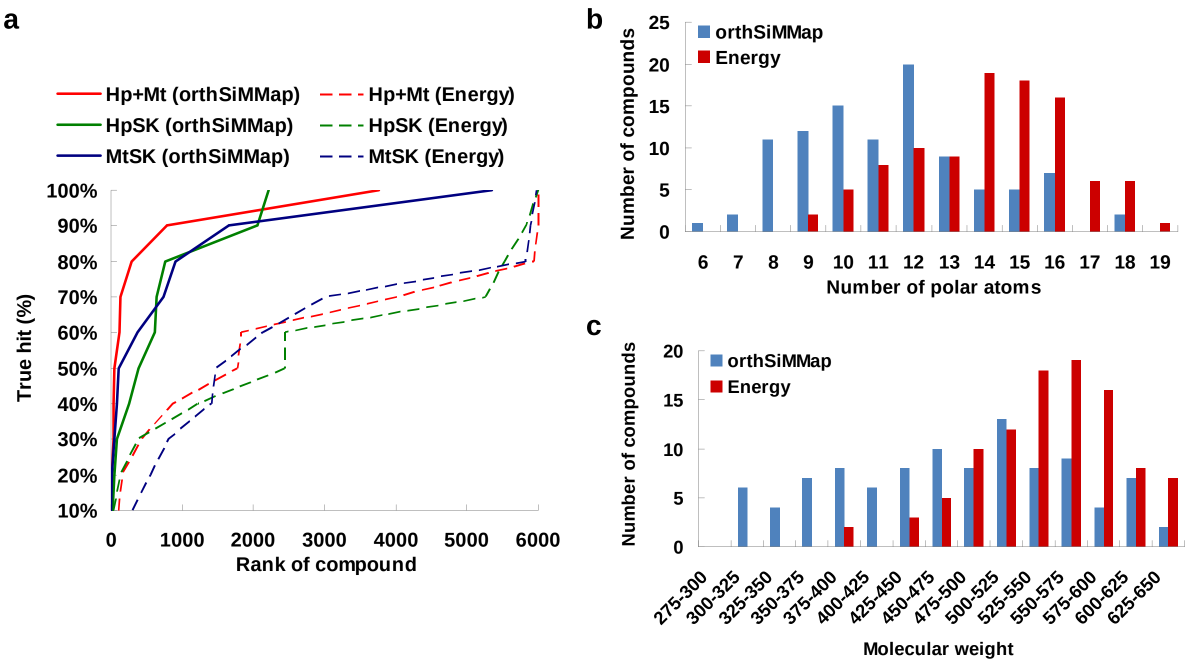
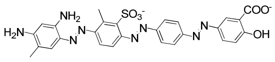
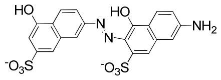
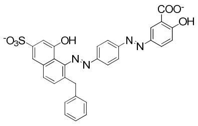
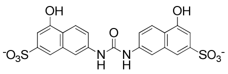
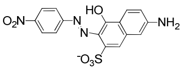
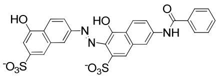
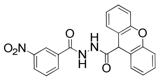
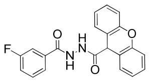
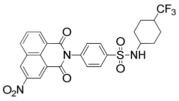
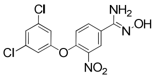


Figure 4.8. Performance of the orthSiMMap method on apo-form HpSK and MtSK. (a) The true-hit rates of energy-based and orthSiMMap scoring approaches. The orthSiMMap scores (solid line) of adaptive inhibitors significantly outperform energy-based scores (dashed line) using the top ranked 6000 compounds by combining the Maybridge and NCI databases. (b) Distribution of number of polar atoms, and (c) molecular weight of top 100 compounds from orthSiMMap scores and energy-based score.

Table 4.4. The ranks of active compounds using orthSiMMap, energy-bases, and combination scoring methods for apo and closed forms of HpSK and MtSK

Compound ID	Compound structure	Apo form		Closed form		RCS [a]
		orthSiMMap	Energy	orthSiMMap	Energy	
NSC45611		48	435	515	827	96
NSC162535		1	1821	242	253	25
NSC45612		32	106	238	229	31
NSC45174		38	162	737	110	147

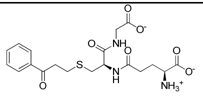
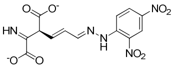
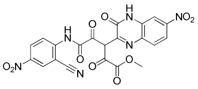
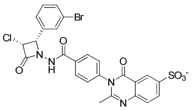
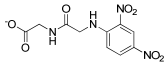
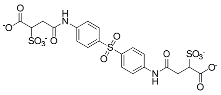
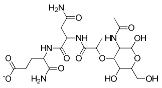
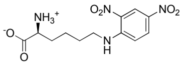
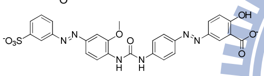
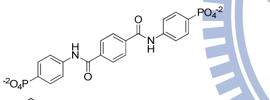
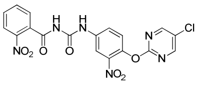
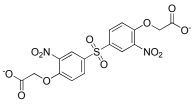
NSC45547		130	5999	308	1540	67
NSC45609		18	4017	5	17	3
RH00037		786	876	891	1549	371
RH00016		3765	6000	1219	5824	1837
GK01385		286	1774	730	49	199
SPB01099		117	5940	68	2871	19

[a] The rank is the rank combination of orthSiMMap and energy.

4.4 Summary

The largest obstacle by far in structure-based drug discovery is the relatively low hit rates in scoring methods due to the lack of adequate quantities of binding partners for a given target. In other words, there is no adequate training set to establish the veracity or utility of an algorithm. Under these circumstances, the accuracy of a given individual scoring function is generally unknown and/or cannot be evaluated at a critical point. The current emphasis of the orthSiMMap scoring developed here thus provides a useful index to improve the screening accuracy for identification of adaptive inhibitors when the target proteins shared conserved binding sites. Through the employment of this developed method, we successfully found six new potent inhibitors (<8.0 μM) of HpSK and MtSK. Two of the 65 kinase inhibitors were also found to inhibit both HpSK and MtSK activity. The finding that NSC45611, NSC162535, and NSC45612 were competitive inhibitors of ATP and shikimate suggests that they belong to a novel class of shikimate kinase inhibitors. Based on the novel inhibitor - NSC162535, the inhibitor complex crystal structure, E114A·162535, was determined by Dr. Wang's group of National Tsing Hua University. These results illustrate a robust orthSiMMap-based approach to identify selective kinase inhibitors.

Table 4.5. Some selected top-ranked compounds using orthSiMMap, energy-bases, and combination scoring methods for apo and closed forms of HpSK and MtSK

Compound ID	Compound structure	Apo form		Closed form		RCS [a]	Bioassay
		iSiMMap	Energy	iSiMMap	Energy		
NSC131133		2	2153	7	456	1	-[b]
NSC407257		3	91	17	317	2	-
NSC644745		430	1	3152	745	1085	-
NSC714539		431	3	1	9	64	Inactive
NSC524127		920	3090	2	189	175	-
NSC2460		313	2633	4	2073	49	Inactive
ZINC05823979		1321	578	8	1021	265	-
NSC83262		11	28	21	1	5	-
NSC16220		728	37	10	2	137	Inactive
NSC82523		13	170	15	238	4	Inactive
NSC624285		39	199	22	8	8	-
NSC85597		28	68	31	476	6	-

[a] The rank is the rank combination of orthSiMMap and energy.

[b] The compound is not tested.

The developed orthSiMMap scoring method appears to outperform the energy-based method (Tables 4.4 and 4.5). Of six potent inhibitors, it was interesting to find that aside from NSC45609, the others have a higher rank in the apo-form than in the closed-form orthSiMMap scoring analysis. Additionally, the top-ranked inhibitors from the apo-form orthSiMMap scoring analysis often possess larger moieties (e.g. naphthalene or nitrobenzene) at both sides as opposed to those with a relatively small moiety (e.g. amide or aliphatic chain). The closed-form orthSiMMap scoring analysis has, nonetheless, yielded useful hits including NSC45609 and SPB01099.

P-loop kinase fold consists of functionally diverse kinase classes, such as shikimate kinase, NTPases and GTPases[60]. They frequently share conserved binding environments (e.g., P-loop and walker A/B motifs) for interacting with partners (e.g., small compounds and proteins). The molecules inhibit P-loop kinases that play a key role in various diseases, such as cancer, cardiovascular diseases, gastric diseases or infections. Although a number of inhibitors in clinical trials [61-63] or on the market (omeprazole and ciprofloxacin) inhibit the activity of P-loop kinases, few of them bind to the ATP-binding site[64]. Meanwhile, target proteins with dynamic induced-fit forms, like the P-loop SKs, represent a major limitation for the structure-based screening approach. The approach of orthSiMMap designing the competitive ATP inhibitors with specific substrate pocket presents a novel strategy of targeting P-loop kinases.

The developed orthSiMMap method is database independent. Comparable anchors were identified in compounds from the Maybridge and NCI databases. Each of the anchors also included analogous chemical moieties. Nonetheless, the derived proportion of these moieties was different because the Maybridge and NCI databases contain heterogeneous distribution of compounds. For example, the proportion of carboxyl, sulfonate, and phosphate was significantly higher in compounds from the NCI database than in those from the Maybridge database. On the other hand, the derived model was sensitive to binding-site properties, as illustrated by the difference between the apo- and closed-form models (Fig. 4.2). In summary, we anticipate that the orthSiMMap method can be useful in discovering new inhibitors, investigating the binding mechanisms, and guiding the lead optimization for orthologous targets. Additionally, crystal structures reveal the details of ligand binding in the induced-fit P-loop kinases and will be valuable in the development of novel P-loop kinase inhibitors.

Chapter 5
Conclusion

5.1 Summary

Briefly, the major contributions of this thesis can be summarized in the following:

(1) The concept of site-moiety map (SiMMap) was proposed for predicting protein-ligand binding modes and characterizing protein-ligand binding sites in structure-based drug design. SiMMap statistically infers the site-moiety map describing the relationship between the moiety preferences and physico-chemical properties of the binding site. Our experimental results showed that the site-moiety map is useful to reflect biological functions and identify active compounds from thousands of compounds. In addition, the site-moiety map can guide to assemble potential leads by optimal steric, hydrogen-bonding, and electronic moieties.

(2) Members of individual protein families often share a homologous fold and conserved structural features to interact with chemically similar ligands throughout evolution, despite low sequence identity. A structure-based site-moiety screening method, orthSiMMap, was developed to discover the inhibitors for a family of orthologous proteins. Here, we utilized the orthSiMMap to pharmacologically interrogate orthologous shikimate kinases (SKs) from *Mycobacterium tuberculosis* and *Helicobacter pylori*. The derived apo/closed core site-moiety maps and the anchor scores were used to identify six potent inhibitors (<8.0 μM). Site-directed mutagenesis (these studies done by Dr. W.C. Wang of National Tsing Hua University) and analogues studies revealed that critical conserved interacting residues contribute to a given pocket-moiety interaction spot. Crystal structures of HpSK·SO_4, R57A, HpSK·shikimate-3-phosphate·ADP, and E114A·162535 (These structures obtained by Dr. Wang, W.C. of National Tsing Hua University) show a characteristic three-layer architecture and a conformationally elastic region having R57, R116, and R132 occupied by shikimate/inhibitor, locking into an induced-fit form. These results illustrate a robust approach in identifying selective inhibitors and reveal insight to the active site chemistry of SKs and a new induced-fit mechanism by an inhibitor. We believe that the SiMMap is able to provide the biological insights of protein-ligand binding models, enrich the screening accuracy, and guide the processes of lead optimization.

5.2 Future works

The one-disease, one-target, and one-drug philosophy has been the dominating drug discovery approach in the past decades. Drugs against multiple targets may overcome the many limitations of single targets and achieve a more effective and safer control of diseases. [65-66] However, to design a selective drug structure which is able to against multiple targets is still a challenge task. The NAD(P) and ATP related pathways play key roles in various biological functions, such as aromatic amino acid synthesis, pyrimidine metabolism and TCA cycle regulation. In these pathways, the NAD(P) and ATP enzymes usually process a sequentially enzymatic reactions to transform specific substrates to products (Fig. 5.1). For maintaining their catalytic functions, these proteins have a cofactor site (ATP or NAD(P)) next a substrate binding site. From the substrate similarity, ATP and NAD(P) have a similar scaffold, such as adenosine and di-phosphate group. Therefore, It is possible to develop selectively multi-targeted inhibitors on a specific pathway through considering the specific substrate site and the similar cofactor site of ATP and NAD(P) related enzymes.

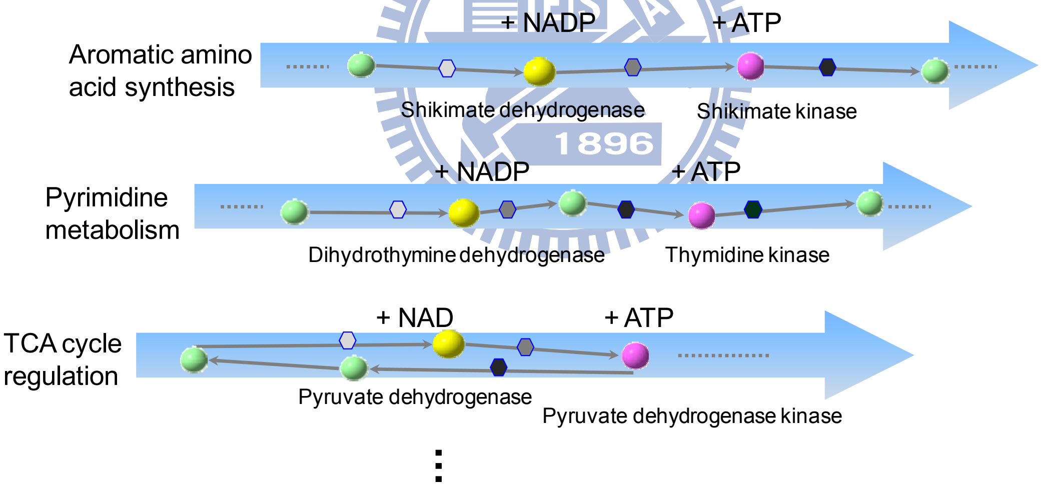


Figure 5.1. The NAD(P) and ATP related pathways play key roles in various biological functions, such as aromatic amino acid synthesis, pyrimidine metabolism and TCA cycle regulation.

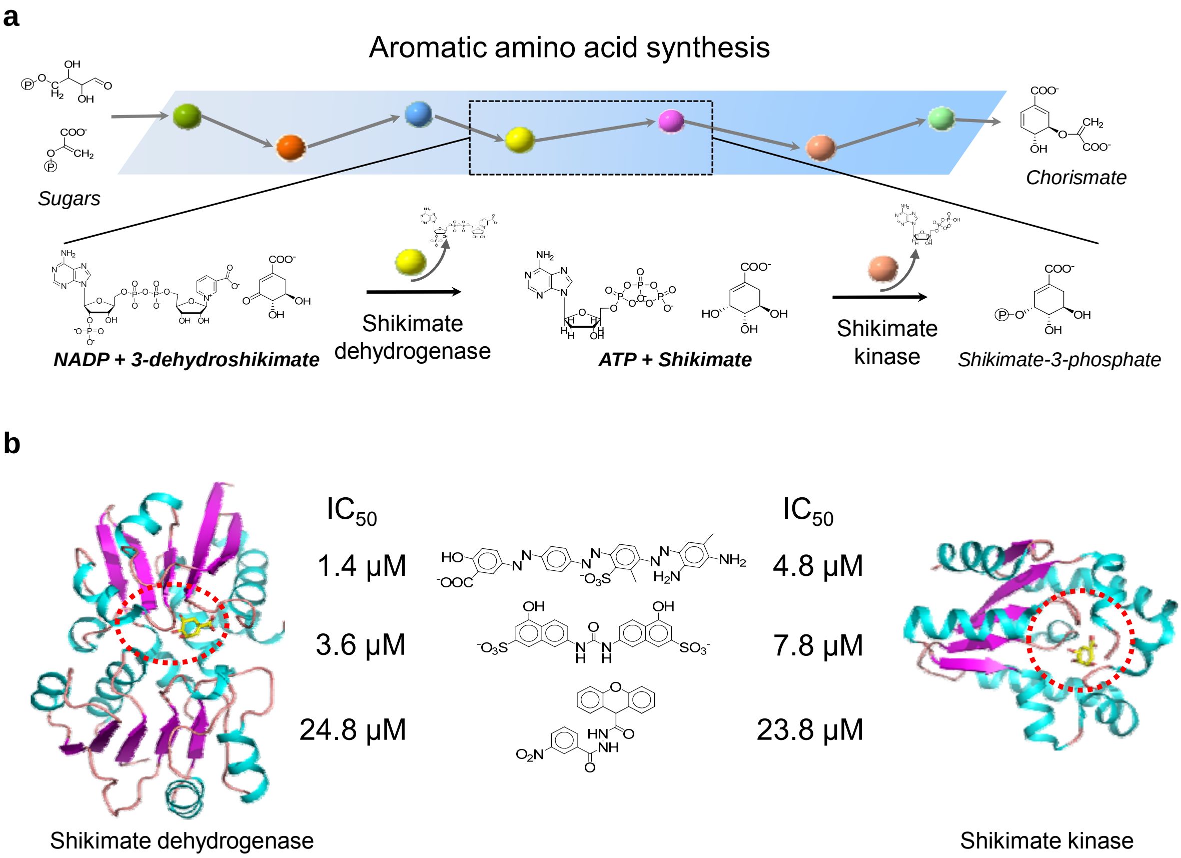


Figure 5.2. Preliminary result of the PathDrug on the shikimate pathway. (a) The shikimate pathway includes seven enzymes to convert erythrose 4-phospate and phosphoenolpyruvate into chorismate. Shikimate dehydrogenase (SDH) and shikimate kinase (SK) are selected as targeting proteins for developing PathDrug. (b) Three compounds structures inhibit both SDH and SK.

To address these issues, we extend our previous studies to propose a new concept, named PathDrug. The core idea of PathDrug is to identify and integrate the consensus binding environments (site-moiety maps) of the enzymes on the same pathway. Using the consensus map of the protein targets on the same pathway to discover the multi-target leads and then guide to optimize the selectivity for a specific pathway. To validate the concept of PathDrug, we cooperate with Dr. W.C. Wang of National Tsing Hua University and select the shikimate dehydrogenase (SDH) and shikimate kinase (SK) of *Helicobacter pylori* to develop the PathDrug. Figure 5.2 shows the preliminary result of dual-targeted inhibitors for SDH and SK. The inhibitory assay was done by Dr. W.C. Wang of National Tsing Hua University. From the PathDrug map of SDH and SK, three inhibitor structures are identified and shown the dual-inhibition. These result preliminarily demonstrate the possibility of PathDrug and provide a potential direction to develop the drug with high affinity and low resistance.

The study of SiMMap and orthSiMMap in this thesis enables us to identify PathDrug, investigate the consensus properties of PathDrug, and discovery pathway-specific inhibitors. In the future, we believe that the designed highly-specific compounds with activity against disease-related pathway can help us to reduce drug resistance and enhance the lead activity.

Appendix A

List of publications

Journal papers

1. **Y.-F. Chen**, K.-C. Hsu, S.-R. Lin, W.-C. Wang, Y.-C. Huang and J.-M. Yang*, " *SiMMap*: a web server for inferring site-moiety map to recognize interaction preferences between protein pockets and compound moieties," Nucleic Acids Research, 2010

2. D. Clinciu, **Y.-F. Chen**, C.-N. Ko, C.-C. Lo and J.-M. Yang*, "TSCC: Two-Stage Combinative Clustering for Virtual Screening Using Protein-ligand Interactions and Physical-Chemical Features, " BMC Genomcis, to be published.

3. K.-C. Hsu, **Y.-F. Chen**, and J.-M. Yang*,"GemAffinity: a scoring function for predicting binding affinity and Virtual Screening", International Journal of Data Mining and Bioinformatics, to be published.

4. H.-C. Hung, C.-P. Tseng, J.-M. Yang, Y.-W Ju, S.-N. Tseng, **Y.-F. Chen**, Y.-S. Chao, H.-P. Hsieh, S.-R. Shih, John T.-A. Hsu, "Aurintricarboxylic acid inhibits influenza virus neuraminidase," Antiviral Research, vol. 81, pp. 123-131, 2009.

5. J.-M. Yang, **Y.-F. Chen**, Y.-Y. Tu, K.-R. Yen, and Y.-L. Yang*, “Combinatorial computation approaches identifying tetracycline derivates as flaviviruses inhibitors,” *PLoS* ONE, pp. e428.1- e428.12, 2007.

6. J.-M. Yang*, **Y.-F. Chen**, T.-W. Shen, B. S. Kristal, and D. F. Hsu, "Consensus Scoring Criteria for Improving Enrichment in Virtual Screening," *Journal of Chemical Information and Modeling*, vol. 45, pp. 1134-1146, 2005

7. J.-M. Yang*, T.-W. Shen, **Y.-F. Chen**, Y.-Y. Chiu, "An evolutionary approach with pharmacophore-based scoring functions for virtual database screening," Lecture Notes in Computer Science, vol. 3102, pp. 481-492, 2004

Conference paper

1. K.-C. Hsu, **Y.-F. Chen**, and J.-M. Yang*, "GemAffinity: a Scoring function for predicting Binding Affinity and Virtual Screening," bibm, pp.309-314, 2009 IEEE International Conference on Bioinformatics and Biomedicine, 2009

References

1 Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews Drug Discovery* **3**, 935-949 (2004).

2 Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discovery Today* **7**, 1047-1055 (2002).

3 Shoichet, B. K., McGovern, S. L., Wei, B. & Irwin, J. Lead discovery using molecular docking. *Current Opinion in Chemical Biology* **6**, 439-446 (2002).

4 Powers, R. A., Morandi, F. & Shoichet, B. K. Structure-based discovery of a novel, noncovalent inhibitor of AmpC beta-lactamase. *Structure* **10**, 1013-1023 (2002).

5 An, J. *et al.* A novel small-molecule inhibitor of the avian influenza H5N1 virus determined through computational screening against the neuraminidase. *Journal of Medicinal Chemistry* **52**, 2667-2672 (2009).

6 Hajduk, P. J. & Greer, J. A decade of fragment-based drug design: strategic advances and lessons learned. *Nature Reviews Drug Discovery* **6**, 211-219 (2007).

7 Hung, H. C. *et al.* Aurintricarboxylic acid inhibits influenza virus neuraminidase. *Antiviral research* **81**, 123-131 (2009).

8 Schapira, M. *et al.* Discovery of diverse thyroid hormone receptor antagonists by high-throughput docking. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 7354-7359 (2003).

9 Tanrikulu, Y. & Schneider, G. Pseudoreceptor models in drug design: bridging ligand- and receptor-based virtual screening. *Nature Reviews Drug discovery* **7**, 667-677 (2008).

10 Yang, J. M., Chen, Y. F., Tu, Y. Y., Yen, K. R. & Yang, Y. L. Combinatorial computational approaches to identify tetracycline derivatives as flavivirus inhibitors. *PloS one* **2**, e428 (2007).

11 Doman, T. N. *et al.* Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *Journal of Medicinal Chemistry* **45**, 2213-2221 (2002).

12 Gohlke, H., Hendlich, M. & Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *Journal of Molecular Biology* **295**, 337-356 (2000).

13 Weiner, S. J. *et al.* A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society* **106**, 765-784 (1984).

14 Gehlhaar, D. K. *et al.* Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chemistry & Biology* **2**, 317-324 (1995).

15 Charifson, P. S., Corkery, J. J., Murcko, M. A. & Walters, W. P. Consensus scoring: A

method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *Journal of Medicinal Chemistry* **42**, 5100-5109 (1999).

16 Ewing, T. J., Makino, S., Skillman, A. G. & Kuntz, I. D. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *Journal of Computer-Aided Molecular Design* **15**, 411-428 (2001).

17 Bissantz, C., Folkers, G. & Rognan, D. Protein-based virtual screening of chemical databases. 1.evaluation of different docking/scoring combinations. *Journal of Medicinal Chemistry* **43**, 4759-4767 (2000).

18 Stahl, M. & Rarey, M. Detailed analysis of scoring functions for virtual screening. *Journal of Medicinal Chemistry* **44**, 1035-1042 (2001).

19 Verdonk, M. L. *et al.* Virtual screening using protein-ligand docking: avoiding artificial enrichment. *Journal of Chemical Information and Computer Sciences* **44**, 793-806 (2004).

20 Wang, R. & Wang, S. How does consensus scoring work for virtual library screening? An idealized computer experiment. *Journal of Chemical Information and Computer Sciences* **41**, 1422-1426 (2001).

21 Fradera, X., Knegtel, R. M. A. & Mestres, J. Similarity-driven flexible ligand docking. *Proteins: Structure, Function, and Bioinformatics* **40**, 623-637 (2000).

22 Hindle, S. A., Rarey, M., Buning, C. & Lengauer, T. Flexible docking under pharmacophore type constraints. *Journal of Computer-Aided Molecular Design* **16**, 129-149 (2002).

23 Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology* **267**, 727-748 (1997).

24 Kramer, B., Rarey, M. & Lengauer, T. Evaluation of the flexX incremental construction algorithm for protein-ligand docking. *Proteins: Structure, Function, and Bioinformatics* **37**, 228-241 (1999).

25 Pan, Y., Huang, N., Cho, S. & MacKerell, A. D., Jr. Consideration of molecular weight during compound selection in virtual target-based database screening. *Journal of Chemical Information and Modeling* **43**, 267-272 (2003).

26 Yang, J.-M. & Shen, T.-W. A pharmacophore-based evolutionary approach for screening selective estrogen receptor modulators. *Proteins: Structure, Function, and Bioinformatics* **59**, 205-220 (2005).

27 Tafi, A. *et al.* Pharmacophore based receptor modeling: the case of adenosine A3 receptor antagonists. An approach to the optimization of protein models. *Journal of Medicinal Chemistry* **49**, 4085-4097 (2006).

28 Pegg, S. C.-H., Haresco, J. J. & Kuntz, I. D. A genetic algorithm for structure-based de

novo design. *Journal of Computer-Aided Molecular Design* **15**, 911-933 (2001).

29 Muegge, I., Martin, Y. C., Hajduk, P. J. & Fesik, S. W. Evaluation of PMF scoring in docking weak lignads to the FK506 binding protein. *Journal of Medicinal Chemistry* **42**, 2498-2503 (1999).

30 Stahl, M. & Schulz-Gasch, T. Practical database screening with docking tools. *Ernst Schering Res Found Workshop*, 127-151 (2003).

31 Bissantz, C., Folkers, G. & Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *Journal of Medicinal Chemistry* **43**, 4759-4767 (2000).

32 Deng, Z., Chuaqui, C. & Singh, J. Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *Journal of Medicinal Chemistry* **47**, 337-344 (2004).

33 Kallblad, P., Mancera, R. L. & Todorov, N. P. Assessment of multiple binding modes in ligand-protein docking. *Journal of Medicinal Chemistry* **47**, 3334-3337 (2004).

34 Amari, S. *et al.* VISCANA: visualized cluster analysis of protein-ligand interaction based on the ab initio fragment molecular orbital method for virtual ligand screening. *Journal of Chemical Information and Modeling* **46**, 221-230 (2006).

35 Yang, J.-M. & Chen, C.-C. GEMDOCK: a generic evolutionary method for molecular docking. *Proteins: Structure, Function, and Bioinformatics* **55**, 288-304 (2004).

36 Morris, G. M. *et al.* Automated docking using a lamarckian genetic algorithm and empirical binding free energy function. *Journal of Computational Chemistry* **19**, 1639-1662 (1998).

37 Yang, J. M. Development and evaluation of a generic evolutionary method for protein-ligand docking. *Journal of Computational Chemistry* **25**, 843-857 (2004).

38 Kussmann-Gerber, S., Kuonen, O., Folkers, G., Pilger, B. D. & Scapozza, L. Drug resistance of herpes simplex virus type 1--structural considerations at the molecular level of the thymidine kinase. *European Journal of Biochemistry* **255**, 472-481 (1998).

39 Wild, K., Bohner, T., Folkers, G. & Schulz, G. E. The structures of thymidine kinase from herpes simplex virus type 1 in complex with substrates and a substrate analogue. *Protein Science* **6**, 2097-2106 (1997).

40 Zhou, H. B. *et al.* Structure-guided optimization of estrogen receptor binding affinity and antagonist potency of pyrazolopyrimidines with basic side chains. *Journal of Medicinal Chemistry* **50**, 399-403 (2007).

41 Maeda, M. The conserved residues of the ligand-binding domains of steroid receptors are located in the core of the molecules. *Journal of molecular graphics & modelling* **19**, 543-551 (2001).

42 Bajorath, J. Integration of virtual and high-throughput screening. *Nature Reviews Drug*

Discovery **1**, 882-894 (2002).

43 Ginsberg, A. M. & Spigelman, M. Challenges in tuberculosis drug research and development. *Nature Medicine* **13**, 290-294 (2007).

44 Lock, R. L. & Harry, E. J. Cell-division inhibitors: new insights for future antibiotics. *Nature Reviews Drug Discovery* **7**, 324-338 (2008).

45 Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **432**, 862-865 (2004).

46 Freire, E. Designing drugs against heterogeneous targets. *Nature Biotechnology* **20**, 15-16 (2002).

47 Wei, D. *et al.* Discovery of multitarget inhibitors by combining molecular docking with common pharmacophore matching. *Journal of Medicinal Chemistry* **51**, 7882-7888 (2008).

48 Zimmermann, G. R., Lehar, J. & Keith, C. T. Multi-target therapeutics: when the whole is greater than the sum of the parts. *Drug Discovery Today* **12**, 34-42 (2007).

49 Chen, Y. F. *et al.* SiMMap: a web server for inferring site-moiety map to recognize interaction preferences between protein pockets and compound moieties. *Nucleic Acids Res.* **38 Suppl**, W424-430 (2010).

50 Roberts, F. *et al.* Evidence for the shikimate pathway in apicomplexan parasites. *Nature* **393**, 801-805 (1998).

51 Cheng, W. C. *Structure-based discovery of Helicobacter pylori and Mycobacterium tuberculosis shikimate kinase inhibitors* Ph.D thesis, National Tsing Hua University, (2009).

52 Wu, C. H. *et al.* The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Research* **34**, D187-D191 (2006).

53 Hartmann, M. D., Bourenkov, G. P., Oberschall, A., Strizhov, N. & Bartunik, H. D. Mechanism of phosphoryl transfer catalyzed by shikimate kinase from Mycobacterium tuberculosis. *Journal of Molecular Biology* **364**, 411-423 (2006).

54 Yang, J.-M., Chen, Y.-F., Shen, T.-W., Kristal, B. S. & Hsu, D. F. Consensus scoring sriteria for Improving enrichment in virtual screening. *Journal of Chemical Information and Modeling*, 1134-1146 (2005).

55 Cheng, W. C., Chang, Y. N. & Wang, W. C. Structural basis for shikimate-binding specificity of Helicobacter pylori shikimate kinase. *Journal of Bacteriology* **187**, 8156-8163 (2005).

56 Gan, J., Gu, Y., Li, Y., Yan, H. & Ji, X. Crystal structure of Mycobacterium tuberculosis shikimate kinase in complex with shikimic acid and an ATP analogue. *Biochemistry* **45**, 8539-8545 (2006).

57 Blum, G., Gazit, A. & Levitzki, A. Development of new insulin-like growth factor-1 receptor kinase inhibitors using catechol mimics. *Journal of Biological Chemistry* **278**,

40442-40454 (2003).

58 Hallak, H. *et al.* Epidermal growth factor-induced activation of the insulin-like growth factor I receptor in rat hepatocytes. *Hepatology* **36**, 1509-1518 (2002).

59 Zahradka, P., Litchie, B., Storie, B. & Helwer, G. Transactivation of the insulin-like growth factor-I receptor by angiotensin II mediates downstream signaling from the angiotensin II type 1 receptor to phosphatidylinositol 3-kinase. *Endocrinology* **145**, 2978-2987 (2004).

60 Leipe, D. D., Koonin, E. V. & Aravind, L. Evolution and classification of P-loop kinases and related proteins. *Journal of Biological Chemistry* **333**, 781-815 (2003).

61 Cheung, A. *et al.* A small-molecule inhibitor of skeletal muscle myosin II. *Nature Cell Biology* **4**, 83-88 (2002).

62 Stewart, A. *et al.* Phase I trial of XR9576 in healthy volunteers demonstrates modulation of P-glycoprotein in CD56+ lymphocytes after oral and intravenous administration. *Clinical Cancer Research* **6**, 4186-4191 (2000).

63 Mayer, T. U. *et al.* Small molecule inhibitor of mitotic spindle bipolarity identified in a phenotype-based screen. *Science* **286**, 971-974 (1999).

64 Chene, P. ATPases as drug targets: learning from their structure. *Nature Reviews Drug Discovery* **1**, 665-673 (2002).

65 Yang, K., Bai, H., Ouyang, Q., Lai, L. & Tang, C. Finding multiple target optimal intervention in disease-related molecular network. *Molecular Systems Biology* **4**, e228 (2008).

66 Morphy, R., Kay, C. & Rankovic, Z. From magic bullets to designed multiple ligands. *Drug Discovery Today* **9**, 641-651 (2004).