

中文摘要

關鍵詞 (keywords)：虛擬藥物篩選, 篩選後分析, QSAR, Pharmacophore model, Binding site prediction, GEMDOCK, iGEMDOCK, 套膜蛋白, 四環黴素類分子, Structural alphabet, Neuraminidase

在本計畫在三年中, 我們發表了八篇國際期刊論文, 獲得 2007 國家新創獎, 在三年中共有 6 位碩士畢業生, 並達成數項具體成果。我們將三年計畫執行成果分為四個部分: 一、虛擬藥物篩選; 二、藥效基團辨識; 三、藥物篩選後分析, 以及四、親和力(binding affinity)預測與 QSAR 分析。在本計畫中, 我們以自行發展的分子鉗合軟體(GEMDOCK)為基礎, 發展了一個圖形介面的藥物篩選工具(iGEMDOCK), 並且配合已發展的藥效基團之新計分函數進行虛擬藥物篩選。篩選後之結果經由群集歸納分析後, 我們可以從每群中取得代表配體進入生物實驗, 提升分析與實驗效率。針對這些代表配體, 我們發展了兩個新的方法 GemAffinity 及 GEMQSAR 預測其親和力, 提升尋得候選藥物的準確率。此四部分的密切結合, 可以構成一個有效率的藥物開發平台, 並於未來投入重要疾病藥物開發。

以下針對本計畫四部分做一說明:

一、虛擬藥物篩選: 我們為虛擬藥物篩選工具(GEMDOCK)加上圖形介面, 使之成為一個便利使用的成熟預測軟體, 並提供免費下載與推廣(<http://gemdock.life.nctu.edu.tw/dock/>)。此外 GEMDOCK 也實際應用到登革熱病毒結構套膜蛋白之藥物篩選, 並篩檢出有效抑制登革熱病毒之小分子抑制劑。經過 BHK-21 哺乳類動物細胞實驗證實其病毒抑制能力分別為 67.1 uM 與 55.6 uM (IC₅₀ 值)。針對胜肽藥物, 我們也發展了預測平台之雛形。該平台主要概念是利用將 3D 胜肽骨幹結構轉換為序列化, 並提供一個快速的大量相似蛋白質結構之搜尋, 提供未來胜肽藥物開發的基礎。

二、藥效基團辨識: 我們發展了一個藥效基團辨識方法, 結合 GEMDOCK, 建構出新的計分函數, 提升藥物篩選之準確度。我們利用人類雌激素受體(ER)與 Bissantz 等人發表之資料¹作為比較。對其 ER 拮抗劑 GEMDOCK 之平均 goodness-of-hit (GH) score 與平均偽陽性率分別為 83%與 0.13%。對 ER 促進劑其值分別為 48%與 0.75%。

三、藥物篩選後分析: 我們發展了一個群集式分析方法, 提升候選配體的分析及實驗效率。其主要概念是透過同時結合蛋白質-配體交互作用力、物理化學與結構特性以歸納分析候選配體。此外, 我們也發展了一套資料融合(data fusion)技術, 用於結合複數以上計分函數, 以各函數間之互補性提升篩選之準確度。

四、親和力(binding affinity)預測與 QSAR 分析: 我們發展了兩個新的方法 GemAffinity 與 GEMQSAR 預測其親和力, 提升尋得候選藥物的準確率。GemAffinity 可以用來預測蛋白質與配體間親和力。GemAffinity 在測試資料之相關係數為 0.58, 優於十二種已發表的計分函數。另一方面, 若已有數個已知實驗活性的配體, 我們可運用本計畫中所發展之 QSAR 分析工具(GEMQSAR)建立 QSAR 模型。此模型更能準確預測配體與對該特定蛋白質之間的結合親和力。

英文摘要 (Abstract)

關鍵詞 (keywords) : Virtual screening, post-analysis, QSAR, pharmacophore model, binding site prediction, GEMDOCK, iGEMDOCK, Envelope Protein, rolitetracycline, structural alphabet, 3D-BLAST, GEMPLS, GEMQSAR, Neuraminidase

We published eight journal papers and won the 2007 national innovation award. 6 graduated master students were supported by this project during 200-2008. We also achieved several important results in this project. Our achievements can be divided into four parts: 1) Virtual screening; 2) Pharmacophore identification; 3) Post-analysis of virtual screening, and 4) Prediction of binding affinity and QSAR analysis. In this project, we developed a graphical-automatic environment, *iGEMDOCK*, based on GEMDOCK. Additionally, we integrated the scoring function of GEMDOCK with a new developed pharmacophore-based scoring function for virtual screening. In post analysis of virtual screening, we developed a new cluster method for clustering candidate compounds and selecting representatives for biological tests. For these representatives, we developed two new methods, GemAffinity and GEMQASR, to measure binding affinities of protein-ligand complexes. These four parts construct an efficient and fast platform for drug discovery.

The four parts are listed as follows:

1. **Virtual screening.** We developed a graphical-automatic environment named *iGEMDOCK* for docking, screening, and post-screening analysis. *iGemdock* is available at <http://gemdock.life.nctu.edu.tw/dock/>. For the application of GEMDOCK, we identified two effective novel inhibitors (IC_{50} : 67.1 μ M and 55.6 μ M) on the propagation of dengue virus type 2 against the viral envelope protein. Moreover, for peptide drug prediction, we presented a novel protein structure database search tool, 3D-BLAST, that is useful for analyzing novel structures and searching peptides that share similar structural motifs. We have combined this tool and GEMDOCK to develop a peptide drug prediction tool.
2. **Pharmacophore identification.** We integrated the scoring function of GEMDOCK with a new developed pharmacophore-based scoring function for virtual screening. This tool has been applied to molecular docking and post-docking analyses for improving screening accuracy. We assessed the accuracy of our approach by using human estrogen receptor (ER) and a ligand database from the comparative studies of Bissantz et al.¹ While using GEMDOCK, the average goodness-of-hit (GH) score was 83% and the average false positive rate was 0.13% for ER antagonists, and the average GH score was 48% and the average false positive rate was 0.75% for ER agonists.

3. **Post-analysis of virtual screening.** We developed a cluster method for post analysis to improve enrichment for virtual screening. The method combined protein-ligand interactions derived from GEMDOCK, physical-chemical features, and structures, for generating profiles of candidate compounds. Based on these profiles, the method clusters candidate compounds and selects representative compounds for biological tests. In addition, we explored consensus scoring criteria and provided a consensus scoring procedure for improving the enrichment in virtual screening using data fusion.
4. **Prediction of binding affinity and QSAR analysis.** We developed two new methods, GemAffinity and GEMQASR, to measure binding affinities of protein-ligand complexes. GemAffinity outperforms 12 comparative scoring functions on a public set. Furthermore, if several compounds with experimental affinities are available, GEMQASR is able to build a QSAR model to predict binding affinities more precisely.

CONTENT

中文摘要.....	I
英文摘要 (Abstract)	II
CONTENT.....	IV
Overview.....	2
1.1 <i>i</i> GEMDOCK: A Graphical-Automatic System for Virtual Screening and Post-Screening Analysis.....	8
1.1.1 Introduction.....	8
1.1.2 Descriptions	9
1.1.3 Results.....	13
1.2 Identifying Two Novel Inhibitors on the Propagation of Dengue Virus Type 2 Using Virtual Screening against the Envelope Protein.....	15
1.2.1 Introduction.....	15
1.2.2 Materials and Methods.....	17
1.2.3 Results and Discussion	19
1.3 Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for fast protein structure database search	31
1.3.1 Introduction.....	31
1.3.2 Materials and Methods.....	32
1.3.3 Results and Discussion	34
Chapter 2: Pharmacophore Identification	36
2.1 A Pharmacophore-Based Evolutionary Approach for Screening Selective Estrogen Receptor Modulators	36
2.1.1 Introduction.....	36
2.1.2 Materials and Methods.....	38
2.1.3 Results and Discussion	48
Chapter 3: Post-analysis of Virtual Screening	62
3.1 Cluster analysis of Structure-based Virtual Screening by Using Protein-ligand Interactions and Compound Structures.....	62
3.1.1 Introduction.....	62
3.1.2 Materials and Methods.....	63
3.1.3 Results and Discussion	69
3.1.4 Conclusions.....	82
3.2 Consensus Scoring Criteria for Improving Enrichment in Virtual Screening	84
3.2.1 Introduction.....	84
3.2.2 Materials and Methods.....	85
3.2.3 Results and Discussion	94

3.2.4. Conclusions.....	108
Chapter 4 Quantitative Structure Activity Relationships.....	110
4.1 Analysis of Protein-ligand Complexes to Predict Binding Affinity.....	110
4.1.1 Introduction.....	110
4.1.2 Materials and Methods.....	112
4.1.3 Results and Discussion	119
4.1.4 Conclusions.....	126
4.2 GEMQSAR: A QSAR Model Using Protein-ligand Interaction Consensus Profiles and Generic Evolutionary Method.....	127
4.2.1 Introduction.....	127
4.2.2 Materials and Methods.....	128
4.2.3 Results and Discussion	142
4.2.4 Conclusions.....	150
References.....	151
計畫成果自評(Self-evaluation of The Project Achievements).....	164

Overview

In the past three years, we published eight journal papers, one conference paper, five poster papers and won the 2007 national innovation award. Our achievements and work are separated into five parts. The relationships of these works are shown as the following figure.

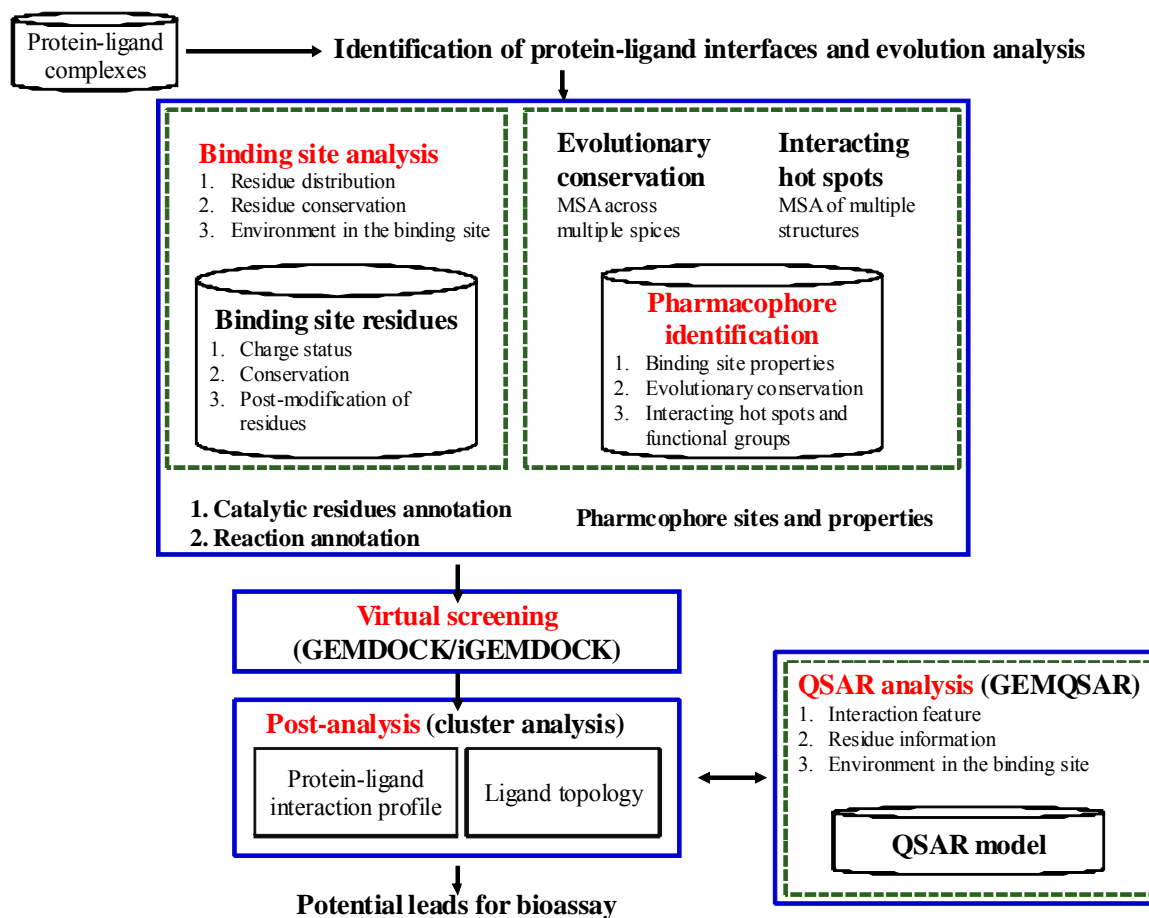


Figure 1. The relationships of our research works in this project.

The core of our works is focused on solving the critical problems in the computer-aided drug discovery. For the automatically virtual screening system, we developed a graphical-automatic environment named *iGEMDOCK* for docking, screening, and post-screening analysis. We also built a web server for freely downloading at <http://gemdock.life.nctu.edu.tw/dock/>. For the application of virtual screening, we have identified two effective novel inhibitors on the propagation of dengue virus type 2 by using virtual screening against the viral envelope protein (DV E protein). Two, rolitetracycline and doxycycline, of these compounds reveal significant inhibition on the DV plaque

formation. Both these compounds are tetracycline derivatives, with estimated IC_{50} values of 67.1 M and 55.6 M, respectively. For peptide drug prediction, we present a novel protein structure database search tool, 3D-BLAST, that is useful for analyzing novel structures and can return a ranked list of alignments. We have combined this tool and our virtual screening tool (GEMDOCK) to develop a peptide drug prediction tool.

For binding site analysis, we developed a new profile method to predict ligand-binding site, called homogenized species-based method. We combined volume information and evolutionary conservation to predict ligand-binding sites. Our method had a better successful rate (75.2%) than Consurf-HSSP (73.1%). A web service for predicting binding sites is served at http://gemdock.life.nctu.edu.tw/cavity_web/.

For pharmacophore identification, we developed a pharmacophore-based evolutionary approach for virtual screening. This tool combines GEMDOCK with a new pharmacophore-based scoring function. The pharmacophore-based scoring function integrates an empirical-based energy function and pharmacological preferences serves as the scoring function for both molecular docking and post-docking analyses to improve screening accuracy. We assessed the accuracy of our approach using human estrogen receptor (ER) and a ligand database from the comparative studies of Bissantz et al.¹ 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.

For post-analysis of virtual screening, we developed a cluster method for post analysis to improve enrichment for virtual screening. The method combines protein-ligand interactions (e.g. hydrogen bonds, electrostatic interactions, and van der Waals), which are generated by our well-developed docking tool (i.e. GEMDOCK), and physical-chemical features and structures for each compound candidate selected by GEMDOCK. For each cluster, this method selected a representative compounds for biological tests and improved the enrichment of virtual screening.

For QSAR analysis, we developed a QSAR methodology associating molecular docking and feature selection with PLS, named GEMPLS. GEMPLS served as feature selection and model building in QSAR analysis. Potential features for contributing inhibition would be selected by evolutionary strategy and built regression by PLS. Due to the low correlation of binding affinity and current scoring functions, we have developed a scoring function, namely GemAffinity, to predict binding affinities of protein-ligand complexes. GemAffinity consists of 5 descriptors including protein-ligand interactions, structural and physicochemical descriptors of ligands, protein properties, metal-ligand bonding, and water effects. The correlation between predicted binding affinities and experimental values is 0.58 and the GemAffinity outperforms 12 comparative scoring functions on this set. GemAffinity will be added into our QSAR method (termed GEMQSAR) to improve the prediction abilities and accuracy.

The overview of achievements in this project is listed as following. We published 8 journal papers, 1 conference paper and 5 posters. The tools developed in this project are also provided web services. Our automatically virtual screening tool, iGEMDOCK also won the prize of 2007 National Innovation Award. Eight masters were supported by this research project.

Research publications

Journal papers:

1. Y.-Y. Chiu, J.-K. Hwang, J.-M. Yang*, "Soft energy function and generic evolutionary method for discriminating native from non-native protein conformations," *Journal of Computational Chemistry*, vol. 29, pp. 1364-1373, 2008 (SCI, IF: 4.89)
2. M.-C. Yang, H.-H. Guan, M.-Y. Liu, Y.-H. Lin, J.-M. Yang, W.-L. Chen, C.-J. Chen, and Simon J. T. Mao*, "Crystal structure of a secondary vitamin D3 binding site of milk β -lactoglobulin," *Proteins: Structure, Function, and Bioinformatics*, vol. 71, pp. 1197-1210, 2008. (SCI, IF: 3.73)
3. Y.Y. Yao, K.L. Shrestha, Y.J. Wu, H.J. Tasi, C.C. Chen, J.-M. Yang, A. Ando, C.Y. Cheng, Y.K. Li*, "Structural simulation and protein engineering to convert an endo-chitosanase to an exo-chitosanase," *Protein Engineering, Design & Selection*, 2008, in press. (SCI, IF: 3.0)
4. C.-H. Tung, J.-W. Huang and J.-M. Yang*, "Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for fast protein structure database search," *Genome Biology*, vol. 8, pp. R31.1~R31.16, 2007. (SCI, IF: 7.17)
5. C.-H. Tung and J.-M. Yang*, "fastSCOP: a fast web server for recognizing protein structural domains and SCOP superfamilies," *Nucleic Acids Research*, pp. W438-W443, 2007. (SCI, IF: **6.31**)
6. J.-M. Yang, Y.-F. Chen, Y.-Y. Tu, K.-R. Yen, and Y.-L. Yang*, "Combinatorial computation approaches identifying tetracycline derivatives as flaviviruses inhibitors," *PLoS ONE*, pp. e428.1-e428.12, 2007.
7. J.-M. Yang* and T.-W. Shen, "A pharmacophore-based evolutionary approach for screening selective estrogen receptor modulators," *Proteins: Structure, Function, and Bioinformatics*, vol. 59, pp. 205-220, 2005. (Times Cited: 13)
8. 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. (Times Cited: 25)

Conferences Papers:

1. K-C Hsu, Y-F Chen, and J-M Yang*, "Binding affinity analysis of protein-ligand complexes," 2nd International Conference on Bioinformatics and Biomedical Engineering, pp. 167-171, 2008.

Posters

1. Y.-F. Chen, L.-J. Chang, J.-M. Yang*, "Integrating GEMDOCK with GEM-PLS and GEM-kNN for QSAR modeling of huAChE and AGHO," in 15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 6th European Conference on

Computational Biology (ECCB), Vienna, Austria, 2007.

2. C.-H. Tung, T.-K. Yang, and J.-M. Yang*, "Structural Binding Pocket Clustering and Protein-Ligand Interaction Analysis for ATP-binding Proteins," in 15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 6th European Conference on Computational Biology (ECCB), Vienna, Austria, 2007.
3. J.-M. Yang, Y.-F. Chen, C.-Y. Chen and Y.-L. Yang, "Identifying Two Tetracycline-Derivates as Effective Novel Inhibitors on the Propagation of Dengue Virus Type 2 Using Virtual Screening against the Envelope Protein", in Annual Conference on Biotechnology, Hsinchu, Taiwan, 2006
4. C.-N. Ko, Y.-F. Chen, Y.-J. Chen and J.-M. Yang, "Cluster analysis of Structure-based Virtual Screening by Using Protein-ligand Interactions and Compound Structures", in Annual Conference on Biotechnology, Hsinchu, Taiwan, 2007
5. Y.-T. Chen and J.-M. Yang, "A New Profile Method for Predicting Protein-ligand Binding Site", in 2008 Annual Conference on Biotechnology, Hsinchu, Taiwan, 2008

Databases and web-based services

GEMDOCK: <http://gemdock.life.nctu.edu.tw/dock/>

Binding site analysis: http://gemdock.life.nctu.edu.tw/cavity_web/

3D-BLAST: <http://3d-blast.life.nctu.edu.tw/>

Awards in the past three years

Table 1. The awards of principal investigators during 2005-2008

Name of PI	Date	Prize
J.-M. Yang	2006	獲得國立交通大學 2006 年傑出人士榮譽獎勵
J.-M. Yang	2007	國家新創獎
J.-M. Yang	2007~	生物資訊協會理事
J.-M. Yang	2005	指導研究生獲資訊學會碩博士論文獎佳作獎

Table 2. The awards of graduate students joined in this project

Student	Professor	Date	Prize
陳佑德	J.-M. Yang	2008	交通大學生物科技學院 2008 生物科技學術壁報競賽優等
陳彥甫 陳右儒	J.-M. Yang	2007	國家新創獎第三名
陳彥甫	J.-M. Yang	2007	2007 年生物科技學術研討會暨壁報比賽 (優等)
董其樺	J.-M. Yang	2007	2007 年生物科技學術研討會暨壁報比賽 (優等)
董其樺	J.-M. Yang	2006	2006 年生物科技學術研討會暨壁報比賽 (優等)
陳彥甫	J.-M. Yang	2006	2006 年生物科技學術研討會暨壁報比賽 (佳作)
黃章維	J.-M. Yang	2006	2006 年生物科技學術研討會暨壁報比賽 (佳作)
董其樺	J.-M. Yang	2005	資訊學會最佳碩博士論文

Accomplishments on education

In the past three years, 8 masters were supported by this research project.

Table 3. Summary of conferences that our students have joined during 2005-2008

Student	Professor	Date	Conference
許凱程	J.-M. Yang	2008/05	The 2 nd International Conference on Bioinformatics and Biomedical Engineering (iCBBE2008)
董其樺	J.-M. Yang	2007/08	The 15 th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB)
陳彥甫	J.-M. Yang	2007/08	The 15 th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB)

Chapter 1 Virtual Screening

The core of our works is focused on solving the critical problems in the computer-aided drug discovery. For the automatically virtual screening system, we developed a graphical-automatic environment named *i*GEMDOCK for docking, screening, and post-screening analysis. We also built a web server for freely downloading at <http://gemdock.life.nctu.edu.tw/dock/>. For the application of virtual screening, we have identified two effective novel inhibitors on the propagation of dengue virus type 2 by using virtual screening against the viral envelope protein (DV E protein). Two, rolitetracycline and doxycycline, of these compounds reveal significant inhibition on the DV plaque formation. Both these compounds are tetracycline derivatives, with estimated IC₅₀ values of 67.1 M and 55.6 M, respectively. For peptide drug prediction, we present a novel protein structure database search tool, 3D-BLAST, that is useful for analyzing novel structures and can return a ranked list of alignments. We have combined this tool and our virtual screening tool (GEMDOCK) to develop a peptide drug prediction tool.

In this section, our works have published four journal papers, two post papers and won one prize. The detail lists as below.

Awards:

1. 2007 National Innovation Award, Yen-Fu Chen, Yu-Ju Chen, and Jinn-Moon Yang, "GEMDOCK: An Integrated Environment for Computer-aided Drug Design and Its Applications", Taiwan

Journal papers:

1. Y.-Y. Chiu, J.-K. Hwang, J.-M. Yang*, "Soft energy function and generic evolutionary method for discriminating native from non-native protein conformations," *Journal of Computational Chemistry*, vol. 29, pp. 1364-1373, 2008 (SCI, IF: 4.89)
2. M.-C. Yang, H.-H. Guan, M.-Y. Liu, Y.-H. Lin, J.-M. Yang, W.-L. Chen, C.-J. Chen, and Simon J. T. Mao*, "Crystal structure of a secondary vitamin D3 binding site of milk β -lactoglobulin," *Proteins: Structure, Function, and Bioinformatics*, vol. 71, pp. 1197-1210, 2008. (SCI, IF: 3.73)
3. Y.Y. Yao, K.L. Shrestha, Y.J. Wu, H.J. Tasi, C.C. Chen, J.-M. Yang, A. Ando, C.Y. Cheng, Y.K. Li*, "Structural simulation and protein engineering to convert an endo-chitosanase to an exo-chitosanase," *Protein Engineering, Design & Selection*, 2008, in press. (SCI, IF: 3.0)
4. J.-M. Yang, Y.-F. Chen, Y.-Y. Tu, K.-R. Yen, and Y.-L. Yang*, "Combinatorial computation approaches identifying tetracycline derivatives as flaviviruses inhibitors," *PLoS ONE*, pp. e428.1-e428.12, 2007.

Posters

1. C.-H. Tung, T.-K. Yang, and J.-M. Yang*, "Structural Binding Pocket Clustering and Protein-Ligand Interaction Analysis for ATP-binding Proteins," in 15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 6th European Conference on Computational Biology (ECCB), Vienna, Austria, 2007.

2. J.-M. Yang, Y.-F. Chen, C.-Y. Chen and Y.-L. Yang, "Identifying Two Tetracycline-Derivates as Effective Novel Inhibitors on the Propagation of Dengue Virus Type 2 Using Virtual Screening against the Envelope Protein", in Annual Conference on Biotechnology, Hsinchu, Taiwan, 2006

1.1 *i*GEMDOCK: A Graphical-Automatic System for Virtual Screening and Post-Screening Analysis

1.1.1 Introduction

As significantly increasing in the number of protein crystal structures, molecular docking and virtual screening (VS) are emergency issues for structural-based drug design (SBDD). Many tools (e.g. GEMDOCK ², DOCK ³, Autodock ⁴, GOLD ⁵) have been developed for VS and successfully applied to identify lead compounds for target proteins from thousands of compounds. However, the inaccuracy of their scoring methods, that is, inadequately predicting the true binding affinity of a ligand for a receptor, is probably the major weakness for VS. To improve the hit rate, some methods have utilized compound structure similarity ⁶ and protein-ligand interactions ^{7, 8} for post-screening analysis.

We have developed a docking/screening tool (GEMDOCK) which achieved high accuracy on some benchmarks ^{2, 9, 10} and successfully identified novel substrates or inhibitors for some targets ^{11, 12}. The GEMDOCK used a soft energy function and a generic evolutionary method for flexible docking. In general, a docking tool for VS consists of four steps: the preparations of the binding site and ligand; molecular docking; and post-screening analysis. For preparations of the binding site, most of docking tools need to add hydrogen atoms (e.g. GOLD, Autodock, and DOCK) and grid (e.g. Autodock and DOCK) the binding site via a command mode. In addition, users require other tools to analyze docked complexes to enrich the hit rate in manual for the post-analysis. These procedures are often time consumed and a high wall for the entry-level end-users. Therefore, some docking programs provided a graphic user interface (GUI), such as ADT of Autodock and GoldMine of GOLD, to reduce the ill effects. The ADT analyzes many docked conformations of a compound and the GoldMine executed post-analysis using docked scores (e.g. van der Waals and hydrogen bonding) of screened compounds. However, these tools do not consider compound structures and protein-ligand interactions for post-screening analysis.

Here, we developed a GUI environment, named *i*GEMDOCK, by integrating docking and virtual screening tool (GEMDOCK), post-screening analysis methods, and visualization tool (RasMol ¹³). For post-screening analysis, we developed two modules (i.e. *mod_ac* and *mod_kc*) to cluster large docked complexes using compound structure properties and protein-compound interactions. The *i*GEMDOCK keeps the following advantages: (1) a friendly GUI for preparing the binding site (*mod_cav*), selecting compounds (*mod_lig*), and docking parameters; (2) grouping the docked compounds; (3) visualizing a docked complex or similar complexes of a group.

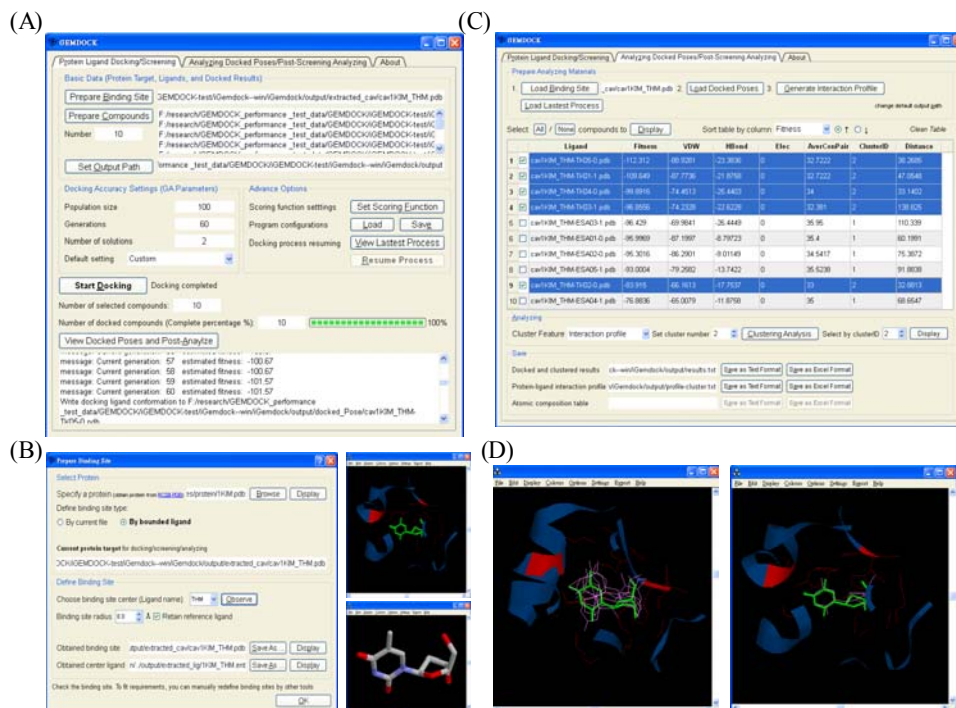


Figure 1.1.1. The framework of the *iGEMDOCK*. (A) The interface of the docking/screening consists of three parts: preparing binding sites and screening compounds; docking parameters; and the progress of working procedure. (B) The interface of the binding site preparation provides users to set the binding site and visualize protein and ligand structures. (C) The interface of the post-screening analysis shows energies of docked compounds and provides the compound groups. (D) The visualization interface indicates a docked pose or similar docked poses in a group.

1.1.2 Descriptions

The *iGEMDOCK* is an automatic and easy-to-used screening GUI environment for molecular docking and post-screening analysis (Figure 1.1.1). To describe *iGEMDOCK*'s functions, we employed herpes simplex virus type 1 thymidine kinase (TK) complex¹⁴ as the target protein and 10 screening compounds, including five TK inhibitors and five estrogen receptor (ER) agonists (Figure 1.1.2). The main parts of *iGEMDOCK* for molecular docking/VS includes binding site preparation module (*mod_cav*); compound selection module (*mod_lig*); docked parameters; and screening progress status (Figure 1.1.1A). The *iGEMDOCK* provides a straightforward method to derive the binding sites and bounded ligands from protein structures in Protein Data Bank (PDB)¹⁵. For the target TK, the PDB code is 1kim and bounded ligand is aciclovir (Figure 1.1.1B). Please note that the *iGEMDOCK* automatically considers the effect of hydrogen atoms. Users are able to visualize both the ligand and binding site structures to select or refine the suitable binding site based on their requirements. The *iGEMDOCK* also provides an interactive interface to select compounds for docking/screening. When the binding site and compounds are preparations, the *iGEMDOCK* can suggest docking parameters (i.e. population size and number of generations) of the GEMDOCK and display the progress of screening status.

For post-screening analysis, the *i*GEMDOCK used compound structures (i.e. atom composition) and protein-ligand interactions to analyze numerous docked poses generated by GEMDOCK. The post-analysis module (Figure 1.1.3) of *i*GEMDOCK facilitates users for several tasks: clustering compounds; mining common structures and interactions; visualizing docked clustering results. The module sorts the compounds based on docked energy terms (e.g. hydrogen bonding interactions, electrostatic energy, van der Waals contact energy, or total energy) (Figure 1.1.1C). It also identifies similar poses and compounds by using k-means and hierarchical clustering methods according to properties of protein-ligand interactions and atomic compositions. Atomic composition, which is similar to the amino acid composition of a protein sequence, is a new concept for measuring compound similarity (Table 1.1.1). For the TK screening, the post-analysis module clusters 10 screening compounds into two groups (i.e. ER and TK groups) using both docked poses and compound structures (Figures 1.1.4 and 1.1.5 in supporting material, respectively). According to protein-ligand interactions (Figure 1.1.4), the interactions between ER and TK compounds are significantly different on six residues: H58, Y101, Q125, Y132, R222, and R225. These clusters let user easily observe docked poses of the individual protein-ligand complex or similar complexes in a group (Figure 1.1.1D). In addition, users can directly download the properties of protein-ligand interactions and atomic compositions for post-screening analysis.

Table 1.1.1. Ten atom types of the atom composition for describing a compound structure

Atom Types	Descriptions
C.ring	Number of Carbon on the ring
C.other	Number of Carbon not on the ring
N.ring	Number of Nitrogen on the ring
N.other	Number of Nitrogen not on the ring
O.ring	Number of Oxygen on the ring
O.other	Number of Oxygen not on the ring
P	Number of Phosphorous
S.ring	Number Sulfur
X	Number of other atoms
Ring#	Number of chemical ring

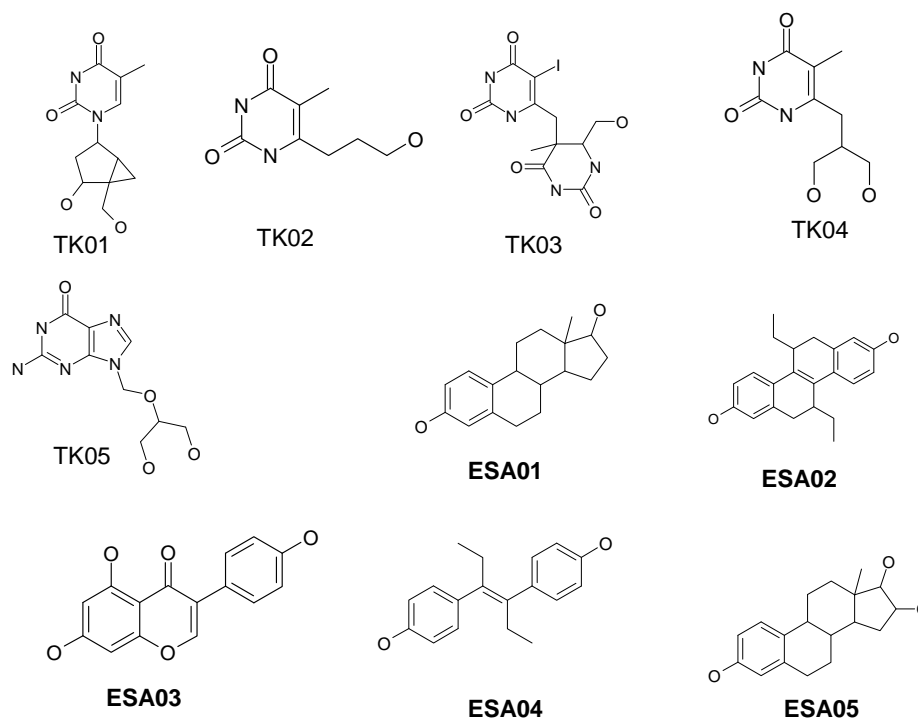


Figure 1.1.2. Ten compound structures consist of five thymidine kinase (TK) inhibitors (TK01, TK02, TK02, TK04, and TK05) and five estrogen receptor (ER) agonists (ESA01, ESA02, ESA03, ESA04, and ESA05).

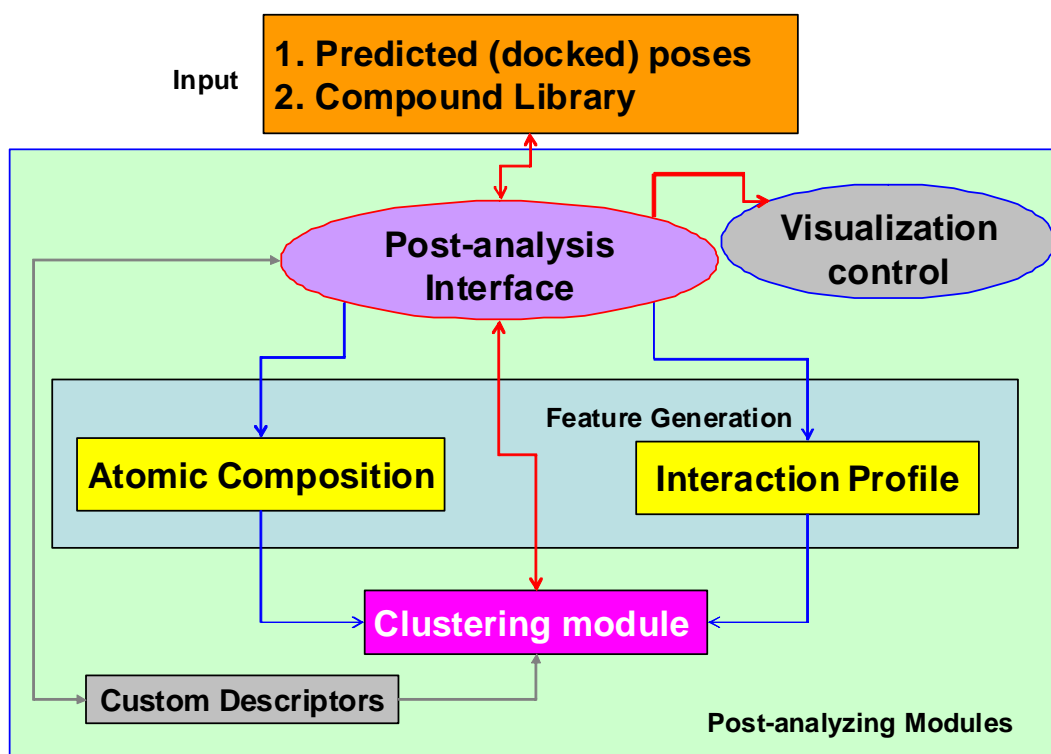


Figure 1.1.3. The architecture of the post-screening analysis of the *i*GEMDOCK

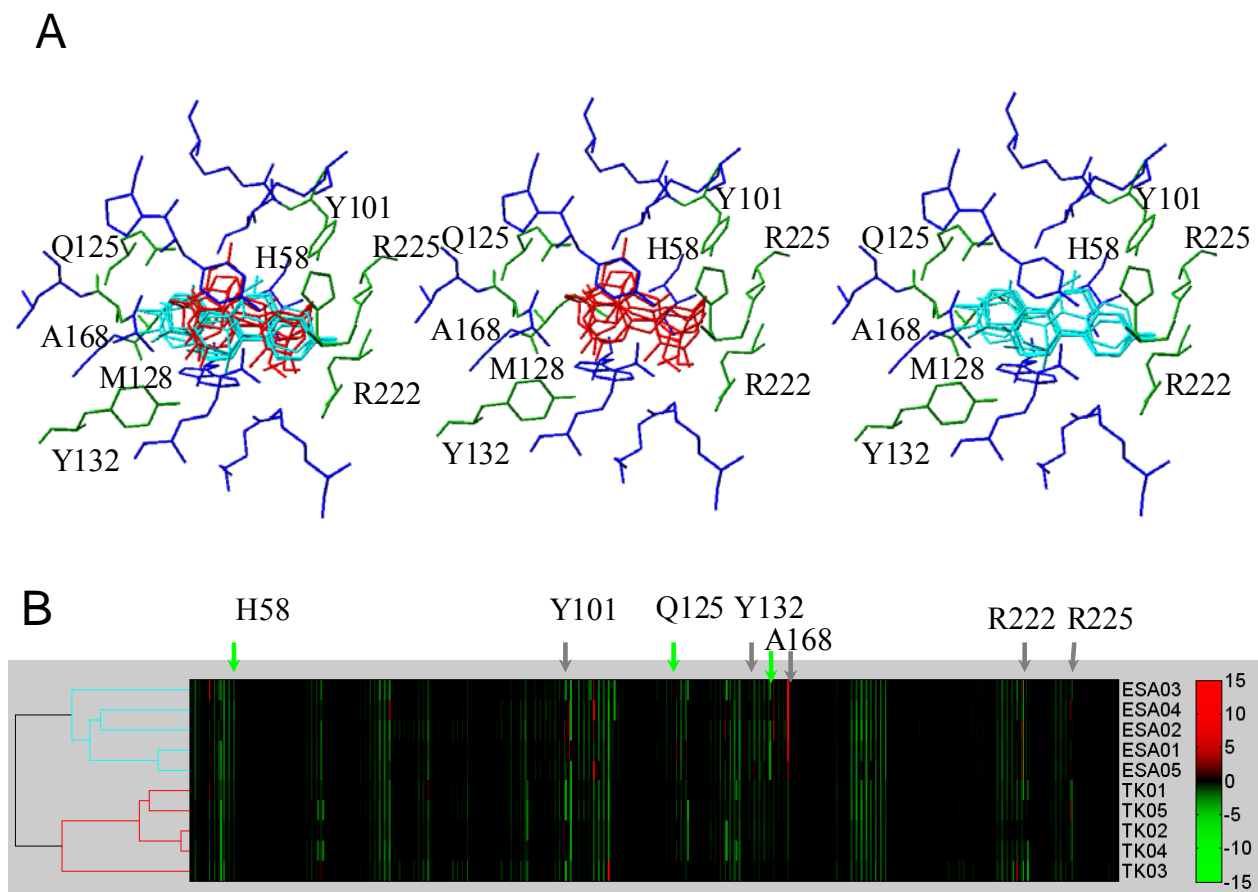


Figure 1.1.4. (A) The docked poses and (B) protein-ligand interactions of 10 compounds (Figure 1.2.3) docked into the herpes simplex virus type 1 thymidine kinase (TK) complex. The docked poses of TK and ER compounds are colored as red and cyan, respectively. The TK and ER are clustered into 2 groups based on protein-ligand interactions. The interactions between ER and TK compounds are significantly different on six residues: H58, Y101, Q125, Y132, R222, and R225.

A

#Compound	C.ring	C.other	N.ring	N.other	O.ring	O.other	P	S	X	#.of.Ring
ESA05	17	1	0	0	0	3	0	0	0	4
ESA04	12	6	0	0	0	2	0	0	0	2
ESA03	15	0	0	0	1	4	0	0	0	3
ESA02	18	4	0	0	0	2	0	0	0	4
ESA01	17	1	0	0	0	2	0	0	0	4
TK05	5	4	4	1	0	4	0	0	0	2
TK04	4	5	2	0	0	4	0	0	0	1
TK03	8	4	4	0	0	5	0	0	0	2
TK02	4	4	2	0	0	3	0	0	0	1
TK01	10	2	2	0	0	4	0	0	0	3

B

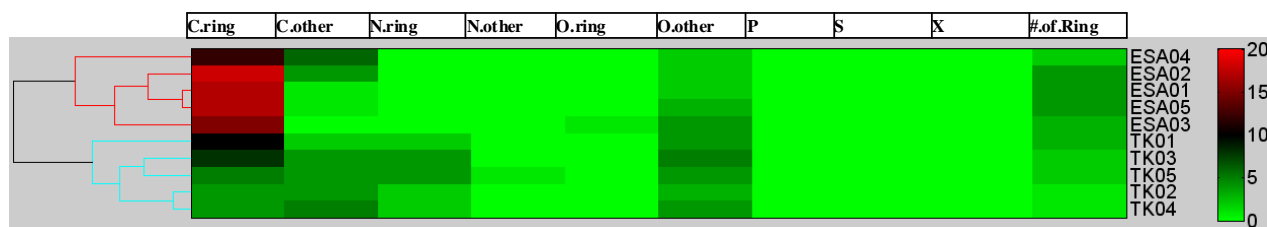


Figure 1.1.5. (A) The values of 10 properties and (B) cluster results of 10 compounds (Figure 1.1.2) of atomic compositions. The TK and ER are clustered into 2 groups based on protein-ligand the properties of the atomic compositions.

1.1.3 Results

The *i*GEMDOCK was evaluated and compared with other methods on CCDC/Astex set (i.e. 305 protein-ligand complexes¹⁶) and two targets (i.e. TK and ER¹⁷) for the docking and VS, respectively. Experimental results reveal that *i*GEMDOCK is robust and comparative to several programs (i.e. GOLD, DOCK, and FlexX) on these sets (Tables 1.1.2 and 1.1.3). For 305 complexes, the successful rates (i.e. the RMSD between a docked pose and the ligand structure is less than 2.0 Å) of *i*GEMDOCK and GOLD are 78% and 68%, respectively. For two screening targets, *i*GEMDOCK is also better than DOCK, GOLD and FlexX. In summary, the *i*GEMDOCK is a useful graphical-automatic environment for docking, screening, and post-screening analysis. It provides post-screening analysis tools by utilizing protein-ligand interactions and compound structures. We believe that *i*GEMDOCK is useful for structure-based drug design.

Table 1.1.2. Comparing GEMDOCK with GOLD on the CCDC/Astex set using the success rate. For each complex, GEMDOCK and GOLD generated 10 docked poses and the one with the lowest energy is selected as the docked solution.

RMSD	iGEMDOCK				GOLD ^a			
	All entries ^b	Clean list ^c	Clean list with R<2.5Å ^c	Clean list with R<2.0 Å ^c	All entries	Clean list ^c	Clean list with R<2.5 Å ^c	Clean list with R<2.0 Å ^c
<0.5 Å	14 %	16 %	17 %	22 %	14 %	17 %	19 %	19 %
<1.0 Å	51 %	55 %	55 %	57 %	44 %	50 %	51 %	56 %
<1.5 Å	71 %	73 %	73 %	75 %	59 %	65 %	66 %	72 %
<2.0 Å	78 %	82 %	83 %	83 %	68 %	72 %	73 %	78 %
<2.5 Å	84 %	86 %	86 %	87 %	75 %	78 %	80 %	85 %
<3.0 Å	86 %	89 %	89 %	89 %	80 %	82 %	83 %	88 %

^a The results of the GOLD directly summarized from Nissink *et al* ¹⁶.

^b The CCDC/Astex consists of 305 complexes.

^c The clean list is divided into three subsets: all clean complexes (224 complexes), the complexes (180 complexes) with resolution < 2.5 Å, and the complexes (92 complexes) with resolution < 2.0 Å.

Table 1.1.3. Comparing GEMDOCK with other methods on screening the ER antagonists and TK inhibitors by false positive rates (%)

Target protein	True positive (%)	iGEMDOCK	DOCK ^a	FlexX ^a	GOLD ^a
ER-antagonists	80	1.72 (17/990) ^b	13.3	57.8	5.3
	90	2.32 (22/990)	17.4	70.9	8.3
	100	5.15 (48/990)	18.9	- ^c	23.4
Thymidine kinase	80	4.75 (51/990)	23.4	8.8	8.3
	90	8.89 (54/990)	25.5	13.3	9.1
	100	9.7 (97/990)	27.0	19.4	9.3

^a Directly summarized from ¹⁸.

^b The false positive rate from 990 random ligands.

^c FlexX could not calculate the docked solution for EST09.

1.2 Identifying Two Novel Inhibitors on the Propagation of Dengue Virus Type 2 Using Virtual Screening against the Envelope Protein

1.2.1 Introduction

Dengue virus (DV) belongs to the *Flavivirus* family, and has become a serious global threat to public health, especially in tropical and subtropical regions because of the rising population density and changes in the environment. DV has four serotypes, all of which are transmitted by *Aedes* mosquitoes and threatens around 2.5 billion people worldwide. Other well-known members of the flavivirus family include the yellow fever, Japanese encephalitis, West Nile and Murray Valley encephalitis viruses¹⁹. Patients with DV infection show various clinical symptoms, ranging from no significant illness, through mild fever to life-threatening dengue hemorrhagic fever and dengue shock syndrome. Currently, only supportive treatment is available for DV, and although considerable research has been directed towards the development of a safe, effective DV vaccine over the past 50 years, no approved commercial product is presently available²⁰. Therefore, one approach to combating the disease is to develop novel strategies for the discovery of leads of antiviral agents for both prevention and treatment.

The DV genome contains a single positive-stranded RNA that encodes a single polyprotein. After processed by proteases encoded by DV and the host, the polyprotein produces three structural proteins, namely capsid, membrane protein (M) and envelope (E) protein, and seven nonstructural proteins, NS1, NS2A, NS2B, NS3, NS4A, NS4B and NS5²¹. The nonstructural proteins are needed for replication, but little is known about their functions. The structural proteins form the basic physical organization of the virion, which includes a viral genome, and is covered by an envelope membrane.

The DV E protein, which is 495 amino acids in length, forms oligomers, and along with the M protein, forms most of the accessible virion surface on the envelope membrane. The E protein initiates the “membrane fusion” process, which is the central molecular event during the entry of enveloped viruses into host cells. The dengue virus enters a host cell when the E protein binds to the receptor on the cellular surface, and subsequently responds by conformational rearrangement, transforming the dimeric prefusion form to change into the trimeric postfusion state. This irreversible conformational change induces fusion of viral and host cell membranes²², and enables the entry to be completed.

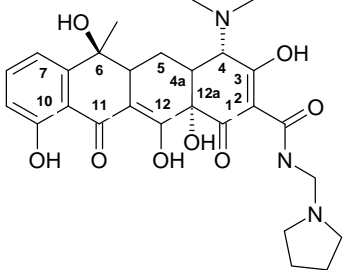
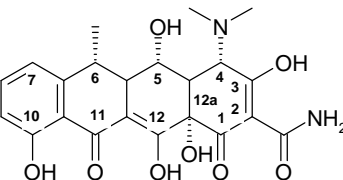
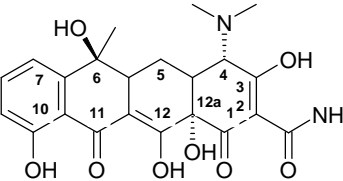
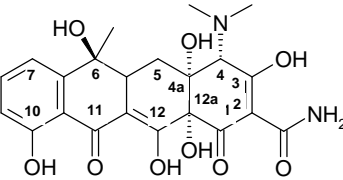
The crystal structures of the E protein of DV type 2 in both the presence (prefusion) and the absence (postfusion) of ligand binding are published in the Protein Data Bank (PDB codes 1oke²⁰ and 1ok8²², respectively). The critical differences between these two structures are a local rearrangement of the “kl” β -hairpin (residues 268-280), and the concomitant opening up a hydrophobic pocket for ligand binding, for instance, by a detergent molecule of

n-octyl- β -D-glucoside (BOG). Mutations affecting the pH threshold for membrane fusion also map to the hydrophobic pocket^{23; 24}. Hence, Modis et al. have concluded that the threshold is a hinge point in the fusion-activating conformational change, and have suggested that this detergent binding pocket could be the target for developing small-molecule fusion inhibitors^{20; 22} by disrupting or even blocking the conformational changes that are needed for DV entry. This concept has opened the possibility of adopting structure-based virtual screening (VS) to identify inhibitors of DV E proteins.

VS is an emerging and promising strategy for discovering novel lead compounds against viral proteins with known structures^{25; 26; 27}. Given the structure of the vicinity around the active site of a target protein and a potential small ligand database, VS predicts the binding mode and the binding affinity for each ligand and ranks a series of candidates. The procedure of VS generally has four phases, including target protein preparation, compound database preparation, molecular docking and post-docking analysis²⁵. In the preparation phase, the structural data of the protein and the compounds are formatted into acceptable formats for a docking program. Molecular docking is then adopted to screen the compound library for potential leads that can dock onto the target protein, while post-docking analysis is used to improve the hit rate. The VS method helps to reduce the burden of high-throughput screening by decreasing the number of compounds requiring processing in an activity assay.

To our knowledge, no drug target site on DV E protein was defined until Modis et al.^{20; 22} proposed the BOG binding site, which is a hydrophobic pocket, as the putative molecular target. This study adopts a well-developed docking tool, GEMDOCK^{2; 9}, to screen the Comprehensive Medicinal Chemistry (CMC) database for substances whose structures could dock into this hydrophobic pocket of E proteins²⁰. In summary, ten compounds were then selectively tested for the inhibitory effect on DV propagation. The derivatives of a compound showing inhibitory effects are also tested for the activities. Two tetracycline derivatives (Table 1.2.1) whose presence in cell cultures exhibited a strong inhibitory effect on the propagation of the DV type 2 PL046 strain were successfully identified. A potential model for the inhibitory action on the DV E protein based on the docked conformations of both active and inactive compounds, which may provide the future direction for the lead optimization, is presented.

Table 1.2.1. Chemical structures and IC₅₀ for the tetracycline derivatives

Compound	Structure	Name	IC ₅₀
1		Rolitetracycline	67.1μM
2		Doxycycline	55.6μM
3		Tetracycline	not applied
4		Oxytetracycline	not applied

1.2.2 Materials and Methods

Preparations of the target protein and screening set

The compound set was prepared by selecting them from the CMC database in May 2004 based on two criteria: (1) molecular weight ranging between 200 and 800, and (2) no compounds with multiple components. A set comprising 5,331 compounds was eventually obtained.

The structure of the BOG binding pocket on DV E protein was isolated and prepared for the GEMDOCK. The structure of the binding pocket in the BOG-bounded conformation (PDB code 1oke²⁰, [Figure 1.2.1A](#)), including amino acids enclosed within a 10 Å radius sphere centered on the bound ligand, was used ([Figure 1.2.1B](#)). The coordinates of protein atoms were taken from the PDB for the screening processing. GEMDOCK docked each compound in the screening set against this binding cavity, and ranked each compound by docked energy of the docked conformation. According to the ranking, compound structures and interactions between compounds and residues in the binding site, ten compounds were chosen for *in vivo* biological activity tests to validate their

inhibitory activities.

Docking method and scoring function

Our previous works ^{2,9} have showed that the docking accuracy of GEMDOCK was better than some well-known docking tools, such as GOLD ⁵ and FlexX ³, on a diverse data set of 100 protein-ligand complexes proposed by Jones et al.⁵ The screening accuracy of GEMDOCK were also better than GOLD, FlexX, and DOCK on screening the ligand database from Bissantz et al. (2000) for the thymidine kinase ²⁸ and the estrogen receptor ⁹. In this study, GEMDOCK parameters in the flexible docking included the initial step sizes ($\sigma=0.8$ and $\psi=0.2$), family competition length ($L = 2$), population size ($N = 300$), and recombination probability ($p_c = 0.3$). For each ligand screened, GEMDOCK optimization stopped either when the convergence was below a certain threshold value or the iterations exceeded the maximal preset value of 60. For the latter case, GEMDOCK will produce 800 solutions in one generation and terminated after it exhausted 48,000 solutions for each compound in the screening set.

The screening quality of docking methods using energy-based scoring functions alone is often influenced by the structure of the ligand being screened (e.g., the numbers of charged and polar atoms). These methods are often biased toward charged polar compounds due to the pair-atom potentials of the electrostatic energy and hydrogen-bonding energy. In order to reduce this ill effect, GEMDOCK could evolve the pharmacological preferences from a number of known active ligands or from domain knowledge 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 ⁹. GEMDOCK could use either a purely empirical scoring function ² or pharmacophore-based scoring function ⁹. When GEMDOCK used a pharmacophore-based scoring function, some known active ligands (more than two) or domain knowledge are required for evolving the pharmacological consensus according to our previous results. The empirical-binding energy (E_{bind}) is the sum of the intermolecular (E_{inter}) and intramolecular energies (E_{intra}), respectively ². The pharmacophore-based energy function ⁹ is the sum of three energy items, including the empirical binding energy (E_{bind}), the energy of binding site pharmacophores (E_{pharma}), and a penalty value (E_{ligpre}) if a ligand does not satisfy the ligand preferences. E_{pharma} and E_{ligpre} are especially useful in selecting active compounds from hundreds of thousands of non-active compounds by excluding ligands that violate the characteristics of known active ligands (or domain knowledge).

Plaque formation assay for the inhibitory effects of compounds on DV2 propagation

A local DV type 2 strain, PL046, was used to infect the mosquito C6/36 cells for the production of DV type 2 virions. Mammalian BHK-21 host cells were cultured at 37°C with 5% CO₂ in MEM medium (Gibco) supplemented with 0.22% of sodium bicarbonate and 10% of fetal bovine serum (FBS) (Gibco). C6/36 cells were grown at 28°C in MEM medium (Gibco) supplemented with 0.11% of sodium bicarbonate and 10% of FBS ²⁹. BHK-21 were passaged at 4×10^5 cells per well in 6-well

plates and incubated at 37°C with 5% CO₂ for 48 hours. Different dilutions of drug compounds were added to the 6-well plates followed by 0.5 ml of medium containing DV type 2 PL046 strain in the amount of 80 to 200 plaque forming unit (pfu) per well. The mixtures were mixed gently and then incubated at 37°C with 5% CO₂ for 1 hour. After aspirating the supernatant, 1 : 1 mixture of MEM medium and 2% methylcellulose were added to the well and incubated further at 37°C with 5% CO₂ for 7 days. The medium was aspirated before the cells were fixed with 3.7% formaldehyde. After 30 minutes, the fixing solution was removed and the cells were stained with 1% crystal violet in 3.7% formaldehyde. The plates were washed with 3.7% formaldehyde before the plaque numbers was scored²⁹.

1.2.3 Results and Discussion

Virtual screening for the inhibitors of E protein

The docking accuracy of GEMDOCK for the DV E protein was first evaluated by docking the BOG into the binding site (Figure 1.2.1). The docked conformation of the BOG (Figure 1.2.2A) with the lowest scoring value was compared with the crystal structure of the BOG based on the root-mean-square deviation (RMSD) of heavy atoms. The average RMSD of ten independent runs was less than 1.20 Å. The molecular recognition on the E protein was also studied to determine the preferred ligand constraints and pharmacophores in the virtual screening. This detergent binding pocket, located between the joint of domains I and II of the E protein, was hydrophobic in the cave and hydrophilic at both sides on the protein surface, while the binding site favored wide range of high-molecular weight and hydrophilic compounds.

GEMDOCK was then adopted to perform virtual screening on the DV E protein on a screening set including 5,331 molecules chosen from the CMC database. Because the binding site of the DV E protein was a hydrophobic pocket, we set the electrostatic constraint, based on the upper bound of the number of charged atoms to 0, and the hydrophilic constraint, based on the upper bound of the fraction of polar atoms, to 0.3, to reduce the undesired effect of the bias in GEMDOCK toward charged polar compounds. The ligand preference acted as a hydrophilic filter, and gave a penalty to highly hydrophilic compounds. The scoring values of both the empirical and pharmacophore-based scoring functions were adopted as ranking conditions to identify the inhibitor candidates.

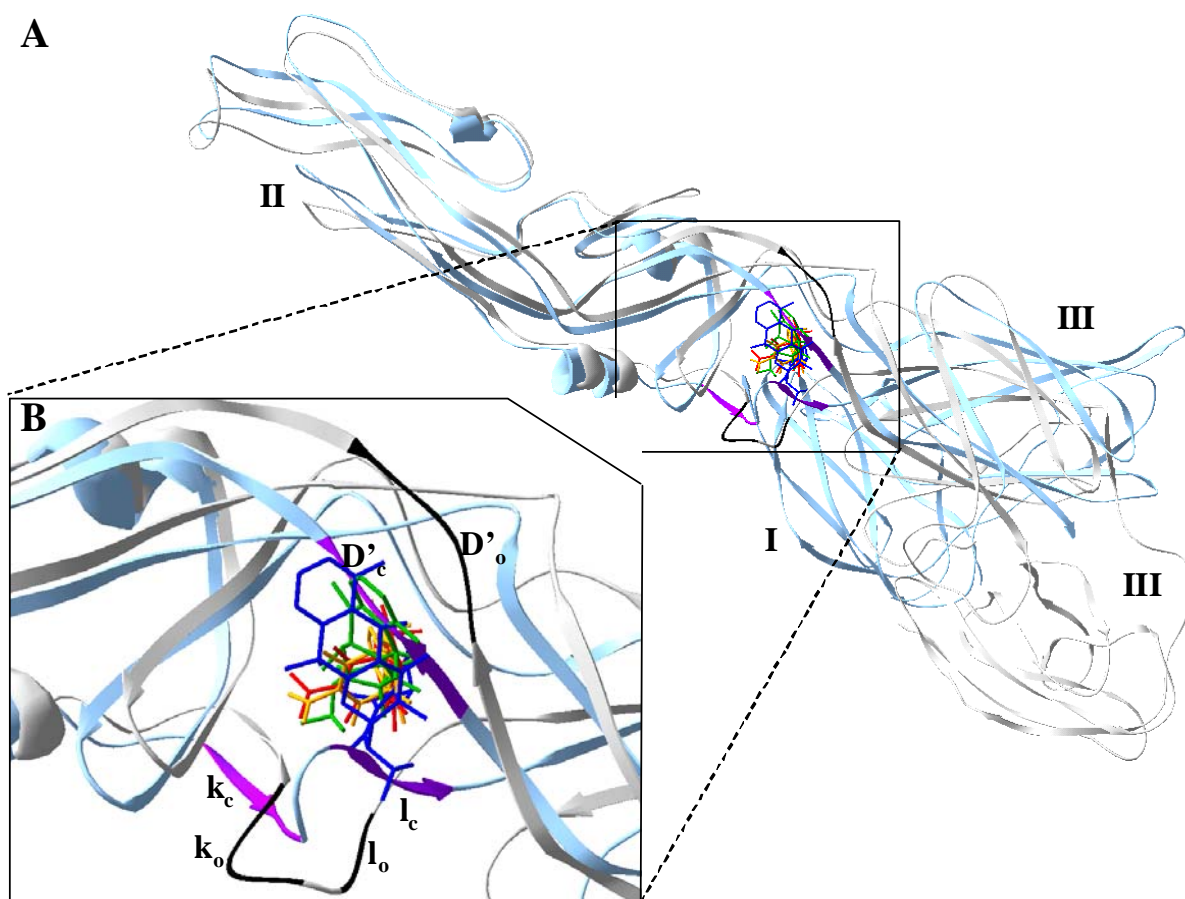


Figure 1.2.1. Prefusion (PDB code 1oke) and Postfusion (PDB code 1ok8) conformations of Dengue E protein and the ligand-binding pocket for virtual screening. (A) Dengue E protein structures with prefusion (gray) and postfusion (blue) and the position of the binding regions (black representing *D*, *k*, *l* in prefusion state, colored representing postfusion state)). (B) The interactions of four compounds docked inside the binding areas: doxycycline (green), rolitetracycline (blue), tetracycline (orange), and oxytetracycline (red). The volumes of the binding site in the prefusion and postfusion forms are significantly different. The critical difference between the two structures is a local rearrangement of the *D_o* segment and the *kl* hairpin, of which *k_o* and *l_o* are the prefusion conformations, while *D_c*, *k_c* and *l_c* are the postfusion conformations. The prefusion conformation is regarded as the bind site for screening. The inhibitory compounds consistently occupy the positions of significant residues in postfusion. The secondary and higher-order structures, and domains I, II and III, correspond to those defined by Modis et al ²⁰.

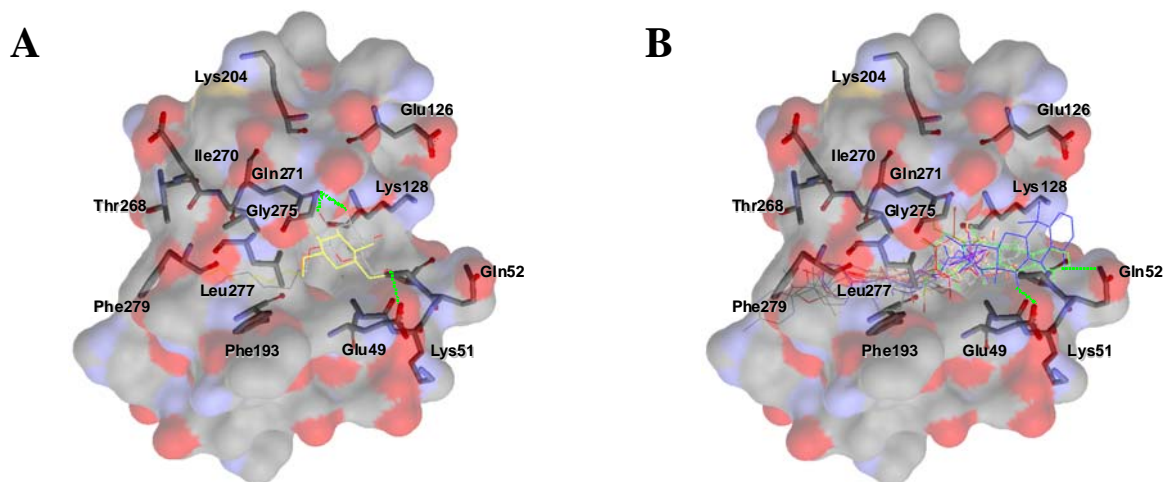


Figure 1.2.2. Docked conformations and screening results against dengue E protein using GEMDOCK. The residues affecting the pH threshold of fusion, and those forming the hydrogen bonds (dash with green line), are indicated. (A) The crystal conformation is in CPK model (i.e., oxygen atoms are red, nitrogens are blue, carbons are gray), and docked conformation (yellow) of the BOG compound: The RMSD of the conformations is 1.20 Å, and both conformations form hydrogen bonds with Glu 49 and Gln 271. (B) The docked conformations of the 10 selected compounds: the four tetracycline derivatives are colored (doxycycline (green), rolitetracycline (blue), tetracycline (orange), and oxytetracycline (red)), and other compounds are shown in the CPK model. The inhibitory compounds (doxycycline and rolitetracycline) are docked in the vicinity of these residues (Thr48, Glu49, Ala50, Lys51 and Gln52), in which the prefusion and postfusion conformations are significantly different. Residues affecting the pH threshold of fusion are marked.

The top ranking 3% compounds (150 compounds) were selected for post analysis to enhance the hit rate after GEMDOCK has screened 5,331 molecules. These selected compounds were clustered by a hierarchical cluster method based on two-dimensional compound structures³⁰ and protein-ligand interactions, as in Jain¹⁸. The atom environments³⁰ were adopted as a two-dimensional compound structural representation to measure compound similarities, and the protein-ligand interactions were used to identify the docked poses and hot spots. According to the structural similarity, docked poses, protein-ligand interactions and the limitation of commercial availability, ten compounds (Figure 1.2.3) were selected for *in vivo* plaque formation assay for their inhibitory effects on DV propagation in cultured cells. Figure 1.2.2B illustrates docked conformations of these selected compounds, and the two candidates of tetracycline derivatives (Table 1.2.1) along with two other derivatives are marked as blue (rolitetracycline), green (doxycycline), orange (tetracycline), and red (oxytetracycline).

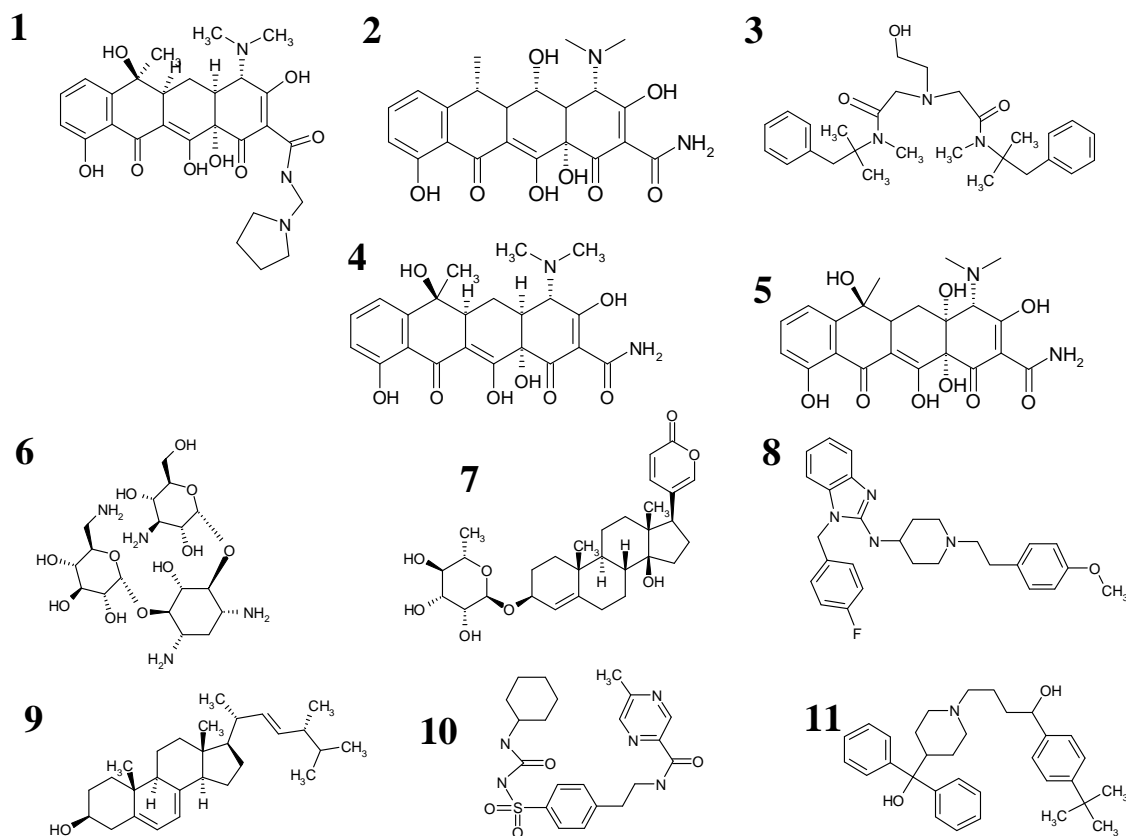


Figure 1.2.3. Eleven selected compounds for competitive blocking assay. Compounds 1 to 4 are shown in [Table 1.2.1](#) and compounds 1, 2 and 3 have inhibitory activities.

In vivo plaque formation assay

To assess whether those individual compounds obtained by screening can affect the biological function of E proteins as predicted, different concentrations of the compounds were added separately to the BHK-21 cell culture, followed immediately by addition of DV type 2 PL046 strain in fixed plaque forming unit (PFU). If the compounds can bind to the E proteins as predicted by the program, then they may interfere with the receptor binding and conformational change necessary for the viral entry, thus reducing the frequency of BHK-21 cells infected by the DV virions, and consequently reducing the number of the plaque formed. Because each plaque represents one infection event, the number of plaques in an assay plate denotes the number of successful infection events. Since the fixed PFU was originally added to the culture, the reduction in the number of plaques reflected the portion of virion infection inhibited by the presence of that compound. Hence, the relative percentage of PFU was calculated, where PFU value from plates with no added compound set to 100%. Among the 10 compounds, rolitetracycline and doxycycline ([Figure 1.2.4](#)) showed dramatic inhibitory effects on DV propagation, yielding an experimental hit ratio of 20%. This demonstrates the utility and economy of computer-aided drug discovery in searching for new bioactive compounds for a

putative molecular target. Additionally, oxethazaine showed a mild inhibitory effect. A 12% reduction (down to 88%) in PFU was observed when the concentration of the compound increased from 200 μM to 500 μM . At the beginning, rolitetracycline showed almost no effect on the DV plaque formation at the concentration of 10 μM . However, the inhibitory effect on DV propagation rose as the concentration was increased. Compared to the mock treatment, the PFU decreased to only 20% at 100 μM and about 5% at 300 μM , yielding an estimated IC_{50} value of 67.1 μM (Figure 1.2.4). Additionally, less than 3% of the PFU remained at a concentration of compound of 500 μM or above. Doxycycline retained 87% PFU at 10 μM , but the PFU decreased to only 14% at 100 μM . Only 1% PFU was left when the concentration of doxycycline reached 500 μM , giving an IC_{50} value of 55.6 μM . Notably, neither tetracycline nor oxytetracycline showed any effect on the DV propagation at concentrations from 10 μM up to 10mM (data not shown), even though they have a similar fused ring structure to rolitetracycline and doxycycline. Table 1.2.1 shows the molecular structures and IC_{50} values of the two active leads and the two inactive tetracycline derivatives, which, with the exception of doxycycline, showed no cellular toxicity effect within the range of tests judged from the cellular morphology and growth. In the case of doxycycline, there was a mild reduction of the cell density when the concentration was 300 μM or higher.

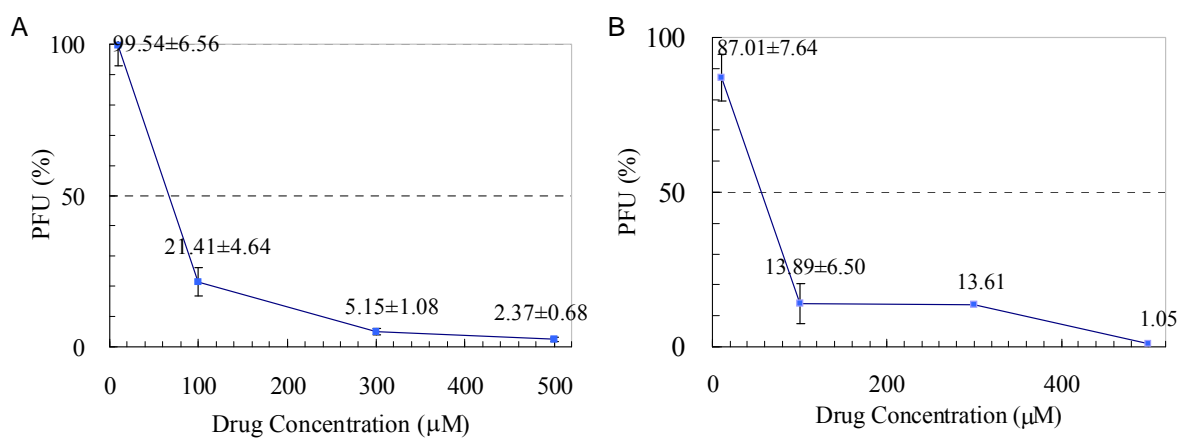


Figure 1.2.4. Effect of (A) doxycycline and (B) rolitetracycline on the plaque formation of dengue virus type 2 on BHK-21 mammalian cells. The IC_{50} values of rolitetracycline and doxytetracycline are 67.1 μM and 55.6 μM . The X axis is the percentage of plaque formation when compared to the control. The Y axis denotes the drug concentration.

Analysis of the Inhibitor-E protein interactions

The docked conformations of the two tetracycline-derivate inhibitors (Table 1.2.1) were consistently different from those of the eight non-inhibitory compounds (Figures 1.2.2B, 1.2.5 and

1.2.6). The inhibitors, doxytetracycline (green) and rolitetracycline (blue), were docked at the outlet of binding pocket and extent into pocket, while the rest (CPK model) were docked inside the pocket. The inhibitory compounds docked near a stretch of residues, namely the Thr48, Glu49, Ala50, Lys51, and Gln52, in the D'_o segment, of which the conformations of the prefusion and postfusion forms were significantly different (Figures 1.2.1B and 1.2.6). The compounds docked very close to or at the space of D_c, which is part of the same stretch in the postfusion state. The functional groups of those compounds may interact with the stretch differently, as well as potentially causes stereo hindrance. Residues in this stretch and several others in the vicinity were revealed to affect the pH-dependent membrane fusion process. As revealed in Figure 1.2.2A, the BOG was docked into the pocket, and was situated in the center between the residues of Gly275, Lys128, Leu277 and Gln52.

Figure 1.2.5 shows the hydrogen-bond networks and orientations of the four tetracycline derivatives in both the prefusion and postfusion forms of E proteins. These derivatives can be divided into two groups according to their docked locations. The two with inhibitory effects, namely rolitetracycline (Figure 1.2.5A) and doxycycline (Figure 1.2.5B), form hydrogen-bond networks mainly with the residues of D_o, which are Thr48, Glu49, Ala50, Lys51 and Gln52, and with additional residues of Gln271 and Gln200. The compounds docked in the position leaning on and interacting with the stretch made up of the residues 48-52. Conversely, the other two compounds, tetracycline (Figure 1.2.4C) and oxytetracycline (Figure 1.2.6D), formed hydrogen bonds mainly with residues Phe279, Thr280, Gln271 and Gln200, and interacted with stretch 48–52 only at Ala50, while leaned away from it. Additionally, the four rings of the two inhibitors were docked at the outlet of the binding pocket, whereas, those of the inactive compounds were inside the pocket. Restated, the inhibitors bound to the outlet of the binding pocket and extended into pocket, while the non-inhibitors bound entirely inside the pocket. GEMDOCK yielded lower binding energies for the two inhibitors than for the inactive compounds. The energy minimization process of SYBYL 6.9 also indicates that the predicted complexes of the inhibitors had lower energies than the non-inhibitors. SYBYL 6.9 computed the energies of rolitetracycline, doxytetracycline, tetracycline and oxytetracyclines as -395.2, -398.7, -356.8, and -371.8kcal/mol.

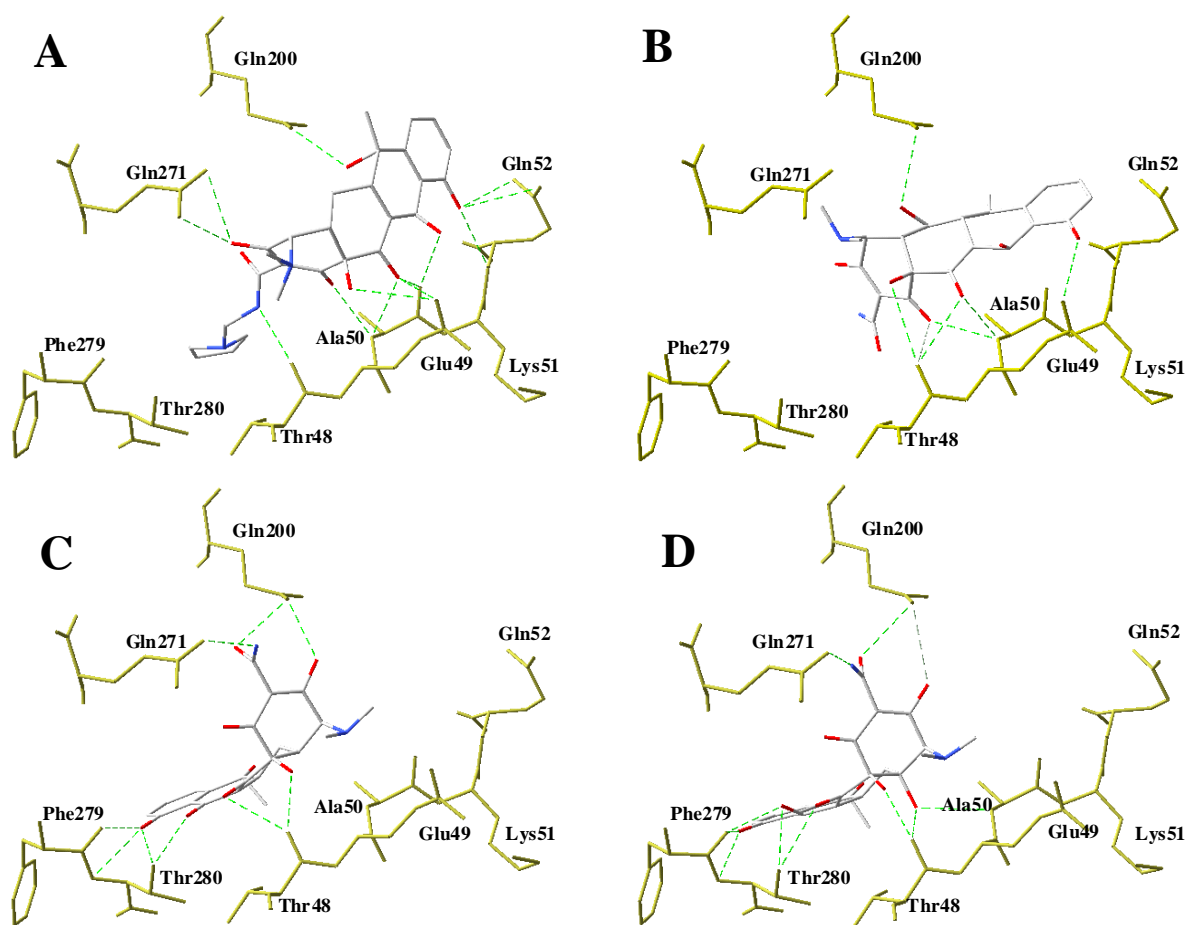


Figure 1.2.5. Docked conformations of (A) rolitetracycline, (B) doxycycline, (C) tetracycline, and (D) oxytetracycline in the binding site of the E protein. Atoms of the E protein are in yellow, and compound ligands are in the CPK model. The hydrogen bonds are represented as the green dash lines. Not all residues are displayed for the sake of clarity. The two inhibitors, rolitetracycline (67.1 M) and doxycycline (55.6 M), form hydrogen-bond networks with Thr48, Glu49, Ala50, Lys51 and Gln52, but have no hydrogen bonds with Phe289 and Thr280. Conversely, tetracycline and oxytetracycline prefer to forms hydrogen bonds with Phe279 and Thr280 than with the other five residues. The docked conformations of these two groups are around 180° to each other based on the positions of the four rings. The Thr48, Glu49, Ala50, Lys51 and Gln52 are in the D_0 segment ([Figure 1.2.5B](#)); Gln271 and Phe279 are in the k_o and l_o segments, respectively.

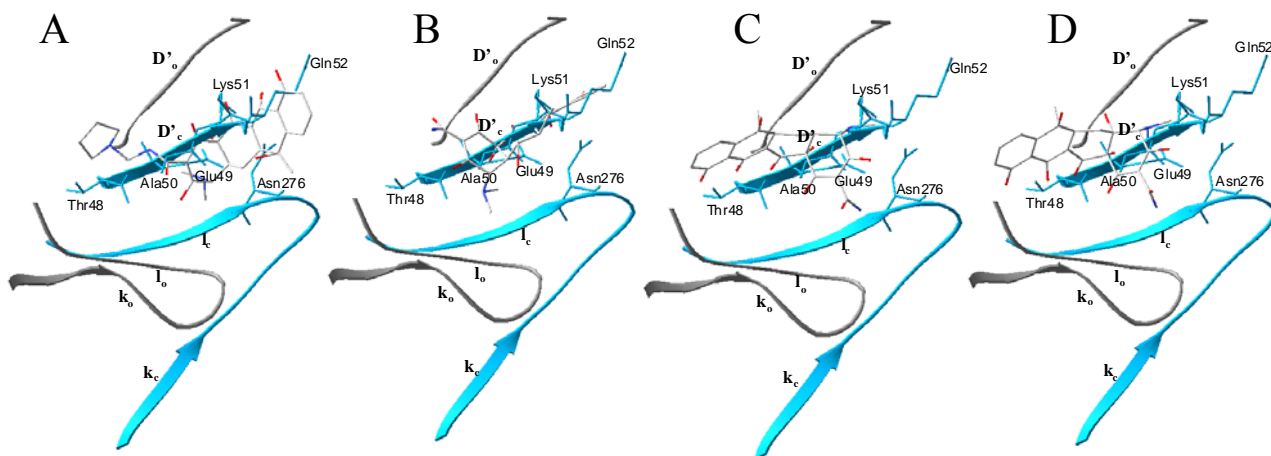


Figure 1.2.6. Docked conformations of (A) rolitetracycline, (B) doxycycline, (C) tetracycline, and (D) oxytetracycline in the binding sites in the prefusion (gray) and postfusion (blue) states. Atoms within compounds are displayed using the CPK model; i.e., oxygen atoms are in red, nitrogens are in blue and carbons are in gray. The side-chains of some residues that overlap with compounds are displayed. Most atoms of the two inhibitors, rolitetracycline and doxycycline, collide with Thr48, Glu49, Ala50, Lys51 and Gln52. By contrast, tetracycline and oxytetracycline overlap slightly with these five residues. The segments D'_o , k_o and l_o belong to the prefusion conformation, while D'_e , k_e , and l_e belong to the postfusion conformation.

For the eradication of enveloped virus infections, identification of compounds that can interfere with the function of viral envelope proteins to prevent viral entry of host cells has been a long-last idea in the field. However, mass high-throughput screening is now considered as costly. Moreover, proper target sites are always hard to identify when structure-based virtual screening approaches are applied. This study applied the VS method to discover potential lead compounds with an inhibitory effect on DV propagation, following the work of Modis et al.^{20; 22}, in which they revealed the structural detail of the DV 2 envelope proteins. Modis et al. concluded that the hydrophobic detergent binding pocket on the DV E protein observed in their structural study is suitable as the target for developing small-molecule inhibitors blocking the process of viral-host membrane fusion, which would interrupt the viral entry and thereby stop the infection. Compounds inserted at this position may hinder conformational change of E proteins, thereby interfering with the fusion transition^{20; 22}. Consistently, mutations on DV E protein mapped to this pocket could indeed influence the pH threshold of fusion^{20; 24; 31; 32; 33; 34; 35} (Figure 1.2.1A). Therefore, this hydrophobic BOG binding pocket was selected as the target site to discover novel lead compounds for developing DV inhibitors.

As noted previously, two tetracycline derivatives, rolitetracycline and doxycycline, exhibited an inhibitory effect on DV propagation. Rolitetracycline and doxycycline were able to inhibit the propagation of dengue virus type 2 among ten selected compounds under the experimental conditions. Significantly, only these two out of the four tested tetracycline derivatives exhibited

inhibitory activity. Therefore, computational modeling was performed in an attempt to provide an explanation of this finding for further investigation. [Figures 1.2.5](#) and [1.2.6](#) reveal that the docked conformations of these two active compounds were significantly different from those of the inactive compounds. Moreover, the atomic interaction behaviors of the two active tetracycline derivatives were different from those of the other two tetracycline derivatives, even though their structures are very similar.

Tetracycline derivative is a group of broad-spectrum antibiotics, which was first discovered in the 1940s ³⁶. The action mechanism of antibiotic tetracycline on bacteria inhibits protein synthesis by preventing aminoacyl-tRNA from attaching to the ribosomal acceptor (A) site ^{36; 37; 38}. In this study, four tetracycline derivatives were selected and subjected to *in vivo* testing for the inhibitory effects on DV propagation. Notably, although tetracycline and oxytetracycline have a similar fused ring structure to rolitetracycline and doxycycline, they exhibited no inhibitory effects. A molecule of a tetracycline-related compounds comprises a linear fused tetracyclic core to which various functional groups are attached ³⁶. Tetracycline is thus the minimum common structure of the four tetracycline-related molecules. Therefore, this common structure does not possess the inhibitory effect on DV propagation. Rather, the substituted functional group confers the activity.

To assess the effect of binding tetracycline derivatives to DV E proteins, the BOG binding site of DV E protein was compared with the tetracycline binding site on the tetracycline repressor (TetR). TetR regulates the resistance mechanism against the antibiotic tetracycline in gram-negative bacteria. The tetracycline binding site of TetR protein has been defined, and the structures have been determined by crystalline ³⁹. This study found that TetR protein has similar characteristics to E protein in their binding sites for tetracycline derivatives. First, the volumes of the binding areas are similar. The volume of the binding sites of TetR ranges from 359Å³ to 495Å³, whereas the binding site on the E protein is 481Å³, according to the Q-SiteFinder tool ⁴⁰ (the first column of [Table 1.2.2](#)). Therefore, into the binding site there is sufficient space to accommodate tetracycline derivatives. Second, in the pockets of both binding sites they exhibit hydrophobic surfaces ([Figure 1.2.7](#)). Third, a cross-docking test performed for TetR and the four tetracycline derivatives indicated that both the hydrophobic binding sites of the DV E and TetR proteins enabled the binding of the four tetracycline derivatives. Additionally, the hydrogen bonds between the four tetracycline derivative compounds and DV E protein are similar to those between TetR protein and its tetracycline-derived ligand. Therefore, tetracycline derivatives should bind DV E protein at the BOG pocket reasonably well.

Table 1.2.2. Comparisons the hydrogen bonds between five compounds between dengue E protein and TetR protein

	E protein ^a				TetR protein ^b
	Tetracycline	Oxytetracycline	Rolitetraycline	Doxycycline	Chlortetracycline
Total Number of HB	7	9	8	8	6
Backbone	4	5	6	5	1
Side chain	3	4	2	3	5
	Ala50O ^c -4N ^d , 3.59 ^e	Thr48O-4N, 3.62	Ala50N-1O, 2.77	Thr48O-1O, 3.62	Thr103O-10OH, 3.64
	Phe279O-10OH, 2.61	Ala50N-4αOH, 3.01	Thr48O-2N, 3.17	Ala50N-1O, 2.76	
Hydrogen bonding to Backbone	Phe279O-11O, 3.46	Thr48O-4αOH, 3.14	Ala50O-10OH, 3.49	Ala50N-12OH, 3.20	
	Thr48O-12αOH, 2.37	Thr48O-12OH, 2.33	Ala50O-11O, 3.22	Ala50N-12αOH, 3.09	
		Phe279O-11O, 2.62	Ala50O-12OH, 2.97	Thr48O-12αOH, 3.29	
			Ala50N-12O, 3.16		
	Gln271OE-2N, 2.62	Gln271OE-2N, 2.89	Gln200NE-6OH, 3.06	Gln200NE-6OH, 3.14	Gln116NE2-2O, 3.27
Hydrogen bonding to Side chain	Gln200NE-3OH, 2.67	Gln200NE-3OH, 3.33	Glu49OE2-12αOH, 3.20	Glu49OE2-10OH, 3.14	His64NE2-3OH, 2.71
	Thr280OG-11O, 2.68	Thr280OG-11O, 3.01		Glu49OE2-11O, 3.57	Asn82ND2-3OH, 2.82
		Thr280OG-12OH, 2.92			Asn82OD2-4N, 2.70
					His100NE2-12OH, 2.92

^a The docked conformation against E protein (PDB entry 1OKE²⁰)

^b Tet-repressor protein, PDB entry 2TRT³⁹

^c Atom of residues, the number denotes the residue number in target protein

^d Atom of ligand, the number denotes the atom number in tetracycline-related compounds

^e Distance of hydrogen bonding

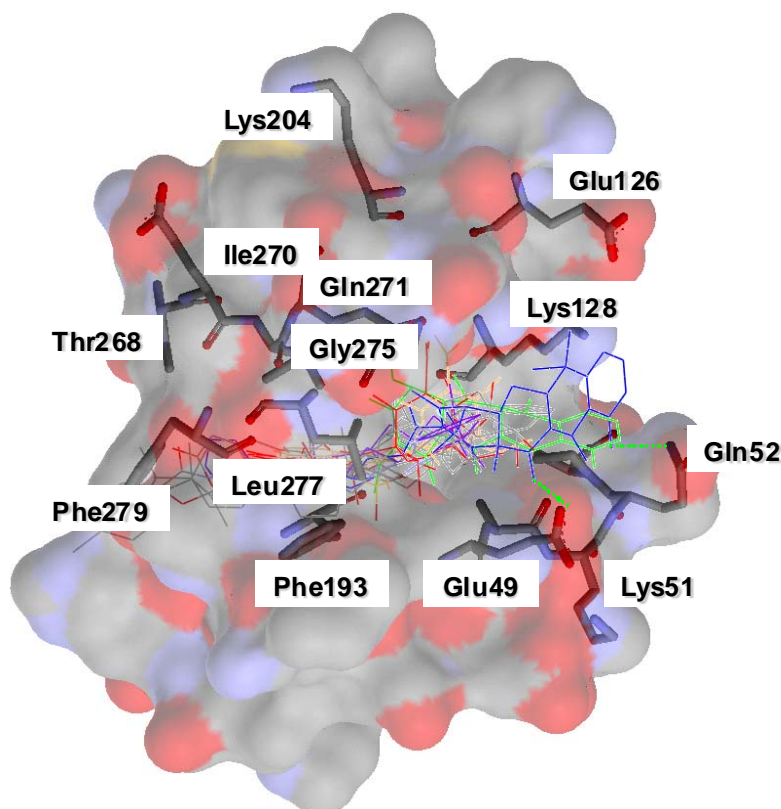


Figure 1.2.7. Docked conformations of 11 selected compounds shown in [Figure 1.2.3](#). Except four tetracycline-like compounds, Doxycycline (green), rolitetracycline (blue), tetracycline (orange), and oxytetracycline (red)), oxetacaine (purple), other compounds are colored with cpk model.

This study proposes an explanation for the inhibition mechanism on DV E protein as a foundation for further investigation. [Figures 1.2.5](#) and [1.2.6](#) indicate the predicted positions of the tetracycline derivatives against E protein. The fused tetracyclic rings of rolitetracycline and doxycycline bind along with the D'₀ strain of the E protein, and occupy the outlet of the binding site. Rolitetracycline and doxycycline both interact with Thr48, Glu49, Ala50 and Gln200 with hydrogen bonds. Such a hydrogen-bonding network provides strong attractive forces, stabilizing the binding of rolitetracycline and doxycycline between the D'₀ strain and kl β-hairpin. By contrast, although tetracycline and oxytetracycline have the same tetracyclic core structures, they showed no inhibitory effect. Both compounds form hydrogen-binding networks with Gln200, Gln271, Phe279 and Thr280. The predicted positions of tetracycline and oxytetracycline are buried deeply in the binding site. Additionally, and the moieties of the tetracyclic ring are docked toward the bottom of the binding site, and contact the surrounding hydrophobic residues via van der Waals interaction. Indeed, the inhibitors never docked to the E protein at the position to which the non-inhibitors docked, revealing that the selected inhibitors indeed possess a plausible binding specificity to the inhibitory location.

In the E protein-host membrane fusion process, the structures of the three domains of E protein are significantly reconfigured to increase the fusion peptide of E protein from the viral membrane for

proper interactions with the host membrane. This conformational modification of E protein plays a critical role in DV infection (Figure 1.2.1). The rearrangement of kl β -hairpin and D'₀ segment (Thr48, Glu49, Ala50, Lys51, Gln52) accompanied the opening of the putative binding site (Figures 1.2.1B and 1.2.6). The docked poses of inhibitors occupied the space of the D'_c strain and kl β -hairpin in the postfusion state, and formed a stable hydrogen-bonding network (Figure 1.2.5). They filled the displacement space, becoming rigid barriers when the rearrangement of β -hairpin and D'₀ strain were blocked (Figure 1.2.6). This network also made the ring moieties of rolitetracycline and doxycycline bind stably to the surface of domain I (Figure 1.2.1A), which may in turn block the rearrangement of domains II and I. Therefore, in addition to the rearrangement region on the domain II, the docked conformation of the inhibitors suggests that the residue region 48–52 is another significant region on the E protein for contacting inhibitors. This finding is consistent with previous reports that Gln52 might affect the pH threshold of fusion in flaviviruses²⁰. Therefore, residues 48–52, as well as being important to inhibitor binding, may also directly affect the fusion of flaviviruses

This study discovered novel inhibitors of the propagation of DV type 2 by performing computer-aided screening against the E glycoprotein, followed by a biological activity assay on the candidates in a cell culture system. The docked conformations indicate that both rolitetracycline and doxycycline block the rearrangement of critical residues involved in the pH threshold of E protein fusion. These compounds can act as the basis for developing new treatment against DV propagation with lesser side effect, and the binding states of these inhibitors can also provide valuable clues for further optimizing E protein inhibitors. This work also notes the additional properties of tetracycline derivatives for being effectively against DV propagation in mammalian cells, which would enable the proposed method to be refined further.

1.3 Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for fast protein structure database search

1.3.1 Introduction

The method of fast peptide and protein structure search is developed in order to build a prototype of peptide drug prediction system. We first developed a web service, termed 3D-BLAST, for the protein structure search according to transforming 3D protein structures into 1D sequences. The core concept of 3D-BLAST is that we aimed to use **structural alphabet** to represent pattern profiles of the backbone peptide fragments by clustering the accumulated (κ , α) plot. The previous results demonstrates the robustness and feasibility of the (κ , α) plot derived structural alphabet for developing a small set of sequence-structure fragments and a fast one-against-all structure database search tool.

A major challenge facing structural biology research in the post-genomics era is to discover the biological functions of genes identified from large-scale sequencing efforts. As protein structures become increasingly available and structural genomics provides structural models in a genome-wide strategy⁴¹, proteins with unassigned functions are accumulating, and the number of protein structures in the Protein Data Bank (PDB) is rapidly rising¹⁵. This current structure-function gap clearly reveals the requirement for powerful bioinformatics methods to discover the structural homology or family of a query protein by known protein sequences and structures.

We have developed a novel kappa-alpha (κ , α) plot derived structural alphabet and a novel BLOSUM-like substitution matrix, called SASM (structural alphabet substitution matrix), for BLAST⁴², which searches on an SADB database. This structural alphabet is valuable for reconstructing protein structures from only a small number of structural fragments, and for developing a fast structure database search method called 3D-BLAST. This tool is as fast as BLAST, and provides the statistical significance (*E*-value) of an alignment to indicate the reliability of a protein structure. To scan a large protein structure database, 3D-BLAST is fast and accurate and will be useful for the initial scan for similar protein structures which are able to be refined by detailed structure comparison methods (e.g. CE and MAMMOTH).

To our best knowledge, 3D-BLAST is the first tool to provide fast protein structure database searching with the *E*-value by using the BLAST, which searches on an SADB database with an SASM matrix. The SADB and the SASM matrix improve the ability of BLAST to search structural homology of a query sequence to a known protein structure or a family of proteins. This tool searches for the structural alphabet high-scoring segment pairs (SAHSPs) existing between a query structure and each structure in the database. Experimental results reveal that the search accuracy of 3D-BLAST is significantly better than that of PSI-BLAST⁴² at $\leq 25\%$ sequence identity.

1.3.2 Materials and Methods

A pair database comprising 674 structural pairs, each with a high structural similarity and low sequence identity, was derived from the SCOP classification database⁴³ for the (κ, α) plot. Each structure in this database (1348 proteins) was divided into a series of 3D protein fragments (225,523 fragments), each five residues long, using kappa (κ) and alpha (α) angles. The angle κ , ranging from 0° to 180° , of residue i is a bond angle formed by three C_α atoms of residues $i - 2$, i , and $i + 2$. The angle α , ranging from -180° to 180° , of a residue i is a dihedral angle formed by the four C_α atoms of residues $i - 1$, i , $i + 1$, and $i + 2$. Each structure has a specific (κ, α) plot when governed by these two angles. When the angles of (κ, α) are divided by 10° , this matrix has 648 cells (36×18). The fragment frequency of each cell in this matrix is unbalanced because the protein structures are significantly conserved with regard to α -helix (82,843 fragments) and β -strand structures (52,371 fragments). Of these helix fragments, 71.1% (58,897 fragments) are located in four cells that contain 22,310, 15,736, 13,013, and 7,838 fragments.

To identify a set of 3D peptide fragments (a structural alphabet), we developed a novel nearest-neighbor clustering (NNC) method to cluster 225,523 peptide fragments in the accumulated (κ, α) plot into 23 groups. The steps of NNC is as follows: (1) identifying a representative peptide fragment for each cell in this matrix; (2) clustering 648 representative peptide fragments into 23 groups by grouping similar representative peptide fragment into individual clusters; (3) in each cluster, identifying a representative peptide fragment and assigning it to a structural alphabet; (4) obtaining a composition of 23 structural alphabets that is similar to the 20 common amino acids. According to the restriction parameter γ , the cell with the highest number of fragments (22,310) in the accumulated (κ, α) plot should be divided into two sub-cells by equally separating the κ and α angles: one is located in $100^\circ \leq \kappa < 115^\circ$ and $40^\circ \leq \alpha < 45^\circ$, and the other is in $105^\circ \leq \kappa < 120^\circ$ and $45^\circ \leq \alpha < 50^\circ$. These two sub-cells were labeled as structural alphabets A and Y, respectively. The NNC method was then applied to cluster the remaining 203,213 fragments into 21 groups.

A representative peptide fragment of each cell in the accumulated (κ, α) plot was first determined. For each cell, a peptide fragment distance matrix (d), stored with the rmsd values by computing all-against-all fragments, was created, and the size was $N \times N$, where N is the total number of the fragments in a cell. An entry (d_{ij}) , which represents the structural distance of fragments i and j , is computed by the rmsd of five C_α atom positions and is given as

$$\left\{ \sum_{k=1}^5 \left[(X_k - x_k)^2 + (Y_k - y_k)^2 + (Z_k - z_k)^2 \right] / 5 \right\}^{1/2} \quad (1.3.1)$$

where (X_k, Y_k, Z_k) and (x_k, y_k, z_k) are the coordinates of the k th atom of the fragments i and j , respectively. For each fragment i , the sum of distance (d_i) between the fragment i and the other fragments in this cell is $\sum_{m=1}^N d_{im}$. The fragment with the minimum sum of distance is selected as

the representative peptide fragment of a cell. After the representative peptide fragment of each cell is identified, a distance matrix (D) is stored with the rmsd values by computing all-against-all representative fragments for these 647 fragments. Each entry (D_{ij} , $1 \leq i, j \leq 647$) is a measure of structural similarity, as defined in Equation 1.3.1, between representative fragments i and j . In order to ensure that the 3D conformations of the fragments clustered in the same group are similar, an rmsd threshold (ε) of the structural similarity is set to 0.5.

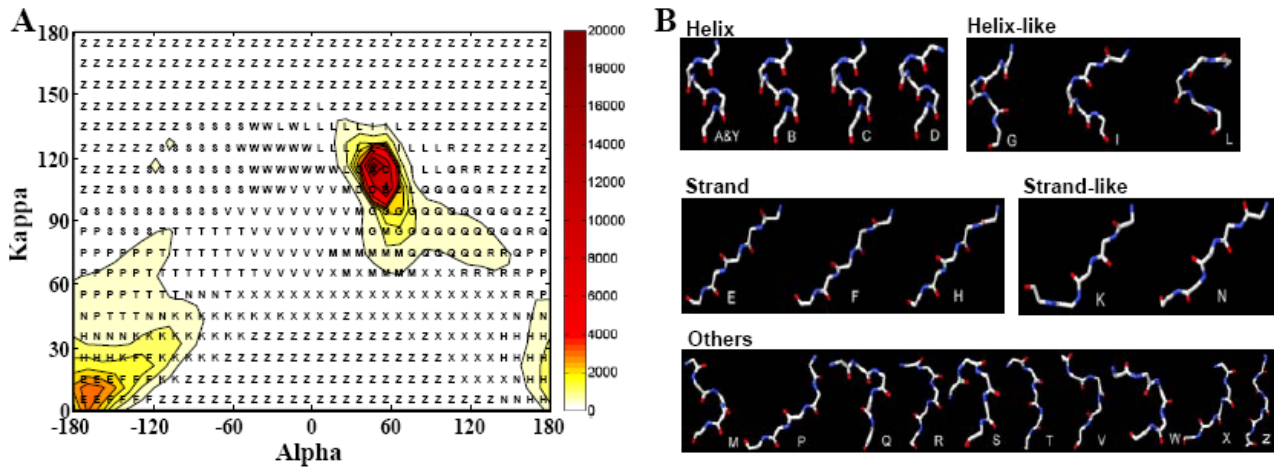


Figure 1.3.1. The distribution and conformation of the 23-state structural alphabet. (A) The distribution of accumulated (κ , α) plot of 225,523 peptide fragments derived from the pair database with 1,348 proteins. This plot, which comprises 648 cells (36×18), is clustered into 23 groups, and each cell is assigned a structure letter. (B) The three-dimensional (3D) peptide conformations of the five main classes of the 23-state structural alphabet are including helix letter (A, Y, B, C, and D), helix-like letters (G, I, and L), strand letters (E, F, and H), strand-like letters (K and N), and others.

Based on the distance matrix D and restriction parameters (ε and γ), the NNC method works as follows: (1) Create a new cluster (C_i , $1 \leq i \leq 20$) by first selecting an unlabeled cell (a) with the maximum number of fragments. Label this cell as C_i . (2) Add an unlabeled cell, which is the nearest neighbor (i.e., a minimum rmsd value in row a of matrix D) of the cell a, into this cluster if this rmsd value is less than ε , and the sum of fragments in this cell is less than γ . Label this cell as C_i . Repeat this step until an added cell violates the restriction thresholds, ε or γ . (3) Repeat steps 1 and 2 until the number of clusters equals 20 or all of the cells are labeled. (4) Assign all of the remaining unlabeled cells to a cluster C_{21} .

Finally, we determined a representative peptide fragment and assigned a structural alphabet for each cluster. For each cell i in a cluster, we defined its weight as $w_i = \frac{1/S_i}{\sum_{i=1}^M 1/S_i}$, where S_i is the number of fragments in cell i and M is the number of cells in this cluster. The sum of distance (D_i) of this fragment i with all of the other cells in the same cluster is equal to $\sum_{m=1}^M w_i w_m D_{im}$, where D_{im} is

the structural distance between representative peptide fragments i and m of the cells i and m , respectively. The fragment with the lowest sum of distance is selected as the representative peptide fragment of this cluster. We sequentially assigned a structural alphabet for each cluster except J , O , and U , since these three letters are not used in BLAST. Figure 1.3.1A shows the distribution of these 23 clusters and structural alphabets on 648 cells in the (κ, α) plot. Figure 1.3.1B shows the 3D conformation of each peptide fragment.

1.3.3 Results and Discussion

A greedy algorithm and the evaluation criteria (global-fit score) presented by Kolodny *et al.*⁴⁴ were applied to measure the performance of 23-state structural alphabet (structural segments) in reconstructing the α - β -barrel protein (PDB code 1TIM-A^{44;45}) and 38 structures selected from the SCOP-516 set, which comprises 516 proteins. This greedy algorithm reconstructs the protein for increasingly large segments of the protein using the best structural fragment, i.e. the one whose concatenation produces a structure with the minimum rmsd from the corresponding segment in the protein from 23 structural segments. No energy minimization procedure was utilized to optimize the reconstructing structures in this study. The global rmsd values were from 0.58 Å to 2.45 Å, and the average rmsd value was 1.15 Å for these 38 proteins.

The 23-state structural alphabet should be able to represent more biological meaning than standard 3-state secondary structural alphabets. First, the classic regular zones of 3-state secondary structures are flexible structures. For instance, α -helices may be curved⁴⁶ and more than one-quarter of them are irregular⁴⁷, and the Φ and Ψ dihedral angles of β -sheets are widely dispersed. The proposed 23-state alphabet describes the α -helices with 8 segments (5 helix letters and 3 helix-like letters) and β -sheets with 5 segments. The 23 structural segments performed well performance in reconstructing protein structures, particularly in the structure segments of classic α -helices and β -sheets. Second, the 3-state secondary structure cannot represent the large conformational variability of coils. Nonetheless, some similar structures can be identified for many of the protein fragments, such as β -turns⁴⁸, π -turns, and β -bulges⁴⁹. Here, 10 structural segments in the 23-state alphabet were utilized to describe the loop conformations. An analysis using the PROMOTIF⁵⁰ tool reveals that most of the segments (>80%) in the letter “W” are β -turns.

An SADB database was easily derived from a known protein structure database based on the (κ, α) plot and the structural alphabet. We have created five SADB databases derived from the following protein structure databases PDB; a non-redundant PDB chain set (nrPDB); all domains of SCOP1.69⁴³; SCOP1.69 with <40% identity to each other, and SCOP1.69 with <95% identity to each other.

The SCOP-516 query protein set, which has a sequence identity below 95% selected from the SCOP database⁴³, was chosen to measure the utility of 3D-BLAST for the discovery of homologous proteins of a query structure. This set contains 516 query proteins that are in SCOP 1.69 but not in

SCOP 1.67, and the search database was SCOP 1.67 (11,001 structures). The total number of alignments was 5,676,516 ($516 \times 11,001$). For evolutionary classification, the first position of the hit list of a query was treated as the evolutionary family/superfamily of this query protein. To compare with several related works on fast database search, 3D-BLAST was also tested on a data set of 108 query domains, termed SCOP-108, proposed by Aung and Tan⁵¹. These queries, which have <40% sequence homology to each other, were chosen from medium-sized families in SCOP. The search database (34,055 structures) represents most domains in SCOP 1.65. Finally, 3D-BLAST was analyzed on 319 structural genomics targets and the search database was the SCOP 1.69 with <95% identity to each other.

Chapter 2: Pharmacophore Identification

2.1 A Pharmacophore-Based Evolutionary Approach for Screening Selective Estrogen Receptor Modulators

2.1.1 Introduction

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^{25; 26}. 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^{25; 52}; and 4) The contribution of VS to the reduction in the cost of drug discovery. VS generally encompasses four phases based both on high-throughput molecular docking methods and the crystal structures of the target protein. These include target protein preparation, compound database preparation, molecular docking, and post-docking analysis²⁵. The molecular docking method screens the compound library to find lead compounds for the target protein, whereas post-docking analysis enriches the hit rate and optimizes the confirmed lead molecules through structure-activity relationship⁵³.

The VS computational method involves two basic critical elements: efficient molecular docking and a reliable scoring method. A molecular docking method for VS should be able to screen a large number of potential ligands with reasonable accuracy and speed. The many molecular docking approaches that have been developed can be roughly categorized as rigid docking⁵⁴, flexible ligand docking^{3; 5}, and protein flexible docking. Most current VS methods employ flexible docking tools, such as incremental and fragment-based approaches (DOCK⁵⁵ and FlexX³) and evolutionary algorithms (GOLD⁵, AutoDock⁵⁶, and GEMDOCK²).

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-based⁵⁷, physics-based⁵⁸, and empirical-based⁵⁹ scoring functions. The performance of these scoring functions is often inconsistent across different systems from a database search^{1; 60}. It has been proposed that combining multiple scoring functions (consensus scoring) improves the enrichment of true positives^{1; 60}.

While the field of VS may be maturing^{25; 26; 52}, and many good VS methods have been proposed, the promise of the virtual compound library⁶¹ to rapidly increase the number of candidate ligands demands further improvement in terms of the computational efficiency of flexible docking

algorithms^{3; 5; 56}. In addition, some 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^{62; 63}. 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. Finally, 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^{64; 65}) and charged polar compounds (due to the pair-atom potentials of the electrostatic energy and hydrogen-bonding energy).

To address the above issues, we developed a new VS method, termed GEMDOCK (Generic Evolutionary Method for molecular DOCKing), modified from our previous studies^{2; 66}. GEMDOCK is an evolutionary-based approach, which was applied in some fast VS algorithms^{5; 56}. Our approach uses multiple operators (e.g., discrete and continuous genetic operators) that cooperate using family competition (similar to a local search procedure) to balance exploration and exploitation. Like some VS methods^{63; 67; 68}, 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. However, unlike existing pharmacophore-based docking methods, we developed and incorporated a new scoring function that evolves a pharmacological consensus (e.g., hot spots) and ligand preferences using the target protein and known active ligands. This scoring function not only serves as the basis for molecular docking but also ranks the screened ligands prior to post-docking analysis by reducing the deleterious effect of certain structural features within some of the ligands.

While GEMDOCK is generally applicable, in particular it has been validated by its application to the docking of a number of selective estrogen receptor modulators (SERMs) that are of great interest in cancer chemotherapy as well as estrogen replacement therapy in postmenopausal women^{69; 70; 71}. To evaluate the strengths and limitations of GEMDOCK and to compare it with several widely used methods (DOCK, GOLD, and FlexX), we evaluated the screening utility of GEMDOCK by testing human estrogen receptor (ER) with the ligand data set, as proposed by Bissantz et al.¹ We also assessed whether our new scoring function was applicable to both the molecular docking and ligand scoring during virtual screening. The screening performance of GEMDOCK on this ligand data set is superior to that of the best available methods, and the docking accuracy is also comparable. Thus, GEMDOCK constitutes a rapid method that reduces the number of false positives during the screening of large databases when both pharmacological interactions and ligand preferences are mined from known active compounds. When known active ligands were not available, the screening accuracy of GEMDOCK is somewhat influenced and is comparable to that of comparative methods on this ligand data set.

2.1.2 Materials and Methods

GEMDOCK was modified and enhanced from our previous tool ² for VS (Figure 2.1.1). GEMDOCK can be sequentially applied to prepare target proteins and ligand databases, predict docked conformations and binding affinity using flexible ligand docking, and rank a series of candidates for post-docking analysis. In this section, we give details of the ligand database and target protein preparations, outline the scoring function used in this study, describe details of mining binding-site pharmacological interactions (e.g., hot spots) and ligand preferences, and briefly describe the docking method.

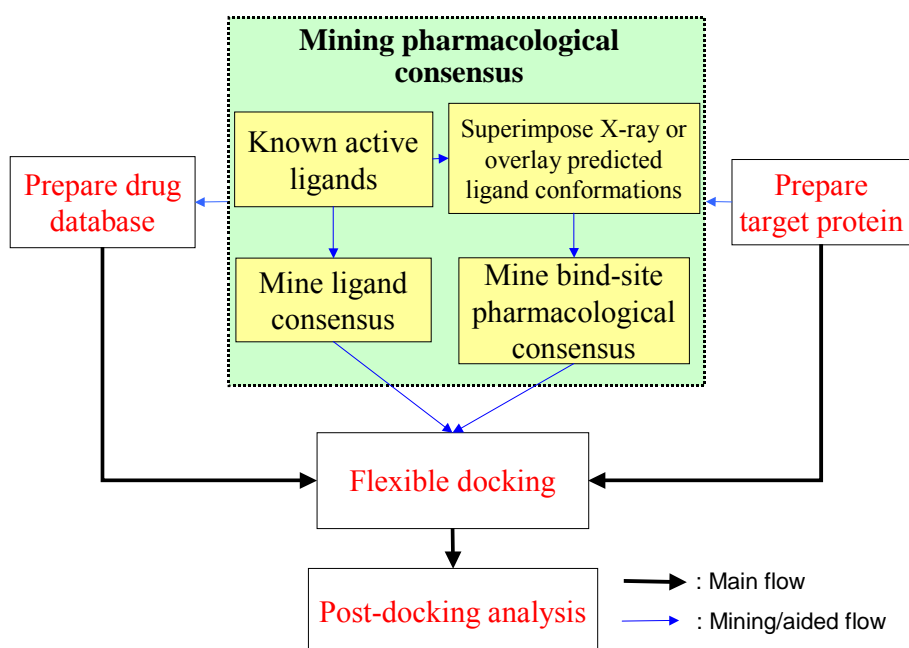


Figure 2.1.1. The main steps of GEMDOCK for virtual database screening, including the target protein and compound database preparation, flexible docking, and post-docking analysis. GEMDOCK mines a pharmacological consensus from the target protein and known active ligands when available.

Preparations of ligand databases and target proteins

SERMs exert their physiological effects by binding to the two currently known estrogen receptors ($ER\alpha$ or $ER\beta$), which are members of the nuclear receptor superfamily of ligand-dependent transcription factors; moreover, SERMs display tissue-selective estrogen agonistic or antagonistic profiles ^{69; 70; 71}. SERMs often beneficially affect the cardiovascular and central nervous systems and exert significant estrogen-like effects on some estrogen targets such as bone, lipid, breast, and uterine cells. Despite of the benefits of SERMs, long-term treatment with SERMs is often limited by

intolerable side effects, such as benign and malignant uterine lesions. Therefore, the design of new SERMs has become a challenging task.

We used the ligand data set and initial ligand conformation from the comparative studies of Bissantz et al. ¹ (e.g., DOCK, FlexX, and GOLD) to evaluate the screening accuracy of GEMDOCK using the ER antagonists. The ligand data set included the 10 known active compounds (EST01–10) listed in [Figure 2.1.2](#) and 990 randomly chosen compounds from the Available Chemical Directory (ACD). The data set is available on the Web at <http://gemdock.life.nctu.edu.tw/dock/download.php>. For screening ER agonists, a set of 10 known ER agonists ([Figure 2.1.3](#), ESA01–10) used in this study was identical to that reported earlier ⁷². In total, the database used for screening ligands against the ER-antagonist complex (PDB code 3ert ⁷¹) and ER-agonist complex (PDB code 1gwr ⁷³) contained 1,000 molecules; that is, 990 random compounds were the same for the two screens. In addition, three ER-antagonist complexes (PDB codes: 1err, 3ert, and 1hj1) and four ER-agonist complexes (PDB codes: 1gwr, 1l2i, 1qkm, and 3erd) with experimentally determined X-ray structures from the PDB were selected to evaluate not only the docking accuracy but also the pharmacological consensuses evolved from known active ligands (i.e., [Figures 2.1.2](#) and [2.1.3](#)) and reference proteins ([Figure 2.1.4](#)). Each ligand from the PDB was represented systematically by four characters followed by three characters. For example, in the ligand “3ert.OHT”, “3ert” denotes the PDB code and “OHT” is the ligand code in the PDB. These ligand structures are shown in [Figure 2.1.2](#) (e.g., EST01, EST02, and EST03) and [Figure 2.1.3](#) (e.g., ESA01, ESA02, ESA03, and ESA04).

The ER-antagonist complex (PDB code: 3ert) and ER-agonist complex (PDB code: 1gwr) were selected as reference proteins for virtual screening. These complexes were reasonable choices because their ligand-binding cavities are wide enough to accommodate a broad variety of ligands and therefore did not require binding site modifications. As shown in [Figure 2.1.4](#), the structures of these two reference proteins complexed with tamoxifen (3ert) or estradiol (1gwr) show that both ligands bind at the same site within the core of the ligand-binding domain and that each ligand induces a different conformation of helix 12 (H12). Comparison of the structures of these two complexes reveals that the H12 (blue) sits above the ligand-binding cavity in the ER-agonist complex (1gwr), thereby forming a lid. In contrast, the side chains of antagonists (e.g., tamoxifen and raloxifene) in the ER-antagonist complexes prevent the agonist-like-induced conformational change of H12 (green), projecting out of the ligand-binding pocket. When preparing the size and location of the ligand-binding site, we considered the protein atoms located less than 10 Å from each ligand atom. The metal atoms were retained and all structured water molecules were removed from the active site. GEMDOCK then assigned a formal charge and atom type for each protein atom based on our previous study ².

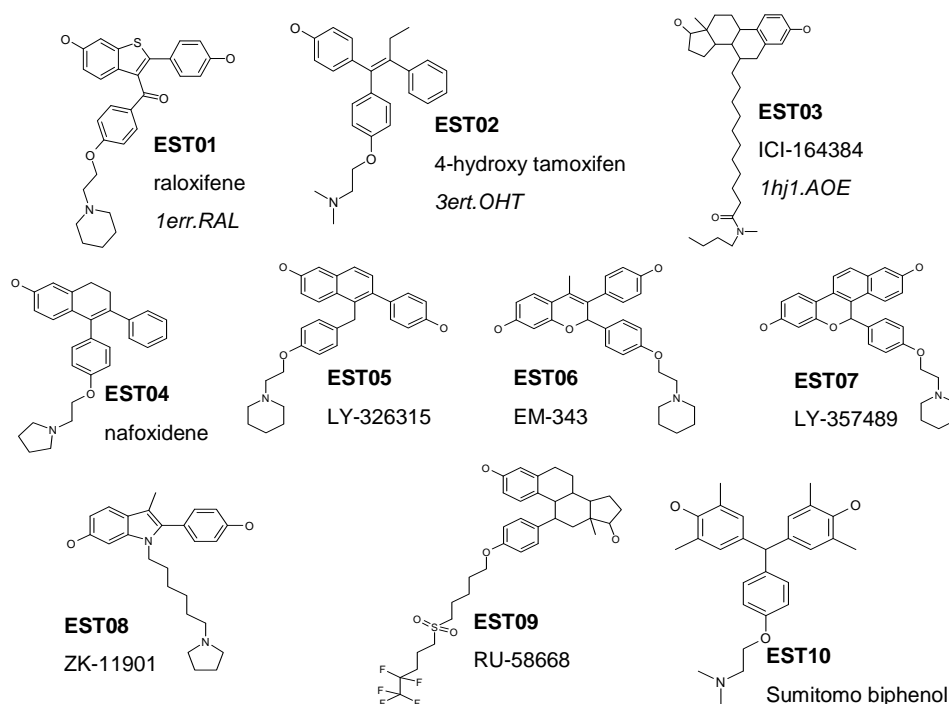


Figure 2.1.2. Ten known ER antagonists are studied with respect to evolving the pharmacological consensus and docking against the ER-antagonist complex. Three ligands, EST01–03, are obtained from the PDB and each ligand is denoted by four characters followed by three characters, as in the PDB (e.g., 3ert.OHT, “3ert” denotes the PDB code and “OHT” is the ligand name in the PDB).

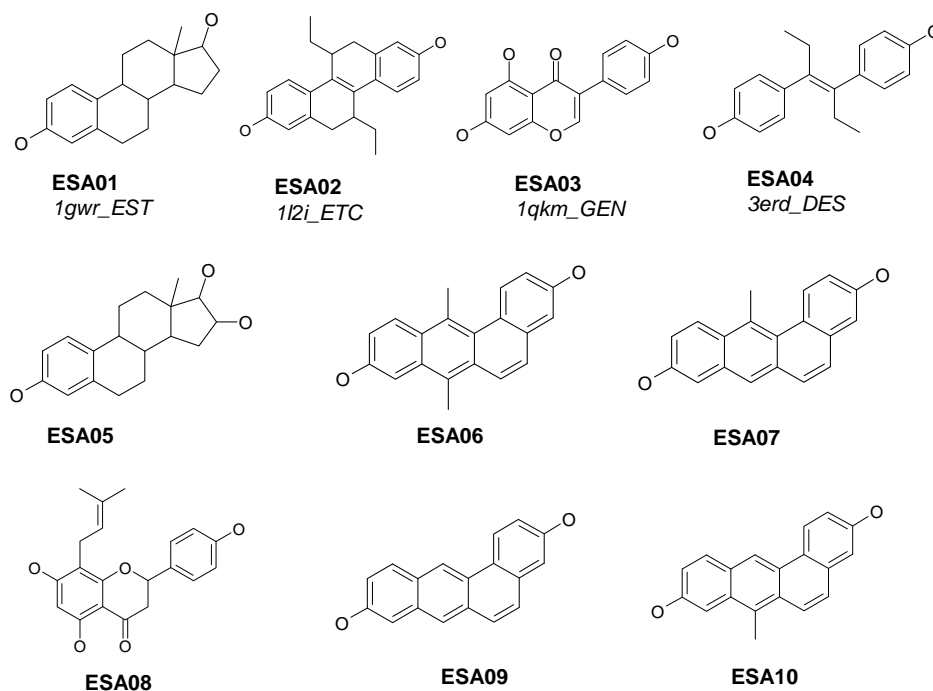


Figure 2.1.3. Ten known ER agonists are docked against the ER-agonist complex (PDB code 1gwr), and the pharmacological consensus is evolved. Four ligands, ESA01–04, are obtained from the PDB and each ligand is represented by four characters followed by three characters in the PDB.

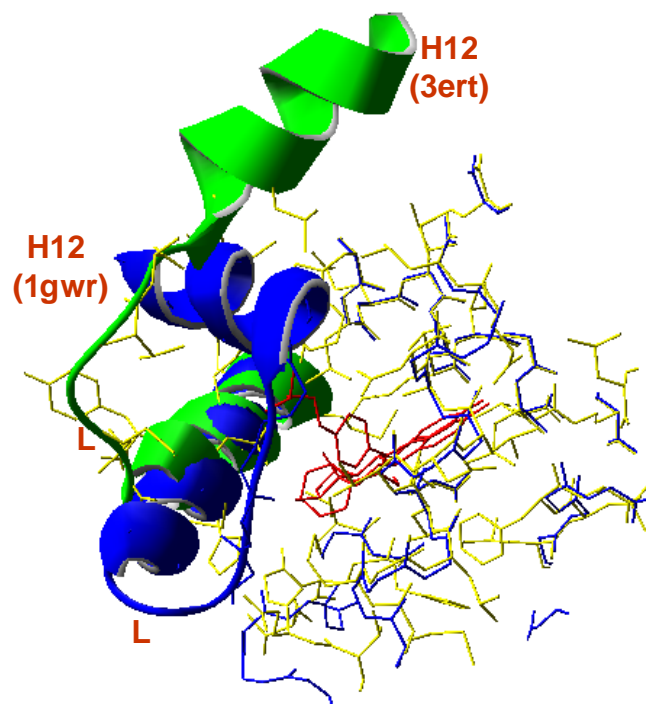


Figure 2.1.4. Comparing the binding sites of the ER reference proteins by superimposing the complexes of the ER agonists (yellow, PDB code: 1gwr) and ER antagonists (blue, PDB entry: 3ert). The bound ligands (estradiol and tamoxifen are shown in red. In the ER-agonist complex, helix 12 (H12) (blue) sits above the ligand-binding cavity, forming a lid. H12 in the ER-antagonist complex protrudes from the pocket.

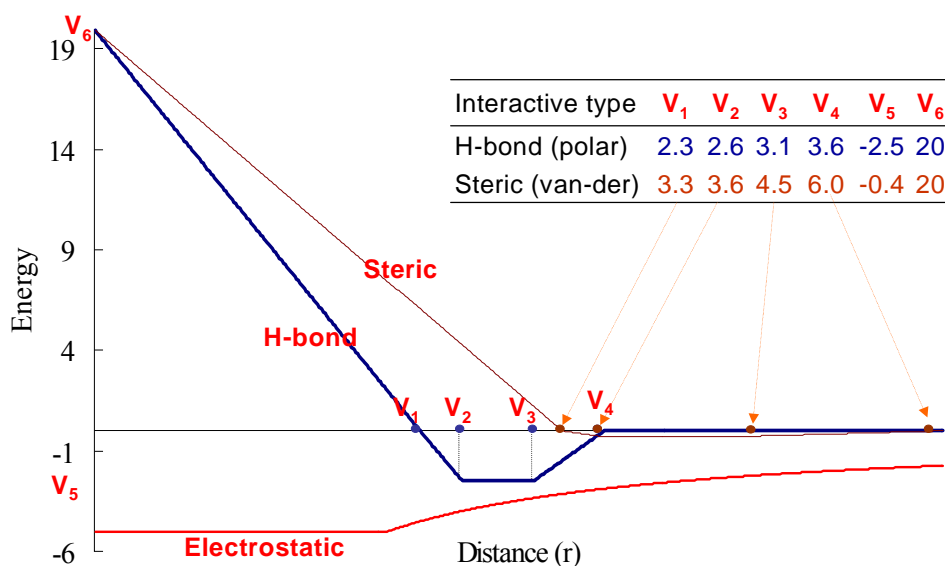


Figure 2.1.5. The linear energy function of pair-wise atoms for steric interactions (light line), hydrogen bonds (bold line), and electrostatic potential in GEMDOCK.

Scoring function

We developed a new scoring function that simultaneously serves as the scoring function for both molecular docking and the ranking of screened compounds for post-docking analysis. This function consists of a simple empirical binding score and a pharmacophore-based score to reduce the number of false positives. The energy function can be dissected into the following terms:

$$E_{tot} = E_{bind} + E_{pharma} + E_{ligpre} \quad (2.1.1)$$

where E_{bind} is the empirical binding energy, E_{pharma} is the energy of binding site pharmacophores (hot spots), and E_{ligpre} is a penalty value if a ligand does not satisfy the ligand preferences. E_{pharma} and E_{ligpre} are especially useful in selecting active compounds from hundreds of thousands of non-active compounds by excluding ligands that violate the characteristics of known active ligands, thereby improving the number of true positives. The values of E_{pharma} and E_{ligpre} are determined according to the pharmacological consensus derived from known active compounds and the target protein. In contrast, the values of E_{pharma} and E_{ligpre} are set to zero if active compounds are not available.

The empirical-binding energy (E_{bind}) is given as

$$E_{bind} = E_{inter} + E_{intra} + E_{penal} \quad (2.1.2)$$

where E_{inter} and E_{intra} are the intermolecular and intramolecular energies, respectively, and E_{penal} is a large penalty value if the ligand is out of the range of the search box. For our present work, E_{penal} was set to 10,000. The intermolecular energy is defined as

$$E_{inter} = \sum_{i=1}^{lig} \sum_{j=1}^{pro} \left[F(r_{ij}^{B_{ij}}) + 332.0 \frac{q_i q_j}{4r_{ij}^2} \right] \quad (2.1.3)$$

where r_{ij} is the distance between the atoms i and j ; q_i and q_j are the formal charges and 332.0 is a factor that converts the electrostatic energy into kilocalories per mole. The *lig* and *pro* denote the numbers of the heavy atoms in the ligand and receptor, respectively. $F(r_{ij}^{B_{ij}})$ is a simple atomic

pair-wise potential function (Figure 2.1.5), as defined in our previous study² where $r_{ij}^{B_{ij}}$ is the distance between atoms i and j with interaction type B_{ij} formed by pair-wise heavy atoms between ligands and proteins, B_{ij} is either a hydrogen bond or a steric state. In this atomic pair-wise model, these two potentials are calculated by the same function form but different parameters, V_1, \dots, V_6 given in Figure 2.1.5. The energy value of a hydrogen bonding should be larger than that for steric potential. In this model, atoms are divided into four different atom types²: donor, acceptor, both, and nonpolar. A hydrogen bond can be formed by the following pair-atom types: donor-acceptor (or acceptor-donor), donor-both (or both-donor), acceptor-both (or both-acceptor), and both-both. Other pair-atom combinations are used to form the steric state. We used the atom formal charge to

calculate the electrostatic energy ², which is set to 5 or -5, respectively, if the electrostatic energy is more than 5 or less than -5. These parameters, V_1 to V_6 , and the maximum electrostatic energy were refined according to the docking accuracies of our previous work ² on a highly diverse dataset of 100 protein-ligand complexes proposed by Jones et al.⁵

The intramolecular energy of a ligand is

$$E_{intra} = \sum_{i=1}^{lig} \sum_{j=i+2}^{lig} \left[F(r_{ij}^{B_{ij}}) + 332.0 \frac{q_i q_j}{4r_{ij}^2} \right] + \sum_{k=1}^{dihed} A [1 - \cos(m\theta_k - \theta_0)] \quad (2.1.4)$$

where $F(r_{ij}^{B_{ij}})$ is defined as for Equation 2.1.3 except the value is set to 1000 when $r_{ij}^{B_{ij}} < 2.0$ Å, and *dihed* is the number of rotatable bonds in a ligand. We followed the work of Gehlhaar et al.⁵⁹ to set the values of A , m , and θ_0 . For the sp^3 - sp^3 bond, $A = 3.0$, $m = 3$, and $\theta_0 = \pi$; for the sp^3 - sp^2 bond, $A = 1.5$, $m = 6$, and $\theta_0 = 0$.

Mining pharmacological consensuses

GEMDOCK evolves the binding-site pharmacological consensus and ligand preferences from both known active ligands and the target protein to improve screening accuracy. We used the premise that previously acquired interactions (hot spots) between ligands and the target protein can be used to guide the selection of lead compounds for subsequent investigation and refinement. When known active ligands were available, GEMDOCK used a pharmacophore-based scoring function (Equation 2.1.1). On the other hand, LP_{elec} and LP_{hb} were set to zero and GEMDOCK used a purely empirical-based scoring function (Equation 2.1.2) if known active compounds were not available.

For each known active ligand, GEMDOCK first yielded 5 docked ligand conformations by docking the ligand into the target protein, and only the docked ligand conformation with the lowest energy was retained for pharmacological consensus analysis. The protein-ligand interactions were extracted by overlapping these lowest-energy docked conformations, and the interactions were classified into two different types, including hydrogen bonding and hydrogen-charged interactions. After all of the protein-ligand interactions were calculated, the atom interaction-profile weight of the target protein representing the pharmacological consensus of a particular interaction was given as

$$Q_j^k = \frac{f_j^k}{N} \quad (2.1.5)$$

where N is the number of known active compounds and f_j^k is the total interaction number of an atom j (in a protein) interacting with an atom of known active ligands with the interaction type k (e.g., hydrogen bonding or hydrogen-charged interactions). In this work, an atom j (in a protein) was considered to interact with an atom i (in a ligand) if the distance between the atoms j and i ranges from $(V_1 + V_2)/2$ to $(V_3 + V_4)/2$, where V_1, \dots, V_4 are given in Figure 2.1.5. An atom j in the reference protein was considered a hot-spot atom when Q_j^k was more than 0.5.

The pharmacophore-based interaction energy (E_{pharma}) between the ligand and the protein is calculated by summing the binding energies of all hot-spot atoms:

$$E_{pharma} = \sum_{i=1}^{lig} \sum_{j=1}^{hs} CW(B_{ij}) F(r_{ij}^{B_{ij}}) \quad (2.1.6)$$

where $CW(B_{ij})$ is a pharmacological-weight function of a hot-spot atom j with interaction type B_{ij} , $F(r_{ij}^{B_{ij}})$ is defined as in Equation 2.1.3, lig is the number of heavy atoms in a screened ligand, and hs is the number of hot-spot atoms in the protein. The $CW(B_{ij})$ is given as

$$CW(B_{ij}) = \begin{cases} 1.0 & \text{if } Q_j^k \leq 0.5 \text{ or } B_{ij} \neq k \\ 1.5 + 5(Q_j^k - 0.5) & \text{if } Q_j^k > 0.5 \text{ and } B_{ij} = k \end{cases} \quad (2.1.7)$$

Q_j^k is the atomic pharmacological-profile weight (Equation 2.1.5) and k is the interaction type of the hot-spot atom j .

We evolved the ligand preferences (E_{ligpre}) from known ligands to reduce the deleterious effects of screening ligand structures that are rich in charged or polar atoms. Docking methods using energy-based scoring functions are often biased toward such compounds, which abound with charged and polar atoms (i.e., hydrogen donor or acceptor atoms) because the pair-atom potential of the electrostatic energy and hydrogen bonding energy is always larger than the steric energy. For example, the atomic pair-wise potential energies of the electrostatic, hydrogen bond, and steric potential were set to -5 , -2.5 , and -0.4 in this work. The ligand preference (E_{ligpre}) is a penalty value for those screened ligands that violate the electrostatic or hydrophilic constraints. The E_{ligpre} is given as

$$E_{ligpre} = LP_{elec} + LP_{hb} \quad (2.1.8)$$

where LP_{elec} and LP_{hb} are the penalties for the electrostatic (i.e., the number of charged atoms of a screened ligand) and hydrophilic (i.e., the fraction of polar atoms in a screened ligand) constraints, respectively. LP_{elec} is defined as

$$LP_{elec} = \begin{cases} 10NA_{elec} & \text{if } NA_{elec} > UB_{elec} \\ 0 & \text{if } NA_{elec} \leq UB_{elec} \end{cases} \quad (2.1.9)$$

$$\text{where } UB_{elec} = \theta_{elec} + \sigma_{elec}$$

, NA_{elec} is the number of charged atoms of a screened ligand and UB_{elec} is the upper bound number of charged atoms derived from known active compounds. θ_{elec} is the maximum number of charged atoms among known active compounds, and σ_{elec} is the standard derivation of the charged atoms of known active compounds. LP_{hb} is defined as

$$LP_{hb} = \begin{cases} 5NA_{hb} & \text{if } r_{hb} > Ur_{hb} \\ 0 & \text{if } r_{hb} \leq Ur_{hb} \end{cases} \quad (2.1.10)$$

$$\text{where } r_{hb} = \frac{NA_{hb}}{NA_t} \text{ and } Ur_{hb} = \theta_{hb} + \sigma_{hb}$$

, r_{hb} is the fraction of polar atoms (i.e., the atom type is both, donor, or acceptor) in a screened ligand and Ur_{hb} is the upper bound of the fraction of polar atoms calculated from known active ligands. NA_{hb} and NA_t are the number of polar atoms and the total number of the heavy atoms of a screened ligand, respectively. θ_{hb} and σ_{hb} are the maximum ratio and the standard derivation of the ratios of polar atoms evolved from known ligands, respectively.

In order to reduce the deleterious effects of biasing toward the selection of high molecular weight compounds, we formulate a normalization strategy defined as

$$E_{bind}^{MW} = \frac{E_{bind}}{(NA_t)^K} \text{ where } K = \begin{cases} 0.5 & \text{if } \mu_{mw} \leq 15 \\ 0.5 - \frac{0.45(\mu_{mw} - 15)}{25} & \text{if } 15 < \mu_{mw} \leq 40 \\ 0.05 & \text{if } \mu_{mw} > 40 \end{cases} \quad (2.1.11)$$

where E_{bind} is the empirical binding energy (Equation 2.1.2), NA_t is the total number of the heavy atoms in a screened ligand, and μ_{mw} is the mean of the number of heavy atoms in known active compounds. When the normalization strategy is applied, the energy function (Equation 2.1.1) is given as

$$E_{tot} = E_{bind}^{MW} + E_{pharma} + E_{ligpre} . \quad (2.1.12)$$

Flexible docking algorithm

Here, we present the outline of our molecular docking method that is a generic evolutionary method enhanced from our original technique². The core idea of our evolutionary approach was to design multiple operators that cooperate using the family competition model, which is similar to a local search procedure. The rotamer-based mutation operator, a discrete operator, is used to reduce the search space of ligand structure conformations. The Gaussian and Cauchy mutations, continuous genetic operators, search the orientation and conformation of the ligand relating to the center of the target protein.

After the ligand database and the target protein were prepared and the pharmacological preferences were evolved, we first specified the crystal coordinates of the protein atoms from the PDB and assigned a formal charge and atom type for each protein atom. GEMDOCK then automatically decides the search cube of a binding site based on the maximum and minimum values of coordinates among these selected protein atoms. For each ligand in the database, the GEMDOCK takes the atomic coordinates from the ligand database and assigns a formal charge and atom type for each atom. It then sequentially predicts the binding conformation and estimates the binding affinity for

each ligand. Finally, GEMDOCK ranks these docked ligand conformations for use in the post-docking analysis.

Our docking method works as follows: It randomly generates a starting population with N docked structures by initializing the orientation and conformation of the ligand relating to the center of the target protein. Each solution is represented as a set of three n -dimensional vectors (x^i, σ^i, ψ^i) , where n is the number of adjustable variables of a docking system and $i = 1, \dots, N$ where N is the population size. The vector x is the adjustable variables, representing a particular orientation and conformation space of a ligand, to be optimized in which x_1, x_2 , and x_3 are the three-dimensional location of the ligand relating to the center of the target protein; x_4, x_5 , and x_6 are the rotational angles of the ligand relating to axes; and from x_7 to x_n are the twisting angles of the rotatable bonds inside the ligand. σ and ψ are the step-size vectors of decreasing-based Gaussian mutation and self-adaptive Cauchy mutation. In other words, each solution x is associated with some parameters for step-size control. The initial values of x_1, x_2 , and x_3 are randomly chosen from the feasible box, and the others, from x_4 to x_n , are randomly chosen from 0 to 2π in radians. The initial step sizes σ is 0.8 and ψ is 0.2. After GEMDOCK initializes the solutions, it enters the main evolutionary loop which consists of two stages in every iteration: decreasing-based Gaussian mutation and self-adaptive Cauchy mutation. Each stage is realized by generating a new quasi-population (with N solutions) as the parent of the next stage. These stages apply a general procedure “FC_adaptive” with only different working population and the mutation operator.

The FC_adaptive procedure employs two parameters, namely, the working population (P , with N solutions) and mutation operator (M), to generate a new quasi-population. The main work of FC_adaptive is to produce offspring and then conduct the family competition. Each individual in the population sequentially becomes the “family father”. With a probability p_c , this family father and another solution that is randomly chosen from the rest of the parent population are used as parents for a recombination operation. Then the new offspring or the family father (if the recombination is not conducted) is operated by the rotamer mutation or by differential evolution to generate a quasi offspring. Finally, the working mutation is operates on the quasi offspring to generate a new offspring. For each family father, such a procedure is repeated L times called the family competition length. Among these L offspring and the family father, only the one with the lowest scoring function value survives. Since we create L children from one “family father” and perform a selection, this is a family competition strategy. This method avoids the population prematureness but also keeps the spirit of local searches. Finally, the FC_adaptive procedure generates N solutions because it forces each solution of the working population to have one final offspring. In the following, genetic operators are briefly described. We use $a = (x^a, \sigma^a, \psi^a)$ to represent the “family father” and $b = (x^b, \sigma^b, \psi^b)$ as another parent. The offspring of each operation is represented as $c = (x^c, \sigma^c, \psi^c)$. The symbol x_j^s is used to denote the j th adjustable optimization variable of a solution s , $\forall j \in \{1, \dots, n\}$.

Recombination operators. GEMDOCK implemented modified discrete recombination and intermediate recombination. A recombination operator selected the “family father (a)” and another solution (b) randomly selected from the working population. The former generates a child as follows:

$$x_j^c = \begin{cases} x_j^a & \text{with probability 0.8} \\ x_j^b & \text{with probability 0.2} \end{cases}$$

The generated child inherits genes from the “family father” with a higher probability 0.8. Intermediate recombination works as:

$$w_j^c = w_j^a + \beta(w_j^b - w_j^a)/2$$

where w is σ or ψ based on the mutation operator applied in the FC_adaptive procedure. The intermediate recombination only operated on step-size vectors and the modified discrete recombination was used for adjustable vectors (x).

Mutation operators. After the recombination, a mutation operator, the main operator of GEMDOCK, is applied to mutate adjustable variables (x). Gaussian and Cauchy Mutations are accomplished by first mutating the step size (w) and then mutating the adjustable variable x :

$$\begin{aligned} w_j' &= w_j A(\cdot) \\ x_j' &= x_j + w_j' D(\cdot) \end{aligned}$$

where w_j and x_j are the i th component of w and x , respectively, and w_j is the respective step size of the x_j where w is σ or ψ . $A(\cdot)$ is evaluated as $\exp[\tau'N(0, 1)+N_j(0, 1)]$ if the mutation is a self-adaptive mutation, where $N(0, 1)$ is the standard normal distribution, $N_j(0, 1)$ is a new value with distribution $N(0, 1)$ that must be regenerated for each index j . When the mutation is a decreasing-based mutation $A(\cdot)$ is defined as a fixed decreasing rate $\gamma = 0.95$. $D(\cdot)$ is evaluated as $N(0, 1)$ or $C(1)$ if the mutation is, respectively, Gaussian mutation or Cauchy mutation. For example, the self-adaptive Cauchy mutation is defined as

$$\psi_j^c = \psi_j^a \exp[\tau'N(0,1) + \tau N_j(0,1)],$$

$$x_j^c = x_j^a + \psi_j^c C_j(t)$$

We set τ and τ' to $(\sqrt{2n})^{-1}$ and $(\sqrt{2\sqrt{2n}})^{-1}$, respectively, according to the suggestion of evolution strategies. A random variable is said to have the Cauchy distribution ($C(t)$) if it has the

density function: $f(y; t) = \frac{t/\pi}{t^2 + y^2}$, $-\infty < y < \infty$. In this paper t is set to 1. Our decreasing-based Gaussian mutation uses the step-size vector σ with a fixed decreasing rate $\gamma = 0.95$ and works as

$$\sigma^c = \gamma\sigma^a \quad \text{and} \quad x_j^c = x_j^a + \sigma^c N_j(0,1)$$

Our rotamer mutation is only used for x_7 to x_n to find the conformations of the rotatable bonds inside the ligand. For each ligand, this operator mutates all of the rotatable angles according to the rotamer

distribution and works as $x_j = \gamma_{ki}$ with probability p_{ki} , where γ_{ki} and p_{ki} are the angle value and the probability, respectively, of i th rotamer of k th bond type including sp^3-sp^3 and sp^3-sp^2 bond. The values of γ_{ki} and p_{ki} are based on the energy distributions of these two bond types.

2.1.3 Results and Discussion

Parameters of GEMDOCK

In our studies, GEMDOCK parameters in the flexible search phase included the initial step sizes ($\sigma=0.8$ and $\psi=0.2$), family competition length ($L = 2$), population size ($N = 200$), and recombination probability ($p_c = 0.3$). For each ligand screened, GEMDOCK optimization stopped either when the convergence was below a certain threshold value or the iterations exceeded the maximal preset value of 60. Therefore, GEMDOCK generated 800 solutions in one generation and terminated after it exhausted 48000 solutions for each docked ligand. The average GEMDOCK docking run took 135 s using a Pentium 1.4-GHz personal computer with a single processor.

Mining the pharmacological consensus

Figure 2.1.6 and Table 2.1.1 show the pharmacological interaction preferences (hot-spot atoms), and Table 2.1.2 shows the ligand preferences. We evolved these pharmacological consensuses and steric binding interactions by overlapping the docked ligand conformations, yielded by GEMDOCK, of all known active compounds. Figures 2.1.6(a) and 2.1.6(b) show the overlap of ten docked poses of ten known active ligands in the vicinity of the ER-antagonist target protein and ER-agonist target protein, respectively. The dashed lines indicate the hydrogen bonds formed between the ligand and the reference proteins. For the ER-antagonist target protein, four binding-site pharmacological interactions were identified and circled as A (hydroxyl group^{71; 74; 75; 76; 77}), B (hydroxyl group^{71; 74; 75; 76}), and C (dimethylamino group^{71; 76} or piperidine nitrogen^{74; 75}). These interactions, evolved from docked conformations, are consistent with the interactions evolved from superimposing three X-ray structures with that from related studies^{71; 74; 75; 76}. As shown in Table 2.1.1, the pharmacological weights ($CW(B_{ij})$) defined in Equation 2.1.7) and the interaction type for the ER-antagonist complex included E353-OE2 (3.0), R394-NH2 (2.9), H524-ND1 (2.4), and D351-OD1 (2.4). For the ER-agonist target protein, two binding-site pharmacological interactions were identified (e.g., A' hydroxyl group and B' hydroxyl group). The pharmacological weights and the interaction type for the ER-agonist complex included E353-OE2 (3.1), R394-NH2 (3.1), and H524-ND1 (3.4). These interactions are also consistent with those evolved by superimposing four X-ray structures (Figure 2.1.6(b)).

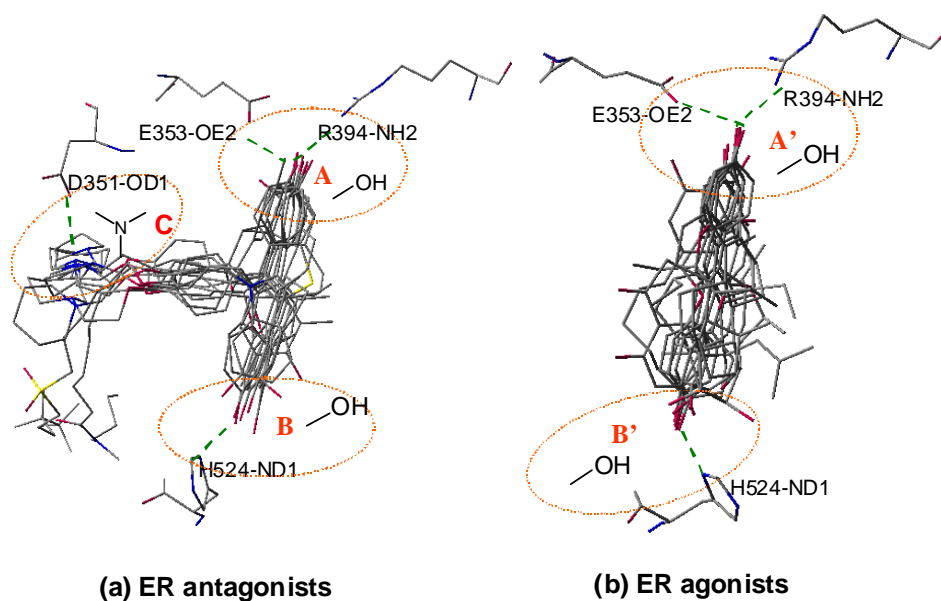


Figure 2.1.6. The binding-site pharmacological consensus are identified by overlapping the docked conformations of (a) ten known ER antagonists and (b) ten known ER agonists against the reference proteins 3ert and 1gwr, respectively. (a) Four pharmacological interactions were identified and circled as 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.

Table 2.1.1. Pharmacological weights of hot-spot atoms of the ER-antagonist and ER-agonist complexes are evolved by overlapping docked conformations of known active ligands.

Residue Id ^a	Atom Id ^b	Hot-spots weight ($CW(B_{ij})$)		Interaction type (Hot spots)
		ER-antagonist complex	ER-agonist complex	
E353	OE2	3.0	3.1	H-bond (OH ↔ O) (phenolic hydroxyl) ^{71; 74; 75; 76; 77}
R394	NH2	2.9	3.1	H-bond (OH ↔ N) (phenolic hydroxyl) ^{71; 74; 75; 76; 77}
H524	ND1	2.4	3.4	H-bond (OH ↔ N) ^{71; 74; 75; 76; 77}
D351	OD1	2.2	- ^c	H-bond (N ↔ O) (dimethylamino group and piperidine nitrogen) ^{74; 75; 76}

^a One-letter amino acid code with the residue sequence numbered as in the PDB.

^b The atom name in the PDB.

^c D351-OD1 is not a hot-spot atom in the ER-agonist reference complex.

Table 2.1.2. Ligand preferences evolved from known active ligands are used to screen lead compounds for the ER-antagonist and ER-agonist complexes

Ligand name	Electrostatic preferences (Equation 2.1.9)			Hydrophilic preferences (Equation 2.1.10)			Molecular weight (Equation 2.1.11)	
	θ_{elec}	σ_{elec}	UB_{elec}	θ_{hb}	σ_{hb}	Ur_{hb}	μ_{mw}	K
ER antagonist	2.0	0.63	2.63	0.15	0.02	0.17	34	0.16
ER agonist	0	0	0	0.25	0.06	0.31	21.4	0.38

For screening ER antagonists and agonists, Table 2.1.2 shows the parameter values of ligand preferences evolved from known ER antagonists (Figure 2.1.2) and agonists (Figure 2.1.3). These ligand preferences improve the screening accuracy by reducing the deleterious effects of ligand molecular weights and ligand structures that are rich in charged or polar atoms. The electrostatic parameter values (see Equation 2.1.9) for ER antagonists included the maximum number of charged atoms ($\theta_{elec}=2.0$), standard derivation of the charged atoms ($\sigma_{elec}=0.63$), and upper bound number of charged atoms ($UB_{elec}=2.63$). For the hydrophilic preferences (see Equation 2.1.10), the maximum ratio (θ_{hb}) was 0.15, the standard derivation (σ_{hb}) of the ratios was 0.02, and the upper bound ratio (Ur_{hb}) of polar atoms was 0.17. For molecular weight (see Equation 2.1.11), the mean of heavy atoms (μ_{mw}) was 21.6 and linear normalization parameter K was 0.16. In contrast, for ER agonists the values of UB_{elec} and Ur_{hb} were 0 and 0.31, respectively, and K was 0.38.

Evaluation of virtual screening accuracy

Some common factors were used to evaluate the screening quality, including coverage (the percentage of active ligands retrieved from the database), yield (the percentage of active ligands in the hit list), false positive (FP) rate, enrichment, and goodness-of-hit (GH). The coverage (true positive rate) is defined as A_h/A (%), A_h/T_h (%) is the yield (hit rate), and the FP rate is defined as $(T_h-A_h)/(T-A)$ (%). The enrichment is defined as $(A_h/T_h)/(A/T)$. A_h is the number of active ligands among the T_h highest ranking compounds which is called the hit list, A is the total number of active ligands in the database, and T is the total number of compounds in the database. The GH score is defined as⁷⁸

$$GH = \left(\frac{A_h(3A + T_h)}{4T_h A} \right) \left(1 - \frac{T_h - A_h}{T - A} \right). \quad (2.1.13)$$

The GH score contains a coefficient to penalize excessive hit list size and, when evaluating hit lists, is calibrated by weighting the score with respect to the yield and coverage. The GH score ranges from 0.0 to 1.0, where 1.0 represents a perfect hit list (i.e., containing all of, and only, the active ligands). In the data sets for screening the ER agonists or ER antagonists, A and T are 10 and 1000, respectively. Here, we also took the averages of hit rates, enrichments, GH scores, and FP rates. For

example, the averages of the hit rates and enrichments are defined as $(\sum_{i=1}^A i / T_h^i) / A$ and $\{\sum_{i=1}^A (i / T_h^i) / (A / T)\} / A$, respectively, where T_h^i is the number of compounds in a hit list containing i active compounds.

Molecular recognition of ER-antagonist and ER-agonist complexes

We tested GEMDOCK² on a highly diverse data set of 100 protein-ligand complexes proposed by Jones et al.⁵ and on two cross-docking ensembles of protein structures. Upon consideration of the solutions at the first rank, in 79% of these complexes the docked lowest energy ligand structures had root-mean-square derivations (RMSDs) below 2.0 Å with respect to the corresponding crystal structures. The success rate increased to 85% if the structured water molecules were retained. In contrast, GOLD⁵ yielded a 71% success rate in identifying the experimental binding model based on the GOLD assessment categories, and the rate was 66% if based on the top-ranked solutions with RMSD values of less than 2 Å. FlexX³ achieved 70% and 46.5% success rates for solutions at any rank and the first rank, respectively.

The main objective of this study was to evaluate whether the new scoring function was applicable to both molecular docking and ligand scoring during virtual screening. First, GEMDOCK was evaluated by docking each ligand of seven ER complexes in the PDB into its respective complex and into its reference protein. Table 2.1.3 shows the overall predicted accuracy of GEMDOCK and GOLD. Ten independent docking runs were performed for each active compound, and the docked ligand conformation with the lowest energy was used to calculate RMSD values for ligand heavy atoms between the docked conformation and the crystal structure. The RMSD values of seven docked conformations (docking each ligand back into its respective complex) were less than 2.0 Å. When these ligands were docked into the reference protein using GEMDOCK, all docked conformations had an RMSD of less than 2.0 Å except for EST03 and ESA03 (genistein). EST03 docked well in the binding site with the exception of the long acyclic side chain. The agonist ESA03 could not be docked into its corresponding pose in the reference protein (1gwr) due to a fundamental difference between the binding site of ER α (1gwr) and ER β (1qkm). As shown in Table 2.1.3, GEMDOCK and GOLD yielded results of equal quality, and GEMDOCK yielded similar results regardless of whether the pharmacological preferences (i.e., E_{pharma} and E_{ligpre}) were considered.

Table 2.1.3. Comparing GEMDOCK with GOLD with respect to docking seven ligands back into respective complexes and reference proteins

Ligand id	GEMDOCK				GOLD	
	Native protein ^b		Reference Protein ^c		Native protein ^b	Reference protein ^c
	E_{tot} ^d	E_{bind} ^d	E_{tot}	E_{bind}		
EST01 (1err.RAL ^a)	0.66	0.65	1.37	1.36	1.02	1.68
EST02 (3ert.OHT)	0.60	0.75	0.60	0.75	1.15	1.15
EST03 (1hj1.AOE)	1.41	1.05	3.27	3.35	5.07	3.92
ESA01 (1gwr.EST)	0.66	0.64	0.66	0.64	0.54	0.54
ESA02 (112i_ETC)	0.61	0.48	0.62	0.69	0.55	0.76
ESA03 (1qkm.GEN)	0.69	1.53	3.32	4.83	0.24	7.16
ESA04 (3erd.DES)	0.67	0.51	1.44	1.43	1.10	1.76

^a Four characters followed by three characters (separated by a period) denote the PDB code and the ligand name in the PDB, respectively.

^b The RMSD value for docking each ligand back into its respective complex.

^c The RMSD value for docking each ligand into its reference complex, 3ert for ER antagonists (e.g., EST01 ~ EST03) and 1gwr for ER agonists (e.g., ESA01 ~ ESA04).

^d E_{tot} and E_{bind} are defined in Equation 2.1.1.

Virtual screening of ER antagonists and ER agonists

We compared the overall accuracy of GEMDOCK using four variations of energy terms to screen ER antagonists and agonists from a data set of 1,000 compounds proposed by Bissantz et al. ¹ (Figure 2.1.7 and Table 2.1.4). Each variation combined three scoring terms applied in GEMDOCK: binding energy (E_{bind}), pharmacological-interaction preferences (E_{pharma}), and ligand preferences (E_{ligpre}). For example, the approach “Pure binding” used only the binding energy (E_{bind}) as the scoring function; the approach “Interaction preference” integrated E_{bind} and E_{pharma} for the scoring function; “Ligand preference” integrated E_{bind} and E_{ligpre} for the scoring function; and “Both” integrated E_{bind} , E_{ligpre} , and E_{pharma} for the scoring function. The parameter values for interaction preferences (E_{pharma}) and ligand preferences (E_{ligpre}) are shown in Tables 2.1.1 and 2.1.2, respectively. The various ranks of ten known active ligands in the ligand screening database are shown in Table 2.1.5, and the comparison of results obtained with other methods is shown in Table 2.1.6.

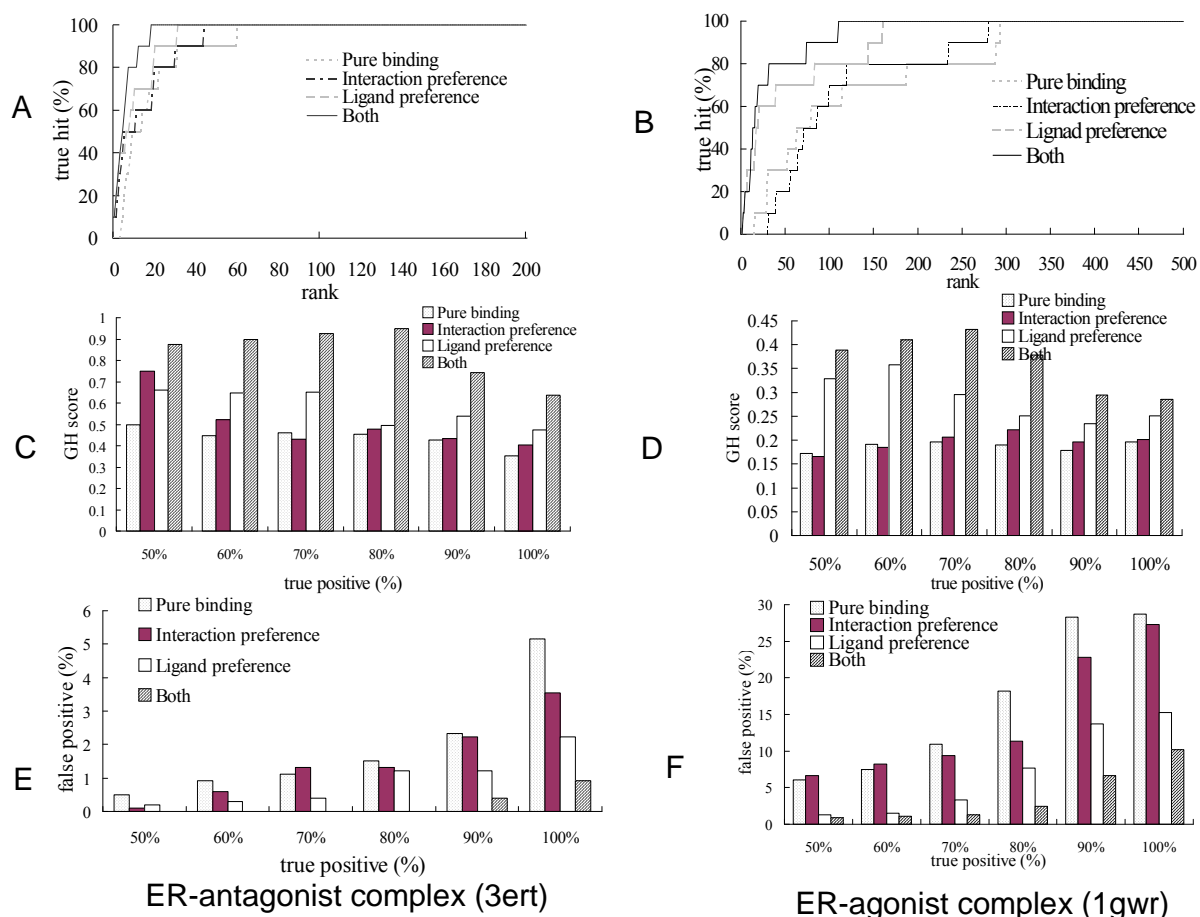


Figure 2.1.7. GEMDOCK screening accuracy of ER antagonists and ER agonists assessed by (A and B) true hits, (C and D) GH scores, and (E and F) the false positive rates against different true positive rates ranging from 50% to 100%. The performance of GEMDOCK was consistently superior when using both ligand preferences and pharmacological-interaction preferences.

Table 2.1.4. GEMDOCK screening accuracy using different combinations of pharmacological preferences on the data set proposed by Bissantz et al.¹

Measure factor	ER antagonists (reference protein: 3ert)				ER agonists (reference protein: 1gwr)			
	Pure binding ^a	Interaction preference ^b	Ligand preference ^c	Both ^d	Pure binding ^a	Interaction preference ^b	Ligand preference ^c	Both ^d
Average hit rate (%)	34.88	57.93	71.58	92.19	6.94	7.52	25.02	45.66
Average enrichment	34.88	57.93	71.58	92.19	6.94	7.52	25.02	45.66
Average false positive rate (%)	1.32	0.94	0.56	0.13	7.83	6.34	2.56	0.75
Average GH score	0.39	0.57	0.67	0.83	0.17	0.18	0.32	0.48

^{a,b,c,d} Using E_{bind} , $E_{bind} + E_{pharma}$, $E_{bind} + E_{ligpre}$, and E_{tot} , respectively, for the scoring function. These energy terms are defined in Equation 2.1.1.

Table 2.1.5. The ranks of ten known ER antagonists and ten known ER agonists using GEMDOCK with different combinations of pharmacological preferences on the data set proposed by Bissantz et al.¹

ER antagonists (reference protein: 3ert)					ER agonists (reference protein: 1gwr)				
Ligand id ^a	Pure binding ^b	Interaction preference ^c	Ligand preference ^d	Both ^e	Ligand id ^f	Pure binding	Interaction preference ^e	Ligand preference ^e	Both
EST01	9	3	3	3	ESA01	87	57	33	8
EST02	23	31	21	13	ESA02	25	49	7	6
EST03	10	20	20	8	ESA03	31	32	3	3
EST04	15	12	7	4	ESA04	220	116	99	29
EST05	6	6	1	1	ESA05	128	97	53	20
EST06	7	5	4	6	ESA06	101	73	41	14
EST07	32	21	9	7	ESA07	53	53	16	7
EST08	18	4	11	5	ESA08	45	102	9	26
EST09	5	1	2	2	ESA09	43	38	10	5
EST10	61	45	32	19	ESA10	97	66	37	11

^{a,f} Defined in Figures 2.1.2 and 2.1.3, respectively.

^{b,c,d,e} Using E_{bind} , $E_{bind} + E_{pharma}$, $E_{bind} + E_{ligpre}$, and E_{tot} , respectively, for the scoring function. These energy terms are defined in Equation 2.1.1.

Table 2.1.6. Comparing GEMDOCK with other methods on screening the ER antagonists by false positive rates (%) on the data set proposed by Bissantz et al.¹

True positive (%)	GEMDOCK ^a	GEMDOCK ^b	Surflex ^c	DOCK ^c	FlexX ^c	GOLD ^c
80	1.5 (15/990) ^d	0.0 (0/990)	1.3	13.3	57.8	5.3
90	2.3 (23/990)	0.4 (4/990)	1.6	17.4	70.9	8.3
100	5.2 (51/990)	0.9 (9/990)	2.9	18.9	- ^e	23.4

^a GEMDOCK without pharmacological-interaction and ligand preferences (e.g., E_{bind} for the scoring function).

^b GEMDOCK with pharmacological interactions and ligand preferences (e.g., E_{tot} for the scoring function).

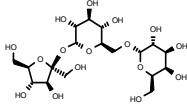
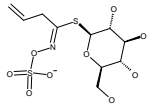
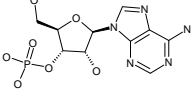
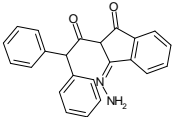
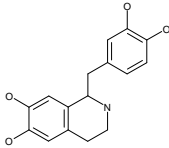
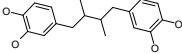
^c Directly summarized from the references^{18; 79}

^d The false positive rate from 990 random ligands (percentage).

^e FlexX could not calculate the docked solution for EST09.

As shown in Table 2.1.4 and Figure 2.1.7, GEMDOCK generally improves the screening quality when both interaction preferences and ligand preferences are considered. The latter was more important than the former for this data set. For the ER antagonists that were screened, average hit rates were 92.19% (Both), 71.58% (Ligand preference), 57.93% (Interaction preference), and 34.8% (E_{bind}). The average GH scores were 0.83 (Both), 0.67 (Ligand preference), 0.57 (Interaction preference), and 0.39 (E_{bind}). Figures 2.1.7C and 2.1.7E show that the GH scores and FP rates of the true positive rates ranged from 50% to 100%. For the ER agonists that were screened, average hit rates were 45.66% (Both), 25.02% (Ligand preference), 7.52% (Interaction preference) and 6.94% (E_{bind}). The average GH scores were 0.48 (Both), 0.32 (Ligand preference), 0.18 (Interaction preference), and 0.17 (E_{bind}). Figures 2.1.7D and 2.1.7F show the GH scores and FP rates with different true positive rates ranging from 50% to 100%.

Table 2.1.7. GEMDOCK ranks using different combinations of pharmacological preferences for some typical ligands on screening ER agonists on the data set proposed by Bissantz et al. ¹

Ligand id in ACD	Ligand structure	NA_{elec}^a	r_{hb}^b	Pure binding ^c	Interaction preference ^d	Ligand preference ^e	Both ^f
MFCD00006630		0.00	0.47	5	172	911	850
MFCD00006616		3.00	0.45	3	1	900	828
MFCD00005746		3.00	0.52	4	2	925	889
MFCD00003783		0.00	0.15	54	270	13	165
MFCD00012742		0.00	0.24	10	6	2	1
MFCD00002206		0.00	0.18	13	11	1	4

^a The number of charged atoms in a screened ligand (Equation 2.1.9).

^b The fraction of polar atoms in a screened ligand (Equation 2.1.10).

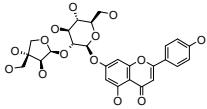
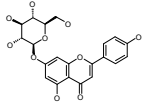
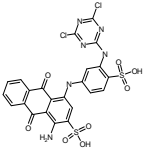
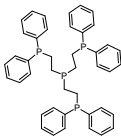
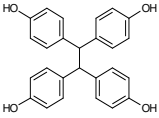
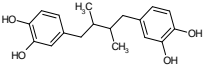
^{c,d,e,f} Using E_{bind} , $E_{bind} + E_{pharma}$, $E_{bind} + E_{ligpre}$, and E_{tot} , respectively, for the scoring function. These energy terms are defined in Equation 2.1.1.

The screening accuracy of GEMDOCK for ER antagonists was better than that of ER agonists on this data set. These results might be caused by using the same 990 random compounds proposed by Bissantz et al. ¹ for these two screens. When they prepared the random ligand set, only the chemical reagents of the ER-antagonist complex were eliminated and therefore the ER-agonist-like compounds might be selected. For example, GEMDOCK screened two ligands, MFCD00012742 and MFCD00002206 (Table 2.1.7), which are similar in structures to ESA03 and ESA04 (Figure 2.1.4), respectively. At the same time, the numbers of the ligands, which violate the ligand preferences, e.g., LP_{elec} and LP_{hb} shown in Table 2.1.2, of ER antagonists and ER agonists, are 400 and 289 compounds, respectively. The MFCD compounds were the random ligands in the data set.

GEMDOCK was superior to other approaches (Surflex, DOCK, FlexX, and GOLD) for screening the ER antagonists (Table 2.1.6). All of these methods were tested using the same reference protein and screening database with true positive rates ranging from 80% to 100%. When the true positive

rate was 90%, the FP rates were 2.3% (GEMDOCK without pharmacological preferences), 0.4% (GEMDOCK with pharmacological preferences), 1.6% (Surflex), 17.4% (DOCK), 70.9% (FlexX), and 8.3% (GOLD).

Table 2.1.8. GEMDOCK ranks using different combinations of pharmacological preferences for some typical ligands when screening ER antagonists on the data set proposed by Bissantz et al. ¹

Ligand id in ACD	Ligand structure	NA_{elec} ^a	r_{hb} ^b	Pure bindin g ^c	Interaction preference ^d	Ligand preference ^e	Both ^f
MFCD00016941		0	0.35	8	2	661	260
MFCD00016787		0	0.32	51	8	747	319
MFCD00001218		6	0.34	13	17	954	937
MFCD00010009		0	0.00	88	430	5	57
MFCD00002371		0	0.13	40	19	16	12
MFCD00002206		0	0.18	37	30	46	20

^a The number of charged atoms in a screened ligand (Equation 2.1.9).

^b The fraction of polar atoms in a screened ligand (Equation 2.1.10).

^{c,d,e,f} Using E_{bind} , $E_{bind} + E_{pharma}$, $E_{bind} + E_{ligpre}$, and E_{tot} , respectively, for the scoring function. These energy terms are defined in Equation 2.1.1.

The influences of pharmacological preferences

When using interaction energy scoring alone for choosing ligands, docking methods (e.g., GEMDOCK and GOLD) favor the selection not only of highly charged polar compounds but also high molecular weight compounds. Figures 2.1.8 and 2.1.9 show the influences of the ligand structures and molecular weight, respectively, when the binding scoring (E_{bind}) alone was used in GEMDOCK. The docking energy of a ligand with charged or polar atoms is often lower than the energy of a non-charged ligand when the docked conformations are similar. For example, the docking energies are -76.86 for ESA01-C (r_{hb} is the smallest), -91.32 for ESA01, and -99.64 for ESA01-COO (with charged atoms, and r_{hb} is the largest) when the docked positions of these ligands are similar (Figure 2.1.8). At the same time, ESA01 and ESA01-COO form the pharmacological interactions shown in Figure 2.1.6(b) (e.g., A' phenolic hydroxyl group and B' phenolic hydroxyl group). In contrast, ESA01-C has no polar atoms to form these pharmacological interactions. We obtained these ligand structures (EAS01-C and ESA01-COO) using the 3-dimensional structure generator CORINA⁸⁰.

Tables 2.1.7 and 2.1.8 show the effect of pharmacological preferences of some typical ligand structures on screened ER agonists and antagonists, respectively. When the binding energy (E_{bind}) alone was used to screen ER agonists, GEMDOCK selected two ligands, MFCD00012742 (1st) and MFCD00002206 (4th), which are similar in structure to ESA03 and ESA04, respectively, and satisfy the ligand preferences. Due to higher numbers of polar atoms at critical sites, these ligands formed greater numbers of pharmacological interactions compared with known active ligands. At the same time, GEMDOCK was able to exclude ligands such as MFCD00006630 ($r_{hb} = 0.47$), MFCD00006616 ($r_{hb} = 0.45$ and $NA_{elec} = 3$), and MFCD00005746 ($r_{hb} = 0.52$ and $NA_{elec} = 3$) that violate the ligand preferences of known ER agonists (Table 2.1.2). For example, their r_{hb} values were larger than the upper bound ratio ($Ur_{hb} = 0.31$) of polar atoms or the upper bound number ($UB_{elec} = 0$) of charged atoms. When the penalty for the ligand preferences (E_{ligpre}) was considered, the ranks of MFCD00006630 (911th), MFCD00006616 (900th), and MFCD00005746 (928th) lagged substantially. Ligands such as MFCD00003783 lagged (244th) since it is unable to interact with three important residues (Glu353, Arg394, and His524; Figure 2.1.6(b)) in the reference protein.

GEMDOCK yielded similar results when the ER antagonists were screened (Table 2.1.8). When the binding energy (E_{bind}) alone was used, the ranks of ligands MFCD00016941 ($r_{hb} = 0.35$), MFCD00016787 ($r_{hb} = 0.32$), and MFCD00001218 ($r_{hb} = 0.34$) were 8th, 51th, and 13th. When both E_{bind} and ligand preferences (E_{ligpre}) were considered for the scoring function, the ranks of these ligands were 661th (MFCD00016941), 747th (MFCD00016787), and 954th (MFCD00001218) since their r_{hb} values were larger than the upper bound ratio (e.g., $Ur_{hb} = 0.17$ in Table 2.1.2) derived from known ER antagonists. These total scoring values were penalized by hydrophilic preferences (i.e., LP_{hb} in Equation 2.1.10). Ligand MFCD00001218 was also penalized by the electrostatic preferences (i.e., LP_{elec} in Equation 2.1.9) because the number of charged atoms ($NA_{elec} = 6$) was

larger than the upper bound ($U_{r_{elec}} = 2.63$ in Table 2.1.2). The screening of ligand MFCD00010009, which has no polar atoms to form pharmacological interactions (Figure 2.1.6(a)), often fell behind when GEMDOCK used both E_{bind} and E_{pharma} for the scoring function. In contrast, ligands MFCD00002371 and MFCD00002206 yielded good ranks for various combinations of energy terms since they are able to form binding-site pharmacological interactions and satisfy the ligand preferences.

Figure 2.1.9 and 2.1.10 show the effect of molecular weight on screening accuracy. A docking method using energy-based scoring alone is often biased toward large molecular weight ligands because the overall van der Waals interaction energy is summed over all pairs of ligand and target protein atoms within a specified cutoff distance. Figure 2.1.9(a) shows that ESA01 (blue) and EST03 (yellow) have a common group A and that EST03 has an additional substructure group (side chain B). The van der Waals force of a large ligand (e.g., EST03) is often larger than that of a small ligand (e.g., ESA01). In this case, EST03 acquires additional van der Waals force from side chain B as shown in Figure 2.1.9(b). For example, when using E_{bind} alone for docking a ligand into the reference protein (3ert), GEMDOCK yielded docking energies of -127.27 for EST03 and -82.82 for ESA01. Figure 2.1.10 shows the true hits obtained by GEMDOCK when screening ER agonists without (dashed line) or with molecular weight normalization (solid line; defined in Equation 2.1.11). When GEMDOCK applied molecular weight normalization and pharmacological preferences to screen ER agonists, the average hit rate was 45.66%, the average FP rate was 0.75%, and the GH score was 0.48. In contrast, these averages were 21.18%, 2.02%, and 0.29 when molecular weight normalization was not considered.

Figure 2.1.11 shows the true hits of GEMDOCK using the cleaned lists and the original data set proposed by Bissantz et al.¹ For each test case (ER antagonists and ER agonists), we prepared the cleaned list by filtering the original set in order to eliminate the ligands, which violate the electrostatic (LP_{elec}) or hydrophilic constraints (LP_{hb}). These two cleaned lists, including the known active compounds, consist of 590 and 701 compounds for screening the ER antagonists and ER agonists, respectively. As shown in Figure 2.1.11, the true hits (gray lines) of GEMDOCK using E_{bind} (C-Pure binding) and $E_{bind} + E_{pharma}$ (C-Interaction preference) as the scoring functions on the cleaned lists are similar to those (black lines) of GEMDOCK using $E_{bind} + E_{ligpre}$ (W-Ligand preference) and $E_{bind} + E_{ligpre} + E_{pharma}$ (W-Both) as scoring functions, on the original set, respectively. Using GEMDOCK on the cleaned sets, average GH scores were 0.82 (Interaction preference) and 0.66 (Pure binding) for ER antagonists, and average GH scores were 0.41 (Interaction preference) and 0.29 (Pure binding) for ER agonists. These experiments indicated that the pharmacological interaction preferences were able to improve the GH scores for both the cleaned lists and original set; moreover, the ligand preferences might improve the screening accuracy of a scoring function and become the filters to prepare a ligand database.

In summary, we developed a near-automatic tool with a novel scoring function for virtual screening by making numerous modifications and enhancements to our original techniques. By

integrating a number of genetic operators, each having a unique search mechanism, GEMDOCK seamlessly blends the local and global searches so that they work cooperatively. The key aspect of the present work is that our new scoring function uses pharmacological-interaction preferences to select the ligand structures that form pharmacological interactions with target proteins; furthermore, the scoring function applies ligand preferences to select ligand structures that are similar to known active ligands. Our scoring function is indeed able to enhance the accuracy during flexible docking and improves the screening utility by reducing the number of false positives during the post-docking analysis. Our results demonstrate the applicability and adaptability of GEMDOCK for virtual screening.

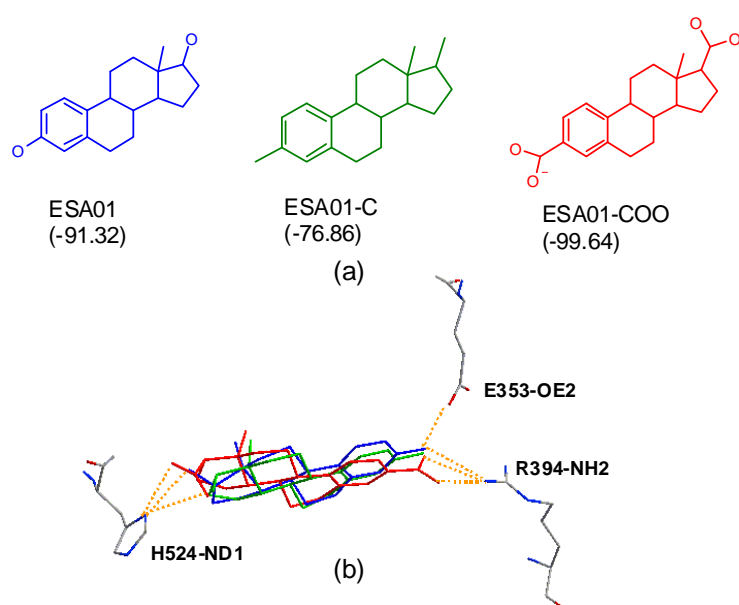


Figure 2.1.8. Docking energy is influenced by ligand structures generated by CORINA. (a) The fraction of polar atoms in ESA01-C is the smallest among these three ligands, whereas that of ESA01-COO is the largest. (b) The docked positions are similar, but the docking energies differ: -91.32 for ESA01, -76.86 for ESA01-C, and -99.64 for ESA01-COO.

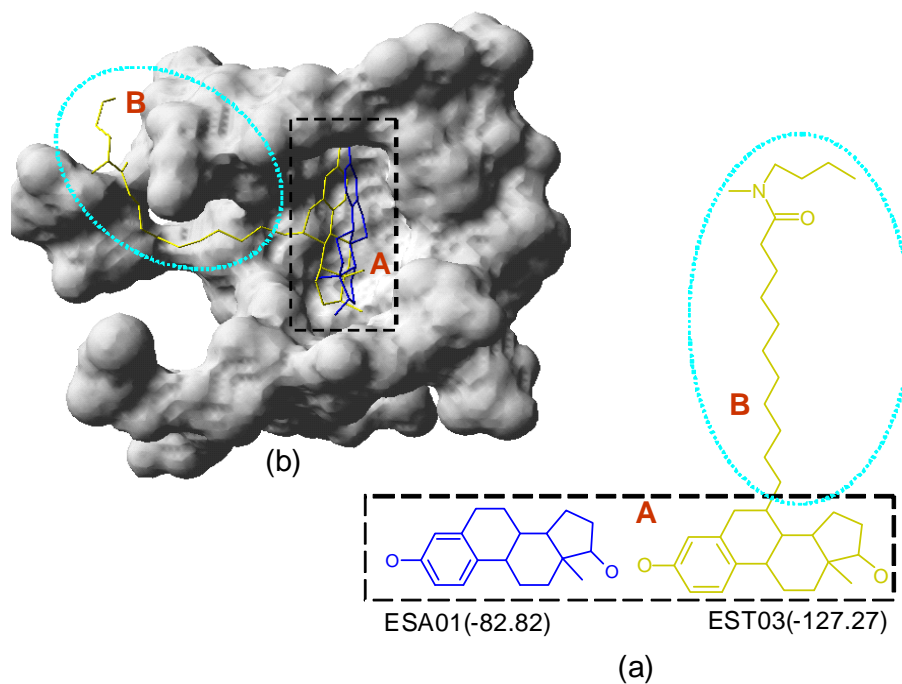


Figure 2.1.9. The influence of molecular weight on docking energy. (a) ESA01 (blue) and EST03 (yellow) have a common group A, and EST03 has an additional substructure group B. (b) The docked conformations (into reference protein 3ert) are similar, and the docking energies are -82.82 for ESA01 and -127.27 for EST03.

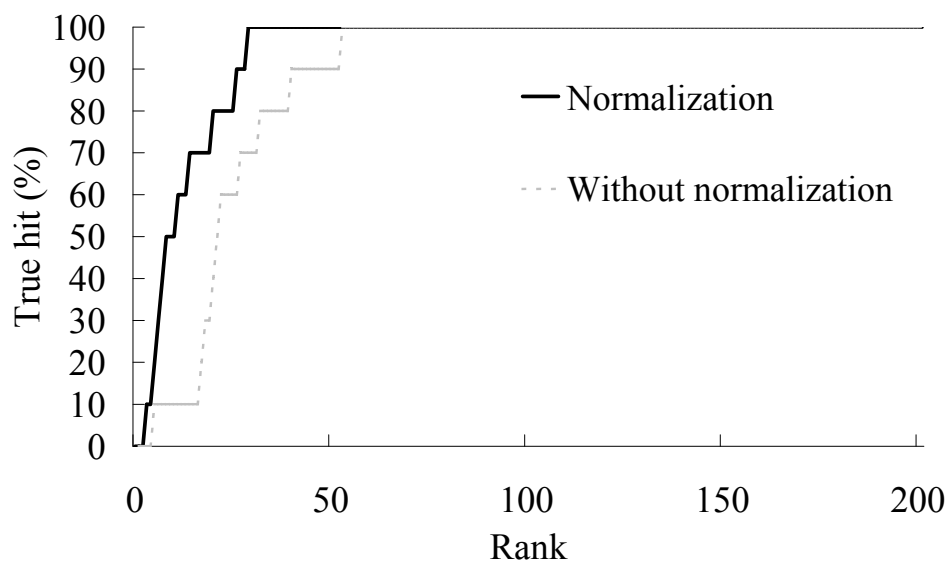
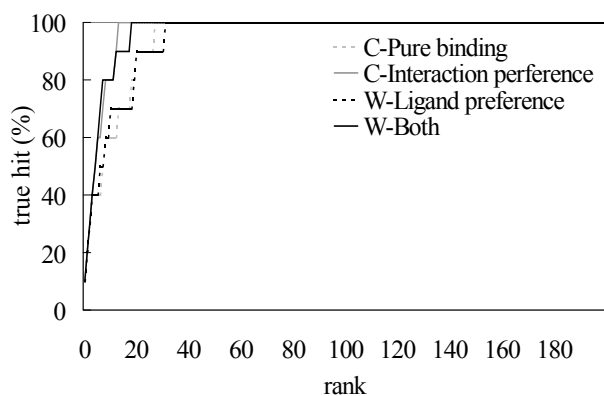
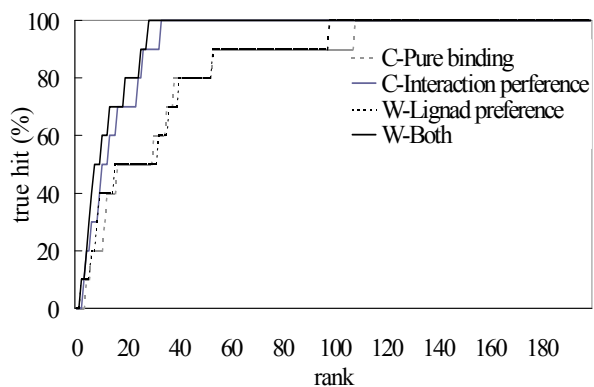


Figure 2.1.10. The accuracy of GEMDOCK for screening ER agonists, assessed using scoring functions with molecular-weight normalization (solid line) and without molecular-weight normalization (dash line).



(a) ER antagonists



(b) ER agonists

Figure 2.1.11. The accuracy of GEMDOCK for screening (a) ER antagonists and (b) ER agonists, assessed using the cleaned ligand sets (C-Pure binding and C-Interaction preference) and the ligand set proposed by Bissantz et al.¹ (W-Ligand preference and W-Both).

Chapter 3: Post-analysis of Virtual Screening

For post-analysis of virtual screening, we developed a cluster method for post analysis to improve enrichment for virtual screening. The method combines protein-ligand interactions (e.g. hydrogen bonds, electrostatic interactions, and van der Waals), which are generated by our well-developed docking tool (i.e. GEMDOCK), and physical-chemical features and structures for each compound candidate selected by GEMDOCK. For each cluster, this method selected a representative compounds for biological tests and improved the enrichment of virtual screening. Our works on the post-analysis have published one journal paper and one poster.

Journal papers:

- 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. (SCI, IF: 3.2) (Times Cited: 21)

Posters

- C.-N. Ko, Y.-F. Chen, Y.-J Chen and J.-M. Yang, "Cluster analysis of Structure-based Virtual Screening by Using Protein-ligand Interactions and Compound Structures", in Annual Conference on Biotechnology, Hsinchu, Taiwan, 2007

3.1 Cluster analysis of Structure-based Virtual Screening by Using Protein-ligand Interactions and Compound Structures

3.1.1 Introduction

With the recent development of high-throughput X-ray crystallography, the total number of structures will grow at an even greater speed⁸¹. And the enormous advances in genomics have resulted in a large increase in the number of potential therapeutic targets that are available for investigation. This growth in potential targets has increased the demand for reliable target validation, as well as technologies that can identify rapidly several quality lead candidates. Virtual screening methods are a primary source for the discovery of lead molecules for drug development, with high-throughput docking algorithms being among the most extensively used of these methods. The application of virtual high-throughput screening^{82; 83}, to the drug discovery process invariably produces a large number of potential lead candidates. And it is well known that current scoring functions used in virtual screening campaigns are often inadequate at predicting the true binding affinity of a ligand for a receptor⁸⁴. These prospective ligands need to be filtered in order to reduce their number for more precise and labor-intensive studies.

The purpose for utilizing post-analysis is to minimize the number of false positives in the selection list and to propagate the true hits to the top of the list. One of the post-analysis methods such as clustering based upon structural similarity can nonetheless generally improve the

performance of the scoring function^{8; 85}. Clustering molecules based upon similarity requires some quantitative measure (descriptor) of the similarity between two molecules. There are many different approaches to generate descriptor, include 2D and 3D methods. Most of the 2D methods have focused on representing a molecule based upon its own structural and chemical composition, like Atom-Pair. But it regardless the information from protein that is important in the field of structure-based drug designs. Deng and co-workers⁸ described a novel approach to representing the properties of a ligand. As opposed to calculating the properties of a ligand from the perspective of its own structural and chemical components, the Structural Interaction Fingerprint (SIFt) method represents a ligand by the interactions it forms in the binding site of a protein. Using seven bits per binding-site residue to represent seven different types of interaction, the SIFt method encoded a ligand-protein interaction into a binary string. The types of interaction that considered are hydrogen bond and physical contact. Recently another approach proposed by Amari et al,⁷ have developed a clustering tool for visualized cluster analysis of protein-ligand interaction (VISCANA) that analyzes the pattern of the interaction of the receptor and ligand on the basis of quantum theory. They applied the ab initio fragment molecular orbital (FMO)⁸⁶ method for represent the interaction between protein and ligand, which used the ab initio electronic structure calculation of proteins and encoding each docked pose into real number string. But the FMO method needed to obtain more reliable descriptions of van der Waals interactions and hydrogen bonds that are important for receptor-ligand binding.

We developed a cluster method for post analysis to improve enrichment for VS. The method combines protein-ligand interactions (e.g. hydrogen bonds, electrostatic interactions, and van der Waals), which are generated by our well-developed docking tool (i.e. GEMDOCK), and physical-chemical features and structures for each compound candidate selected by GEMDOCK. The physical-chemical features of a compound were described by atom pair descriptors (i.e. compound topological similarity) proposed by Carhart et al. Based on these normalized feature profiles, hierarchical clustering methods were used to cluster these compound candidates. For each cluster, this method selected a representative compounds for biological tests. Our method was evaluated on five well-known drug targets, including thymidine kinase (TK), dihydrofolate reductase (DHFR), estrogen receptor agonist (ESA), estrogen receptor antagonists (EST) and neuraminidase (NA). We also practically applied our method for the screening of *Helicobacter pylori* shikimate kinase (HpSK) and the test the inhibitor activities of selected compounds in bioassay.

3.1.2 Materials and Methods

We developed a cluster method for post analysis to improve enrichment for VS (Figure 3.1.1A). The method combines protein-ligand interactions (e.g. hydrogen bonds, electrostatic interactions, and van der Waals), which are generated by our well-developed docking tool (i.e. GEMDOCK^{2; 66}), and physical-chemical features and structures for each compound candidate selected by GEMDOCK. The physical-chemical features of compound structures were described by atom pair descriptors (i.e.

compound topological similarity) proposed by Carhart et al^{87; 88} (shown as Figure 3.1.1B). Based on these feature profiles shown in Figure 3.1.1A, hierarchical clustering methods were used to cluster these compound candidates. For each cluster, this method selected a representative compounds for biological tests.

The interactions of atom pairs on each protein-ligand complex were collected as a real number vector which the length and order were corresponded to atoms on the binding site of target protein (shown as Figure 3.1.1B). The structure of each compound was represented by the atom-pair descriptor which was proposed by Carhart *et al.*,^{88; 89} (shown as Figure 3.1.1B). The atom-pair descriptor approach describes the molecular structure by the bonding interval of all pairs of atom types. Hierarchical clustering analysis was employed for analyzing the profiles of interactions and compound structures on MATLAB^{90; 91}. The post-processing of interaction and structure analyses separates the screening candidates into several meaningful groups and the compound with the lowest docked energy in each group was selected as the representatives for the bioassays (Figure 3.1.1B). For evaluate the method of cluster analysis, we adopt this method for five important drug targets and the results of validation suggested a threshold of cutoff. This analysis process was also applied to the virtual screening of *Helicobacter pylori* shikimate kinase (HpSK). The biological assays reported the discovery of a new inhibitor structure from five candidates.

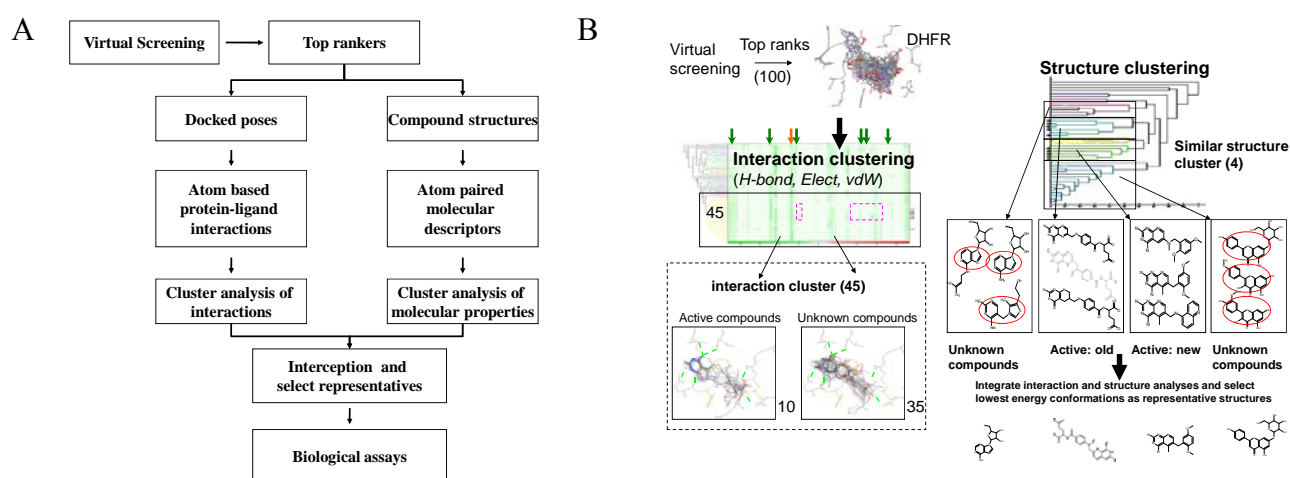


Figure 3.1.1. Cluster analysis for post-processing of virtual screening. (A) The flowchart of cluster analysis. The top ranks in screening are selected for cluster analysis. Each compound is grouped by its interaction and molecular properties. (B) The cluster analyses of interaction and molecular properties. The interaction analysis uses the hydrogen bonding, electrostatic and van der Waals interactions of atom pairs on the protein-ligand complexes to cluster similar docked poses. The molecular property analysis used atom-paired descriptors to cluster similar molecular structures.

Preparations of the target proteins and screening sets

The compound set for virtual screening was prepared by selecting them from the CMC database in May 2004 based on two criteria: (1) molecular weight ranging between 200 and 800, and (2) no compounds with multiple components. A set comprising 5,331 compounds was eventually obtained. The five sets of target specific compounds were from the literatures (Thymidine kinase (TK) and human estrogen receptor alpha (ER α) antagonists¹⁷, ER α agonist⁹², human dihydrofolate reductase (hDHFR)¹⁰ and Neuraminidase (NA)⁹³). The other additional 990 compound set was randomly selected from ACD database for validation^{10;17}. All 3D structures of compound were generated by CORINA3.0 and shown in [Figure 3.1.2](#).

The structure of the protein binding pocket on TK, ER α (for antagonist and agonist), hDHFR, NA and SK were isolated and prepared for the GEMDOCK. The structure of the binding pocket in the pre-described ligand-bound conformation (PDB code 1kim¹⁴, 3ert⁹⁴, 1gwr⁹⁴, 1hfr⁹⁵, 1mwe⁹⁶ and 1zui⁹⁷) including including amino acids enclosed within a 10 Å radius sphere centered on the bound ligand, were used. The coordinates of protein atoms were taken from the PDB for the screening processing. GEMDOCK docked each compound in the data sets against these binding cavities, and ranked each compound by docked energy of the docked conformation. To validate our method, compounds for five validated set were chosen for cluster analysis. According to the ranking, top ranked compounds for SK were chosen for cluster analysis and *in vivo* biological activity tests to validate their inhibitory activities.

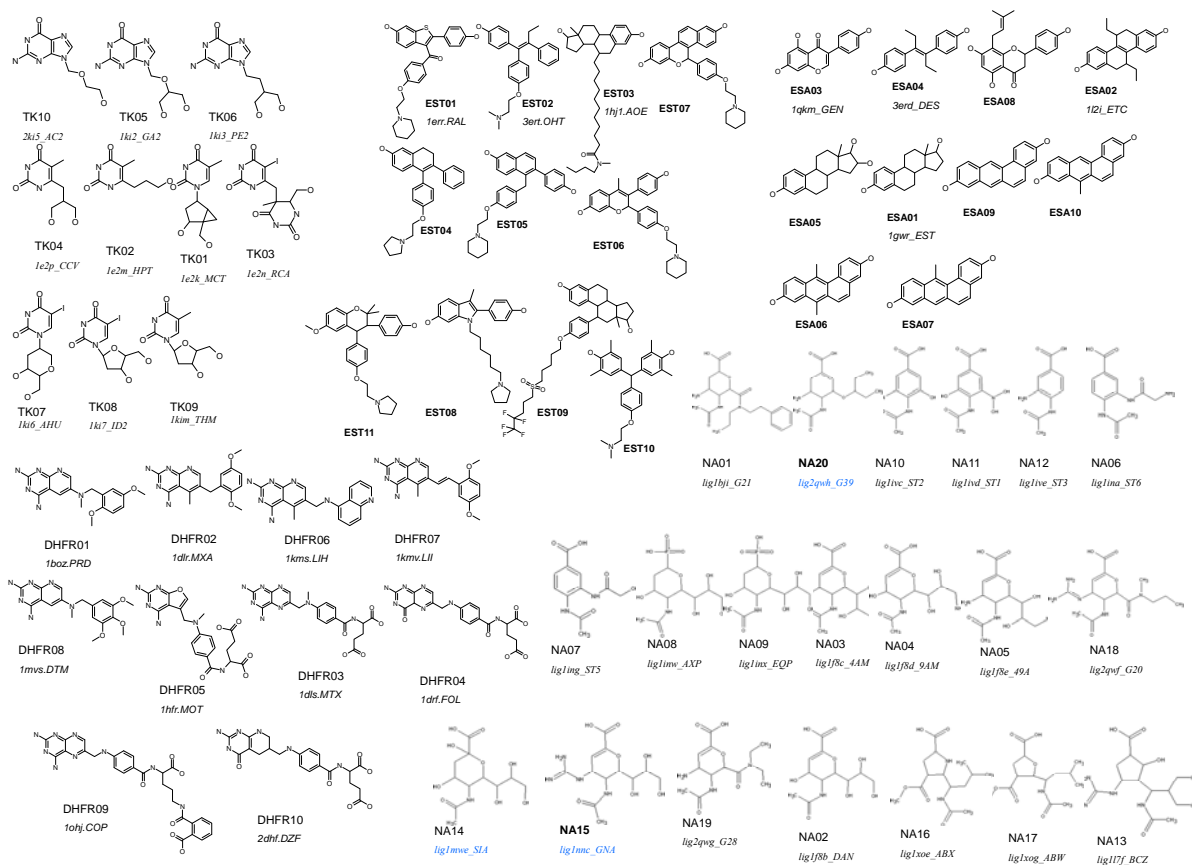


Figure 3.1.2. Active compound structures of validation sets. The compounds of each target are denoted in the abbreviation of protein and numbers. The compound co-crystallized in the complex of PDB is further denoted in the PDB code of the protein and the ligand.

Docking method and scoring function

Our previous works ^{2;9} have showed that the docking accuracy of GEMDOCK was better than some well-known docking tools, such as GOLD ⁵ and FlexX³, on a diverse data set of 100 protein-ligand complexes proposed by Jones et al.⁵ The screening accuracy of GEMDOCK were also better than GOLD, FlexX, and DOCK on screening the ligand database from Bissantz et al. (2000) for the thymidine kinase ²⁸ and the estrogen receptor ⁹. In this study, GEMDOCK parameters in the flexible docking included the initial step sizes ($\sigma=0.8$ and $\psi=0.2$), family competition length ($L = 2$), population size ($N = 300$), and recombination probability ($p_c = 0.3$). For each ligand screened, GEMDOCK optimization stopped either when the convergence was below a certain threshold value or the iterations exceeded the maximal preset value of 60. For the latter case, GEMDOCK will produce 800 solutions in one generation and terminated after it exhausted 48,000 solutions for each compound in the screening set.

The screening quality of docking methods using energy-based scoring functions alone is often influenced by the structure of the ligand being screened (e.g., the numbers of charged and polar atoms). These methods are often biased toward charged polar compounds due to the pair-atom

potentials of the electrostatic energy and hydrogen-bonding energy. In order to reduce this ill effect, GEMDOCK could evolve the pharmacological preferences from a number of known active ligands or from domain knowledge 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 ligand⁹. GEMDOCK could use either a purely empirical scoring function² or pharmacophore-based scoring function⁹. When GEMDOCK used a pharmacophore-based scoring function, some known active ligands (more than two) or domain knowledge are required for evolving the pharmacological consensus according to our previous results. The empirical-binding energy (E_{bind}) is the sum of the intermolecular (E_{inter}) and intramolecular energies (E_{intra}), respectively². The pharmacophore-based energy function⁹ is the sum of three energy items, including the empirical binding energy (E_{bind}), the energy of binding site pharmacophores (E_{pharma}), and a penalty value (E_{ligpre}) if a ligand does not satisfy the ligand preferences. E_{pharma} and E_{ligpre} are especially useful in selecting active compounds from hundreds of thousands of non-active compounds by excluding ligands that violate the characteristics of known active ligands (or domain knowledge).

All ligands are docked into the target protein binding pocket and the atom based protein-ligand interaction descriptors is generated from the docked poses corr. The interactions of atom pairs on each protein-ligand complex were collected as a real number vector which the length and order were corresponded to atoms on the binding site of target protein

Atom pair descriptors

The method of atom-pair descriptor is to describe the molecular topology by counting the shortest path of valence bonds between two atom types. The definition of atom type is show as [Table 3.1.1](#). Atom pair descriptors was generated from the self-developed atom-pair generator program, and the methodology was proposed by Carhart et al.,^{88; 89}. Two major components for constructing a set of atom-pair descriptors include the definition of atom type and the number of distance bins between two atom types. An atom-pair is a simple type of substructure defined in term of the atom types and the shortest path graph distance between two atoms. The graph distance is defined as the smallest number of atoms along the path connecting two atoms in a molecular structure. The formula of an atom-pair is as atom type i —(distance)—atom type j . Where the distance is the valence bond distance between the atoms type i and j in the case of a 2D atom-pair description. We clustered the SYBYL 23 atom types into 10 atom types ([Table 3.1.1](#)) in order to reduce the number of atom-pair descriptors and improve the accuracy. Our settings of atom-pair approach followed the previous research of Hans Matter⁹⁸. The maximum of valence bond distance was set as 14 in this study and a total of 825 (55 x 15) atom-pair descriptors were generated for each molecular structure⁹⁹. The representation of 825 atom-pair descriptors is bit string.

[Table 3.1.1](#). Atom types used in the atom-pair descriptors

Description	Atom type	Atom type in SYBYL Mol2
Aromatic carbons	C.ar	C.ar
Nonaromatic carbons	C.na	C.3 C.2 C.1 C.cat
Aromatic nitrogen	N.ar	N.ar
Nonaromatic nitrogen	N.na	N.3 N.2 N.1 N.am N.4 N.pl3
Oxygen atoms in the sp ³ hybridization state	O.3	O.3
Oxygen atoms in the sp ² hybridization state	O.2	O.2
All sulfur atoms	S	S.3 S.2 S.O S.o2
Phosphorus atoms	P.3	P.3
Halogen atoms	X	F Cl Br I

Hierarchical cluster method

Hierarchical methods have the advantage of building up an interpretable relationship between the clusters. Hierarchical clustering analyses were carried out with MATLAB⁹¹. We removed the column with zero in each ligand before calculating the distance of interaction and structure vectors. The similarity distance of the protein-ligand interaction vectors are measured by the correlation coefficients. Formula was as followed.

$$D_{xy}^{corr} = 1 - \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)S_x S_y} \quad (3.1.1)$$

where D_{xy}^{corr} is the correlation distance between docked poses X and Y . The S_x is the standard deviation of X . X_i is the i th value of X . n is the number of atoms in the binding cavity. The similarity distance of atom-pair descriptor strings is measured through the tanimoto coefficients. The formula was as followed.

$$D_{xy}^{tani} = \frac{|X \cap Y|}{|X \cup Y|} \quad (3.1.2)$$

where D_{xy}^{tani} is the tanimoto distance between X and Y . $|X \cap Y|$ is the number of ON bits common in both X and Y , and the $|X \cup Y|$ is the number of ON bits present in either X or Y . The standard UPGMA clustering method is adapted for calculating the distance between two clusters while constructing the dendrogram. The formula was as followed.

$$D_{n_s}^{cluster} = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D_{r_i y_j}^{sim} \quad (3.1.3)$$

The reference threshold was calculated from verifying dataset by Equation 3.1.4 for determining the number of clusters. D_{xy}^{sim} is the method for measuring similarity. For interaction analysis, D_{xy}^{sim} is D_{xy}^{corr} . For molecular structure, D_{xy}^{sim} is D_{xy}^{tani} . The dendrogram graph was plotted for visualizing the binding mode of multi docked poses by protein-ligand interaction.

3.1.3 Results and Discussion

We applied the cluster analysis method on five datasets for validation. The verifying dataset was constructed by cross-docking of all 61 known active compounds against 5 target proteins. The testing dataset for virtual screening was constructed by docking known active compounds and 990 randomly selected compounds against 5 target proteins. GEMDOCK was adapted to produce the docked prediction and score the docked poses. The molecular recognitions of five classes of active compounds for target protein were shown in Figure 3.1.3. The average RMSD of docked poses and crystal conformations were below 2.0Å and the details were shown in Table 3.1.2. The residues of pharmacological consensus were used in the docking procedures but not in the scoring and clustering procedures.

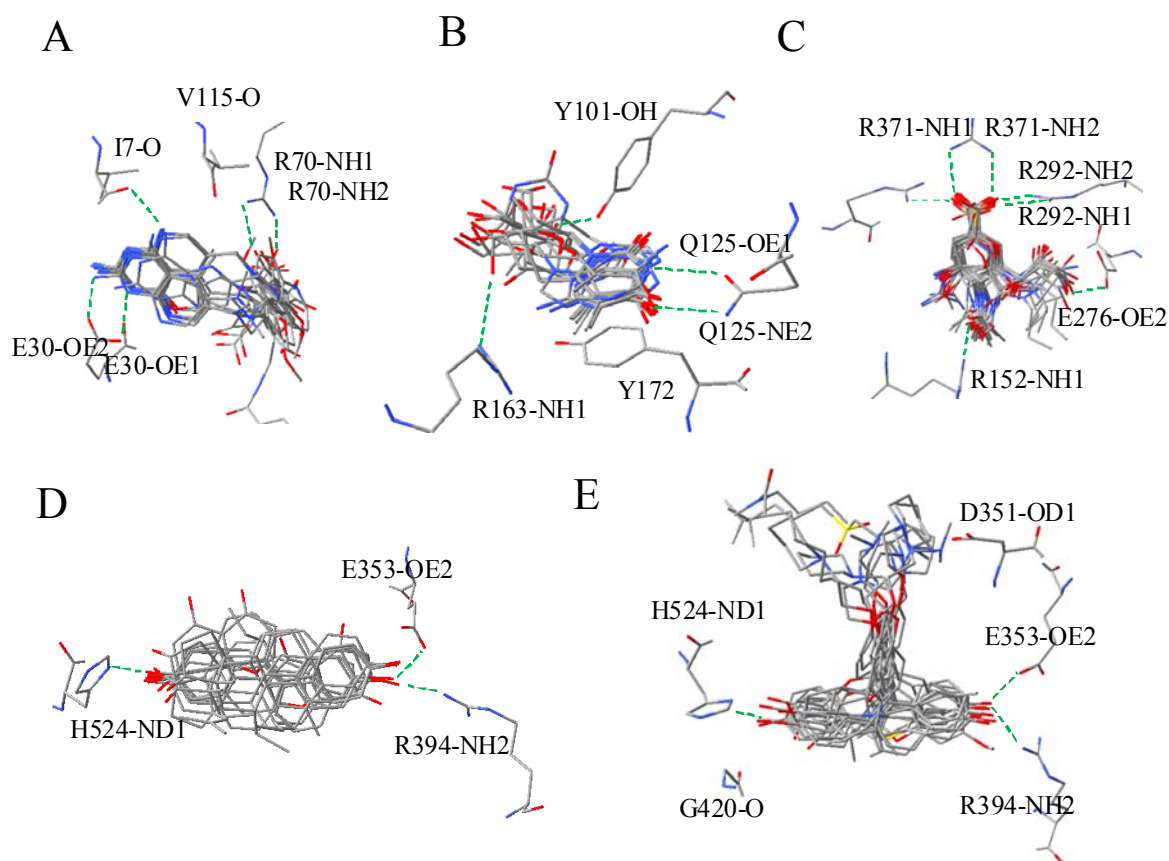


Figure 3.1.3. Molecular recognition on (A) hDHFR (1hfr), (B) TK (1kim), (C) NA (1mwe), (D) ER α -agonist form (1gwr), (E) ER α -antagonist form (3ert). The residues of pharmacological consensus are label with their numbers. The hydrogen bonding of docked pose and residues are denoted as green dash lines.

Table 3.1.2. The RMSD values of docked poses and crystal ligands

TK (1kim)		ER (3ert, 1gwr)		hDHFR (1hfr)		NA (1mwe)	
Complex	RMSD	Complex	RMSD	Complex	RMSD	Complex	RMSD
<i>1e2k.TMC</i>	0.69	<i>1err.RAL^a</i>	1.27	<i>1boz.PRD</i>	1.13	<i>lig1l7f_BCZ</i>	0.88
<i>1e2m.HPT</i>	0.51	<i>3ert.OHT^a</i>	0.71	<i>1dlr.MXA</i>	0.62	<i>lig1nnc_GNA</i>	0.75
<i>1e2n.RCA</i>	1.34	<i>1hj1.AOE^a</i>	3.13	<i>1dls.MTX</i>	1.53	<i>lig2qwf_G20</i>	0.60
<i>1e2p.CCV</i>	0.67	<i>1uom.PTI^a</i>	0.81	<i>1drf.FOL</i>	1.24	<i>lig1bji_G21</i>	0.81
<i>1ki2.GA2</i>	3.04	<i>1gwr.EST^b</i>	0.71	<i>1hfr.MOT</i>	0.51	<i>lig1f8b_DAN</i>	0.64
<i>1ki3.PE2</i>	3.21	<i>1l2i.ETC^b</i>	0.52	<i>1kms.LIH</i>	1.36	<i>lig1f8c_4AM</i>	0.46
<i>1ki6.AHU</i>	0.37	<i>1qkm.GEN^b</i>	2.92	<i>1kmv.LII</i>	0.83	<i>lig1f8d_9AM</i>	0.59
<i>1ki7.ID2</i>	0.49	<i>3erd.DES^b</i>	1.32	<i>1mvs.DTM</i>	0.75	<i>lig1f8e_49A</i>	0.60
<i>1kim.THM</i>	0.41			<i>1ohj.COP</i>	1.27	<i>lig1ina_ST6</i>	0.79
<i>2ki5.AC2</i>	3.14			<i>2dhf.DZF</i>	1.12	<i>lig1ing_ST5</i>	1.03
						<i>lig1inw_AXP</i>	0.93
						<i>lig1inx_EQP</i>	0.92
						<i>lig1ivc_ST2</i>	2.09
						<i>lig1ivd_ST1</i>	1.02
						<i>lig1ive_ST3</i>	1.03
						<i>lig1mwe_SIA</i>	0.52
						<i>lig1xoe_ABX</i>	1.33
						<i>lig1xog_ABW</i>	2.42
						<i>lig2qwg_G28</i>	0.80
						<i>lig2qwh_G39</i>	0.74
<i>Average RMSD</i>	1.58	<i>Average RMSD</i>	1.42	<i>Average RMSD</i>	1.03	<i>Average RMSD</i>	0.95

^a Four antagonists dock into target protein (3ert)

^b Four agonists dock into target protein (1gwr)

Significance test of descriptors

t-test for protein-ligand interaction descriptor

We verified whether the protein-ligand interaction descriptors could identify the similar binding poses from data. The similar binding poses were defined as the poses of active compounds against its target protein, and the dissimilar binding poses was defined as the poses of other compounds docked into the same target protein. For *t*-test, the H_0 is that there are no differences under the representation of interaction descriptors and the results were listed in Table 3.1.3. The *t*-scores are calculated as the standard two sample *t*-test statistics:

$$T = \frac{\mu_1 - \mu_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}} \quad (3.1.5)$$

Where μ is the mean of samples, and

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \mu)^2} \quad (3.1.6)$$

The result of *t*-test showed that the representation of interaction descriptors for docked pose could show the difference of active and inactive ones.

Table 3.1.3. *t*-test for the interaction representations of similar and dissimilar binding poses ($\alpha=0.01$).

Target protein	H_0^a	Similar : Average Distance	Dissimilar : Average Distance	P-value	Similar : Std ^b of Distance	Dissimilar : Std ^a of Distance
DHFR	Reject	0.21	0.50	1.71E-58	0.09	0.13
ESA	Reject	0.25	0.42	7.04E-20	0.13	0.12
EST	Reject	0.31	0.48	7.94E-39	0.09	0.12
NA	Reject	0.17	0.73	0.00E+00	0.07	0.20
TK	Reject	0.19	0.47	3.89E-64	0.08	0.15

^a H_0 is that there are no differences in the representation of descriptors

^a Standard Deviation

For the needs of post-analysis, we further checked whether the interaction representations could discriminate the active compounds and other four groups of non-active compounds in the docked interactions and conformations. Table 3.1.4 showed the details of tests. Docked poses on DHFR and NA fully passed all the tests, but the other poses on TK, ESA, and EST didn't. We checked the docked poses on these proteins and found that the reason maybe came from the fused ring structures of ligands. ESA and EST were specific on the steroid compounds for their hormonal regulation functions; TK was specific on the thymidine base compounds for the kinase functions. But the skeleton of rings on the ligands of them were shared the closed lengths (Figure 3.1.2) and this phenomenon also showed on the distance of atom-pair descriptors.

Table 3.1.4. *t*-test for the interaction representations of five groups of compounds against target proteins($\alpha=0.01$)

Target protein	Compound class	H_0^a	Similar : Average Distance	Dissimilar : Average Distance	P-value	Similar : Std ^b of Distance	Dissimilar : Std ^a of Distance
DHFR	DHFR	Reject	0.21	0.50	1.71E-58	0.09	0.13
	ESA	Reject	0.52	0.58	2.73E-03	0.18	0.12
	EST	Reject	0.52	0.63	7.51E-07	0.21	0.13
	NA	Reject	0.46	0.55	5.34E-23	0.13	0.14
	TK	Reject	0.38	0.51	8.03E-11	0.16	0.13
ESA	DHFR	Pass	0.55	0.62	0.10111	0.28	0.16
	ESA	Reject	0.23	0.48	2.29E-31	0.14	0.14
	EST	Pass	0.67	0.76	0.23105	0.25	0.14
	NA	Reject	0.33	0.59	1.51E-58	0.24	0.20
	TK	Reject	0.46	0.57	0.000121	0.25	0.20
EST	DHFR	Pass	0.55	0.57	4.01E-01	0.21	0.14
	ESA	Reject	0.25	0.42	7.04E-20	0.13	0.12
	EST	Reject	0.31	0.48	7.94E-39	0.09	0.12
	NA	Reject	0.40	0.46	1.46E-09	0.15	0.15
	TK	Reject	0.28	0.43	2.17E-29	0.09	0.15
NA	DHFR	Reject	0.35	0.68	3.46E-25	0.22	0.25
	ESA	Reject	0.59	0.71	2.91E-04	0.28	0.24
	EST	Reject	0.56	0.66	2.46E-04	0.25	0.24
	NA	Reject	0.17	0.73	0.00E+00	0.07	0.20
	TK	Reject	0.48	0.60	3.46E-07	0.18	0.23
TK	DHFR	Reject	0.42	0.62	9.80E-12	0.13	0.10
	ESA	Reject	0.16	0.52	9.99E-62	0.07	0.13
	EST	Pass	0.58	0.65	6.28E-02	0.18	0.14
	NA	Reject	0.40	0.53	2.92E-53	0.11	0.15
	TK	Reject	0.19	0.47	3.89E-64	0.08	0.15

^a H_0 is that there are no difference in the representation of descriptors

^b Standard Deviation

***t*-test for atom-pair descriptor**

The *t*-tests were also used for check the representations of atom-pair descriptors on compound structures. The similar structures were defined as the active compounds against its target protein, and the dissimilar structures were defined as the other compounds against the same target protein. For *t*-test, the H_0 is that there are no differences under the representation of atom-pair descriptors and the results were listed in [Table 3.1.4](#).

Table 3.1.4. *t*-test for the atom-pair representations of similar and dissimilar structures ($\alpha=0.01$).

Target protein	H_0^a	Similar : Average Distance	Dissimilar : Average Distance	P-value	Similar : Std ^b of Distance	Dissimilar : Std ^a of Distance
DHFR	Reject	0.42	0.63	5.84E-23	0.15	0.12
ESA	Reject	0.24	0.66	4.60E-65	0.11	0.14
EST	Reject	0.27	0.63	2.85E-56	0.14	0.14
NA	Reject	0.32	0.65	1.75E-131	0.18	0.17
TK	Reject	0.22	0.63	2.11E-93	0.09	0.19

^a H_0 is that there are no difference in the representation of descriptors

^b Standard Deviation

Cutoff of hierarchical cluster for verifying datasets

The cluster sizes and members of hierarchical cluster method were depended on the determination of cutoff distance. We proposed an approach for cutoff determinate by validation sets. We decided the cutoff according to maximize the true positive rates and minimize the false positive rates in the cluster analyses. In the cluster analyses of validate sets, we defined the true positive as the active compounds for its pharmacological target and the others compound were defined as false positives. The true positives were further denoted as intra set and the false positives were denoted as inter set. To maximize the true positives at a given distance threshold t , we defined an equation as followed

$$\max \left(\left(\frac{C_{intra-d < t}}{C_{intra}} + \frac{C_{inter-d > t}}{C_{inter}} \right) / 2 \right) \quad (3.1.7)$$

where t is the given distance threshold. $C_{intra-d < t}$ is the number of intra active compound pairs which had the distances $<$ threshold t . C_{inter} is the number of compound pairs between active and non-active compounds. The threshold t of interaction and structure clusters were tested from 0 to 1 on five protein targets. For interaction cluster, the distance threshold t at 0.39 of correlation coefficient measurement had the maximum discrimination (88.9%). For structure cluster, the maximum discrimination (91.5%) was at the $t=0.55$ of tanimoto coefficient measurement (Figure 3.1.4)

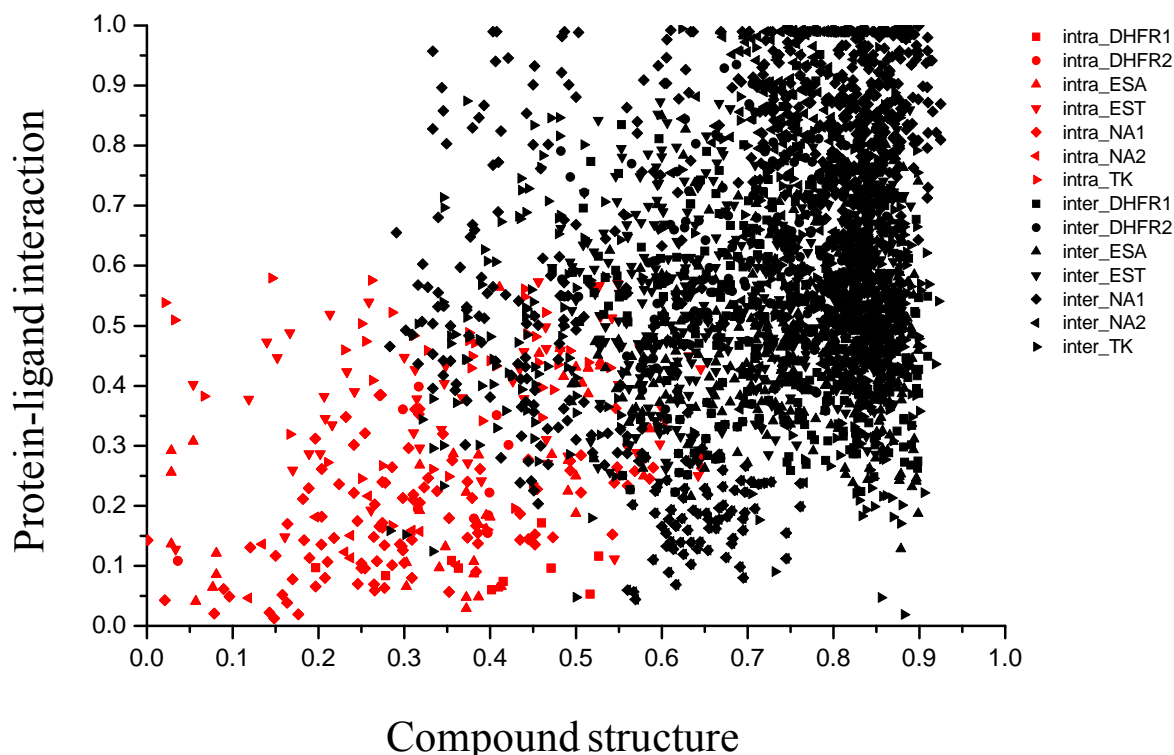


Figure 3.1.4. The normalized interaction and structure distances of five classes of compounds. The distance distributions of intra datasets are labeled in red, and the inter datasets are labeled in black. The cutoff distance of atom-pair descriptor for hierarchical cluster is 0.55 (tanimoto coefficients), and protein-ligand interaction descriptor is 0.39 (correlation coefficients).

Cluster analysis for five classes of protein targets

We demonstrate the characters of the cluster analysis through using the single feature representation.

Cluster of protein-ligand interaction

hDHFR the result of interaction clustering was shown on [Figure 3.1.5](#). Three major clusters were identified as cluster c, d and e. First, all active compounds were group together into the cluster a, ([Figure 3.1.5A](#)), All hDHFR ligands in cluster b had hydrogen-bond (E30-OE1, E30-OE2, V115-O, I7-O), van der Waals force (I60-CAR, F31-RING), the binding interactions of each docked poses within the cluster b were similar. The cluster c contained 6 TK ligands and one NA ligand, and all docked poses of cluster d were NA ligands ([Figure 3.1.5A](#)). Ligands in cluster d had hydrogen-bond (Y121-O, I7-O), and all docked poses with in cluster e had hydrogen-bond (E30-OE1, V8-N). There

were two types of hDHFR active compounds (Figure 3.1.2)⁹, the DHFR03, 04, 05, 09, 10, belonged to old types and had two more carboxylic acid group. The old drugs had different binding affinity comparing with new drugs. The interactions of two types of DHFR drugs were shown in Figure 3.1.6. The old types contain additional hydrogen bonds (R70-NH1, R70-NH2, and N64-ND2) comparing to new types, these were shown on the residue numbers in red of Figure 3.1.6A. The important van der Waals force (I60-van der Waals force, F31-stacking force, F34-stacking force, NAP-stacking force) could easily discriminate in the cluster map and those interacted residues were shown in Figure 3.1.6B and 6C. Visual inspection of hDHFR cluster analysis demonstrated that our method could help to easily identify the difference of these binding interactions and could discriminate the screening results by their differences in the protein-ligand interactions.

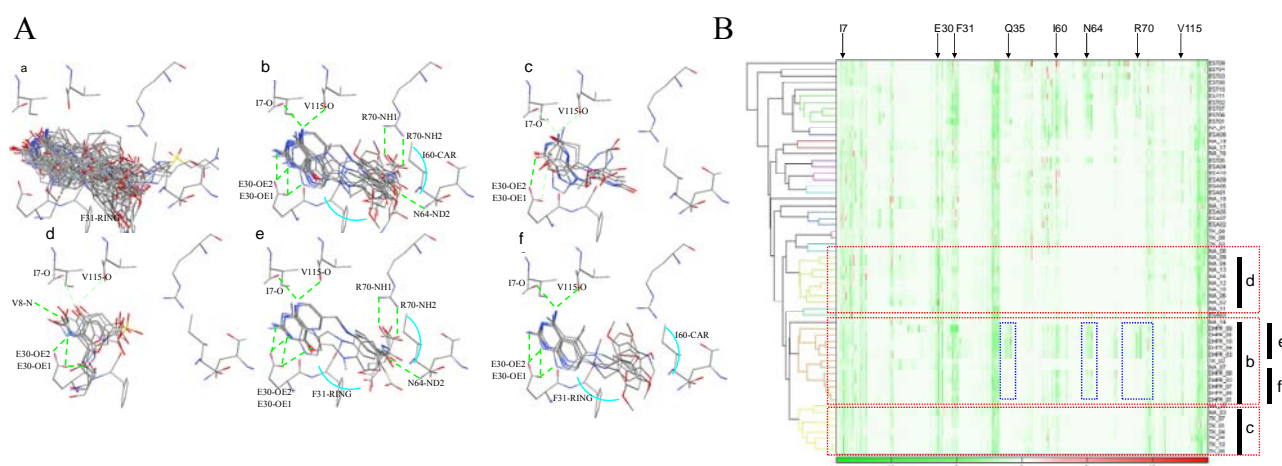


Figure 3.1.5. The interaction cluster for hDHFR. (A) The clusters of 61 known compounds against the target protein hDHFR (PDB code 1hfr). (B) Hierarchical clustering of protein-ligand interaction of 61 docked poses on hDHFR (PDB id: 1hfr). The atom based interactions of each ligand is represented as one row in the heat map. The color gradient from green to red corresponds to energy gradient from low to high. The hierarchical tree shows on the left of heat map. The interacted residues were shown in the top of the heat map. (a) The overall docked poses of 61 compounds on 1hfr. (b) The cluster with most numbers of hDHFR active compounds (c) and (d) show the clusters closed to hDHFR active compounds. (e) and (f) show the sub-clusters of (b). The hydrogen bonding is denoted as dash line and stack force is denoted as curve line.

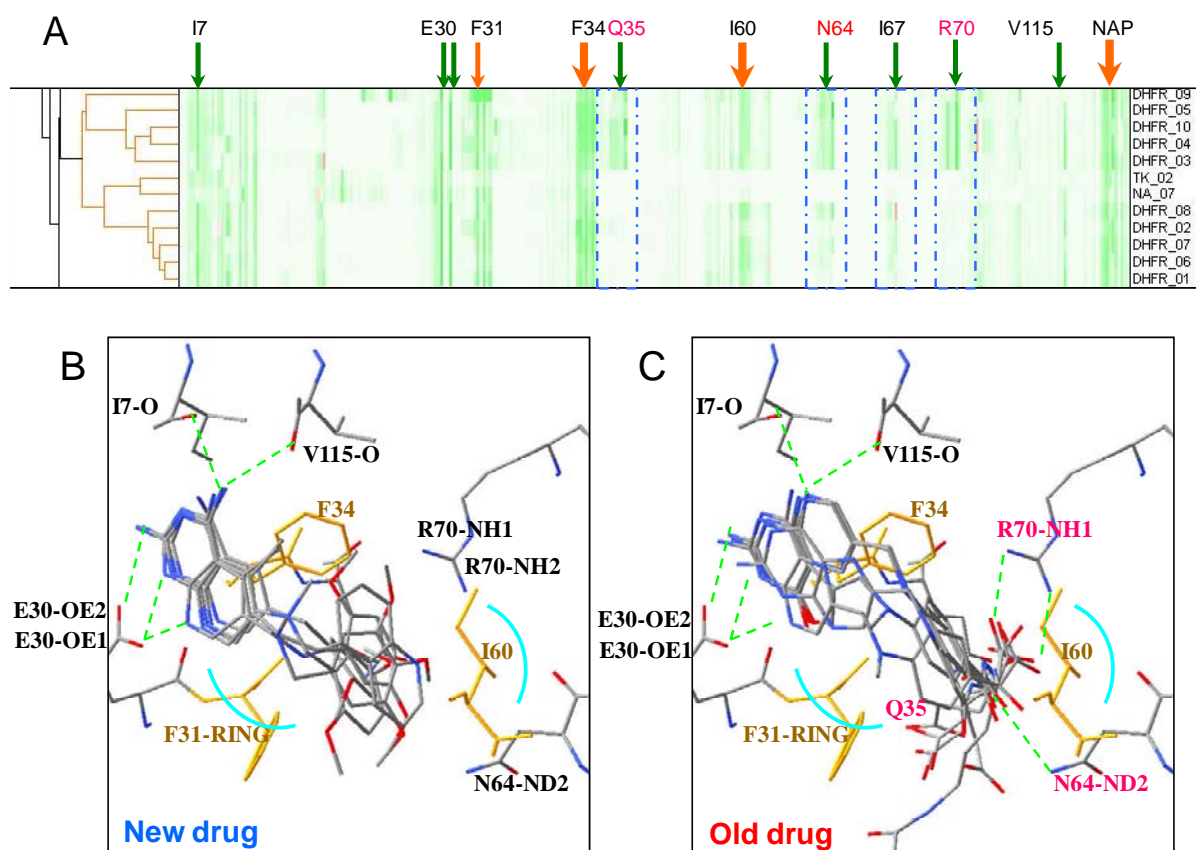


Figure 3.1.6. The sub-cluster of hDHFR active compounds bounded on hDHFR. The cluster includes 5 new drugs (DHFR01, 02, 06, 07, 08) and 5 old drugs (DHFR03, 04, 05, 09, 10). (A) The heat map of hDHFR active compounds. The arrows in green and orange denote the hydrogen bonding and van der Waals interactions, respectively. The residues labeled in red were the differences of interactions (Q35, N64, and R70) in new and old drugs. (B) The interactions of new drugs. (C) The interactions of old drugs.

TK Our cluster method analyze the interactions of TK (1kim) and the result showed in [Figure 3.1.7](#). Two major clusters were identified in the cluster analysis ([Figure 3.1.7A](#)). All docked poses of active compounds were grouped into cluster b. the major differences of interactions between cluster b and c were shown in [Figure 3.1.7B](#). The cluster b contained the hydrogen bond with Q125-NE1 and NE2, the cluster c only interacted with Q125-NE2. The cluster c contained two positive van der Waals force on A167-C, but cluster c had no interaction. That was because the structures of cluster c were slightly larger than the volume of the cavity. Those differences could easily inspect from the heat map of interactions ([Figure 3.1.7B](#)), and it is useful for mining conserved interaction within clusters.

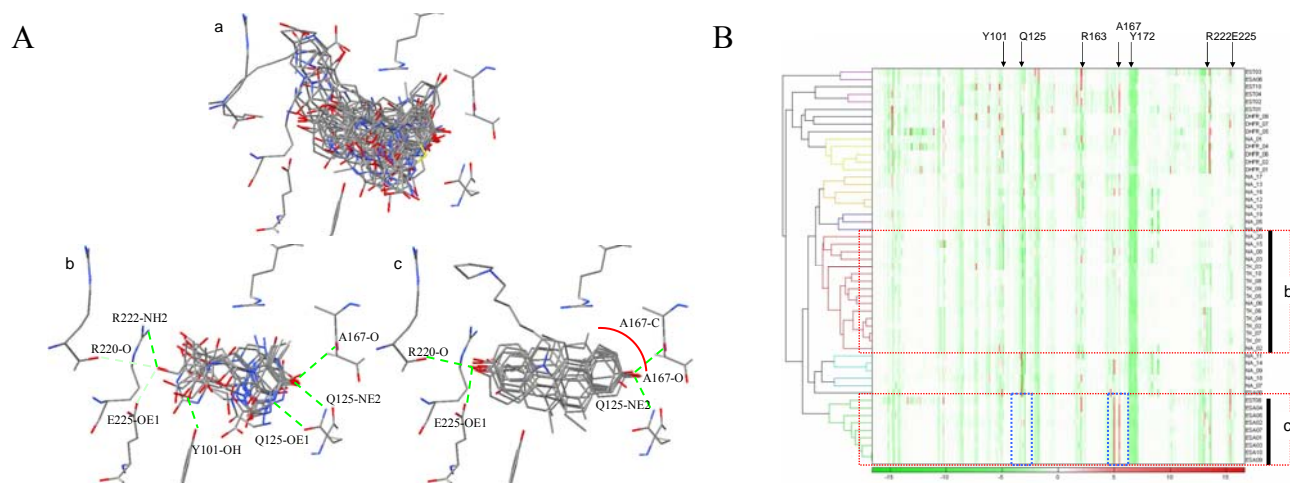


Figure 3.1.7. The interaction cluster for TK. (A) The clusters of 53 known compounds against the target protein TK (PDB code 1kim). (B) Hierarchical clustering of protein-ligand interaction of 53 docked poses on TK (PDB id: 1kim). (a) The overall docked poses of 53 compounds on 1hfr. (b) The cluster with most numbers of TK active compounds (c) show the clusters closed to TK active compounds. The hydrogen bonding is denoted as dash line and stack force is denoted as curve line.

NA, EST and ESA The processes of cluster analysis on these three targets were the same as described above. The results of NA, EST, and ESA were shown in Figure 3.1.8A, 8B and 8C, respectively. The known active compounds against NA were grouped within a cluster (frame in red) and had hydrogen-bond with target protein (R152, E277, R292, and R371). In the part of ER α antagonists (3ert), the active compounds were divided into four sub-clusters on the heat map (Figure 3.1.8B). Two were singleton, one contains 4 inhibitors, and last cluster contained 5 inhibitors and 8 ER α agonists. We could inspect that the positive van der Waals force on (I424, M388, and L349) made EST11 and EST10 different from other inhibitors. In the part of ER α agonists (1gwr), all active compounds except of ESA08 were group into one cluster, and the ESA08 had additional interaction with target protein (T347 and L525).

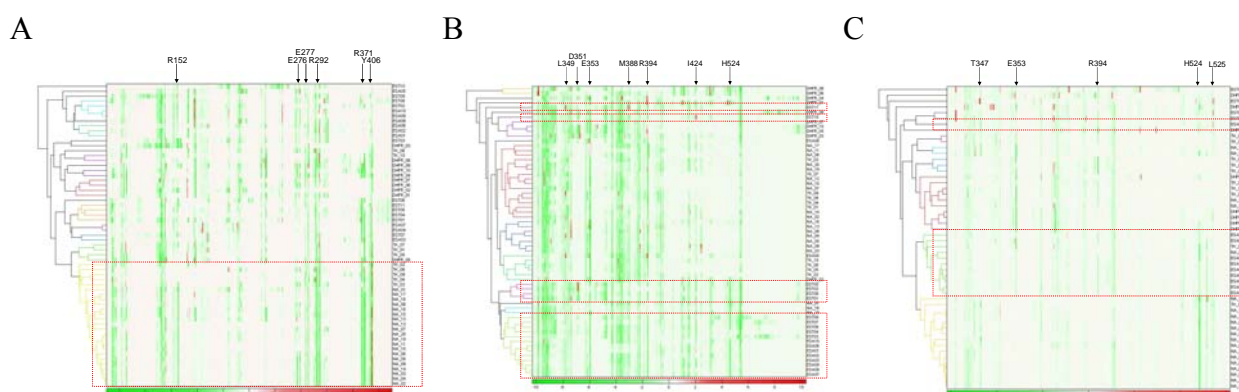


Figure 3.1.8. The interaction cluster for NA (PDB code 1mew), ER α antiaginst (PDB code 3ert) and ER α aginst (PDB code 1gwr). (A) The cluster analysis for NA. All the known active compounds were grouped within a cluster (frame in red) and had hydrogen-bond with target protein (R152, E277,

R292, and R371). (B) The cluster result of ER α antiagonists. The active compounds were divided into four clusters by the red frames on the heat map. The positive van der Waals force on (I424, M388, and L349) made EST11 and EST10 different from other inhibitors. (C) The cluster result of ER α agonists. All active compounds except of ESA08 were grouped into one cluster. ESA08 had additional interaction with target protein (T347 and L525).

Cluster of compound structures

The structure analysis was based on the topology and bonding information and represented by atom-pair descriptors. The hierarchical cluster result of 61 known compound structures was shown in Figure 3.1.9. Under the threshold $t=0.55$, there were three major clusters a, b and c. 10 ER α agonists were grouped in the cluster a and all 11 ER α antagonists were also grouped within in the cluster b. In the cluster c, it contained 10 TK inhibitors and 14 NA inhibitors. The topological structures between TK and NA inhibitors were similar.

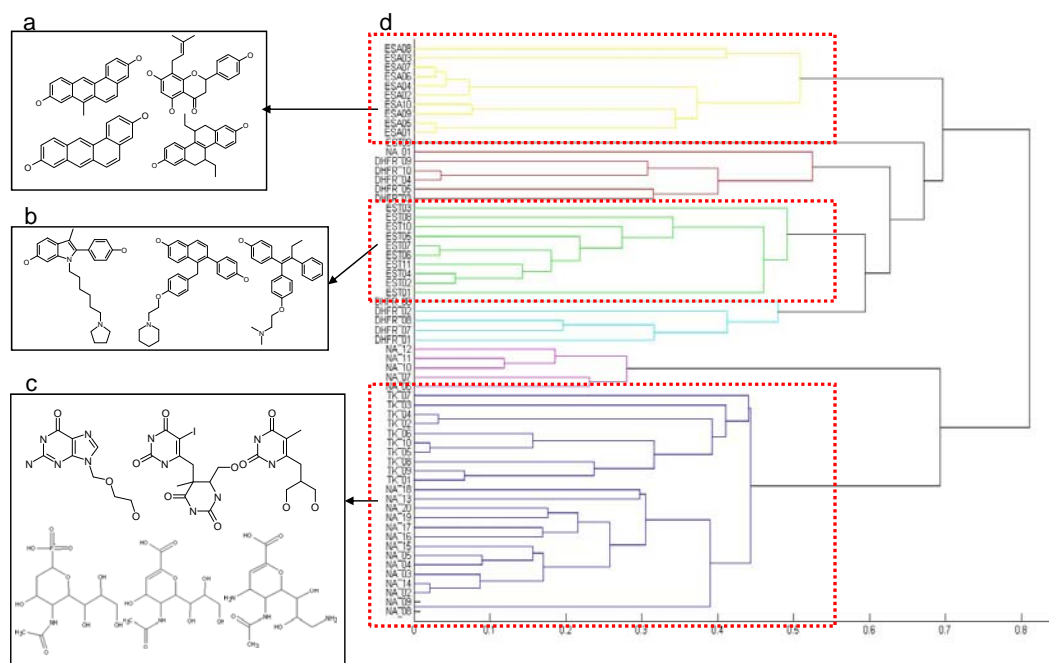


Figure 3.1.9. The hierarchical cluster of 61 known compound structures. The structure was represented by the atom-pair method and the similarity distance measured by tanimoto coefficient. Under the threshold $t = 0.55$, there were three major clusters, (a) (b) and (c). (a) 10 ER α agonists within the cluster. (b) 11 ER α antagonists in a structure cluster. (c) 10 TK inhibitors and 14 NA inhibitors.

Cluster analysis of virtual Screening on hDHFR

GEMDCOK was adapted for the virtual screening test for a set of 10 hDHFR inhibitors and 990 compounds from ACD. At the post processing of virtual screening, we adapted the cluster analysis for the top 100 in hDHFR screening list. The atom based protein-ligand interactions and atom-pair structure descriptors were generated for each ligand in the top 100. Then, the correlation coefficient and tonimoto coefficient were applied for measuring the distances of interactions and structures,

respectively. The result of hierarchical clustering was shown in Figure 3.1.10A represented as a heat map. All hDHFR inhibitors were belonged to the same cluster and this cluster contained 45 compounds included 10 active compounds and 35 unknown compounds. The interaction comparisons of active and unknown compounds were shown in Figure 3.1.10A. The hDHFR active compounds had the hydrogen bonding interactions with target protein at I7-O, V115-O, E30-OE1, E30-OE2, and N64-ND2. The van der Waals contacts formed on F31-stacking force, F31-stacking force, I60-van der Waals contact and NAP-stacking force. 35 unknown compounds had similar hydrogen bonding network to target protein (I7-O, V115-O, E30-OE1, E30-OE2, and N64-ND2) The van der Waals interactions were similar to active compounds (F31-stacking force, F31-stacking force, I60-van der Waals contact, and NAP-stacking force). The binding interactions within the cluster were similar and the most of unknown contained the flavones and purine structures which ring size closed to thymidine base.

The structure analysis result was shown in Figure 3.1.10B. Two groups of compounds were clustered as further two structural differences clusters, respectively. The active compounds were spliced into two clusters, the old types and new types (Figure 3.1.10B) because of the difference of the carboxylic acid group. The unknown structures were also clustered into purine and flavones groups which were labeled in the red circle in Figure 3.1.10B. From the cluster result, we selected the compound with lowest energy in each cluster for representing all compounds of the cluster. Each representative compounds structures were shown as Figure 3.1.10B. Our method was able to identify and classify screening result through structure and interactions of protein and ligand. The representative compound of each cluster presented the characteristics of whole cluster. The bio assay to representative structures could saved expense and improved the efficiency to discover hit from screening results.

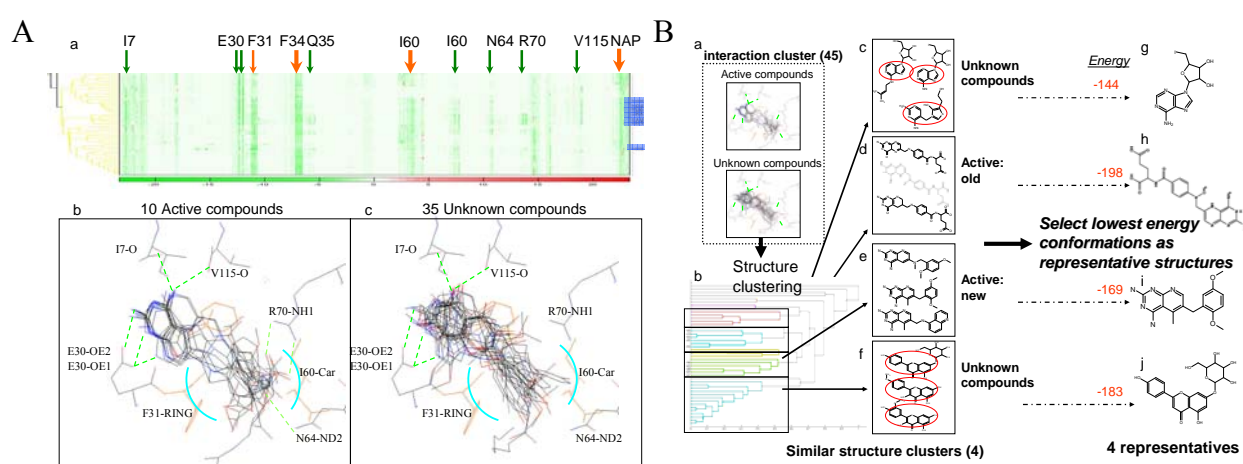


Figure 3.1.10. The cluster analysis of virtual screening on hDHFR (PDB code 1hfr). The screening set was composed of 990 random selected compounds from ACD and 10 hDHFR inhibitors. The top 100 of screening list were selected for clustering analysis. (A) The details of interactions for the

cluster with 10 hDHFR inhibitors. The cluster contained 45 compounds include 10 active compounds and 35 unknown compounds. The important hydrogen bonding and van der Waals interactions were denoted as green and orange arrows, respectively. (B) The structure analysis for the cluster with 10 hDHFR inhibitors. The structures of 45 compounds were represented by atom-pair description and clustered according to tanimoto coefficient. Four major structure clusters were grouped. The active compounds were separated into the old types (d) and the new types (e) because of the difference of the carboxylic acid group. The unknowns were clustered into 2 groups and the structures of 2 groups were different. The compound with lowest energy of each cluster was chosen as the representative for the cluster.

Table 3.1.5. The pharmacophore consensus calculated by superimposing known active compounds used for molecular docking on TK, ER, hDHFR, NA, and HpSK

		Pharmacophore consensus weight ($CW(B_{ij})$)	
Residue Id ^a	Atom Id ^b	hDHFR-ligand	Interaction type
I7	O	3.50	H-bond (NH ↔ O) (NH group)
E30	OE1	4.00	H-bond (NH ↔ O) (NH group)
E30	OE2	4.00	H-bond (NH ↔ O) (NH group)
R70	NH1	1.50	H-bond (O ↔ NH) (carbonyl group)
R70	NH2	1.50	H-bond (O ↔ NH) (carbonyl group)
V115	O	2.50	H-bond (NH ↔ O) (NH group)
ER-antagonist			
E353	OE2	3.0	H-bond (OH ↔ O) (phenolic hydroxyl)
R394	NH2	2.9	H-bond (OH ↔ N) (phenolic hydroxyl)
H524	ND1	2.4	H-bond (OH ↔ N)
D351	OD1	2.2	H-bond (N ↔ O) (dimethylamino group and piperidine nitrogen)
ER-agonist			
E353	OE2	3.1	H-bond (OH ↔ O) (phenolic hydroxyl)
R394	NH2	3.1	H-bond (OH ↔ N) (phenolic hydroxyl)
H524	ND1	3.4	H-bond (OH ↔ N)
		Pharmacophore consensus weight ($CW(B_{ij})$)	
Residue Id ^a	Atom Id ^b	TK-ligand	Interaction type
Q125	OE1	4.00	H-bond (NH ↔ O) (NH group)
Q125	NE2	3.50	H-bond (O ↔ NH) (carbonyl group)

Y101	OH	2.00	H-bond (OH ↔ OH) (hydroxyl group)
R163	NH1	1.50	H-bond (OH ↔ N) (hydroxyl group)
	CG		
	CD1		
Y172	CD2	2.50	van der Waals force (C ↔ C)
	CE1		
	CE2		
	CZ		
NA-ligand			
R371	NH1	2.0	H-bond (NH ↔ O) (NH group)
R371	NH2	2.0	H-bond (NH ↔ O) (NH group)
R292	NH1	1.5	H-bond (NH ↔ O) (NH group)
R292	NH2	1.5	H-bond (NH ↔ O) (NH group)
E276	OE2	1.5	H-bond (OH ↔ OH) (hydroxyl group)
R152	NH1	2.0	H-bond (O ↔ NH) (carbonyl group)
Pharmacophore consensus weight ($CW(B_{ij})$)			
SK-substrate			
D33	OD1	1.5	H-bond (OH ↔ OH) (hydroxyl group)
D33	OD2	1.5	H-bond (OH ↔ OH) (hydroxyl group)
R57	NH1	1.5	H-bond (O ↔ NH) (carbonyl group)
G80	N	1.5	H-bond (NH ↔ O) (NH group)
R132	NH1	1.5	H-bond (NH ↔ O) (NH group)
R132	NH2	1.5	H-bond (NH ↔ O) (NH group)

^a One-code amino acid with the residue sequence number in PDB.

^b The atom name with the atom serial number in PDB.

Table 3.1.6. The ligand preferences calculated from known active compounds used for virtual screening on TK, ER, hDHFR, NA, and HpSK

Ligand name	Electrostatic preferences (Equation 3.1.9)			Hydrophilic preferences (Equation 3.1.10)		
	$elec$	$elec$	UB_{elec}	hb	hb	Ur_{hb}
TK-substrate	0	0	0	0.50	0.05	0.55
ER-antagonist	2.00	0.56	2.56	0.15	0.03	0.18
ER-agonist	0	0	0	0.25	0.06	0.31
hDHFR-ligand	4.00	2.11	6.11	0.40	0.05	0.45
NA-ligand	4.00	0.75	4.75	0.50	0.05	0.55
SK-substrate	2.00	0	2.00	0.42	0	0.42

Cluster analysis for the screening result of HpSK

The post analysis of HpSK screening was performed by cluster analysis and further tested *in vivo*. We selected top 300 for cluster analysis. Compounds were analyzed by protein-ligand interactions and structures. The cutoffs of hierarchical clusters were set manually for giving an appropriated number of clusters. The result of cluster analysis was shown as [Figure 3.1.10](#). There were 8 major interaction clusters and compounds in these clusters were grouped into 23 sub-clusters by their structures. Finally, 23 representative compounds were selected for bioassay. Five of 23 compounds were actually bought and tested the inhibitory activities. The *in vivo* test of Dr. Wang, W.-C. identified the compound code MCMC00000106 (furosemide) shown 36% inhibition on shikimate kinase at the concentration of 625 μm .

3.1.4 Conclusions

We developed a cluster method for post analysis to improve enrichment for VS. The method combines protein-ligand interactions (e.g. hydrogen bonds, electrostatic interactions, and van der Waals), which are generated by our well-developed docking tool (i.e. GEMDOCK), and physical-chemical features and structures for each compound candidate selected by GEMDOCK. The physical-chemical features of a compound were described by atom pair descriptors (i.e. compound topological similarity) proposed by Carhart et al. Based on these normalized feature profiles, hierarchical clustering methods were used to cluster these compound candidates. For each cluster, this method selected a representative compounds for biological tests. This analysis method was validated on five pharmaceutical interest targets, TK inhibitors, DHFR inhibitors, ER agonist, ER antagonists and NA inhibitors. The validation on five targets suggested an approximated threshold for the cutoff of hierarchical clusters. The validation results also showed the power for mining the representatives with the important interactions and diverse structures from the virtual screening data. The practical application for the inhibitor analysis of HpSK reported a new inhibitor structure from the screening data. We screened the CMC database against HpSK and chosen top 300 of screening candidates for post-processing analysis. The analysis presented 23 representative candidates and five of 23 representative candidates were tested *in vivo* by cooperated laboratory of Dr. Wen-Ching Wang. The *in vivo* test identified a new inhibitor structure, furosemide and this candidate inhibited the 36% enzyme activity of HpSK at 625uM.

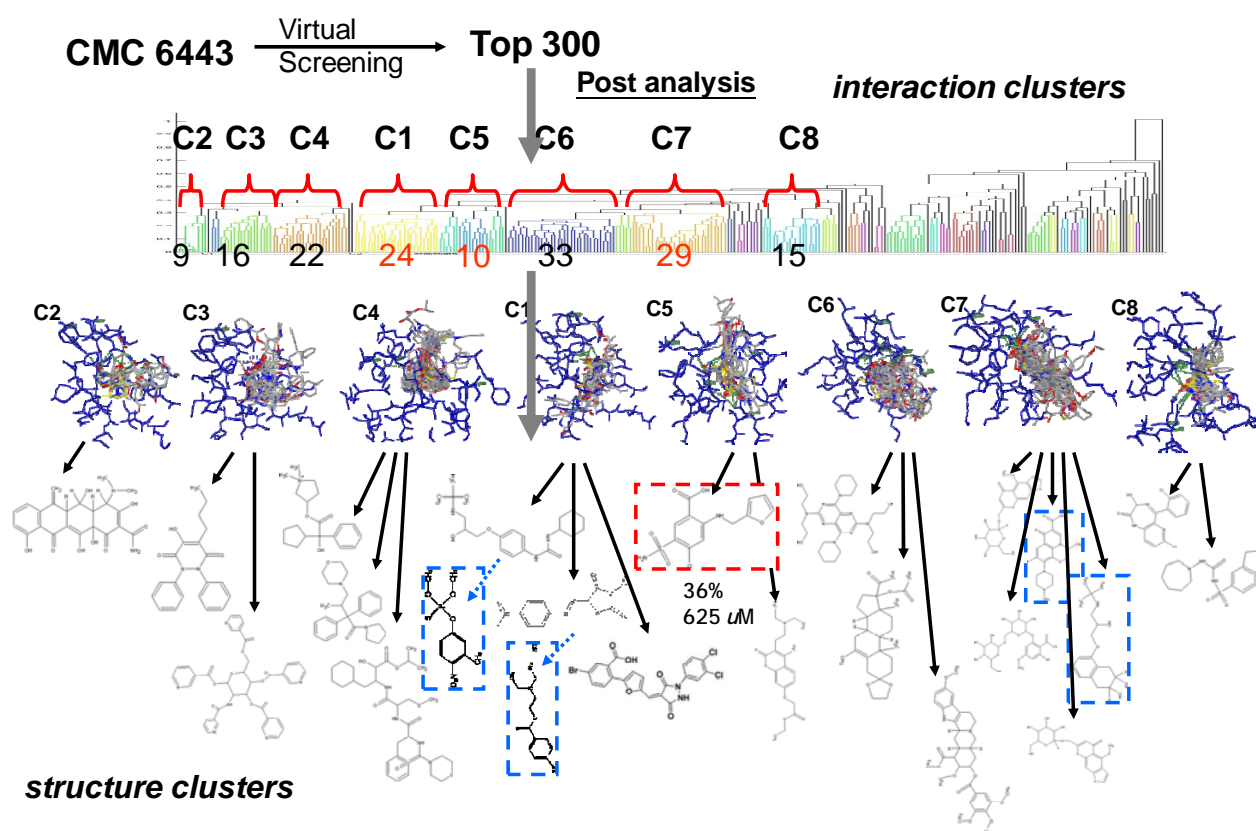


Figure 3.1.11. The cluster analysis for HpSK screening. The top 300 were selected for cluster analysis. The interaction cluster grouped the poses into 8 clusters and the structure cluster further grouped the structures as 23 sub-clusters. Five compounds were tested *in vivo* and one showed the inhibitory activity for HpSK.

3.2 Consensus Scoring Criteria for Improving Enrichment in Virtual Screening

3.2.1 Introduction

The average cost and time of bringing a new drug to market has been estimated to be US\$ 802 million in 2000 US dollars and 12 years¹⁰⁰, respectively. Discovery of novel lead compounds through virtual screening (VS) of chemical databases against protein structures is an emerging and promising step in computer aided drug design^{25; 26; 27; 52}. 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. The VS computational method involves two basic critical elements: efficient molecular docking and a reliable scoring method. A molecular docking method for VS should be able to screen a large number of potential ligands with reasonable accuracy and speed; and 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-based⁵⁷, physics-based⁵⁸, and empirical-based⁵⁹ scoring functions.

The performance of these scoring functions is often inconsistent across different systems from a database search^{17; 60}. 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.¹⁰¹ More recently, the same phenomena has been previously reported in information retrieval (IR) and in molecular similarity measurement^{102; 103; 104; 105; 106; 107}. Charifson *et al.* (1999)¹⁰¹, presented a computational study in which they used an intersection-based consensus approach to combine scoring functions. They 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 (DOCK⁵⁵ and GAMBLER) and thirteen scoring functions. Bissantz *et al.* (2000)¹⁷ used three docking programs (DOCK, FlexX³, and GOLD⁵) in combination with seven scoring functions to assess the accuracy of VS methods against two protein targets (thymidine kinase (TK) and estrogen receptor (ER)). Stahl and Rarey (2001)⁶⁰ presented a study of the performance of four scoring functions for library docking using the program FlexX on seven target proteins. The study in Verdonk *et al.* (2004)¹⁰⁸ addressed a number of issues on the use of VS protein-ligand docking based on VS experiments against four targets (neuraminidase, ptp1b, cdk2, and ER) using the program GOLD and three scoring functions. Wang and Wang (2004)¹⁰⁹ 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 are significant and potentially robust in that the performance results of these 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.

Here we address these issues for improving the enrichment in VS using the concept of data fusion and exploring diversity on scoring characteristics between individual scoring functions. In particular, we use the rank/score function as a scoring characteristic and the variance of the rank/score graph between individual scoring functions as a diversity measurement. Data fusion approaches have been proposed, developed, and implemented in information retrieval^{102; 103; 106; 107}, molecular similarity¹⁰⁵, and microarray gene expression analysis¹¹⁰, where the following two general criteria have been identified for potential improvement: (a) each of the individual scoring functions has to have a relatively good performance, and; (b) the scoring characteristics of each of the scoring functions have to be different. In viewing CS as a problem of data fusion, we investigate these two criteria, using the performance ratio P_l/P_h (P_l and P_h are the high and low performance of a pairing combination, respectively) as the relative performance measurement and the rank/score graph as the scoring characteristic, together with rank-based and score-based consensus scoring (RCS and SCS) procedure for improving the enrichment in VS by combining five scoring functions on the four target proteins TK, DHFR, ER-antagonist receptor (ER), and ER-agonist receptor (ERA) using two docking algorithms GEMDOCK^{2; 66} and GOLD⁵. A novel consensus scoring system in VS was then developed and evaluated.

3.2.2 Materials and Methods

Preparations of ligand databases and target proteins

We used the ligand data set from the comparative studies of Bissantz *et al.*¹⁷ to evaluate the screening accuracy of different CS on TK, DHFR, ER, and ERA. The receptors for these screens cover different receptor types and therefore provide a reasonable test of CS. For each target protein, the ligand database included 10 known active compounds and 990 random compounds. In total, the database used for screening ligands against the target proteins contained 1000 molecules, that is, 990 random compounds were the same for each of these screens. For screening TK and ER, the sets of 10 known active compounds were identical to that reported earlier.¹⁷ For screening ER agonists, a set of

10 known agonists was identical to that reported earlier⁷² and the 10 active compounds of DHFR were selected from the Protein Data Bank (PDB).¹¹¹

Four complexes of the target proteins were selected for virtual screening from the PDB: TK complex (PDB code: 1kim), DHFR (PDB code: 1hfr), ER-antagonist complex (PDB code: 3ert), and ER-agonist complex (PDB code: 1gwr). These complexes were reasonable choices because their ligand-binding cavities are wide enough to accommodate a broad variety of ligands and therefore did not require binding site modifications. The active compound set of each target protein, target proteins, and 990 random compounds are available on the Web at <http://gemdock.life.nctu.edu.tw/dock/download.php>.

Docking methods and scoring functions

GEMDOCK docking

Our previous work^{2; 66} showed that the docking accuracy of GEMDOCK was better than comparative approaches, such as GOLD and FlexX, on a diverse data set of 100 protein-ligand complexes proposed by Jones *et al.*⁵. The screening accuracy of GEMDOCK were also better than GOLD, FlexX, and DOCK on screening the ligand database from Bissantz *et al.* (2000) for TK²⁸ and ER- antagonist receptor¹¹². In this study, GEMDOCK parameters in the flexible docking included the initial step sizes ($\sigma=0.8$ and $\psi=0.2$), family competition length ($L = 2$), population size ($N = 200$), and recombination probability ($p_c = 0.3$). For each ligand screened, GEMDOCK optimization stopped either when the convergence was below a certain threshold value or the iterations exceeded the maximal preset value of 60. Therefore, GEMDOCK generated 800 solutions in one generation and terminated after it exhausted 48,000 solutions for each docked ligand.

GEMDOCK used a simple empirical-based scoring function (denoted GEMDOCK-Binding) and a pharmacophore-based scoring function (denoted GEMDOCK-Pharma) that used a simple empirical binding score and a pharmacophore-based score. The empirical-binding energy (E_{bind}) is given as

$$E_{GEMDOCK-Binding} = E_{inter} + E_{intra} \quad (3.2.1)$$

where E_{inter} and E_{intra} are the intermolecular and intramolecular energies, respectively.² The energy function, GEMDOCK-Pharma, can be dissected into the following terms²⁸:

$$E_{GEMDOCK-Pharma} = E_{GEMDOCK-binding} + E_{pharma} + E_{ligpre} \quad (3.2.2)$$

where $E_{GEMDOCK-Bind}$ is the empirical binding energy defined in Equation (1), E_{pharma} is the energy of binding site pharmacophores (hot spots), and E_{ligpre} is a penalty value if a ligand does not satisfy the ligand preferences^{28; 112}. E_{pharma} and E_{ligpre} are especially useful in selecting active compounds from hundreds of thousands of non-active compounds by excluding ligands that violate the characteristics of known active ligands, thereby improving the number of true positives.

GOLD 2.1 Docking.

GOLD⁵ is a widely used and reliable docking tool. Standard parameters of the GOLD program were used in this study. For each of the 10 genetic algorithm (GA) runs, a maximum number of 10,000 operations were performed on a population of 50 individuals. Operator weights for crossover, mutation, and migration were set to 95%, 95%, and 10%, respectively. The maximum distance between hydrogen donors and fitting points was set to 2 Å, and nonbonded van der Waals (vdW) energy was cut off at 4.0 Å. To further speed up the calculation, the GA docking was stopped when the top three solutions were within 1.5 Å rmsd of each other. These parameters are chosen according to the standard default settings recommended by the authors for virtual screening.

GOLD offered two scoring functions that were the GoldScore⁵ and the ChemScore¹¹³. The GoldScore function was made up of three components: protein-ligand hydrogen bond energy ($E_{H_Bond_Energy}$), protein-ligand van der Waals energy ($E_{Complex_Energy}$), and ligand internal van der Waals energy and ligand torsional strain energy ($E_{Internal_Energy}$). Here, the GoldScore function was divided into two kinds of functions (GOLD-GoldScore and GOLD-Goldinter), which were given as 5.

$$E_{GOLD-GoldScore} = -(E_{H_Bond_Energy} + E_{Complex_Energy}) - E_{Internal_Energy} \quad (3.2.3)$$

and

$$E_{GOLD-Goldinter} = -(E_{H_Bond_Energy} + E_{Complex_Energy}) \quad (3.2.4)$$

The ChemScore function was derived empirically from a set of 82 protein-ligand complexes by regression against measured affinity data. The ChemScore function was defined as¹¹³

$$G_{GOLD-ChemScore} = G_0 \cdot G_{hbond} \cdot G_{metal} \cdot G_{lipophilic} \cdot G_{rot} \quad (3.2.5)$$

Each component of this equation is the product of a term dependent on the magnitude of a particular physical contribution to free energy and a scale factor determined by regression. G_{hbond} was the hydrogen bond contribution; G_{metal} and $G_{lipophilic}$ were metal-ligand and lipophilic. Binding contributions, respectively; and G_{rot} was a term, which penalizes flexibility.

Here, two docking methods (GEMDOCK and GOLD) and five scoring functions (GEMDOCK-Binding, GEMDOCK-Pharma, GOLD-GoldScore, GOLD-Goldinter, and GOLD-ChemScore) were used to study the screening performance of data fusion. In order to analyze the performance uniformly, the fitness scores of these five scoring functions were taken as the negative of the sum of the component energy terms, so that larger fitness scores were better.

Performance evaluation

It is important to have objective criteria for evaluating the overall quality (and performance) of a scoring method. Some common factors used for this purpose are false positive (FP) rate, yield (the percentage of active ligands in the hit list), enrichment, and goodness-of-hit (GH score). Suppose that A_h is the number of active ligands among the T_h highest ranking compounds (i.e., the hit list), A

is the total number of active ligands in the database, and T is the total number of compounds in the database. Then A_h/T_h (%) is the hit rate and $(T_h - A_h)/(T - A)$ (%) is the FP rate respectively. The enrichment is defined as $(A_h/T_h)/(A/T)$. The GH score is defined as⁷⁸

$$GH = \left(\frac{A_h(3A + T_h)}{4T_h A} \right) \left(1 - \frac{T_h - A_h}{T - A} \right) \quad (3.2.5)$$

The GH score contains a coefficient to penalize excessive hit list size and, when evaluating hit lists, is calibrated by weighting the score with respect to the yield and coverage. The GH score ranges from 0.0 to 1.0, where 1.0 represents a perfect hit list (i.e., containing all of, and only, the active ligands). Here, we took the averages of FP rates, enrichments, and GH scores. For example, the averages of the FP rate and enrichment are defined as

$$\sum_{i=1}^A (T_h^i - i)/(T - A) \quad (3.2.6)$$

and

$$\left\{ \sum_{i=1}^A (i/T_h^i)/(A/T) \right\} / A, \quad (3.2.7)$$

respectively, where T_h^i is the number of compounds in a hit list containing i active compounds.

Methods of Data Fusion.

Our approach to combination methods and CS in VS is analogous to those used in IR^{102; 106; 107} and in microarray gene expression analysis¹¹⁰. Here we explore the fundamental question, i.e., when and how two scoring functions should be combined in order to achieve a performance higher than both individual scoring functions. Since the number of compounds is in the thousands or even tens of thousands, listing all mathematically possible scoring functions would be a computationally intractable problem. We therefore instead chose to take a combinatorial approach to the problem that focuses on taking a group of m scoring functions and evaluates the performance of all possible combinations, which are $\sum_{k=1}^m \binom{m}{k} = 2^m - 1$ (when m is 5, this number is 31). In addition, when we track the performance of all combinations, we investigate specifically when and why any combinations outperform all individual scoring functions in terms of the performance and the scoring characteristics of each of the individual scoring functions.

A scoring function $S_A(x)$ of the scoring method A is a function which assigns a real number to each compound x in the set of all n compounds $D = \{c_1, c_2, \dots, c_n\}$. Hence, the scoring function $S_A(x)$ is a function from $D \rightarrow \mathfrak{R}$ (the set of real numbers). When treating $S_A(x)$ as an array of real numbers, sorting the array and assigning a rank to each of their compounds would transfer the scoring function $S_A(x)$ to a ranking function $R_A(x)$ from D to N where $N = \{1, 2, \dots, n\}$. In the following, we elaborate on the issues of performance evaluation and methods of combinations.

In order to fairly compare and correctly combine multiple scoring functions, one has to normalize the

scores obtained by different methods. In our approach, we normalize all scoring functions $S_A(x): D \rightarrow \mathfrak{R}$ to the range of x which is less than or equal to 1 but greater than or equal to zero, i.e., $S'_A(x): D \rightarrow [0, 1]$, as follows:

$$S'_A(x) = \frac{S_A(x) - S_{min}}{S_{max} - S_{min}}, x \in D \quad (3.2.8)$$

where S_{max} is the maximum value and S_{min} is the minimum value of $S_A(x_j), 1 \leq j \leq n$, respectively; n is the number of compounds in the list. Here, S_{max} is the first rank and S_{min} the last rank among n compounds.

Methods of Combination

Given a list of m scoring functions, there are several different methods of combinations such as: rank by voting, rank by rank, rank by score, and conditional probability (Bayes' rule). Rank by voting has been reported to have a poor performance¹⁰⁹. The conditional probability (CP) fusion algorithm approaches the consensus scoring problem by weighting each compound in a virtual screening experiment run with an activity-based conditional probability¹¹⁴. Although the combination method CP has been shown experimentally to perform as well or better than the sum-rank (i.e. rank by rank) method, the calculation of the combination scoring $CP(x)$ for each compound x assumes that the individual scoring probabilities represent independent events.

In this paper, we consider two combinations using rank-based consensus scoring (RCS) and score-based consensus scoring (SCS). Since we distinguish the two functions (ranking function $R_A(x)$ and normalized scoring function $S'_A(x)$) for a scoring method A , we calculate the scoring function for RCS and SCS of the m scoring methods $A_k, k = 1, 2, \dots, m$, as follows:

$$S_R(x) = \sum_{k=1}^m R_{A_k}(x) / m, \quad (\text{for RCS}) \quad (3.2.9)$$

$$S_S(x) = \sum_{k=1}^m S'_{A_k}(x) / m. \quad (\text{for SCS}) \quad (3.2.10)$$

When we sort $S_R(x)$ and $S_S(x)$ into ascending and descending order, respectively, the ranking functions $R_R(x)$ and $R_S(x)$ can be obtained for RCS and SCS, respectively. We note that, in the two functions, we simply assign equal weight to each scoring method. Combination methods which give different weights to each individual scoring method have been reported¹¹⁵. The weighting method of scoring functions is a part of our future works.

Rank/Score Graph

In the process of searching for prediction variables or criteria for consensus scoring and method combination, we have defined various performance factors to evaluate a scoring method A and various methods of combinations. In this paper, we explore the scoring characteristics of scoring method A by calculating the rank/score function f_A as follows:

$$f_A(j) = (S'_A \circ R_A^{-1})(j) = S'_A(R_A^{-1}(j)) \quad (3.2.11)$$

where j is the rank of the compound x which has the score $f_A(j)$, i.e., j is in $N = \{1, 2, 3, \dots, n\}$. We note that N is not the set of compounds (which is D), but the set of all positive integers less than or equal to n . In fact, N is used as the index set for the ranking function value. The rank/score function f_A so defined signifies the scoring behavior of the scoring method A and is independent of the compounds. The graph of the rank/score function $y = f_A(x)$ with respect to the scoring method A is the rank/score graph of A. The x -axis and y -axis of a rank/score graph are the rank and the normalized score, respectively. The variation (R/S_{var}) of a rank/score graph and the relative performance measurement (P_l/P_h) of combining two scoring functions A and B are defined as

$$R/S_{var}(f_A, f_B) = \left\{ \sum_{j=1}^n (f_A(j) - f_B(j))^2 / n \right\}^{1/2} \quad (3.2.12)$$

and

$$P_l / P_h = \min(P(A), P(B)) / \max(P(A), P(B)) \quad (3.2.13)$$

where n is the number of compounds in the hit list and j is the rank of the compound with score $f_h(j)$, where $h = A$ or B ; $P(A)$ and $P(B)$ are the performances (measured as GH score and false positive rate) of methods A and B, respectively.

In IR, consensus scoring has been demonstrated to improve the performance when the combinations of the scoring functions involved have high performance (e.g., low FP rates or high GH scores) and their variation of the rank/score graph was large. Here, a new CS index (called CS_{index}), which is an indicative criterion for combining two scoring functions A and B from m ($m \geq 2$) scoring methods, was developed to guide the combinations in VS and defined as

$$CS_{index}(A, B) = g(R/S_{var}(f_A, f_B)) + g(P(A, B)), \quad (3.2.14)$$

where $g(P(A, B)) = g(P(A) + P(B) - 2P_m) + g(P_l / P_h)$.

$g(\cdot)$ is a normalization function (i.e., $g(v) : v \rightarrow [0, 1]$) and CS_{index} ranges between 0 and 2; P_m is the mean performance of m primary scoring functions (i.e., $P_m = \sum_{k=1}^m P(A_k) / m$).

Algorithm

We provided a consensus scoring (CS) procedure for both RCS and SCS to improve the screening accuracy in VS. The flowchart of the algorithm is shown in [Figure 3.2.1](#) and a more detailed description of the algorithm is shown as the following.

The RCS/SCS Algorithm

- 1) Given: A compound set D with n compounds in a compound database (or a hit list), $c_i \in D$, $i = 1, 2, \dots, n$, t receptor targets, performance evaluator P (e.g., the GH score or FP rate), and m scoring methods A_k with scoring functions $S_{A_k}(x)$, $k = 1, 2, \dots, m$.
- 2) Output: The best consensus scoring and combination methods for the t receptor targets and

- the compound set D .
- 3) Step 1: If we knew in advance which scoring function works better for a given target or targets, output this scoring function directly. Otherwise, execute following steps to select the best CS.
 - 4) Step 2: For each receptor target, calculate the scoring functions $S_{Ak}(x)$ using the m scoring methods A_k , $k = 1, 2, \dots, m$. Obtain each ranking function $R_{Ak}(x)$ from each $S_{Ak}(x)$ by ranking the scores in $S_{Ak}(x)$ in descending order. (Note: There are the m single scoring methods).
 - 5) Step 3: Calculate the other $2^m - m - 1$ combinations and consensus scoring using Equations 9 and 10. (Note: These are the $\binom{m}{k}$ k -combinations, $k = 2, 3, \dots, m$, and the scoring functions are all normalized). If the $2^m - 1$ scoring methods can be evaluated (including both rank and score combinations), then go to Step 4. Otherwise, go to Step 5.
 - 6) Step 4: [The performance of the individual scoring function can be evaluated (i.e., the active and inactive compounds are known)]
 - 7) Step 4.1: Evaluate the performance of all of single and combination scoring functions using evaluator P (e.g., GH score or FP rate). Note that these are the ranking functions $R_{Ak}(x)$ and scoring function $S_{Ak}(x)$, $k = 1, 2, \dots, 2^m - 1$. Graph the performance curve for all the single and combination functions using rank and score combinations. Order the performance within each of the m group with $\binom{m}{k}$ combinations where $k = 1, 2, \dots, m$.
 - 8) Step 4.2: For each single scoring method A , obtain rank/score graph using Equation 11.
 - 9) Step 4.3: Search in the space of $2^m - m - 1$ consensus scorings and find any combination method $A_k^{(g)}$ which is the combination of the g single scoring methods $\{A_{k_1}, A_{k_2}, \dots, A_{k_g}\}$ where $2 \leq g \leq m$ and $k_j \in [1, m]$ so that (a) $P(A_{k_j})$ have high performance (e.g., high GH scores or lower FP rates), (b) $f_{A_{k_j}}$ and $f_{A_{k_i}}$ are dissimilar and complementary for any i, j and $i \neq j$ in $[1, g]$ (i.e., $(R/S_{\text{var}}(f_{A_{k_j}}, f_{A_{k_i}}))$ is large), and (c) $P(A_k^{(g)})$ is better than or as good as $P(A_k)$, where A_k are the single scoring functions and $k = 1, 2, \dots, m$. The consensus scorings are often to improve the screening accuracy when the value CS_{index} (Equation 13) is more than 1.2.
 - 10) Step 4.4: The combination method $A_k^{(g)}$ is the desired consensus scoring method that we seek for the receptor target and the compound set D . Go to Step 6.

- 11) Step 5: [The performance of the individual scoring function is unknown (i.e., the active and inactive compounds are unknown)]
- 12) Step 5.1: For each single scoring method A , obtain the rank/score graph using Equation 11.
- 13) Step 5.2: Search in the space of the m single scoring functions. Find any group of g single scoring functions $A^{(g)} = \{A_{k_1}, A_{k_2}, \dots, A_{k_g}\}$, where $2 \leq g \leq m$ and $k_j \in [1, m]$ so that $f_{A_{k_j}}$ and $f_{A_{k_i}}$ are dissimilar and complementary for any i, j and $i \neq j$ in $[1, g]$ (i.e., $R/S_{\text{var}}(f_{A_{k_j}}, f_{A_{k_i}})$ is large).
- 14) Step 5.3: The combination method $A_k^{(g)}$ of g single scoring methods is the desired combination method for the receptor target and the compound set D .
- 15) Step 6: Output the $A_k^{(g)}$ which is the desired combination methods.

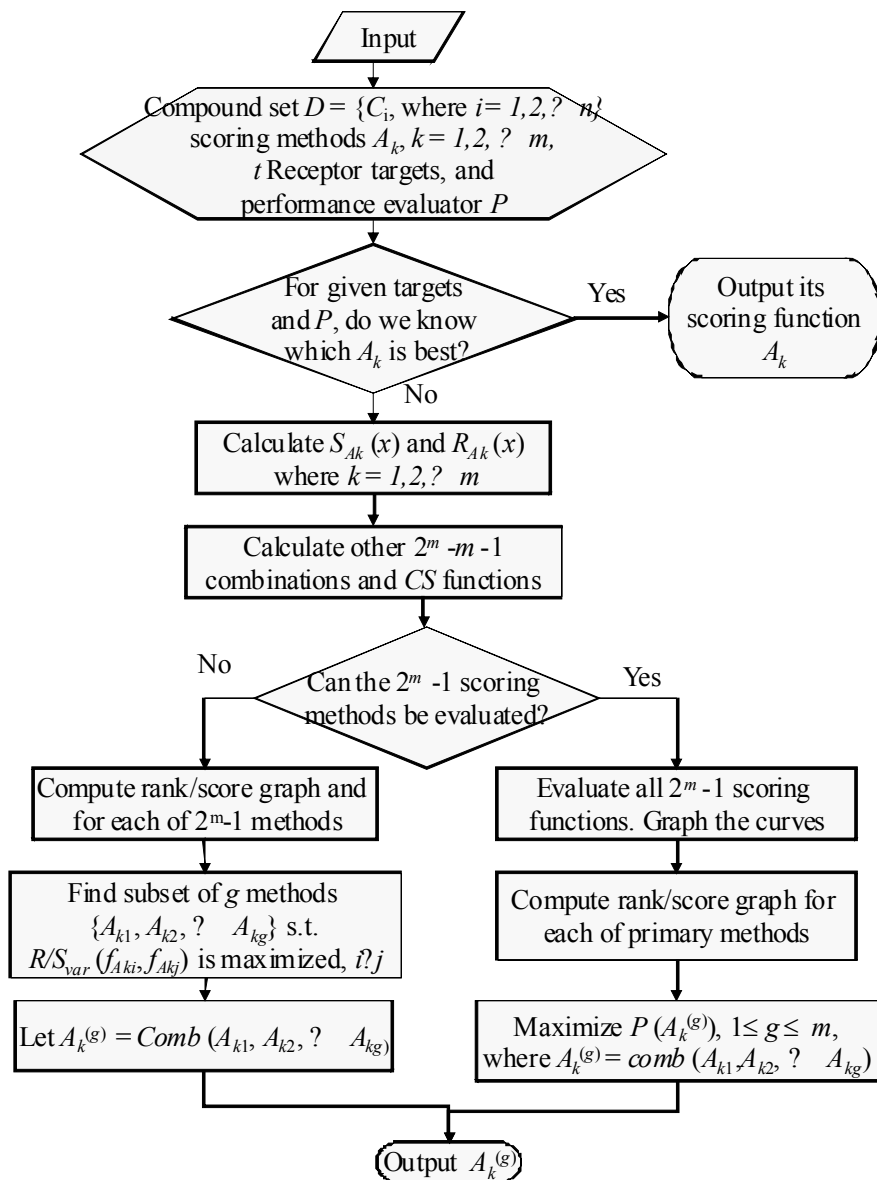


Figure 3.2.1. The Flowchart of RCS/SCS Algorithm

3.2.3 Results and Discussion

Table 3.2.1 shows the overall accuracy of using two docking programs (GEMDOCK and GOLD) and five scoring functions to assess the accuracy of VS methods against four protein targets (TK, DHFR, ER, and ERA). These scoring methods, defined in Equations 3.2.1 to 3.2.5, were termed as GEMDOCK-Binding (Method A), GEMDOCK-Pharma (Method B), GOLD-GoldScore (Method C), GOLD-Goldinter (Method D), and GOLD-ChemScore (Method E). For each method, the first term denotes the docking tool and the second term represents the scoring function used. For example, Method A uses GEMDOCK as the docking tool and Equation 3.2.1 as the scoring function; Method E uses GOLD as the docking tool and Equation 3.2.5 as the scoring function. The average FP rate (Equation 3.2.6) and average GH score were used to evaluate the screening accuracy. Among these five methods, the accuracy of GEMDOCK-Pharma was the best for TK and both ER receptors, and GOLD-Goldinter outperformed other methods on DHFR.

Table 3.2.1. Screening accuracy of five methods on screening TK, DHFR, ER-antagonist receptor, and ER-agonist receptor. The bold case is the best score. GEMDOCK-Pharma with the pharmacophore-based scoring function and GOLD with the Goldinter score are superior to others

Target protein ^a	Measure factor	GEMDOCK -Binding (Method A ^b)	GEMDOCK -Pharma (Method B ^c)	GOLD -GoldScore (Method C ^d)	GOLD -Goldinter (Method D ^e)	GOLD -ChemScore (Method E ^f)
TK	Average Enrichment	12.29	42.27	10.34	7.09	1.32
	Average FP rate (%)	4.11	0.82	5.04	7.61	38.48
	Average GH Score	0.22	0.45	0.20	0.17	0.08
DHFR	Average Enrichment	29.57	70.21	29.40	90.64	1.17
	Average FP rate (%)	3.24	0.32	15.49	1.48	50.04
	Average GH Score	0.35	0.66	0.32	0.81	0.05
ER-antagonist receptor (ER)	Average Enrichment	34.88	92.19	34.07	75.14	67.14
	Average FP rate (%)	1.32	0.13	20.44	0.88	1.17
	Average GH Score	0.39	0.83	0.34	0.70	0.64
ER-agonist receptor (ERA)	Average Enrichment	6.94	45.66	3.50	15.21	25.09
	Average FP rate (%)	7.83	0.75	21.67	6.40	5.24
	Average GH Score	0.17	0.48	0.12	0.23	0.31

^a TK: HIV-1 thymidine kinase (PDB code: 1kim); DHFR: human dihydrofolate reductase (PDB code: 1hfr); ER-antagonist receptor: estrogen receptor of antagonists (PDB code: 3ert); and ER-agonist receptor: estrogen receptor of agonists (PDB code: 1gwr).

^b Method A uses GEMDOCK as the docking tool and Equation (1) as the scoring function.

^c Method B uses GEMDOCK as the docking tool and Equation (2) as the scoring function.

^d Method C uses GOLD as the docking tool and Equation (3) as the scoring function.

^e Method D uses GOLD as the docking tool and Equation (4) as the scoring function.

^f Method E uses GOLD as the docking tool and Equation (5) as the scoring function.

Table 3.2.2 shows FP rates of GEMDOCK and four comparative approaches: Surflex ¹⁸, DOCK ⁵⁵, FlexX ³, and GOLD ⁵) for screening the ER and TK. All of these methods were tested using the same reference protein and screening database with true positive rates ranging from 80% to 100%. GEMDOCK-Pharma (GEMDOCK with pharmacological preferences) was superior to the comparative approaches and GOLD-Goldscore (GOLD used Equation 3 as the scoring function) was better than FlexX and DOCK, two widely used docking tools. For example, the FP rates were 2.3% (GEMDOCK-Binding), 0.4% (GEMDOCK-Pharma), 1.6% (Surflex), 17.4% (DOCK), 70.9% (FlexX), and 8.3% (GOLD-GlodScore) when the true positive rate was 90% for ER antagonists.

Table 3.2.2. Comparing GEMDOCK with other methods on screening the ER antagonists and TK inhibitors by false positive rates (%) on the same data set proposed by Bissantz et al.¹⁷ The bold case is the best score

Target protein	True positive (%)	GEMDOCK -Binding ^a	GEMDOC K -Pharma ^b	Surflex ^c	DOCK ^c	FlexX ^c	GOLD ^c
ER-antagonists	80	1.5 (15/990)^d	0.0 (0/990)	1.3	13.3	57.8	5.3
	90	2.3 (23/990)	0.4 (4/990)	1.6	17.4	70.9	8.3
	100	5.2 (51/990)	0.9 (9/990)	2.9	18.9	- ^e	23.4
Thymidine kinase	80	4.7 (47/990)	0.6 (6/990)	0.9	23.4	8.8	8.3
	90	8.9 (88/990)	1.3 (13/990)	2.8	25.5	13.3	9.1
	100	9.7 (96/990)	2.9 (29/990)	3.2	27.0	19.4	9.3

^{a,b} GEMDOCK uses Equations 1 and 2 as scoring functions, respectively.

^c Directly summarized from¹⁸.

^d The false positive rate from 990 random ligands.

^e FlexX could not calculate the docked solution for EST09.

Our consensus scoring methods consist of rank combinations and score combinations on five methods, including Method A, B, C, D, and E (Table 3.2.1). Result statistics of VS in TK, DHFR, and ER, and ERA are summarized in Figures 3.2.1 and 3.2.2 and Tables 3.2.3 and 3.2.4. Figures 3.2.1 and 3.2.2 plot the average FP rates and average GH scores, respectively, of all 31 possible combinations including the five individual scoring functions. The y -axis values for each combination (including the single case) are sorted in ascending order in each group of k -combinations, $k = 1, 2, 3, 4, \text{ and } 5$, respectively. A k -combination method means that it combines k methods. For example, the number of 2-combination methods is 10 (i.e., $C_5^2=10$) in this paper. The Method BD is the combination of Methods B and D; and the Method CDE is the combination of Methods C, D, and E. Tables 3.2.3 (RCS) and 3.2.4 (SCS) give average FP rates and average GH scores of five kinds of k -combination methods for screening four targets. According to these experimental results, the behavior of RCS and SCS is similar. Therefore, we focus on the analysis of RCS in the following.

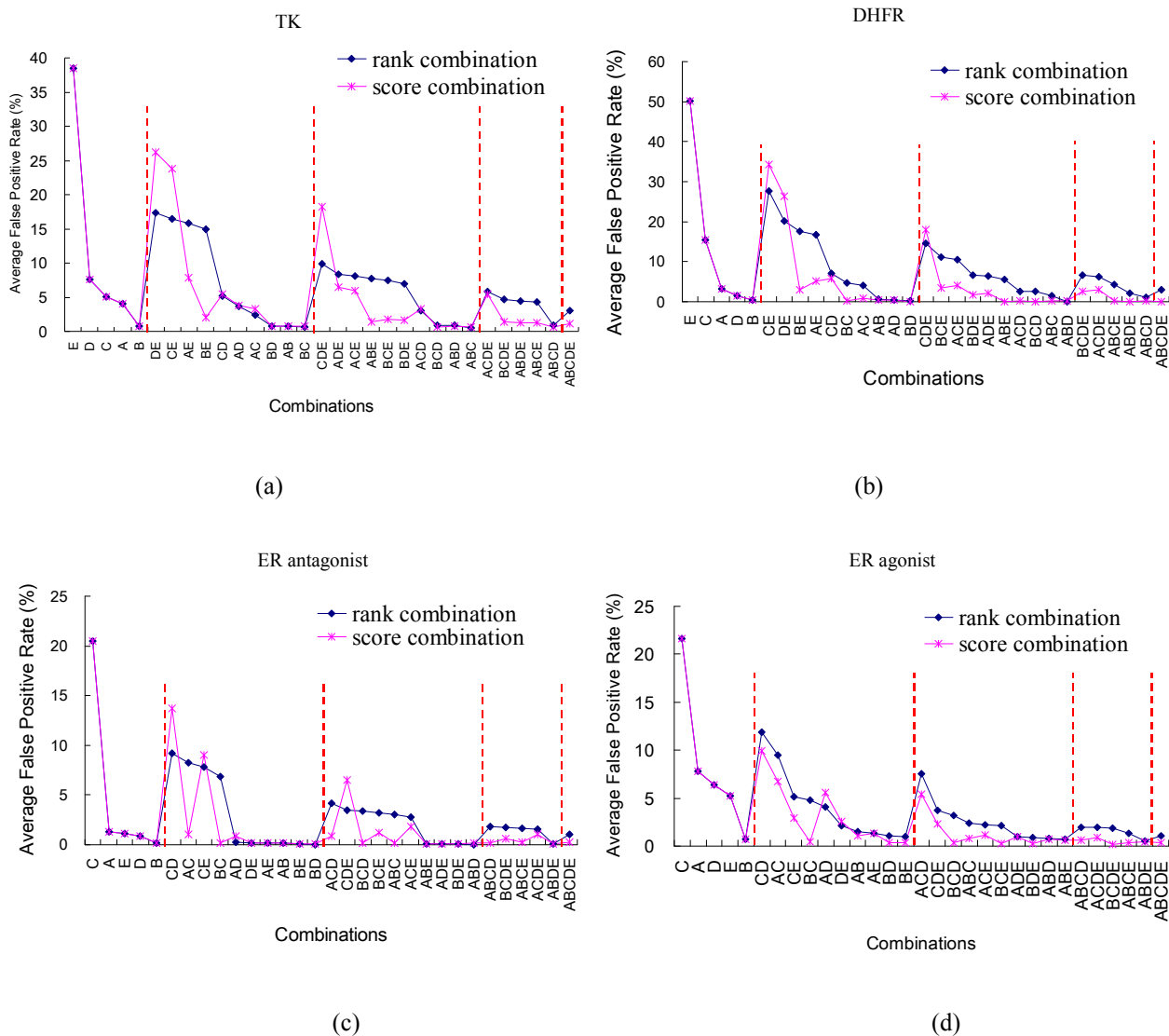
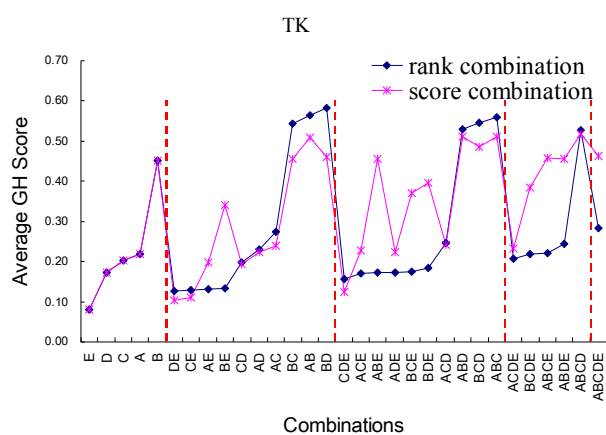
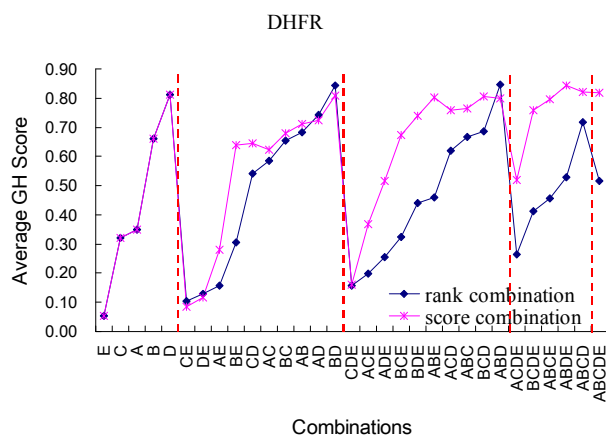


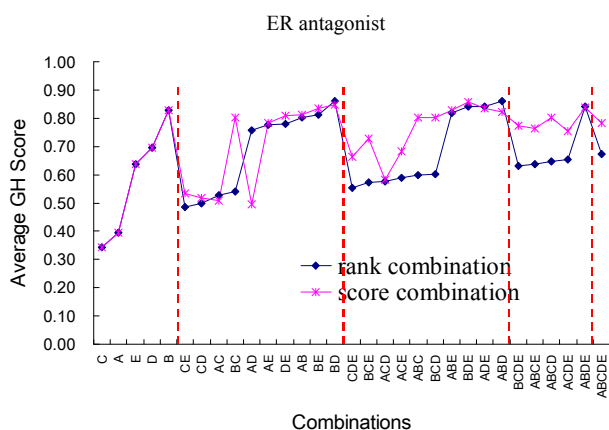
Figure 3.2.2. Average false positive rates of 31 various rank combinations and scoring combinations of five methods for four virtual screening targets: (a) TK, (b) DHFR, (c) ER-antagonist receptor, and (d) ER-agonist receptor. These five methods (i.e., A, B, C, D, and E) are defined in Table 3.2.1.



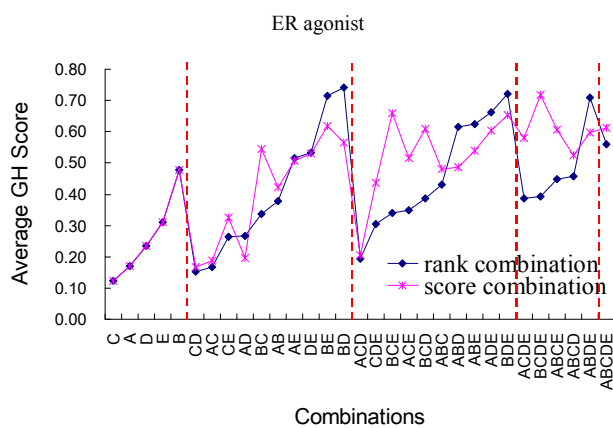
(a)



(b)



(c)



(d)

Figure 3.2.3. Average GH scores of 31 various rank combinations and scoring combinations of five methods for four virtual screening targets: (a) TK, (b) DHFR, (c) ER-antagonist receptor, and (d) ER-agonist receptor. These five methods (i.e., A, B, C, D, and E) are defined in [Table 3.2.1](#).

Table 3.2.3. Screening accuracies of different rank combinations of five methods on screening four targets: TK, DHFR, ER, and ERA. The methods and targets are defined in [Table 3.2.1](#) and the bold case is the best score

Measurement factors	Single (5) ^a				2-Com (10) ^b				3-Com (10) ^c				4-Com (5) ^d				5-Com (1) ^e			
Average false positive rate (%)	TK	DHFR	ER	ERA	TK	DHFR	ER	ERA	TK	DHFR	ER	ERA	TK	DHFR	ER	ERA	TK	DHFR	ER	ERA
Average	11.21	14.12	4.79	8.38	7.77	9.91	3.31	4.24	5.40	6.16	2.04	2.46	4.04	4.11	1.39	1.52	- ^f	- ^f	- ^f	- ^f
SD	15.44	20.98	8.76	7.89	7.34	9.80	4.09	3.76	3.63	4.73	1.72	2.07	1.90	2.44	0.74	0.61	- ^f	- ^f	- ^f	- ^f
Maximum value	38.48	50.04	20.44	21.67	17.35	27.56	9.22	11.86	9.92	14.64	4.16	7.55	5.88	6.69	1.86	1.98	3.10	3.00	1.04	1.08
Minimum value	0.82	0.32	0.13	0.75	0.58	0.14	0.04	0.99	0.53	0.07	0.04	0.72	0.83	1.07	0.08	0.55	3.10	3.00	1.04	1.08
Average GH score	TK	DHFR	ER	ERA	TK	DHFR	ER	ERA	TK	DHFR	ER	ERA	TK	DHFR	ER	ERA	TK	DHFR	ER	ERA
Average	0.23	0.44	0.58	0.26	0.29	0.47	0.68	0.41	0.29	0.47	0.69	0.46	0.28	0.48	0.68	0.48	- ^f	- ^f	- ^f	- ^f
SD	0.14	0.30	0.21	0.14	0.19	0.28	0.15	0.21	0.18	0.23	0.13	0.18	0.14	0.17	0.09	0.13	- ^f	- ^f	- ^f	- ^f
Maximum value	0.45	0.81	0.83	0.48	0.58	0.84	0.86	0.74	0.56	0.85	0.86	0.72	0.53	0.72	0.84	0.71	0.28	0.52	0.67	0.56
Minimum value	0.08	0.05	0.34	0.12	0.13	0.10	0.48	0.15	0.16	0.16	0.55	0.19	0.21	0.27	0.63	0.39	0.28	0.52	0.67	0.56

^a Five individual methods

^b Combination of two selected methods, ten compositions.

^c Combination of three selected methods, ten compositions.

^d Combination of four selected methods, five compositions.

^e Combination of five selected methods and only one composition.

^f Average and standard deviation could not be calculated when one value exists.

Table 3.2.4. Screening accuracies of different score combinations of five methods on screening four targets: TK, DHFR, ER, and ERA. The methods and targets are defined in Table 3.2.1 and the bold case is the best score

Measurement factors	Single (5) ^a	2-Com (10) ^b	3-Com (10) ^c	4-Com (5) ^d	5-Com (1) ^e
Average false positive rate (%)	TK DHFR ER ERA	TK DHFR ER ERA	TK DHFR ER ERA	TK DHFR ER ERA	TK DHFR ER ERA
Average	11.21 14.12 4.79 8.38	7.46 7.67 2.57 3.13	4.07 3.02 1.12 1.28	2.01 1.15 0.44 0.50	- ^f - ^f - ^f - ^f
SD	15.44 20.98 8.76 7.89	9.51 12.25 4.80 3.26	5.40 5.47 2.00 1.57	1.92 1.45 0.41 0.27	- ^f - ^f - ^f - ^f
Maximum value	38.48 50.04 20.44 21.67	26.16 34.31 13.74 9.91	18.18 18.00 6.54 5.42	5.40 2.91 1.08 0.92	1.14 0.09 0.22 0.34
Minimum value	0.82 0.32 0.13 0.75	0.73 0.10 0.07 0.33	0.64 0.10 0.07 0.23	0.69 0.05 0.10 0.19	1.14 0.09 0.22 0.34
Average GH score	TK DHFR ER ERA	TK DHFR ER ERA	TK DHFR ER ERA	TK DHFR ER ERA	TK DHFR ER ERA
Average	0.23 0.44 0.58 0.26	0.28 0.53 0.69 0.41	0.36 0.64 0.76 0.52	0.41 0.75 0.79 0.60	- ^f - ^f - ^f - ^f
SD	0.14 0.30 0.21 0.14	0.15 0.27 0.16 0.17	0.14 0.22 0.09 0.13	0.11 0.13 0.03 0.07	- ^f - ^f - ^f - ^f
Maximum value	0.45 0.81 0.83 0.48	0.51 0.81 0.85 0.62	0.51 0.81 0.86 0.66	0.52 0.84 0.84 0.72	0.46 0.82 0.78 0.61
Minimum value	0.08 0.05 0.34 0.12	0.10 0.08 0.50 0.17	0.12 0.16 0.58 0.20	0.23 0.52 0.75 0.52	0.46 0.82 0.78 0.61

^a Five individual methods

^b Combination of two selected methods, ten compositions.

^c Combination of three selected methods, ten compositions.

^d Combination of four selected methods, five compositions.

^e Combination of five selected methods and only one composition.

^f Average and standard deviation could not be calculated when one value exists.

As shown in [Figure 3.2.2](#) and [Figure 3.2.3](#) and [Table 3.2.3](#), the average accuracy improved with the increase of fused methods. Five individual methods on screening TK, found that the best GH score and best false positive rate are 0.23 and 11.21%, respectively ([Table 3.2.3](#)). When method fusions with rank combinations were carried out by combining a pair of methods one by one, the accuracy improved from 0.23 to 0.29 in average of overall GH score and the average of false positive rates dropped from 11.21% to 7.77% ([Table 3.2.3](#)). Fusing three and four selected methods maintained mean GH scores at 0.29 and 0.28. The overall false positive rates fell in average value of 5.40% and 4.04%, respectively.

The effectiveness of fused methods may be influenced by the performance of primary methods. For example, the best GH scores of primary methods are 0.45, 0.81, 0.83, and 0.48, respectively, for TK, DHFR, ER, and ERA ([Table 3.2.1](#)), and the improvements of fused methods for TK and ERA are significantly better than the ones of DHFR and ER. With the increase of fused methods, the average accuracy as measured by means GH scores and false positive rates. Other entries (i.e., DHFR, ER, and ERA) in [Table 3.2.3](#) show similar trends of these promotions in average GH scores and false positive rates.

Although [Table 3.2.3](#) shows the average accuracy level improving with number of fused methods, the unique contribution of data fusion is most clearly observed when one considers individually the results obtained with each of the possible combinations. Specifically, comparing within [Figure 3.2.2](#) and [Table 3.2.4](#), the maximum accuracy always occurs in the combination of a pair of methods. In all of the screening sets in this paper, the best composition consistently appeared with the combination of Methods B and D.

For ER antagonists, the GH scores of Method A and Method C were 0.39 and 0.34 and the other three methods (Methods B, D, and E) had good GH scores with 0.83, 0.70, and 0.64, respectively ([Table 3.2.1](#)). As shown in [Figure 3.2.2\(c\)](#) and [Figure 3.2.3 \(c\)](#), combinations with Method A or Method C may reduce the performance of an individual method. For example, Methods CD and BC performed worse than Methods B and D. One possible reason is that these less accurate methods are predominantly adding noise that overwhelms the correction ability of fusion. On the other hand, combinations with Method B or Method D performed comparatively better than the other Method combinations. The Method BD had the highest value (0.86) in the GH score and lowest value in the false positive rate (0.04%). Other targets (i.e., TK, DHFR, and ERA) in [Table 3.2.3](#) and [Figure 3.2.2](#) show similar results. This phenomenon indicated data fusion could improve the quality of screening if each of the combination methods has relatively high performance.

[Figure 3.2.4](#) shows the rank/score graphs of five individual scoring methods and [Table 3.2.5](#) shows the variances of rank/score graphs of 10 compositions combining two methods for four screening targets. The scoring value showing in [Figure 3.2.4](#) was normalized through [Equation 3.2.8](#). The variation of rank/score graph of the Method AB, on average, is the smallest (i.e., the rank score graphs are the most similar) because Methods A and B used the same docking tool and the similar scoring function. The Method CD has the similar phenomenon. [Table 3.2.1](#) shows that Method B

consistently outperformed Method A and [Figure 3.2.3](#) shows that the fusion methods combining with Method B are consistently better than the methods combining with Method A in four test cases. For DHFR, ER, and ERA, Method D is better than Method C and the fusion methods with Method D consistently outperformed than the fusion methods with Method C. According to these observations, we could divide these five methods into three groups. The first group consisted of Methods A and B, the second group included Methods C and D, and the final group is GOLD-ChemScore (Method E).

Table 3.2.5. The relationships between the GH-score improvement with the performance ratio (PI/Ph), CSindex and the variance (R/Svar) of rank/score graph of 10 pairing combinations of five methods on four virtual screening targets

Target protein*		AB ^a	AC	AD	AE	BC	BD	BE	CD	CE	DE
TK	$g(P_l/P_h)^b$	0.41	1.00	0.82	0.26	0.36	0.27	0.00	0.91	0.30	0.39
	$g(R/S_{var})^c$	0.34	0.64	0.62	0.39	1.00	0.97	0.74	0.00	0.19	0.17
	CS_{index}^d	1.34	1.64	1.37	0.39	1.92	1.74	1.03	0.80	0.19	0.19
	RCS ^e	0.11	0.06	0.01	-0.09	0.09	0.13	-0.32	0.00	-0.07	-0.05
	SCS ^f	0.06	0.02	0.00	-0.02	0.01	0.01	-0.11	-0.01	-0.09	-0.07
DHFR	$g(P_l/P_h)^b$	0.54	1.00	0.43	0.10	0.49	0.88	0.02	0.39	0.12	0.00
	$g(R/S_{var})^c$	0.04	0.91	0.88	0.32	1.00	0.97	0.41	0.00	0.45	0.46
	CS_{index}	0.61	1.56	1.46	0.32	1.52	1.97	0.53	0.54	0.45	0.65
	RCS	0.02	0.23	-0.07	-0.19	0.01	0.03	-0.36	-0.27	-0.22	-0.68
	SCS	0.05	0.28	-0.09	-0.07	0.02	0.00	-0.02	-0.17	-0.23	-0.70
ER antagonists											
(ER)	$g(P_l/P_h)^b$	0.12	0.92	0.30	0.41	0.00	0.86	0.71	0.16	0.25	1.00
	$g(R/S_{var})^c$	0.00	0.82	0.46	0.29	1.00	0.68	0.47	0.21	0.28	0.03
	CS_{index}	0.15	1.11	0.62	0.47	1.01	1.68	1.30	0.21	0.30	0.96
	RCS	-0.03	0.13	0.06	0.14	-0.29	0.03	-0.01	-0.20	-0.15	0.08
	SCS	-0.01	0.11	-0.20	0.15	-0.03	0.02	0.01	-0.18	-0.10	0.11
ER agonists											
(ERA)	$g(P_l/P_h)^b$	0.21	0.91	0.96	0.59	0.00	0.47	0.80	0.53	0.27	1.00
	$g(R/S_{var})^c$	0.40	0.49	0.57	0.10	0.88	1.00	0.39	0.00	0.33	0.45
	CS_{index}	0.70	0.77	1.07	0.43	0.93	1.61	1.39	0.08	0.33	1.22
	RCS	-0.10	0.00	0.03	0.20	-0.14	0.26	0.24	-0.08	-0.05	0.22
	SCS	-0.06	0.02	-0.04	0.20	0.07	0.09	0.14	-0.07	0.01	0.22

* Four target proteins (TK, DHFR, ER, and ERA) are defined in [Table 3.2.1](#).

^a There are 10 compositions of combining pair methods from 5 primary scoring methods (A, B, C, D, and E) defined in [Table 3.2.1](#).

^b the normalization performance ratio (Equation 13) of a pair-combination method.

^c the normalization variance (Equation 12) of a rank/score graph of a pair-combination method.

^d a performance indicator (Equation 14) of a pair-combination method.

^{e,f} the GH-score improvements of rank-based consensus scoring and score-based consensus scoring, respectively.

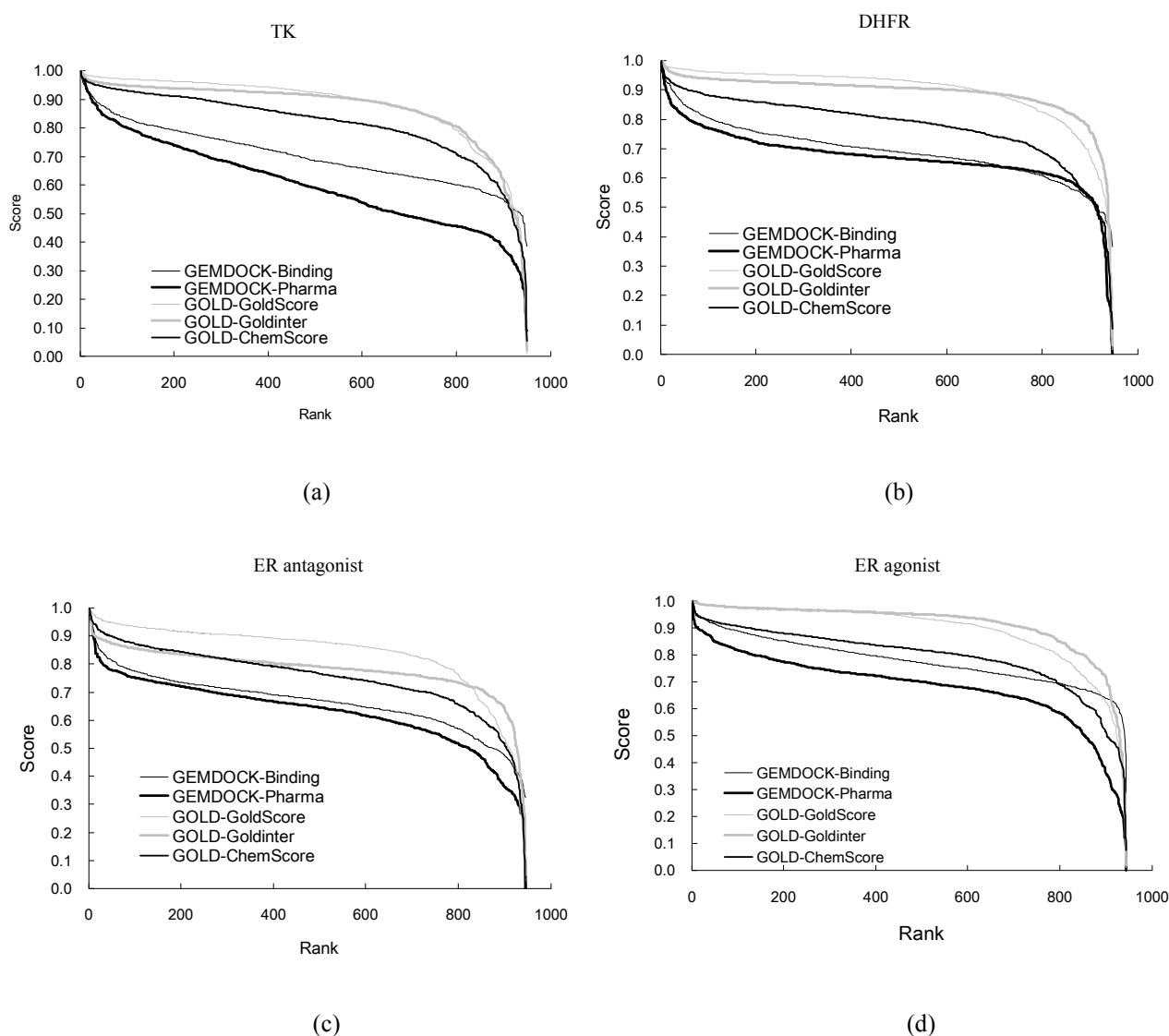


Figure 3.2.3. Rank/score curves of five methods (defined in [Table 3.2.1](#)) for four virtual screening targets: (a) TK, (b) DHFR, (c) ER-antagonist receptor, and (d) ER-agonist receptor.

Analyzing the relation between [Figure 3.2.2](#) and [Figure 3.2.3](#) revealed a possible mode of the fusion performance for VS according to the observation of data fusion in IR that the fusion

performance corresponds to the comparability of individual performance and the graphical variation of individual rank/score graph. The most variation of rank/score graphs was the Method BD among 10 pair combinations (Figure 3.2.4 and Table 3.2.5) and the Method BD also brought the best GH score for all test cases (Figure 3.2.3). In Figure 3.2.3b (DHFR), Methods B and D had the highest GH score (0.66 and 0.81) among primary methods (Table 3.2.1) and the combination of these two methods had the best GH score (0.84) and the lowest false positive rate (0.14%) among the combinations with two methods. A similar phenomenon occurred in the ER antagonist study (Figure 3.2.3c), On the other hand, Methods A and B had the highest GH score (0.22 and 0.45) among primary methods in TK (Figure 3.2.3a) but the best combination was the Method BD among 10 pair combinations. Figure 3.2.4 shows that the variation between Methods B and D is larger than the rank/score variation of Methods A and B.

These experimental results using the BD model implied that the variation of rank/score graph might be useful to improve the screening accuracy in both VS and IR. This concept is supported by observations of similar phenomenon occurring in ER agonists (Figure 3.2.3d). Specifically, Methods B and E that had the highest GH scores but their rank/score variation is smaller than the variance of Methods B and D. The performance of Method BD was better than Method BE for ER agonists.

Figure 3.2.5 and Figure 3.2.6 and Table 3.2.5 are the results of the Algorithm (Figure 3.2.1) when $g = 2$ where pairing combinations were considered and $R/S_{var}(f_A, f_B)$ was used to calculate the bi-diversity of methods A and B. Figure 3.2.5a shows the relationship between the GH-score improvement and the variance (R/S_{var} , Equation 12) of rank/score graphs of 10 pairing combinations for each target protein. Figure 3.2.5b indicates the relationship between the GH-score improvement and the relative performance measurement (P_l/P_h , Equation 3.2.13).

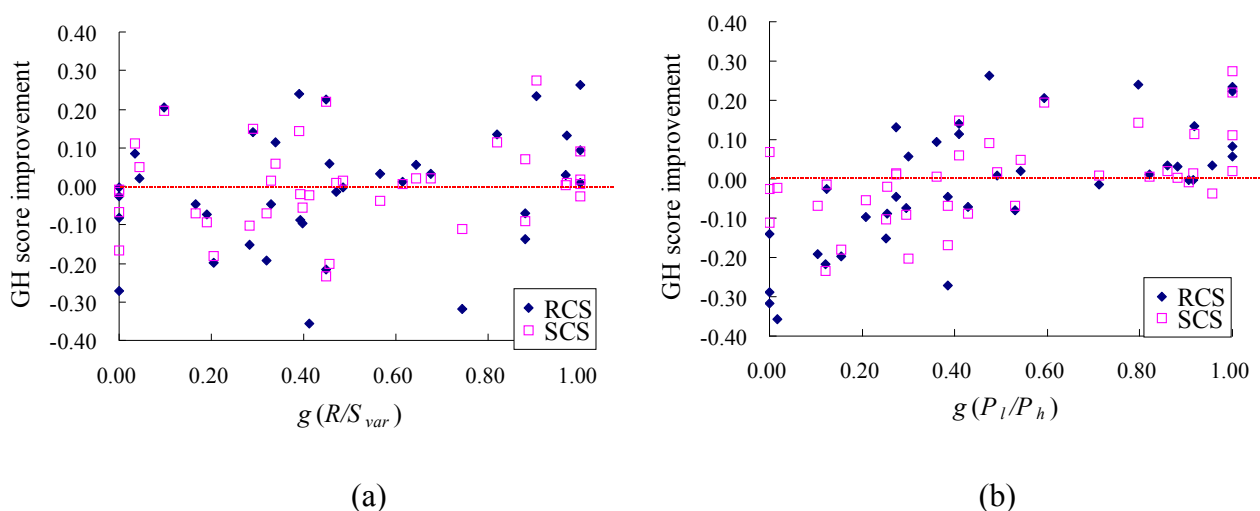


Figure 3.2.5. The relationships between the GH-score improvement with (a) normalized value of variance of rank/score graph (Equation 3.2.12) and (b) normalized value of P_l/P_h of 40 pairing combinations of five methods for four virtual screening targets (defined in Table 3.2.1).

Figure 3.2.6 and Table 3.2.5 show that a pairing combination is able to improve the performance

if the normalized R/Svar and PI/Ph of a combining method are considered simultaneously and their values are more than 0.5. These results implied that consensus scoring yielded improved screening accuracy if the multiple scoring functions involved have high performance and their rank/score variation was large. A similar phenomenon was also found in data fusion in IR^{106,107}. The CSindex (Equation 3.2.14) is used to integrate these two criteria (R/Svar and PI/Ph). Figure 3.2.7 shows the relationship between GH score improvement and the CSindex of 10 pairing combinations for each target protein. A pairing combination often improves screening accuracies when its CSindex was more than 1.5.

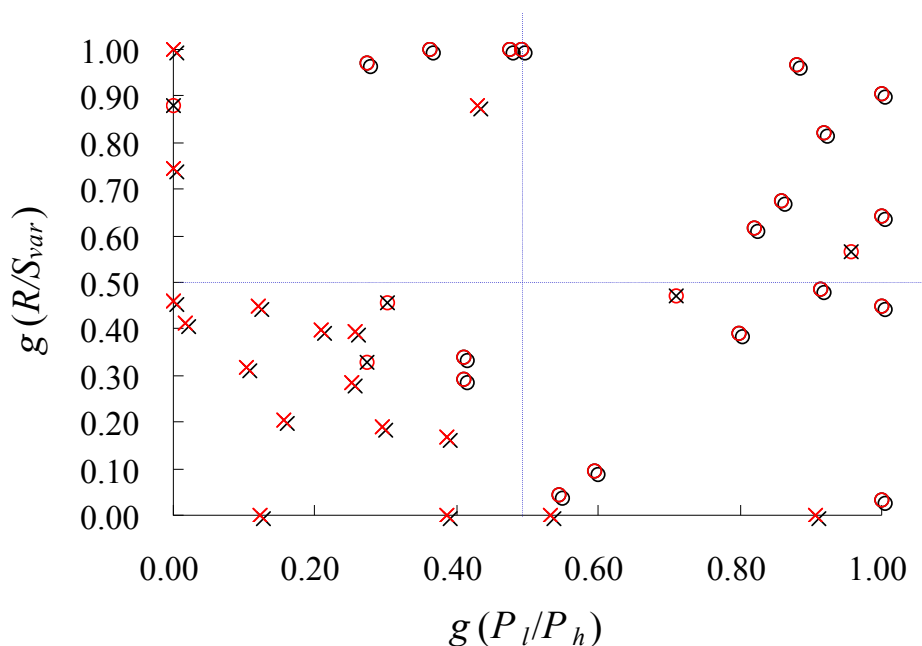


Figure 3.2.6. The GH-score improvements with normalized variances of rank/score graphs (R/Svar) and normalized relative performance measurement (PI/ Ph) of 40 RCS and SCS pairing combinations of five methods for four virtual screening targets. The positive and negative GH-score improvements are denoted with circle and cross, respectively.

Consensus scoring is a popular strategy for solving the scoring inaccuracy problem in virtual screening. In this study, our consensus scoring methods consist of rank combinations (Figure 3.2.2 and Table 3.2.3) and score combinations (Figure 3.2.3 and Table 3.2.4) on five scoring functions related to two docking algorithms. When we associated individual ranks into different combinations, the accuracy (in terms of average false positive rates and average GH scores) of some of these combinations was better than each individual. From this study of data fusion on screening four cases of receptor targets, it demonstrates that a fusion method is able to improve the screening accuracy in VS only when (a) each of the individual scoring function has a relatively good performance (P_l/P_h) and (b) the scoring characteristics of each of the scoring functions are quite different (R/S_{var}). The observations of RCS and SCS are summarized as follows:

- a) Figure 3.2.6 shows that combining multiple scoring functions improves enrichment of true positives only if both $g(P_l/P_h) > 0.5$ and $g(R/S_{var}) > 0.5$. These two prediction indicators can be combined into a single indicator $CS_{index} > 1.5$ (Figure 3.2.7). For example, in ER-antagonists receptor, the GH scores of Methods B (0.83) and D (0.70) (Table 3.2.1) are the best and the Method BD (0.86) is the best among 31 combinations (Table 3.2.3). The CS_{index} of the Method BD is 1.68 (Table 3.2.5).

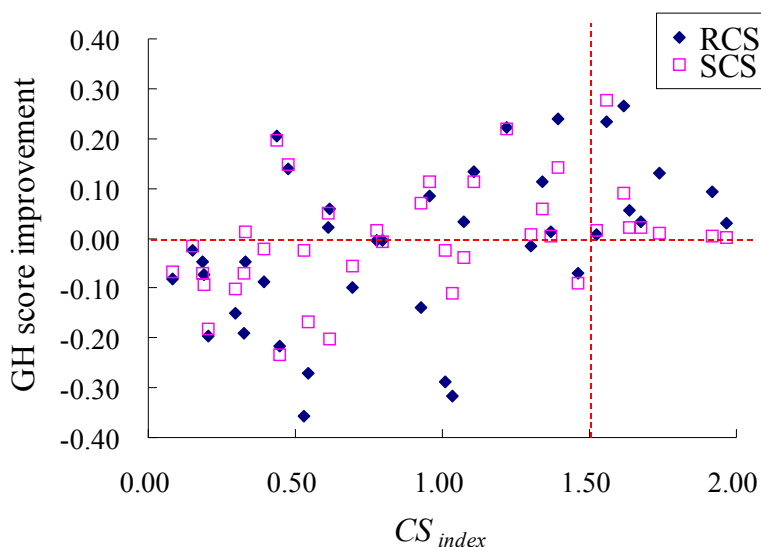


Figure 3.2.7. The relationships between the GH-score improvement with the CS_{index} (Equation 3.2.14) of 40 pairing combinations of five methods for four virtual screening targets.

- b) These statistics revealed that the accuracy of consensus scoring was improved by increasing scoring methods for both RCS and SCS (Tables 3.2.3 and 3.2.4). However, the combination of all scoring methods did not display the best possible performance observed. For RCS, The performance of 2-combination or 3-combination methods outperformed 4-combination or 5-combination methods.
- c) The variance, R/S_{var} , of a pair of rank/score graphs is a useful index to improve the screening accuracy for combining two individual methods when the individual scoring functions are quite different (or complementary, e.g., normalized $R/S_{var} > 0.5$). This criterion is useful and important because very often the performance of the individual scoring function is not known or cannot be evaluated at the juncture. Our approach can be used in different situations, whether it is running a truly blind screen, a combination screen coupling a blinded set with partial analysis and subsequent use of previous hits as a training set, or a screen with a true training set. Our approach also reveals that approaches that yield the best average GH score/FP (i.e., SCS) which are relevant for screens without training sets, are different from those approaches that optimize individual GH score (i.e., RCS), which are applicable when a training set is available.
- d) The best GH scores of RCS are consistently superior to SCS for these four target proteins

(Figure 3.2.3). Table 3.2.3 shows that RCS yielded its best individual GH scores: 0.58 (Method BD in TK), 0.85 (Method ABD in DHFR), 0.86 (Methods BD and ABD in ER), and 0.74 (Method BD in ERA). Table 3.2.4 shows that SCS obtained its best GH scores: 0.52 (Method ABCD in TK), 0.84 (Method ABDE in DHFR), 0.86 (Methods BDE in ER), and 0.72 (Method BCDE in ERA).

- e) The best average GH scores and best average FP rates of SCS are significantly superior to RCS on all target proteins (Table 3.2.3 and Table 3.2.4). For example, in TK, the best average GH scores are 0.41 (SCS) and 0.29 (RCS), and the best average FP rates are 2.01% (SCS) and 4.04% (RCS). In ER antagonists, the best average GH scores are 0.79 (SCS) and 0.69 (RCS), and the best average FP rates are 0.44% (SCS) and 1.39% (RCS).
- f) For RCS methods, the moderate number of scoring functions, two or three, are the best and sufficient for the purpose of consensus scoring (Figure 3.2.2). In contrast, the number of combining methods is three or four to achieve the best performance for SCS methods (Figure 3.2.3). This phenomenon was also found in data fusion in IR and was consistent with the previous findings in CS¹⁰⁹.
- g) When combining methods with highly differential performance, Figure 3.2.2 (FP rates) and 3.2.2 (GH scores) show that SCS works better than RCS. For example, the combinations of BE (in TK) and ABE (in DHFR) where Methods B and E are the best and the worst, respectively; among five primary methods. In ER and ERA, the combinations of BC (ER) and BCE (ERA) have the similar results.

3.2.4. Conclusions

It has been shown that consensus scoring improves VS and has become a robust scoring method because each individual scoring function has strengths and weakness with respect to docking algorithms, receptor targets, and the database sets. It appears that on average the consensus scoring does perform better than the average performance of the individual scoring methods, but does not perform better than the best of the individual scoring function. In our experiment on the four receptors TK, DHFR, ER, and ERA, the two docking algorithms we used (GEMDOCK and GOLD) have been shown to be very good. Although performances (measured as GH score and false positive rate) of each individual scoring function do vary within each of and among the receptor targets, interesting patterns do stand out where we showed that combinations of two scoring function leads to significant improvement on average GH score and average FP rate.

We summarize and state the two consensus scoring criteria, which would serve as two predictive variables for improving enrichment in VS: a consensus scoring which combines multiple scoring functions should only be used when (a) the scoring functions involved have high performance and (b) the scoring characteristic of each of the individual scoring functions are quite different. These two CS criteria also work for different performance between SCS and RCS. It has been reported that, on average, score combination is more effective than rank combination. However, we have demonstrated that in a majority of cases under the two CS criteria, rank combination does perform better or as good as score combination. This is analogous to the results reported in IR^{102; 106; 107}. Our second criterion calculates the rank/score function of each scoring function and then computes the differences between the rank/score functions of the scoring methods involved. Our second criterion

does not involve performance evaluation of the combined methods. This criterion is useful because very often the performance of individual scoring functions is not known or cannot be evaluated. We believe that our rank-based and score-based consensus scoring (RCS and SCS) procedure and consensus criteria for improving the enrichment in VS should be useful to researchers and practitioners in virtual screening.

Our work has provided a framework to study consensus scoring criteria and a procedure (the Algorithm) for both rank-based and score-based consensus scoring to improve the hit rates, FP rates, the enrichment, and the GH score. The procedure is computationally efficient, able to adapt to different situations, and scalable to a large number of compounds and a greater number of combinations. Moreover, we have shown the power of two-combinations (pairing combinations) and used the rank/score graph to assess the bi-diversity between two scoring methods used. Our current work represents the first of a series of investigations to explore consensus scoring criteria for improving enrichment in VS. It also engenders a whole school of issues and directions worthy of further study:

- 1) We will study the extension to three and higher number of combinations of scoring functions using the rank/score graph variation (R/S_{var}) as a diversity measurement for the scoring methods involved.
- 2) In this paper, we use rank/score function f_A as the scoring characteristic for the scoring method A. Then we use the variation on the rank/score function (R/S_{var}) to characterize the scoring diversity between two scoring methods A and B. Other parameters such as the difference between the score functions S_A and S_B and the difference between the rank functions R_A and R_B can be also used to distinguish the scoring diversity. The rank/score graphs (Figure 3.2.4) have provided a clear visualization for characterizing the scoring diversity between individual scoring functions.
- 3) In our combination (RCS and SCS) of scoring functions, we use averages to compute the scores for the rank and score combination. Combination using different weighting schemes can be used. Hsu and Palumbo¹¹⁵ presented work on combination of two scoring methods using weighty scheme with a step of one tenth as a proportion.
- 4) The two docking algorithms we used, GEMDOCK and GOLD with five scoring functions, were superior to other evolutionary algorithms on two receptor targets (Table 3.2.2). A more diverse set of docking methods, scoring functions, and receptor targets will be studied to determine the advantages of SCS and RCS for VS.

Chapter 4 Quantitative Structure Activity Relationships

For QSAR analysis, we developed a QSAR methodology associating molecular docking and feature selection with PLS, named GEMPLS. GEMPLS served as feature selection and model building in QSAR analysis. Potential features for contributing inhibition would be selected by evolutionary strategy and built regression by PLS. Due to the low correlation of binding affinity and current scoring functions, we also analyzed the factors, which affect binding affinities of protein-ligand complexes, from five dimensions including protein-ligand interactions, protein properties, structure and chemical-physical descriptors of ligands, metal-ligand bonding, and solvent effects. The correlation between predicted binding affinities and experimental values are 0.612 and 0.601 on the training set (891 protein-ligand complexes) and on testing set (98 protein-ligand complexes), respectively. These seven factors will be added into our QSAR method (termed GEMQSAR) to improve the prediction abilities and accuracies. The works in this part published one conference paper and one international poster paper

Conferences Papers:

- K-C Hsu, Y-F Chen, and J-M Yang*, "Binding affinity analysis of protein-ligand complexes," 2nd International Conference on Bioinformatics and Biomedical Engineering, pp. 167-171, 2008.

Posters

- Y.-F. Chen, L.-J. Chang, J.-M. Yang*, "Integrating GEMDOCK with GEM-PLS and GEM-kNN for QSAR modeling of huAChE and AGHO," in 15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 6th European Conference on Computational Biology (ECCB), Vienna, Austria, 2007.

4.1 Analysis of Protein-ligand Complexes to Predict Binding Affinity

4.1.1 Introduction

One of the important issues in computer-aided drug discovery is to predict the binding affinity between a compound and a target protein. Compounds with high binding affinities to a target protein are often considered as the potential inhibitors that slow or block physiological, chemical, or enzymatic actions of this protein. Binding affinities are usually determined by some experiments such as microcalorimetry^{116; 117}, ELISA assays¹¹⁸, NMR spectrometry¹¹⁹, and surface plasmon resonance¹²⁰. The high-throughput screening method¹²¹ is used to detect the binding affinities of a large number of compounds to identify lead compounds for a target protein. However, these experiment methods are often labor-intensive, time-consuming, and expensive. Therefore, many computational methods have been proposed to discover the lead compounds by predicting the binding affinities between compounds and the target proteins.

The structure-based virtual screening is one of the computational methods to identify lead compounds for a specific receptor from thousands of compounds. Each compound in the database is docked into a target protein and assigned a binding score for measuring binding affinities based on scoring functions. The scoring methods for virtual screening 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 compounds with higher scores are considered as potential compounds and selected for biological assay which is the measurement of the pharmacological activity of a compound. The scoring functions that calculate the binding free energy mainly include knowledge-based⁵⁷, physics-based⁵⁸, and empirical⁵⁹ scoring functions. Physics-based energy functions are based on physical mechanisms and often derived from *ab initio* quantum-mechanical calculations according to the principles of physics. One advantage of physics-based energy functions is the lucid physical meaning of each individual term, but physics-based energy functions often requires the high-computation cost and their energy landscapes are often very rugged. In general, an empirical scoring function has simplified energy terms based on physical mechanisms. Knowledge-based scoring functions are derived from energy-like functions by considering the distributions of inter-atomic distances in a set of crystal structures of protein–ligand complexes.

In practice, the performance of a scoring function is limited by our incomplete understanding of the complex issues involved in chemical interactions. 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 virtual screening. Some knowledge-base scorings, such as X-SCORE¹²², ChemScore¹¹³, DrugScore⁵⁷, and PLD¹²³, applied regression techniques to predict binding affinities by deriving from a set of protein-ligand complexes with experimental binding affinities. The deficiency of current scoring functions is mainly due to the inadequate descriptions of the interactions between ligands and proteins²⁵. For example, most scoring functions use simple model to handle metal-ligand and water-ligand interactions and consider hydrogen-bonding interactions as the same even if some interactions are essential for chemical reactions.

Here, we address these issues by deriving 87 descriptors from 891 protein-ligand complexes selected from PDB according to five dimensions, including protein-ligand interactions, structural and physicochemical descriptors of a compound, protein binding site evolution, metal-ligand interactions, and water effects. Based on these 87 descriptors, we applied a stepwise regression method to select top five descriptors, which are highly correlated to experimental affinities, for developing the GemAffinity. The GemAffinity was then used to predict binding affinities on an independent set and it outperforms 12 comparative scoring functions on this set. For post-screening analysis, we applied The GemAffinity to score the docked complexes generated by GEMDOCK², which is well-developed molecular docking tool, on four targets. The GemAffinity is able to enrich the

prediction accuracy of GEMDOCK on these four targets.

4.1.2 Materials and Methods

Overview

Figure 4.1.1 shows the overview of deriving the descriptors and developing GemAffinity for predicting binding affinities. With the aim to identify descriptors for the binding process in the molecular recognition, we first selected the 989 protein-ligand complexes with experimental binding affinities from PDB. We derived 87 descriptors from protein functions and evolution, ligand structures and physicochemical aspects, and molecular interactions. Furthermore, a stepwise regression was applied to find the relationship between descriptors and binding affinities. Some key selected descriptors, which significantly reflect the experimental binding affinities, are used to develop GemAffinity for the prediction of binding affinities. Finally, the GemAffinity was evaluated on an independent testing set and post-screening analysis on four target proteins.

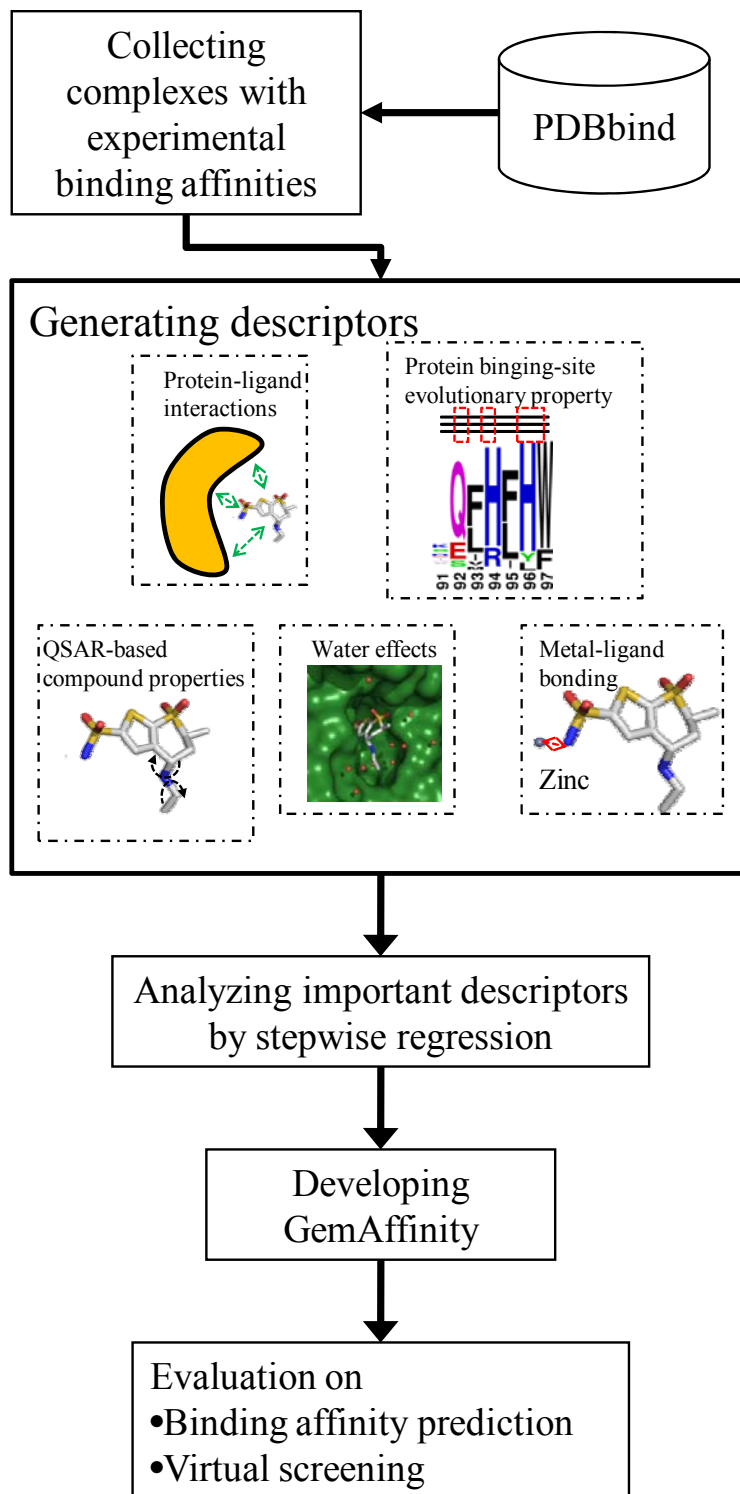


Figure 4.1.1. The overview of analyzing protein-ligand complexes to develop the GemAffinity for predicting binding affinities.

Protein-ligand complex dataset

We collected 1,091 protein-ligand complexes with experimental binding affinities from the PDBbind¹²⁴. Among these 1091 complexes, 989 complexes are selected by excluding 80 complexes (missing atoms in binding sites) and 22 complexes (no HETATM records of water atoms). The descriptors of the excluded complexes are incorrect due to the miss data and may bias the results. The remaining complexes were randomly divided into a training set (891 complexes) and a testing set (98 complexes). For each complex, we used the negative logarithm of K_d or K_i (i.e. $-\log K_d$ or $-\log K_i$) as its binding affinity.

Virtual screening dataset

A widely used approach to test scoring functions for the virtual screening is to rank the docked poses generated by docking tools to identify inhibitors (substrates) for a target receptor from thousands of compounds. We selected four target proteins with ~1000 compounds to evaluate the GemAffinity performance and to compare it with other methods. These four target proteins include thymidine kinase (TK), estrogen receptor antagonist (ER), estrogen receptor agonist (ERA), and human carbonic anhydrase II (HCAII). The receptors for these screens cover different receptor types and therefore provide a reasonable test of scoring functions. Four complexes of the target proteins were selected for virtual screening from the PDB: TK complex (PDB code: 1kim¹⁴), ER-antagonist complex (PDB code: 3ert⁷¹), ER-agonist complex (PDB code: 1gwr¹²⁵), and HCAII complex (PDB code: 1cil¹²⁶). These complexes were reasonable choices because their ligand-binding cavities are wide enough to accommodate a broad variety of ligands and therefore did not require binding site modifications.

For each target, the screening compound data set proposed by Bissantz *et al.*¹²⁷ consists of 990 random compounds. In addition, the set includes 10 known active compounds² for TK, ER, and ERA; and 20 known active compounds^{126; 128; 129; 130; 131} for HCAII target. The active compound set of each target protein, target proteins, and 990 random compounds are available on the Web at <http://gemdock.life.nctu.edu.tw/dock/download.php>.

Descriptors for binding affinity

We derived 87 descriptors from 891 complexes described in the protein-ligand complex dataset according to five dimensions: protein-ligand interactions, protein binding-site conserved properties, QSAR-based (quantitative structure activity relationship) compound properties, water effects, and metal-ligand bonding. These dimensions are described as follows:

Protein-ligand interactions: We considered three protein-ligand interaction types, including van der Waals interactions, electrostatic interactions, and hydrogen-bonding interactions. The van der Waals force consists of a piece-wise linear potential (PLP), generated by GEMDOCK ² and Lennard-Jones potential. The hydrogen-bonding force includes a PLP ² and a non-linear potential generated by AutoDock ⁴. The electrostatic force was measured by the GEMDOCK. In addition, we also regarded the numbers of protein-ligand interactions within different cutoffs for each force type. For van der Waals force, the distance of an atomic pair between a protein and its ligand was divided into 10 cutoffs which are 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0, 6.5, and 7.0 Å. For hydrogen-bonding and electrostatic interactions, the distance was divided into 10 cutoffs, including 2.5, 2.6, 2.7, 2.8, 2.9, 3.0, 3.1, 3.2, 3.3, and 3.4 Å. In total, there are 35 descriptors for protein-ligand interactions.

Protein binding-site evolutionary property: Conserved residues in a binding site of a protein often highly correlated to the biological functions. For example, the catalytic charged residues of a protein are often conserved across many species. They polarize substrates to stabilize transition states ¹³². If these conserved residues mutate, the protein may lose their functions or execute different biological functions. Based on these observations, we used the number of highly conserved residues, which form hydrogen bonds between a protein and its ligand, as the descriptor of protein binding-site evolutionary property in the binding site.

Here, we developed a conserved score method, derived from our previous work on protein-protein interactions ¹³³, to measure the conservation degree of a residue in the binding site. The amino acid sequence of a protein-ligand complex was first subjected to PSI-BLAST ⁴² by searching on UniRef90 ¹³⁴. The *E*-value cutoff of PSI-BLAST was set to 10^{-5} and the number of the iterations was 3. Based on the position specific scoring matrix (PSSM) automatically generated by PSI-BLAST, the conserved score of each residue was defined as

$$C_i = M_{ir} - K_{rr} \quad (4.1.1)$$

where M_{ir} is the value in the PSSM for the residue type r at the position i , and K_{rr} is the diagonal value of BLOSUM62¹³⁵ for the residue type r . The residue-conserved descriptor was defined as

$$R_i = 1 \quad \text{if } C_i > 0 \text{ and number of hydrogen bonds} \geq 1$$

$$H = \sum_i R_i$$

where R_i is 1 if C_i is greater than 0 and at least one hydrogen bond is formed at the position i between protein and its bound ligand. H is the number of highly conserved residues with hydrogen bonds of a protein. Among 891 complexes in the training set, 571 complexes consist of highly conserved residues forming hydrogen bonds between proteins and their bound ligands.

We used HCAII (PDB code: 1cil ¹²⁶) as an example to describe the steps of deriving the highly conserved residues forming hydrogen bonds with its bound ligand (Figure 4.1.2). First, the multiple

sequence alignment (MSA) of the query sequence and the homologous sequences are derived by PSI-BLAST (Figure 4.1.2a). Based on the MSA and Equation 4.1.1, conserved score of each residue can be calculated. Here, the residue is considered as a highly conserved residue (e.g. GLN92, HIS94, HIS96, HIS119, and THR199) when its score is greater than or equal to 1.0 (Figure 4.1.2b). Finally, we count the number of highly conserved residues which form hydrogen bonds with bound ligand. In this example, the number of highly conserved residues with hydrogen bonds is 5.

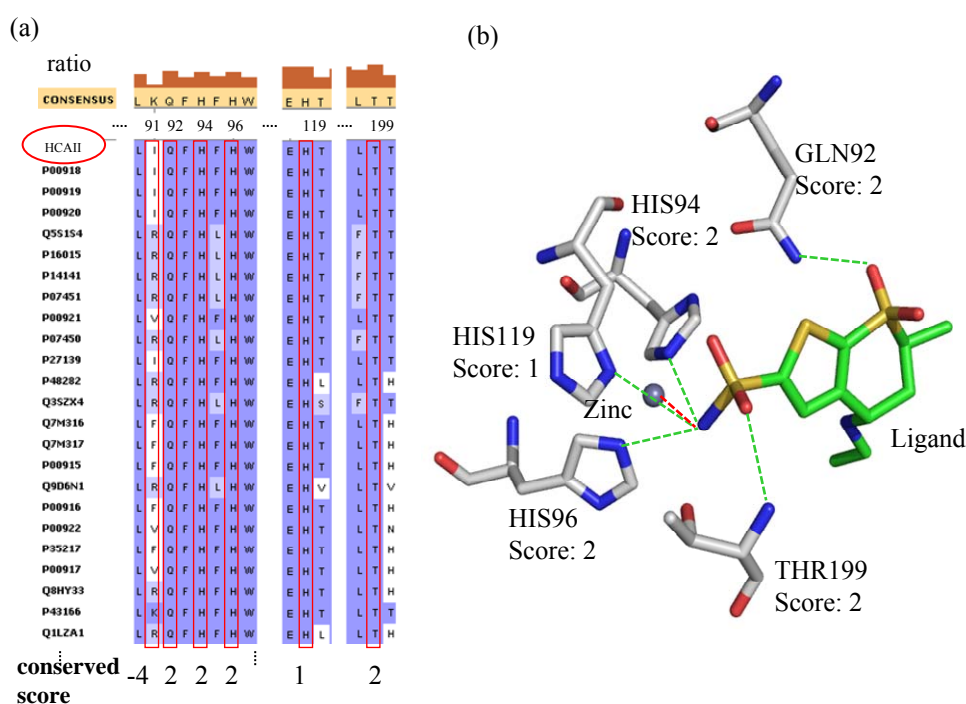


Figure 4.1.2. An example of describing metal-ligand bonds and highly conserved residues forming hydrogen bonds. (a) The multiple sequence alignment of the query sequence, human carbonic anhydrase II (HCAII, PDB code: 1cil), and its homologous sequences by PSI-BLAST. The E -value cutoff of PSI-BLAST was set to 10^{-5} and the number of the iterations was 3. The conserved scores of part residues are given in the bottom. (b) The residues that are highly conserved and form hydrogen bonds with the ligand are GLN92, HIS94, HIS96, HIS119, and THR199. In this example, the number of highly conserved residues with hydrogen bonds is 5. In addition, the ligand also forms a metal-ligand bond (red line) with the Zinc (the bond length is 1.96 angstrom). This metal-ligand bond is one of the primary binding interactions in HCAII.

QSAR-based compound properties: The QSAR methods^{53; 136} demonstrated that physical and biological properties of a ligand are useful for predicting binding affinities of a group similar compounds. We used the QSAR module of the Cerius2 to generate 26 compound descriptors, such as spatial, conformational, electronic, structural, and thermodynamic terms.

Water effects: In general, a compound should be solvated before it binds to its receptor. The water effects play a crucial role in mediating the interactions between proteins and their bound ligands¹³⁷. The interactions between the bound ligand and structural waters in a complex were used to measure the water effects. We calculated the number of structural water molecules which are within a specific distance cutoff around a ligand. The distance bins are classified into 12 cutoffs, including 2.5, 2.6, 2.7, 2.8, 2.9, 3.0, 3.1, 3.2, 3.3, 3.4, 3.5, and 3.6 Å.

Metal-ligand bonding: The metal atoms in an active site often play the key role for stabilizing ligands and reaction catalysis. However, many scoring functions consider the interaction between metal atoms and ligands as a simple hydrogen-bonding interaction. In this study, we separated metal-ligand bonding interactions from hydrogen-bonding interactions by considering the metal atoms which are within a specific distance from a ligand. In the 891 complexes, 99 and 152 complexes have metal atoms which are within 2.2 Å and 3.6 Å distance from ligand, respectively. Figure 4.1.3 shows that most of metal-ligand bonding distances were less than 2.8 angstrom. It meant that the metal-ligand bonding distances were almost shorter than the normal hydrogen bonding distances. The metal-ligand bonding interaction is 1 if a metal atom is less than a specific distance; conversely, the value is 0. Here, we considered 13 distance cutoffs which are 2.2, 2.3, 2.4, 2.5, 2.6, 2.7, 2.8, 2.9, 3.0, 3.1, 3.2, 3.3, and 3.4 Å. In the HCAII, the ligand forms a metal-ligand bond (red line) with the Zinc and the bond length is 1.96 (angstrom)Å (Figure 4.1.2b). This metal-ligand bond is one of the primary binding interactions in HCAII¹²⁶.

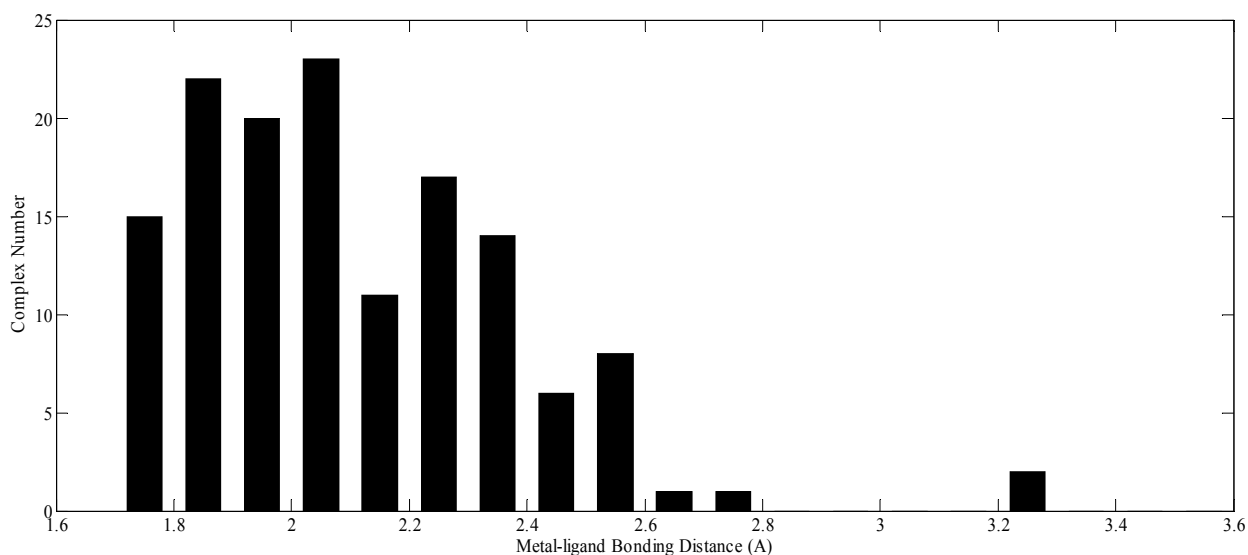


Figure 4.1.3. The distribution of metal-ligand bonding distances in the 891 complexes in the training set.

Stepwise regression analysis

The stepwise regression¹³⁸ method was used to select descriptors which are highly correlated to protein-ligand binding affinities one by one in the order of descending importance. Here, we employed the simple stepwise regression method to develop the scoring methods. This method is often used to avoid the ill effects (such as over fitting and the loss of biological/physical/chemical meaning.) of the machine learning approaches, such as the support vector machines, genetic algorithms, and neural networks. This model first selected the descriptor with the highest Pearson's correlation coefficient into the predicted model. The model sequentially added the other descriptor which is able to improve the correlation between predicted binding affinities derived by the selected descriptors and experimental binding affinities. The process was stopped if the improved correlation of the added one was less than 0.005 to reduce ill-effect of the overfitting data. Finally, the model selects five descriptors for the GemAffinity to predict binding affinities (Table 4.1.1).

Table 4.1.1. The selected descriptors in the new scoring function

Selected order	Name	r ^a	Descriptors
1	f_{vdw}	0.497	Sum of Lennard-Jones potential between a protein and a ligand
2	V_{Metal}	0.544	The distance between metal ions and a ligand is less than 2.2 Å
3	N_{rotBon}	0.579	Number of rotatable bonds of a ligand
4	N_{water}	0.594	Number of structural waters which are near to a ligand within 3.6 Å
5	N_{conHB}	0.599	Number of highly conserved residues forming hydrogen bonds between proteins and ligands

^a the correlation between the predicted binding affinities by stepwise regression models and experimental binding affinities.

Application on virtual screening

A scoring method should be able to effectively discriminate between correct binding states and non-native docked conformations during the molecular docking phase. For virtual screening, it is important for scoring functions to identify a small number of active compounds from hundreds of thousands of non-active compounds during the post-docking analysis. Here, we evaluated the GemAffinity to score docked protein-ligand complexes generated by the GEMDOCK on four targets (i.e. TK, ER, ERA, and HCAII). Our previous works show that the GEMDOCK is comparative to some approaches (e.g. GOLD⁵ and FlexX¹³⁹) for molecular docking² and virtual screening^{9;10}.

In this study, the GEMDOCK parameters in the flexible docking included the initial step sizes ($\sigma=0.8$ and $\psi=0.2$), family competition length ($L = 2$), population size ($N = 200$), and recombination probability ($p_c = 0.3$). For each ligand screening, GEMDOCK optimization stopped either when the convergence was below a certain threshold value or the iterations exceeded the maximal preset value of 60. Therefore, the GEMDOCK generated 800 solutions in one generation and terminated after it exhausted 48,000 solutions for each docked ligand. Standard parameters of the GOLD⁵ program and its scoring function (i.e. GoldScore) were used in this study. For each of the 10 genetic algorithm (GA) runs, a maximum number of 10,000 operations were performed on a population of 50 individuals. The maximum distance between hydrogen donors and fitting points was set to 2 Å, and nonbonded van der Waals (vdW) energy was cut off at 4.0 Å. To further speed up the calculation, the GA docking was stopped when the top three solutions were within 1.5 Å rmsd of each other. These parameters are chosen according to the standard default settings recommended by the authors for virtual screening.

4.1.3 Results and Discussion

Scoring function

The stepwise regression method selects top five descriptors for predicting binding affinities (Table 4.1.1). These descriptors are the sum of Lennard-Jones potential (f_{vdw}), the number of metal-ligand bonds (V_{Metal}), the number of rotatable bonds of a ligand ($N_{rotBond}$), the number of structural waters which are near to a ligand within 3.6 Å (N_{water}), and the number of highly conserved residues forming hydrogen bonds between ligands and proteins (N_{conHB}). The coefficient of each descriptor and the GemAffinity is given as

$$-\log(K_{d,pred}) = -0.072f_{vdw} + 1.168V_{Metal} - 0.078N_{rotBond} - 0.039N_{water} + 0.102N_{conHB} + 3.430$$

The selected orders of these five descriptors are correlated to the accuracy orders (e.g. 0.497, 0.544, and 0.579 shown in Table 4.1.1), which are the Pearson's correlation coefficient between predicted binding affinities and experimental binding affinities. For example, Pearson's correlation coefficients are 0.497 and 0.544 using only the first descriptor (i.e. f_{vdw}) and the first two descriptors (i.e. f_{vdw} and V_{Metal}). Please note the values of this scoring function and $-\log(K_{d,pred})$ are positive to present a high binding affinity. In general, f_{vdw} is negative and its coefficient (i.e. -0.072) is negative; therefore, the value $-0.072f_{vdw}$ is positive and positively contributes to binding affinities of compounds. V_{Metal} and N_{conHB} are positive and their respective coefficients are also positive values which positively contribute to the binding affinities. On the other hand, N_{water} and $N_{rotBond}$ are positive but their respective coefficients are negative. These two terms negatively contribute to binding affinities.

Selected descriptors

The Lennard Jones 12-6 potential was selected first and its Pearson's correlation coefficient is 0.497 (Table 4.1.1). This result shows that the complementary shape between proteins and their ligands is critical for protein-ligand binding affinities. In addition, Pearson's correlation coefficient between van der Waals forces (i.e. PLP) of the GEMDOCK and the experimental binding affinities is 0.48 on 891 complexes in the training set. Lennard Jones potential function has various parameters of atomic pairs, and the GEMDOCK treats all atomic pairs as the same. This difference may be one of reasons why the performance of Lennard-Jones potential function is slightly better than the one of the GEMDOCK.

The energy function, using the PLP potential to soften the repulsive term of Lennard-Jones potential, of GEMDOCK has a good performance in flexible protein-ligand docking. The short range repulsive interactions (e.g. Lennard-Jones potential) tend to infinity at low interatomic separation leading to rough energy surfaces with high energy barriers. A soft scoring function (e.g. PLP potential used in GEMDOCK) has been applied for softening the repulsive intermolecular potential to decrease the strong sensitivity of interaction energies to local conformation changes. Generally, a soft scoring function has the benefit of being computationally efficient, conversely, it may increase the number of false near-native solutions (structures). The tradeoff of its advantages and limitations can be optimized.

The metal-ligand bonding is the second selected term. The metal-ligand bonding is a metal-ligand interaction with the distance cutoff less than 2.2 angstrom in this study (Figure 4.1.3). A metal-ligand bond is often a strong bonding interaction and is able to stabilize the interactive conformations between a protein and its compound (Figure 4.1.4). In our GemAffinity, we highlighted and considered the metal-ligand bond as a special force because its distance cutoff (i.e. 2.2 angstrom) is much shorter than the general hydrogen-bonding distance (i.e. 2.8 angstrom). This scoring function discriminates between metal-ligand interactions and hydrogen bonds to avoid energy rises drastically when the distance of a metal-ligand bond is shorter than 2.2 angstrom. Figure 4.1.4 shows the effect of the metal-ligand bonds in the cytidine deaminase protein (PDB code 1ctu and 1ctt¹⁴⁰). The binding affinity of the complex with the metal-ligand bond (1ctu) between zinc²⁺ and the bound ligand is 11.92. On the other hand, the binding affinity of the complex (1ctt) without metal-ligand bond (1ctt) is 4.52. The only change of the binding affinities of these two complexes is due to the loss of the metal-ligand bond¹⁴⁰.

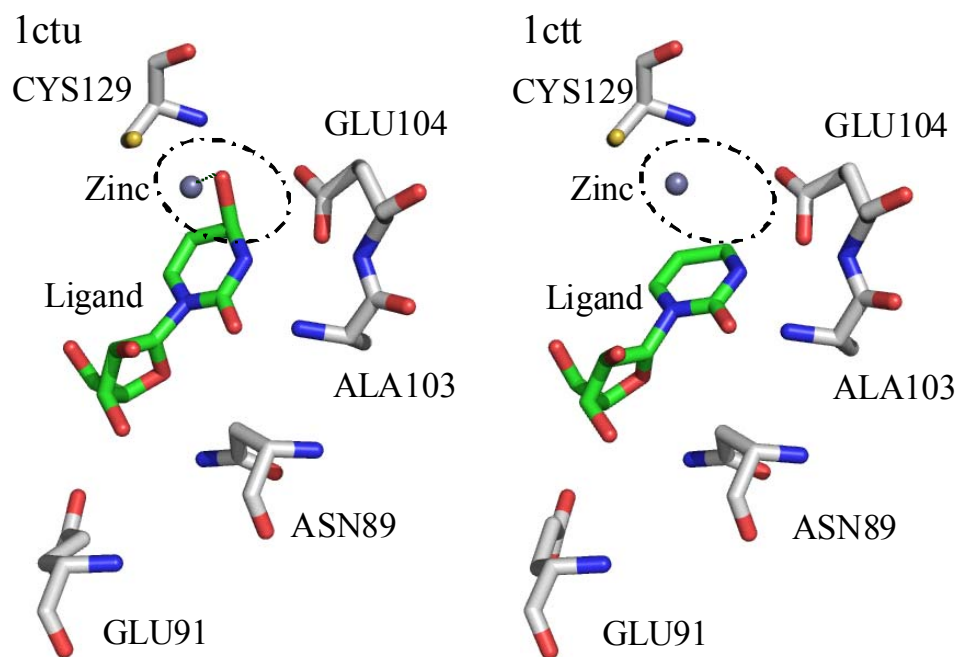


Figure 4.1.4. The effect of the metal-ligand bonds. In the cytidine deaminase protein (PDB code 1ctu and 1ctt), the binding affinities of the complexes with the metal-ligand bond (1ctu) and without metal-ligand bond (1ctt) are 11.92 and 4.52, respectively. The change of the binding affinities of these two complexes is due to the loss of the metal-ligand bond.

The third selected descriptor was the number of rotatable bonds of a ligand. This descriptor is important for molecular mechanics because the freedom of rotatable bonds become lower during the binding process. The number of rotatable bonds gave a measure of the unfavorable torsional entropy loss upon protein-ligand binding affinities. The fourth selected descriptor (i.e. the number of structural water molecules within 3.6 Å distance from the bound ligand) was the water effects. A large amount of water molecules around a ligand within 3.6 angstrom imply that the large volume of the ligand is exposed to the solvent. A ligand buried deeply inside a protein could have larger binding affinities than those were bounded on the protein surface.

The final selected term for the GemAffinity is the number of highly conserved residues with a hydrogen bond (bonds). A highly conserved residue is often highly responsible to maintain biological functions of a protein because the hydrogen bonds on conserved residues were often responsible for stabilizing the ligands and catalyzing the reactions. We used the number of highly conserved residues with hydrogen bonds to measure the binding affinities. The descriptor is generated according to the properties of each binding site, and it distinguishes specific interactions from other hydrogen bonds. The Pearson's correlation coefficient between the number of highly

conserved residues with hydrogen bonds and the experimental binding affinities is 0.15. Conversely, the coefficient between the numbers of hydrogen bonds and the experimental affinities is 0.12. It is reasonable to discriminate the hydrogen-bonds with highly conserved residues from others.

A hydrogen bond formed between highly conserved residue and the ligand often highly influences the binding affinity (Figure 4.1.5). For example, the binding affinity of the beta-glucosidase protein (PDB code 1uz1¹⁴¹ and 2j77) from *thermotoga maritima* drops from 6.89 to 4.89 due to the loss of a hydrogen bonding between the ligand and the Glu166. The Glu166 is a conserved residue and its conserved score is 2 using Equation 4.1.1. Conversely, a hydrogen bond formed between non-conserved residues and the ligand may lightly influence the binding affinity (Figure 4.1.6). The binding affinity of the oligo-peptide binding protein (PDB code 1b58¹⁴² and 1b3h¹⁴³) slightly reduced from 6.58 (1b58) to 6.21 (1b3h) even if two hydrogen bonds between the ligand and two residues Asn436 and Tyr269. These two residues are not conserved and their conserved scores are -3 and -6 based on Equation 4.1.1.

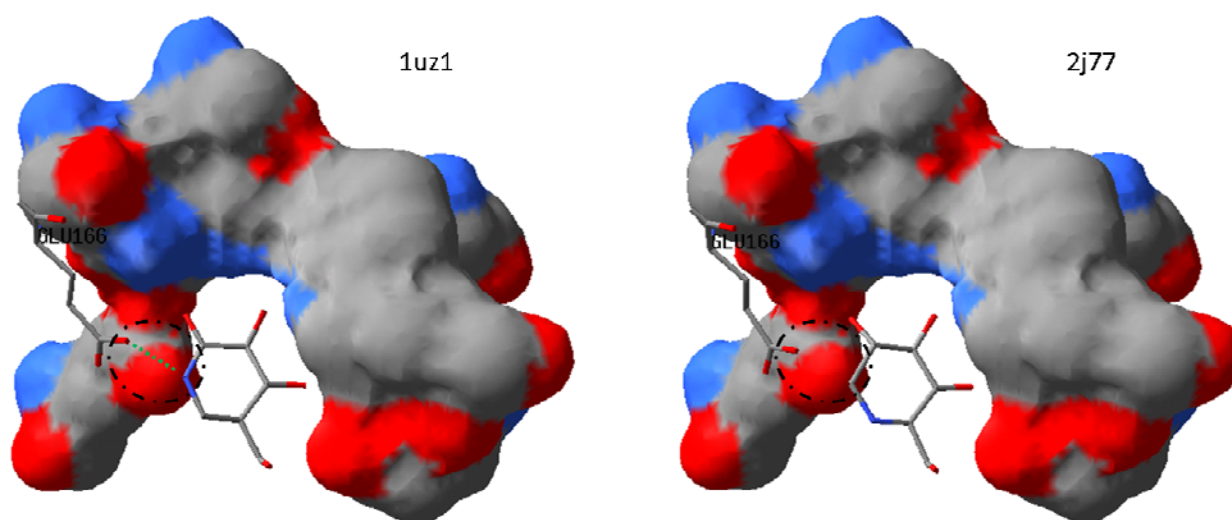


Figure 4.1.5. An example of a hydrogen bond loss in the highly conserved residue. The binding affinities of beta-glucosidase protein (PDB code 1uz1 and 2j77) are 6.89 and 4.89, respectively. For these two complexes, the bound-ligand structures and protein-ligand interacting are very similar and the only difference is the position of Nitrogen atom in the ligand. For the complex 1uz1, the nitrogen forms a hydrogen bond with Glu166; conversely, it is unable to form any hydrogen bond on the complex 2j77. The conserved score of Glu166 is 2.

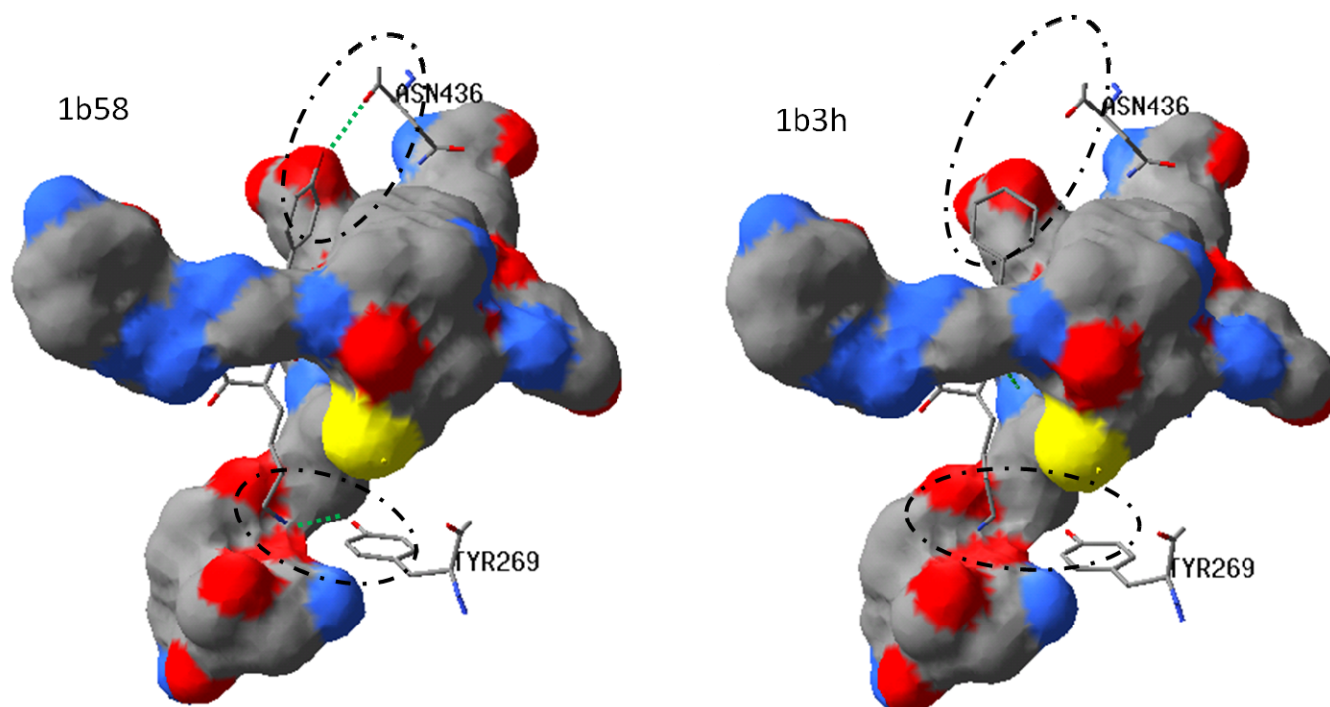


Figure 4.1.6. An example of hydrogen bonds loss in the low conserved residues. For the oligo-peptide binding protein (PDB code 1b58 and 1b3h), the binding affinities of the complexes with hydrogen bonds (1b58) and without hydrogen bonds (1ctt) between bound ligands and residues (i.e. Asn436 and Tyr269) are 6.58 and 6.21, respectively. The conserved scores of Asn436 and Tyr269 are -3 and -6.

Binding affinity prediction

After the GemAffinity was developed, we evaluated it on 98 complexes in the independent test set and compared it with other methods. Pearson's correlation coefficients, yielded by our GemAffinity, on the training set (891 complexes) and the independent test set (98 complexes) are 0.60 and 0.58, respectively (Figure 4.1.7). In general, it is neither straightforward nor completely fair to compare the results of different scoring functions for predicting binding affinities of protein-ligand complexes. Here, we compared the GemAffinity with other 12 scoring functions on the data set (Figure 4.1.8). The correlations of other 12 scoring functions were directly summarized from the previous work¹¹³. The results show that the GemAffinity is the best among 13 comparative scoring functions on this data set. Some scoring functions, the empirical scoring functions (*i.e.* X-SCORE¹²², F-Score¹³⁹, ChemScore¹¹³, LigScore¹⁴⁴, PLP⁵⁹, LUDI¹⁴⁵, and HINT¹⁴⁶), only consist of common types of protein-ligand interactions, such as van der Waals, hydrogen bonding, and electrostatic integrations. These scoring functions are usually usefully in predicting the affinities for most of protein-ligand complexes. However, if some unusual factors (e.g. metal-ligand bonding interactions)

are necessary for the binding process, these scoring functions may fail in predicting the binding affinities. Conversely, the GemAffinity is much better than these comparative functions when these complexes have metal ions interacting to bound ligand within 2.2 Å. In addition, Pearson's correlation coefficient of the GemAffinity is reduced to 0.565 if we considered all of hydrogen bonds.

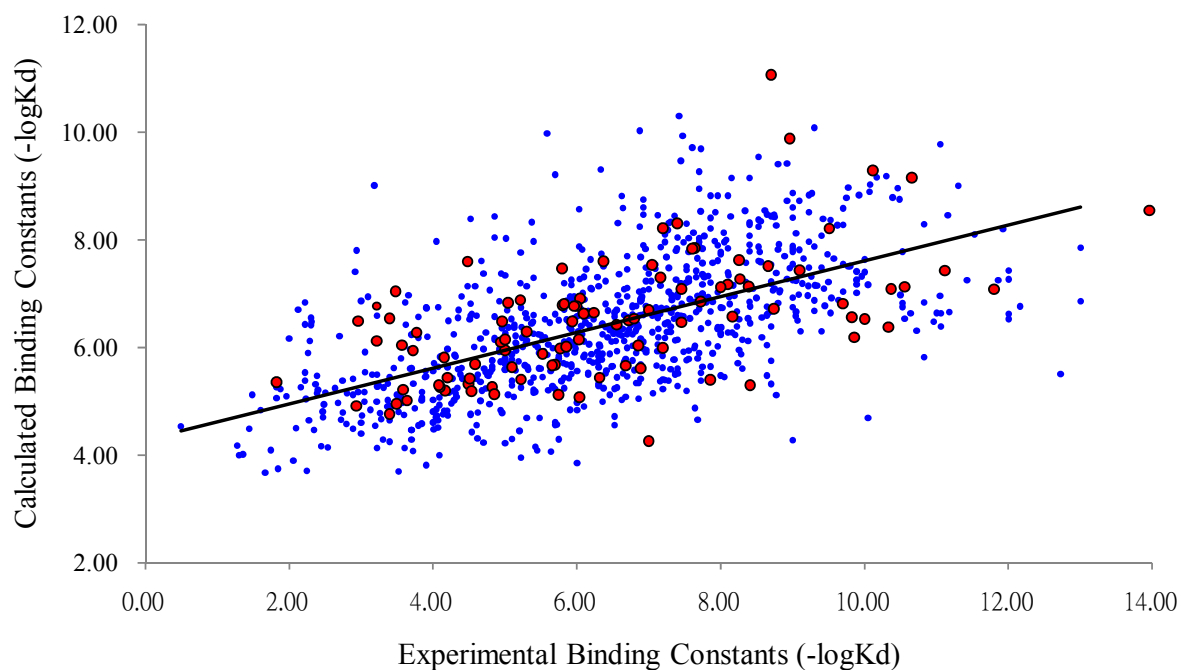


Figure 4.1.7. Pearson's correlation coefficients between experimental binding affinities and predicted binding affinities using the GemAffinity are 0.60 and 0.58, respectively, on training set (●) and the testing set (●).

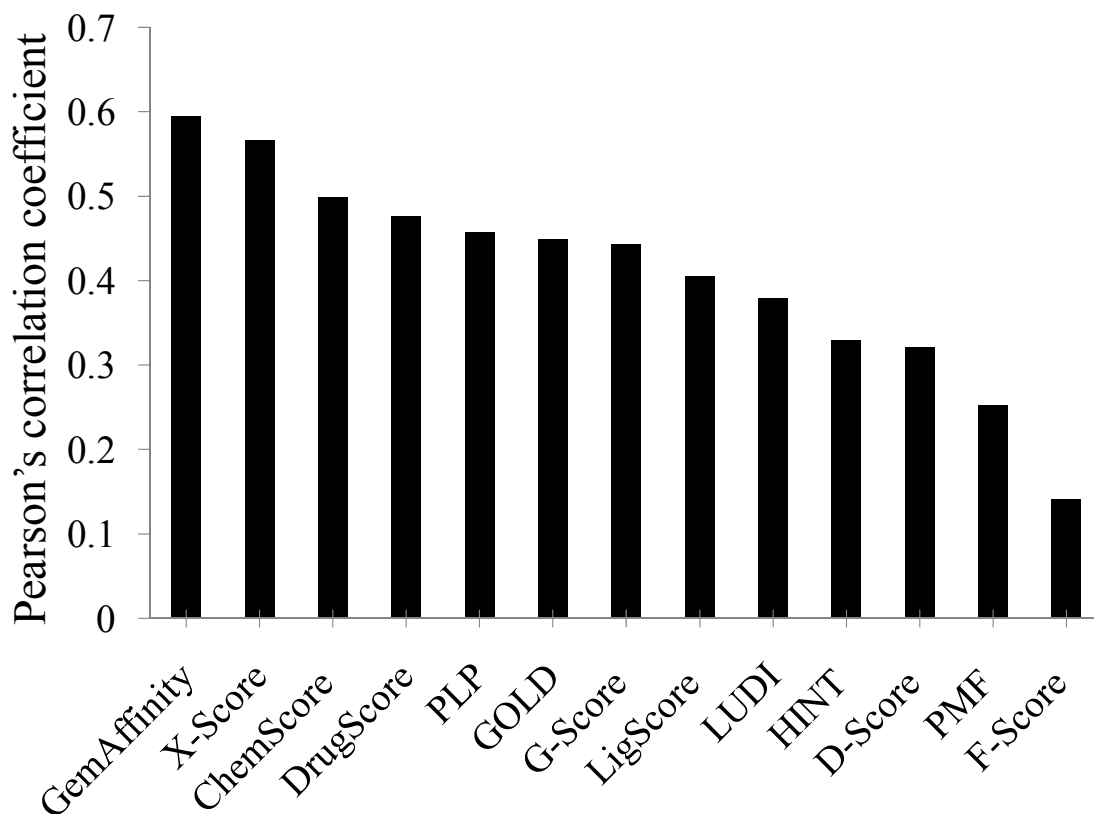


Figure 4.1.8. The comparison of the GemAffinity and other 12 scoring functions on the independent data set with 98 protein-ligand complexes.

Virtual screening

A scoring function for predicting the binding affinity of a protein-ligand complex should be applied to virtual screening for evaluating its screening accuracy. Here, we used two docking programs (GEMDOCK and GOLD) and three scoring functions (PLP in GEMDOCK, GoldScore, and the GemAffinity) to assess the accuracy on four protein targets (TK, ER, ERA, and HCAII) based on the receiver operating characteristic ROC curve ([Figure 4.1.9](#)). All of these methods were tested using the same reference protein and screening database.

Among these scoring methods, Experimental results show that the GemAffinity is the best for targets ERA and HCAII; and is very comparable to other methods on targets TK and ER. For target HCAII ¹²⁶ ([Figure 4.1.2](#)), three residues His92, His96 and His119, which are recognized as highly conserved residues, are metal binding residues and the metal-ligand bonding is one of the primary binding interactions. The catalytic residues (e.g. His 64, Glu106, and Thr199) are also highly conserved residues using [Equation 4.1.1](#). In this target, the GemAffinity significantly outperforms GEMDOCK and GOLD on the target HCAII. The main reason is that the GemAffinity considers the metal-ligand interaction as an individual term and divides the hydrogen bonds into conserved and

non-conserved interactions for calculating binding affinities. In contrast, the scoring functions of both GEMDOCK and GOLD considered the metal-ligand as one kind of hydrogen interacts and considered all hydrogen bonds as the same.

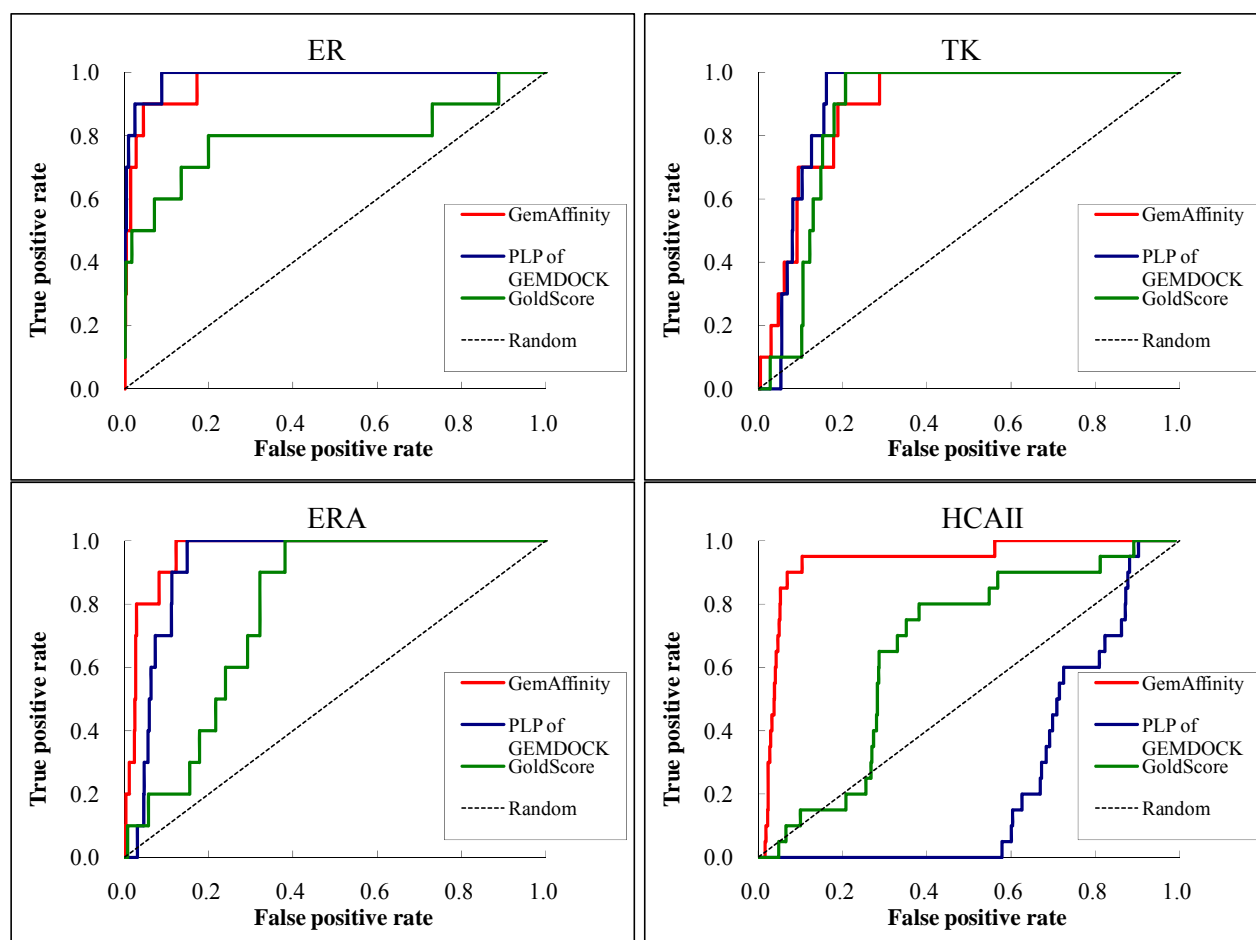


Figure 4.1.9. The ROC curves of the GemAffinity (red), scoring functions of GEMDOCK (blue), and GOLD (green) on the four targets. The curve of random selection was also plotted as the reference.

4.1.4 Conclusions

We developed a scoring function for predicting binding affinity of a protein-ligand complex by analyzing 87 descriptors derived from 891 protein-ligand structures in PDB. This scoring function consists of five selected descriptors, including Lennard-Jones potential, metal-ligand bonding, water effects, deformation penalties upon the binding process, and the number of highly conserved residues with hydrogen bonds, by using a simple stepwise regression method to derive from these 87 descriptors. The GemAffinity is able to reflect the experimental binding affinities and biological-physicochemical meanings in the protein-ligand binding processes. Experiment results show that the GemAffinity is much better than 12 comparative scoring methods on predicting binding affinities on testing structures. For virtual screening, the GemAffinity is very comparable to the scoring functions of GEMDOCK and GOLD on four target proteins. In addition, the GemAffinity outperforms ones of GEMDOCK and GOLD if the target owns metal-ligand

interactions, structural waters, and highly conserved residues with hydrogen interaction. These results demonstrate that the GemAffinity is useful to predict binding affinity and to combine with other scoring functions to improve prediction accuracy. We believe that the GemAffinity is useful for molecular recognition and virtual screening.

4.2 GEMQSAR: A QSAR Model Using Protein-ligand Interaction Consensus Profiles and Generic Evolutionary Method

4.2.1 Introduction

The dynamic increase of three-dimensional structures for drug targets and the rapid advances in technologies of computer-aided drug design provide the foundation for the development and testing of structure-based drug design and quantitative structure activity relationships (QSAR). As a result of the multidisciplinary effort from artificial intelligence, multivariate statistics and applied mathematics researchers, new QSAR methodologies continue to appear in the literature, such as comparative molecular field analysis (CoMFA)¹⁴⁷ or comparative molecular binding energy analysis (COMBINE)^{148; 149}. Although thousands of successful applications have used the above methods of QSAR and hundreds of CoMFA studies validate their approaches, the method suffers several challenges: 3D QSAR analysis such as CoMFA and COMBINE contain problems such as superposition of steric structures or selection of molecular descriptor^{148; 149; 150}.

We addressed such issues by developing a new method naturally integrating a well-developed molecular docking tool (i.e. GEMDOCK) with evolutionary-based QSAR tools, GEM-PLS¹⁵¹ and GEM-kNN, by using protein-ligand interactions. GEMDOCK^{2; 9; 151} was adapted for protein-ligand docking and the protein-ligand interactions were then calculated by the empirical scoring of GEMDCOK. We combined the evolutionary-based PLS (GEMPLS) and kNN (GEMkNN) for optimization and statistic in QSAR analyses. In the process of QSAR analysis, the features selected and reduced for consensus feature set and specific skeleton set were used to identify critical functional groups of compounds and key residues in the protein. To evaluate our method for QSAR analysis, we have verified the QSAR method on the human acetylcholinesterase (huAChE) dataset¹⁵². In addition, we have practically applied the method for the first QSAR model of *Arthrobacter globiformis* histamine oxidase (AGHO).

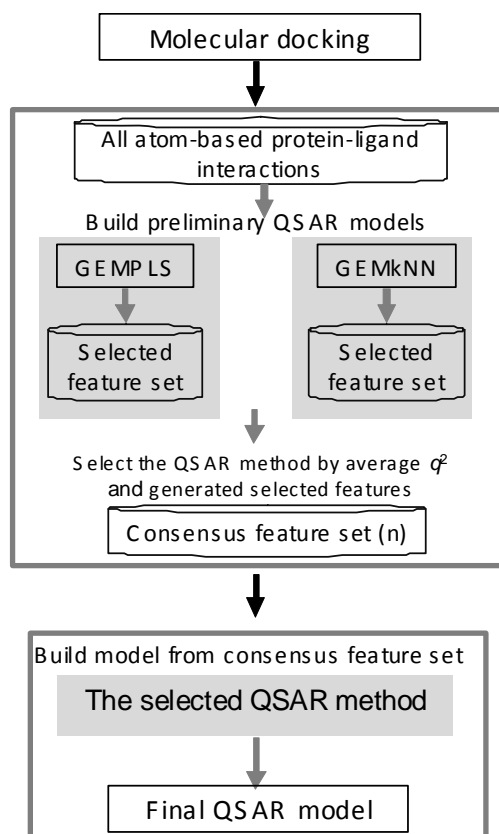


Figure 4.2.1. The main steps of GEMQSAR method. It includes GEMDOCK used for the simulation of protein-ligand complex and generation of protein-ligand interaction profile, GEMPLS and GEMkNN for feature selection and model building. The consensus feature set is generated in the modeling steps.

4.2.2 Materials and Methods

GEMQSAR was created by integrating a well-developed molecular docking tool (i.e. GEMDOCK) with evolutionary-based QSAR tools, GEM-PLS and GEM-kNN, using protein-ligand interactions. To find out the significant hot spots in the binding site, we have focused the features on atom based protein-ligand interactions. In addition, we have adopted the concept of consensus feature set and specific skeleton set to improve the stability and performance of our method. [Figure 4.2.1](#) shows the main step of our method that involves the following steps:

- (a) Generate the protein-ligand complexes through molecular docking tool, GEMDOCK.
- (d) According to the protein-ligand complex, generate the atom based protein-ligand interactions of

each residue in the binding site to be the molecular features set.

- (e) Build preliminary models by GEMPLS and GEMkNN.
- (f) According to the average q^2 value of the leave-one-out cross validation correlation for training set, select the QSAR method with higher average q^2 value for QSAR modeling and generate the consensus feature set.
- (h) Build the QSAR model from the consensus feature set and the select QSAR method.

In order to improve the performance of the method for QSAR model constructing, two feature-filtering steps have been introduced into the procedure of the QSAR modeling. The filtering steps include the following:

- (a) Generation of consensus feature set. In order to improve the stability of model, GEMPLS or GEMkNN was carried out ten times respectively, and ten groups of feature were selected. The consensus features were selected by using [Equation 4.2.1](#).

$$N_i \geq \mu_{total} - \sigma_{total} \quad (4.2.1)$$

If the selected times N of each feature i was greater or equal to the difference between average μ_{total} and standard derivation σ_{total} of selected frequency in total features, the feature i was included into the consensus feature set. In the process of model constructing, we have employed the collection of selected features from GEMPLS or GEMkNN based on which one has higher average q^2 value in validation to generate the consensus feature set. The final QSAR model was built using the consensus feature set and the QSAR method with higher average q^2 value in the preliminary models.

- (b) Generation of specific skeleton set. For the inhibitors or substrates of protein target, they always shared some conserved moieties of molecular structure and these parts often have no substitution groups. Most contribution for the differences of activities comes from the variable substitution groups. In order to mine the activity contribution of the variable substitution groups, we removed the interaction generated from the conserved moieties in the feature set. The definitions of specific skeleton in data set were not shown here ([Figure 4.2.2](#)).

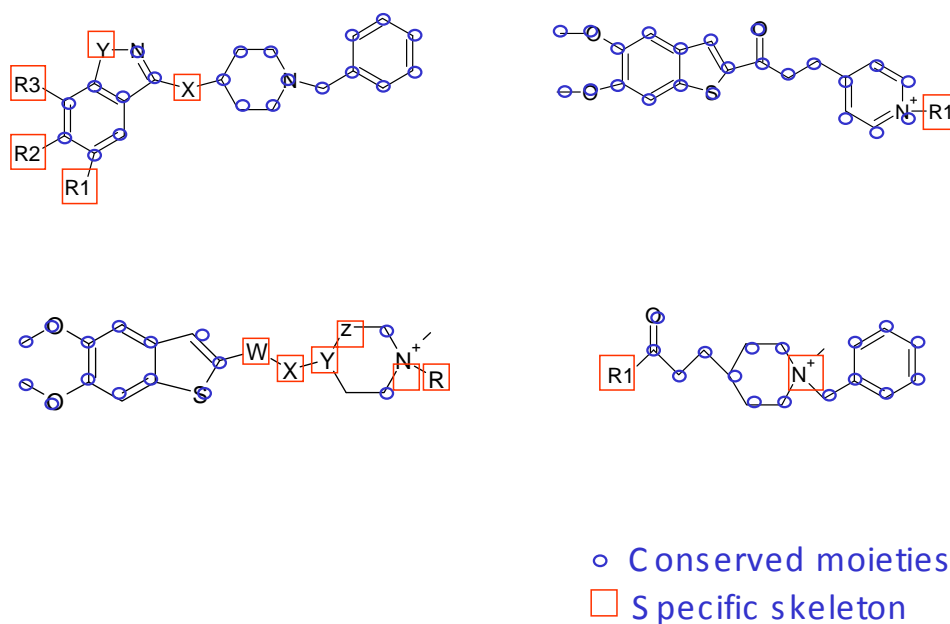


Figure 4.2.2. The definition of the specific skeleton set. In the compound set of *huAChE*, the compounds shared some conserved moieties of molecular structure and these moieties always had no substitution groups. The red labeled parts of the compound are defined as the specific skeleton set in this study.

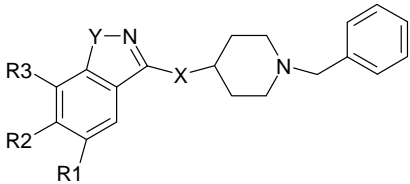
Generation of molecular descriptors

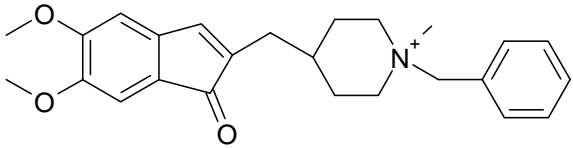
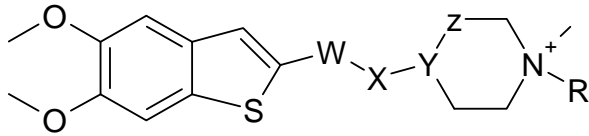
All 3D chemical structures of data sets were generated using CORINA 3.0. The human AChE (*huAChE*) from Guo, J., et al. and the *Arthrobacter globiformis* histamine oxidase (AGHO) compound sets were used to develop the QSAR models. The human AChE (*huAChE*) compound set from Guo, J., et al.,¹⁵² was used to evaluate GEMQSAR method. There are sixty-nine compounds with IC_{50} values measured with *huAChE* assay in the set, and the compounds are divided into four groups mainly. Within the set, fifty-three compounds were selected for the training set and sixteen compounds for the testing set to validate the result of our method (Table 4.2.1 and Table 4.2.2). The AGHO compound set containing twelve compounds with K_m values measured in AGHO enzyme assay (data from Dr. Chiun-Jye Yuan) was selected for the training model (Table 4.2.3). One derivative structure was selected for validation using the AGHO enzyme assay.

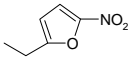

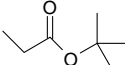
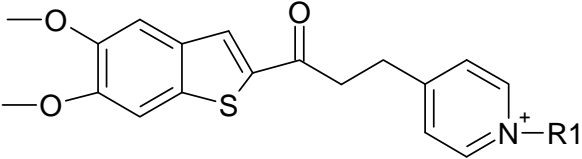


GEMDOCK was adapted to perform the process of *huAChE* and AGHO. GEMDOCK docked each compound in the compound set against the binding cavity, and generated the protein-ligand

interactions (hydrogen bonding, van der Waals and electrostatic interactions) of each compound by empirical scoring function of the docked conformation. The structure of the binding cavity for *huAChE* and AGHO, including amino acids enclosed within a 8 Å radius sphere centered on the bound ligand, was used. Our previous work ^{9, 66} has showed that the docking accuracy of GEMDOCK was better than some other docking tools, such as GOLD ⁵ and FlexX ³, on a diverse data set of 100 protein-ligand complexes proposed by Jones et al⁵. The accuracy of GEMDOCK were also better than GOLD, FlexX, and DOCK on the ligand database from Bissantz et al. (2000) for TK ²⁸ and ER-antagonist receptor ⁹. GEMDOCK parameters in the flexible docking included the initial step sizes ($\sigma=0.8$ and $\psi=0.2$), family competition length ($L = 2$), population size ($N = 300$), and recombination probability ($p_c = 0.3$).

Table 4.2.1. Chemical Structures in *huAChE* training set¹⁵²

							
R1	R2	R3	-X-	-Y-	Ligand ID	IC ₅₀ (nM)	pIC ₅₀
-H	-H	-H	-(CH ₂) ₂ -	-O-	1	55	7.26
-CH ₃	-H	-H	-(CH ₂) ₂ -	-O-	2	7.8	8.11
-CH ₃	-OCH ₃	-H	-(CH ₂) ₂ -	-O-	3	5.8	8.24
-OCH ₃	-H	-H	-(CH ₂) ₂ -	-O-	4	7.2	8.14
-H	-H	-OCH ₃	-(CH ₂) ₂ -	-O-	5	7.1	8.15
-H	-NH-CO-CH ₃	-H	-(CH ₂) ₂ -	-O-	6	2.8	8.55
-H	-NH-SO ₂ -Φ	-H	-(CH ₂) ₂ -	-O-	7	14	7.85
-H	-4-morpholino	-H	-(CH ₂) ₂ -	-O-	8	0.8	9.10
-H	-NH ₂	-H	-(CH ₂) ₂ -	-O-	9	20	7.70
-H	-Br	-H	-(CH ₂) ₂ -	-O-	10	50	7.30
-H	-CN	-H	-(CH ₂) ₂ -	-O-	11	101	7.00

-H	-CO-NH ₂	-H	-(CH ₂) ₂ -	-O-	12	8.8	8.06
-H	-H	-H	-(CH ₂) ₃ -	-O-	13	900	6.05
-H	-H	-H	-O-CH ₂ -	-O-	14	2600	5.59
-H	-H	-H	-NH-CH ₂ -	-O-	15	320	6.49
-H	-H	-H	-(CH ₂) ₂ -	-S-	16	99	7.00
-H	-H	-H	-(CH ₂) ₂ -	-CH=CH-	17	220	6.66
-H	-H	-H	-(CH ₂) ₂ -	-NH-	18	120	6.92
-CH ₂ -CH ₂ -CO-NH-		-H	-(CH ₂) ₂ -	-O-	19	0.57	9.24
-NH-CO-CH ₂ -		-H	-(CH ₂) ₂ -	-O-	20	0.95	9.02
-N(CH ₃)-CO-CH ₂ -		-H	-(CH ₂) ₂ -	-O-	21	0.48	9.32
-H	-NH-CO-CH ₂ -		-(CH ₂) ₂ -	-O-	22	3.6	8.44
					23	250	6.60
							
-W-	-X-	-Y-Z-	R	Ligand ID	IC ₅₀ (nM)	pIC ₅₀	
-(CO)-	-CH ₂ -CH ₂ -	-CH-CH ₂ -	-CH ₂ -Φ	24	8	8.10	
-(CO)-	-CH ₂ -C(OH)-	-CH-CH ₂ -	-CH ₂ -Φ	25	43	7.37	
-(CO)-	-CH ₂ C(OH)CH ₂ CH ₂ -	-CH-CH ₂ -	-CH ₂ -Φ	26	380	6.42	
-(CO)-	-CH ₂ CH ₂ CH ₂ CH ₂ -	-CH-CH ₂ -	-CH ₂ -Φ	27	110	6.96	
-(CO)-	-CH ₂ C(OCH ₃)-	-CH-CH ₂ -	-CH ₂ -Φ	28	120	6.92	
-(CO)-	-CH-	-C-CH ₂ -	-CH ₂ -Φ	29	520	6.28	

-C(OH)-	-	-CH-CH ₂ -	-CH ₂ -Φ	30	19580	4.71
-	-CH-	-C-CH ₂ -	-CH ₂ -Φ	31	2670	5.57
-(CO)-	-CH ₂ -CH ₂ -	-CH-CH ₂ -	-(CH ₂) ₂ OCH ₃	32	53	7.28
-(CO)-	-CH ₂ -CH ₂ -	-CH-CH ₂ -		33	32	7.49
-(CO)-	-CH ₂ -CH ₂ -	-CH-CH ₂ -		34	28	7.55
-(CO)-	-CH ₂ -CH ₂ -	-CH-CH ₂ -		35	79	7.10
-(CO)-	-CH ₂ -CH ₂ -	-CH-CH ₂ -	-CH ₂ CH ₂ -O-Φ	36	390	6.41
-(CO)-	-CH ₂ -CH ₂ -	-CH-CH ₂ -	-CH ₂ -CN	37	1000	6.00
						
-R1				Ligand ID	IC₅₀(nM)	pIC₅₀
-CH ₃				38	900	6.05
-CH ₂ CH ₃				39	280	6.55
-CH ₂ CH=CH ₂				40	540	6.27
				41	110	6.96
				42	40	7.40
-R1				Ligand ID	IC₅₀(nM)	pIC₅₀
-CH ₂ CH ₂ -O-CH ₂ CH ₃				43	7	8.15

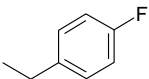
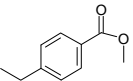
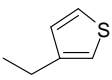
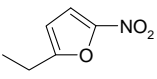
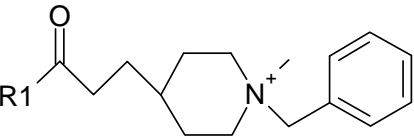
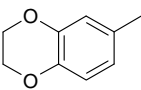
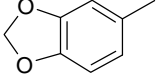
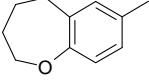
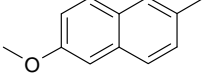
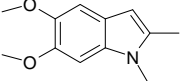
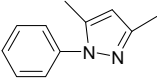
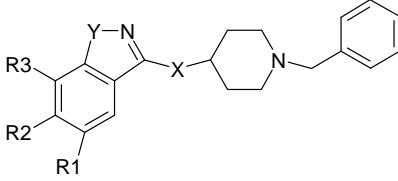
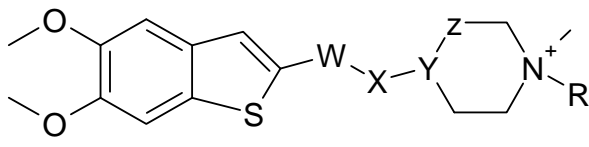
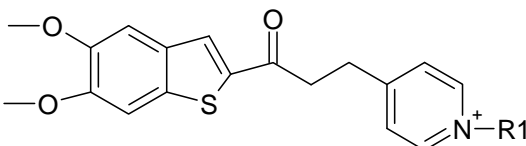
	44	2.6	8.59
	45	1000	6.00
	46	6	8.22
	47	4.5	8.35
			
-R1	Ligand ID	IC₅₀(nM)	pIC₅₀
	48	100	7.00
	49	41.5	7.38
	50	139	6.86
	51	50	7.30
	52	120	6.92
	53	22	7.66

Table 4.2.2. Chemical structures in the huAChE testing set¹⁵²

							
R1	R2	R3	-X-	-Y-	Ligand ID	IC ₅₀ (nM)	pIC ₅₀
-H	-OCH ₃	-H	-(CH ₂) ₂ -	-O-	54	8.3	8.08
-H	-NH-CO-Φ	-H	-(CH ₂) ₂ -	-O-	55	9.4	8.03
-H	-OH	-H	-(CH ₂) ₂ -	-O-	56	26	7.59
-H	-H	-H	-(CH ₂) ₂ -	-O-	57	210	6.68
-H	-H	-H	-NH-(CH ₂) ₂ -	-O-	58	810	6.09
-H	-H	-H	-(CH ₂) ₂ -	-N=CH ₂	59	340	6.47
-CH ₂ CONH-		-H	-(CH ₂) ₂ -	-O-	60	0.33	9.48
							
-W-	-X-	-Y-Z-	R		Ligand ID	IC ₅₀ (nM)	pIC ₅₀
-(CO)-	-CH ₂ C(OH)-	-CH-CH ₂ -	-CH ₂ -Φ		61	190	6.72
-(CO)-	-CH ₂ -	-C(OH)-CH ₂ -	-CH ₂ -Φ		62	90	7.05
-(CO)-	-CH ₂ -	-C=CH-	-CH ₂ -Φ		63	750	6.12
-	-CH ₂ -	-CH-CH ₂ -	-CH ₂ -Φ		64	30000	4.52
-(CO)-	-CH ₂ CH ₂ -	-CH-CH ₂ -	-CH ₂ COOCH ₃		65	54	7.27
							

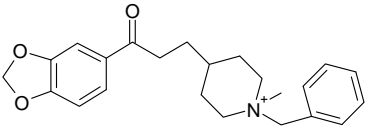
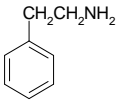
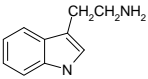
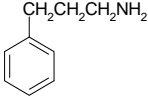
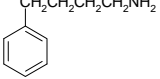
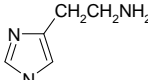
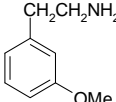
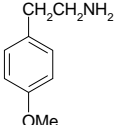
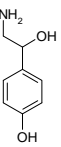
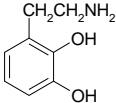
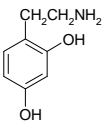
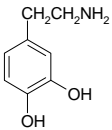
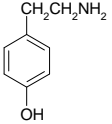
-R1	Ligand ID	IC₅₀(nM)	pIC₅₀
-(CH ₂) ₂ CH ₃	66	2570	5.59
-(CH ₂) ₂ OCH ₃	67	30	7.52
-CH ₂ -Φ	68	4.6	8.34
	69	240	6.62

Table 4.2.3. Chemical structures in the AGHO training set

Ligand	Structure	pK_m^a	Predicted K_m^b
Phenylethylamine		1.77	1.69
Tryptamine		2.44	2.37
Phenylpropylamine		2.06	2.10
Phenylbutylamine		2.60	2.58
Histamine		1.03	1.03
3-Methoxy-phenylethylamine		1.80	1.94
4-Methoxy-phenylethylamine		1.78	1.74
Octopamine		0.91	0.93
2,3-Dihydroxy-phenylethylamine		1.73	1.75
2,4-Dihydroxy-phenylethylamine		1.69	1.64

Dopamine		1.48	1.55
Tyramine		1.77	1.76

^a Values for apparent $\log(1/K_m)$ are expressed in mM.

^b Values for predicted $\log(1/K_m)$ by GEMQSAR analysis.

Structural Model of huAChE

The past QSAR studies of AChE inhibitor many were based on using ligand-based design methods such as CoMFA^{153; 154; 155; 156; 157}. To simulate the protein-ligand interactions, we have modified an induce-fit structure of huAChE from crystallized structure to be the target protein and the docking simulation of protein-ligand complex would be considered for the QSAR model constructing.

The AChE X-ray crystallized structures we used in the study were huAChE (PDB entry 1B41¹⁵⁸) and tcAChE (PDB entry 1EVE¹⁵⁹). The crystallized structure of huAChE (1B41) has no ligand complex with the protein and the structure of tcAChE (1EVE) has the co-crystallized inhibitor, E20 complex with the protein. The sequence identity between the two proteins is 57% and the root mean square deviation (RMSD) between the structures of huAChE and tcAChE are 0.88 Å for the set of all C α atoms in the whole protein. To simulate the binding conformation of the co-crystallized E2020 inhibitor relative to the huAChE structure, we have aligned the huAChE structure to the tcAChE structure by a maximal overlap of C α atoms for the huAChE/tcAChE residues within the proteins. Because the absence of a solid understanding of the roles of solvent molecules in the huAChE active site, we did not take all waters into consideration. After ascertaining the binding conformation of E2020 inhibitor relative to the huAChE structure, the hydrogen atoms were added to the huAChE-E2020 complex via SYBYL7.0. The energy optimized procedure by Tripos was then performed to the huAChE-E2020 complex force field, which had a termination gradient of 0.05 kcal/mol*Å via SYBYL7.0. The resulting structure of protein was extracted for the docking simulation of protein-ligand complex to be considered for the QSAR model construction. Further, the structure of the binding cavity for huAChE, including amino acids enclosed within a 8 Å radius sphere centered on the bound ligand, was used.

Structural Model of AGHO

There is no X-ray crystallized structure of AGHO so far. In order to simulate the protein-ligand interactions in the binding site of AGHO, we have constructed a homology modeling of AGHO. First we obtained the amino acid sequence of AGHO from the SwissProt/TrEMBL. Subsequently the amino acid sequence was used to search for the template by BLAST¹⁶⁰, and we selected the AGAO structure (PDB entry 1IU7¹⁶¹) to be the template. The sequence identity between the AGHO and AGAO is 61%, suggesting a high structural homology.

In preparation for the template structure, we selected the structure of AGAO A chain (PDB entry 1IU7) to be the template, and removed the Cu²⁺ ion and H₂O molecules from the crystal structure. There was a special cofactor, 2,4,5-trihydroxyphenylalanyl quinine (TPQ) in the protein, and it was generated from an intrinsic tyrosine in the amino acid sequence by a self-processing technique that required the Cu²⁺ ion and molecular oxygen¹⁶². We have modified the TPQ to tyrosine by removing the O atoms from the side-chain of TPQ. Subsequently, the homology modeling of AGHO was constructed according to the amino acid sequence of target protein and the structure of template by SWISSMODEL¹⁶³.

The root mean square deviation (RMSD) between the target protein structure and template structure is 0.15 Å for the set of all C_α atoms in the whole protein, indicating the good overall alignment and substantial structural homology. To ascertain the orientation of Cu²⁺ ion relative to the modeling structure, we have aligned the modeling structure to the AGAO (PDB entry 1IU7) structure by a maximal overlap of C_α atoms for the residues within the two proteins. To modify the tyrosine to TPQ, we modified the hydrogen atoms of the side-chain of tyrosine to oxygen in position 2, 4 and 5. Because the absence of a solid understanding of the roles of individual solvent molecules in the AGHO active site, we did not take all waters into consideration. In order to mimic the structural character of AGHO, we aligned the modeling structure to the structure of AGAO A chain (1IU7A) and AGAO B chain (1IU7B) respectively, and then we adopted the relative coordinate after alignment of each monomer.

After the modification, hydrogen atoms were added and the charge of Cu²⁺ was assigned to the structure via SYBYL7.0. The structure of model was then optimized by Tripos force field, having a termination gradient of 0.05 kcal/mol*Å in SYBYL7.0. The resulting structure of protein was extracted for the docking simulation. The structure of the binding cavity for AGHO, including amino acids enclosing within a 8 Å radius sphere centered on the catalytic cofactor (TPQ), was used.

To evaluate the performance of GEMDOCK on CuAOs, the molecular recognition of known substrate, phenylethylamine¹⁶¹ was performed for the active site of AGAO (PDB entry 1IU7). There is no co-crystallized protein-ligand complex of AGAO thus; we defined the binding site by the catalytic cofactors (TPQ). The coordination of amino acids within the sphere of 8 Å relative to the

cofactor, 2,4,5-trihydroxyphenylalanyl quinine (TPQ382) is obtained as the binding site of AGAO. The docked ligand at the active site of AGAO formed the hydrogen bonding interactions with the cofactor TPQ382-O5, D298-OD2 (general base) and I379-O by the function group $-NH_2$ of the ligand. The aromatic ring of the docked ligand stayed in the hydrophobic pocket of AGAO. These interactions correspond to the necessary interactions of ligand binding in CuAOS as described in other literatures. The result of molecular recognition revealed that GEMDOCK could generate reasonably bound poses in CuAOs.

GEMPLS

GEMPLS is a hybrid approach that combines genetic algorithm (GA) as a robust optimization technique with PLS as a powerful statistical technique for the variable selection and model evolution. GA operates on a population of potential solutions applying the principle of survival of the fittest to produce successively better approximations to optimum solution. PLS deals with strongly collinear input data and makes no restriction on the number of variables used. In GEMPLS, the chromosomes consist of some randomly selected features and the latent variables (lv). The squared cross-validated correlation coefficient q^2 in the PLS analysis is used as an objective function to provide a measurement of how the internal predictability with respect to the selected features of the chromosome. And GA will find the fittest features with the highest q^2 in the PLS analysis.

The main steps involved in GEMPLS include the following: (a) initiation and evaluation of the initial population, (b) selection of the reproductive population, (c) crossover and mutate the reproductive population, (d) evaluation of the child population, (e) reinsertion of the child population to form the population in the next generation. The cycle of above four steps (from step (b) to (e)) is repeated until the number of generations reaches the possible maximum. In order to improve the performance of GEMPLS for QSAR model building, a number of refinements have been introduced into GEMPLS. The refinements include the following:

- (a) An extra bit lv , representing the number of latent variables, is appended to the original chromosome of GA and expected to efficiently solve the problem of the optimum number of latent variables through evolutionary process
- (b) Adopt Mahalanobis distance to discriminate significant features. Mahalanobis distance is a very useful way of determining the deviation of a sample from the mean of the distribution in multivariable calculus. Therefore, the Mahalanobis distance is adopted to identify significant features from all of those.

$$M^2 = (v - \mu)' \Sigma^{-1} (v - \mu) \quad (4.2.2)$$

M is the Mahalanobis distance from the feature vector v (column vector of data matrix here) to the mean vector μ , where Σ is the covariance matrix of the features.

- (c) Cooperate with biased mutations to lead the evolution. We have recommended that uniform mutation is cooperating with biased mutation to lead the evolution of GA toward a significant feature set and to reduce the interference of noise features.

$$P_i = P_{\min} + (P_{\max} - P_{\min}) \times \left(\frac{N_s - p_i}{N_s - 1} \right) \quad (4.2.3)$$

P_i is the probability of setting feature bit i to 1, p_i is the position of feature i in the descending order of Mahalanobis distance of all features, P_{\min} and P_{\max} are the minimum and maximum values of P_i , and N_s is the number of significant features. P_i is derived from p_i only when p_i is ahead of N_s , otherwise P_i is set to P_{\min} . In other words, the significant feature i with higher Mahalanobis distance will obtain the higher P_i . In this study, the corresponding parameters are defined as: $P_{\max} = 0.8$, $P_{\min} = 0.2$

GEMkNN

GEMkNN is a hybrid approach that combines GA as a robust optimization tool with kNN as a pattern recognition method to evaluate the discriminative ability of the subset (kNN is a conceptually simple, nonlinear approach to pattern recognition problems). In GEMkNN, the chromosomes consist of some randomly selected features and the number of selected similar molecules (k). The similarities between compounds are evaluated by Euclidean distance. The squared cross-validated correlation coefficient q^2 in the kNN analysis is used as an objective function to provide a measure of how the internal predictability with respect to the selected features of the chromosome. The GA will find the fittest features with the highest q^2 in the kNN analysis.

The main steps involved in GEMkNN are the same as in GEMPLS, and the same refinements of GEMPLS included (a) adopt Mahalanobis distance to discriminate significant features (Equation 4.2.2) and (b) cooperate with biased mutation to lead the evolution (Equation 4.2.3) also have been introduced into GEMkNN.

Model evaluation

The predictability of QSAR model was assessed by the conventional cross-validated correlation coefficient (q^2), the cross-validated *SDEP* ($SDEP_{cv}$), and external *SDEP* ($SDEP_{ex}$):

$$q^2 = 1 - \frac{\sum (y_i - y_{pred_i})^2}{\sum (y_i - \bar{y})^2} \quad (4.2.4)$$

$$SDEP = \sqrt{\frac{\sum (y_i - y_{pred_i})^2}{N}} \quad (4.2.5)$$

where y_i is the observed biological activity of compound i , $y_{pred,i}$ is the predicted biological activity of compound i in the validation set, \bar{y} is the average biological activities of the data set, and N is the total number of compounds.

After deciding the optimum number of latent variables, the corresponding highest q^2 , lowest *SDEP* can be used to assess the predictability of QSAR model, i.e. the model with more remarkable predictability can provide the higher q^2 and the lower *SDEP* between the observed and predicted biological activities.

4.2.3 Results and Discussion

The GEMQSAR method combined GEMDOCK and QSAR tools, GEMPLS and GEMkNN was validated with 69 *huAChE* inhibitors and practically applied in the analysis of protein-ligand interactions on AGHO. The consensus feature profile and specific feature set were adapted for identifying critical functional groups of compounds and key residues in the protein.

QSAR model of *huAChE*

Evaluation of GEMDOCK on *AChE*

In order to evaluate GEMDOCK on *AChE*, we have docked the crystallized ligand (E20) of *tcAChE* into its reference protein (1EVE). The RMSD value between the docked conformation and the crystal structure of reference protein (1EVE_E20) is 1.73 Å. To evaluate the performance of docking tool on modeled *huAChE*, we compared the docked poses of 1EVE and modeled *huAChE*. In the complex of *tcAChE* (1EVE), the ligand E20 forms a stable stack force with W84, W279 and F330. The nearest distance between the atom N of ligand and the water is 2.90 Å. The docked ligand forms a stable stack force with W84, W279 and F330, and the nearest distance between the atom N of ligand and the water is 3.69 Å.

Validation QSAR model of *huAChE*

GEMPLS and GEMkNN have been employed in the raw feature sets and the specific feature set that removed the conserved moieties for *huAChE* inhibitors. In the whole feature set, the result of cross-validated correlation coefficient (q^2) in GEMkNN (0.66) is better than GEMPLS (0.63) (shown in [Table 4.2.4](#)). The number of total features reduced from 223 to 217 with specific skeleton set and the q^2 value of GEMkNN improved to 0.74 better than GEMPLS. The r^2 value of testing set improved in both GEMPLS and GEMkNN. The average q^2 value of cross-validated correlation coefficient of GEMkNN is better than GEMPLS in the preliminary *huAChE* models. Thus, GEMkNN has been employed in the QSAR model construction of *huAChE*.

Table 4.2.4. The performance of GEMPLS and GEMkNN relative to different protein-ligand interactions profiles in the huAChE set

	Whole Interaction Profile ^a		Specific Interaction Profile ^b	
	GEMPLS	GEMkNN	GEMPLS	GEMkNN
No. of features (atoms) ^c	223	223	217	217
Average of q^2 ^d	0.627	0.657	-2.607	0.737
Average of r^2 ^e	0.402	0.123	0.466	0.724
Standard derivation of q^2 ^f	0.015	0.018	0.093	0.018
Standard derivation of r^2 ^g	0.125	0.095	0.050	0.119
No. of selected features (atoms)	36.2	36.1	33.8	34.7

^a The interaction profile between protein and whole atoms of ligand

^b The interaction profile between protein and the specific skeleton of ligand

^c The number of feature in origin molecular feature set.

^d The average q^2 values in 10 times in training set.

^e The average r^2 values in 10 times in testing set.

^f The standard deviation of q^2 values in 10 times in training set.

^g The standard deviation of r^2 values in 10 times in testing set.

With the method flow shown in [Figure 4.2.1](#), we constructed the QSAR model of huAChE by GEMQSAR. First, we rebuild the QSAR model from the specific feature set and consensus features selected in the preliminary models. The average q^2 values of leave-one-out cross validation is 0.82 and the average correlation of $r^2 = 0.72$ between the predicted values and the experimental values. For constructing a specific QSAR model, we have adopted the one that the q^2 value is most close to the average q^2 value in 10 times of model training. This is because we hope to select a steady model and to avoid over-fitting in QSAR model building. At last we adopted the model with a leave-one-out cross validation of $q^2 = 0.82$ and a correlation of $r^2 = 0.78$ between the predictive values and experimental values (the values shown in [Table 4.2.5](#)).

[Table 4.2.6](#) shows the comparison of our method and previous research of Guo *et. al.* Guo *et. Al*, who generated their model by the PLS method from the predicted protein-ligand complexes docked by GOLD having q^2 and r^2 values in their training set of 0.72 and 0.63. GEMQSAR used PLS and kNN to generate the relationships of the atom-based protein-ligand interactions of huAChE. The

generic evolutionary method performed feature selection in the model building improved the accuracy of the training model and the q^2 and r^2 values in the training set of our method were 0.82 and 0.72 respectively. Without the feature selection step, our method only generated a model with q^2 and r^2 values of 0.74 and 0.72 which are close to the results of Guo *et. al.*. The feature selection step improved the QSAR method not only as shown in training but also as shown in prediction ability. The r^2 values between the experimental and the predicted values of huAChE testing set of GEMkNN and Guo *et. al.* are 0.78 and 0.69 (Figure 4.2.3)

Table 4.2.5. The performances of GEMQSAR with different features in the huAChE set

	Interaction Profile ^a		Consensus Feature Profile _b	
	Whole ^c	Specific ^d	Whole	Specific
No. of features (atoms)	223	217	156	92
Average of q^2	0.66	0.74	0.70	0.82
Average of r^2	0.12	0.72	0.06	0.72
Standard derivation of q^2	0.018	0.018	0.009	0.006
Standard derivation of r^2	0.095	0.119	0.050	0.056
No. of selected features (atoms)	36.1	34.7	29.8	23.5

^a The feature set of all the protein-ligand interactions.

^b The consensus protein-ligand interaction profile. The feature set that consensus from the features of preliminary models

^c The interaction profile between the protein and all ligand atoms

^d The interaction profile between the protein and the specific skeleton of ligand

Table 4.2.6. Comparison of GEMQSAR methods and results of Guo *et. al.*

	GEMQSAR	Guo <i>et. al.</i>
Docking Tool	GEMDOCK	GOLD
Basis	Atom-based interactions ^c	Residue-based interactions ^d
q^2 ^a	0.82	0.72
r^2 ^b	0.72	0.63

^a The mean q^2 value of 20 independent QSAR models in the training set .

^b The mean r^2 value of 20 independent QSAR models in the testing set .

^c The interactions containing electrostatic, hydrogen bonding, and van der Waals interactions .

^d The interactions containing electrostatic and van der Waals interactions.

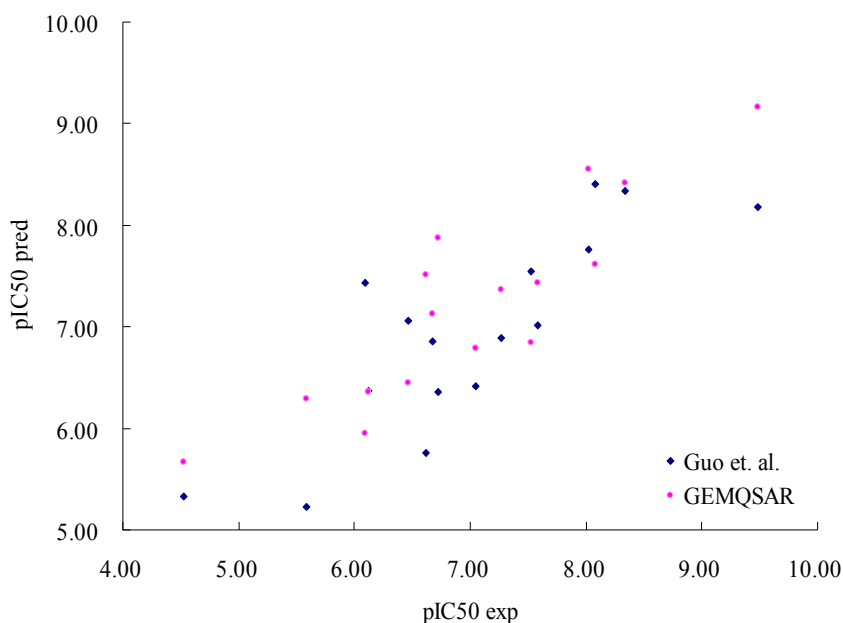


Figure 4.2.3. The comparison of the experimental IC_{50} values and predicted IC_{50} value of the testing set for huAChE QSAR model. The r^2 of the prediction of testing set for huAChE QSAR models generated by GEMQSAR and Guo *et. al.* are 0.78 and 0.69, respectively.

The comparisons of the result of the huAChE model in the whole features and consensus features are shown in [Table 4.2.5](#). GEMkNN was selected to build the final huAChE model because it generated a higher accuracy than GEMPLS on the huAChE data set. The consensus features selected by GEMkNN are 156 interactions in huAChE. Compared to the average q^2 values when considering all interactions versus only consensus interactions, GEMkNN built the better QSAR model at considering only the consensus interactions. The model built from the consensus features has higher correlation with the training set ($q^2 = 0.70$) and more consistent in model building (standard derivation of $q^2 = 0.009$). When considering the interactions from specific atoms in the ligand structure, the training q^2 and the consistency of model building further improve to 0.82 and 0.006 (standard derivation). These evidences demonstrated that GEMQSAR method with consensus features could generate a more consistent and better predicting QSAR model on huAChE.

[Table 4.2.7](#) listed the selected features and their roles for the binding of huAChE. Several residues have been found very important in previous studies^{152; 153}. Residue Y72 could form a wall to stabilize ligand and W86 forms π - π interaction with choline. N87 and Y337 contribute to the electrostatic forces in the active site. Residues Y124 and F338 provide hydrophobic contacts with ligand. S203 and H447 are significant in huAChE, which are the catalytic triad in the enzyme catalysis. In previous studies¹⁵³, residue H287 was found to possibly affect the binding affinity of

AChE inhibitors and W286 might play the same role as H287. Residue Y341 forms the local pocket in the active site. The QSAR analysis for huAChE demonstrated that our method could mine significant features for influencing the binding affinity. The comparisons of previous works also indicated that the QSAR model built by our method not only has well predicted performances but also generates consistent models.

Table 4.2.7. Important atoms in the huAChE by the GEMQSAR

Residues	Atom Type	Description ^{152; 153}
TYR72	CD1	Forms a wall to stabilize ligand ring
TRP86	CH2	Forming π - π interaction with choline
ASN87	CA	Electrostatic contributors in the gorge area
TYR119	CA	- ^a
GLY120	N	-
TYR124	CE1、CE2	Provide hydrophobic contacts
GLY126	CA	-
SER203	CB	Catalytic triad
TRP286	CG、NE1、CZ3	Probably helpful in enhancing the activity of ligand with polar groups
SER293	O	-
ARG296	N、CA、C	-
TYR337	C、O、CD2、CE1	Electrostatic contributors in the gorge area
PHE338	C	Provide hydrophobic contacts
TYR341	CB	The residue in the local pocket
HIS447	CD2	Catalytic triad
GLY448	O	-

^a not available.

The application of AGHO QSAR model

Our method has been employed to practical application of QSAR analysis for AGHO. AGHO is one of CuAOs (EC 1.4.3.6). In prokaryotic organisms, these enzymes are utilized for growth on amine. In human, these enzymes have been found to be correlated with heart failure¹⁴⁷ and chronic medical condition in diabetic patients^{164; 165}. The training set of AGHO has twelve known substrates with K_m values (Table 4.2.3).

GEMDOCK generated the poses of 12 compounds for modeled AGHO binding sites and generate 131 raw interaction features (included hydrogen bonding, van der Waals and electrostatic

interactions). Then GEMPLS and GEMkNN were employed for building ten preliminary QSAR models, respectively. On the all protein-ligand features of AGHO, the average q^2 of cross-validation for GEMPLS and GEMkNN is 0.98 and 0.82 in the training set, respectively (Table 4.2.9). When applying the specific skeleton set for QSAR analysis, the average q^2 of cross-validation of GEMPLS and GEMkNN is 0.89 and 0.38. We selected GEMPLS for generating the consensus feature set and building the AGHO model from the preliminary QSAR analyses. The average q^2 values of leave-one-out cross validation of training set for AGHO is 0.98 and standard deviation of the q^2 values 0.001 by GEMPLS method. After rebuilding ten consensus models by GEMPLS, we adopted the AGHO model with a leave-one-out cross validation of $q^2 = 0.98$. The predicted K_m values show in the Table 4.2.10. Analyzing the relationships of substitution groups and affinity, the affinity of AGHO substrates is related to the hydrophobicity of compound structures. To explore the relationship between affinity and substitution groups, we created a set of derivative structures from 12 known compounds and used the AGHO QSAR model to predict the potential affinities. The derivative structures focused on the two main factors for hydrophobicity, (a) the length of substitution group (Figure 4.2.4A) and (b) the size of aromatic ring (Figure 4.2.4B). The predicted values of derivative structures show the trend of affinity would increase with the extension of substitutions and ring size (Table 4.2.11). In order to validate GEMQSAR analysis, two compounds, 2-(1H-benzo-indol-3-yl)ethanamine and phenylmethanamine were selected into the biological assay (Figure 4.2.4). The biological assay identified a new substrate structure, phenylmethanamine with pK_m value 1.26 which is very close to the predicted pK_m value 1.08 (shown in Figure 4.2.4A). The discovery of a new substrate demonstrated the robustness and prediction power of GEMQSAR analysis method.

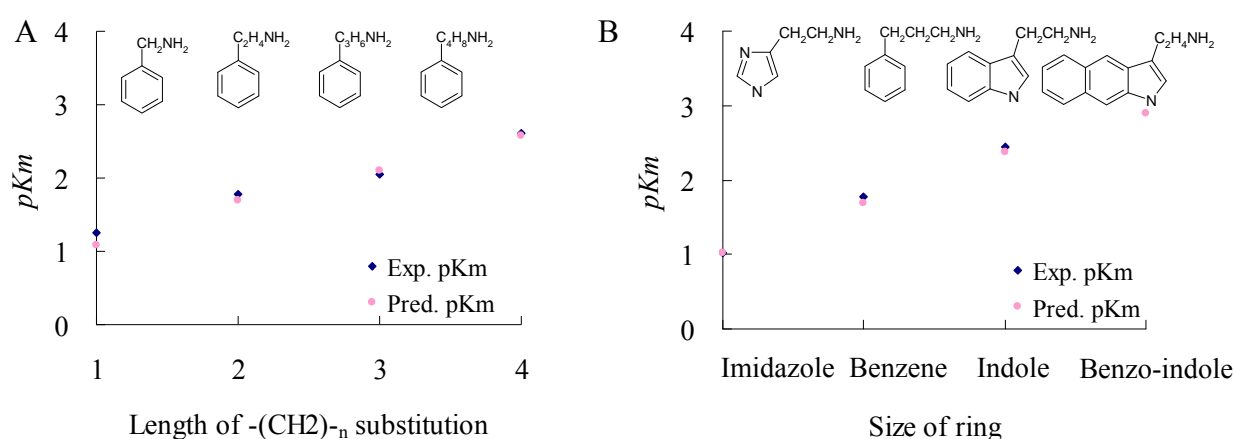


Figure 4.2.4. The relationships of the binding affinity with (A) the extension of substitutions, and (B) size of ring. (A) The relationship of length of $-(CH_2)-$ substitution and affinity predicted by AGHO QSAR analysis (B) The relationship of ring size and affinity predicted by AGHO QSAR analysis.

The biological assay identified a new substrate structure, phenylmethanamine with pK_m value 1.26 which is very close to the predicted pK_m value 1.08.

Table 4.2.9. The performance of GEMPLS and GEMkNN relative to different protein-ligand interactions profiles in the AGHO set

	Whole Interaction Profile ^a		Specific Interaction Profile ^b	
	GEMPLS	GEMkNN	GEMPLS	GEMkNN
Number of features (atoms)	131	131	94	94
Average of q^2	0.98	0.82	0.89	0.38
Standard derivation of q^2	0.001	0.008	0.069	0.031
No. of selected features (atoms)	11	16.5	13	33.5

^a The interaction profile between protein and whole atoms of ligand

^b The interaction profile between protein and the specific skeleton of ligand

Table 4.2.10. The performances of GEMQSAR using four different features in the AGHO set

	Interaction Profile ^a		Consensus Feature Profile ^b	
	Whole ^c	Specific ^d	Whole	Specific
No. of features (atoms)	131	94	34	52
Average of q^2	0.98	0.89	0.98	0.95
Standard derivation of q^2	0.001	0.069	0.001	0.007
No. of selected features (atoms)	11	13	12.9	12.8

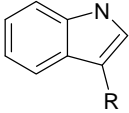
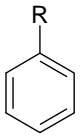
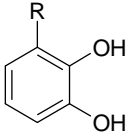
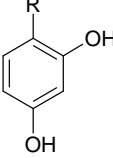
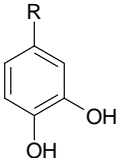
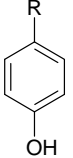
^a The feature set of all the protein-ligand interactions.

^b The consensus protein-ligand interaction profiles derived from 10 preliminary models

^c The interaction profiles between the protein and all atoms of a ligand

^d The interaction profiles between the protein and the atoms on the specific skeleton of a ligand

Table 4.2.11. Chemical structures of AGHO derived substrates with different lengths of substitution groups

Structure	Ligand	R	Predicted pK_m ^a
	Tryptamine	-CH ₂ NH ₂	2.19
		-C ₂ H ₄ NH ₂	2.37
		-C ₃ H ₆ NH ₂	2.81
		-C ₄ H ₈ NH ₂	2.32
	Phenylethylamine	-CH ₂ NH ₂	1.08
		-C ₂ H ₄ NH ₂	1.69
		-C ₃ H ₆ NH ₂	2.10
		-C ₄ H ₈ NH ₂	2.58
	2,3-Dihydroxy-phenylethylamine	-CH ₂ NH ₂	1.32
		-C ₂ H ₄ NH ₂	1.75
		-C ₃ H ₆ NH ₂	2.04
		-C ₄ H ₈ NH ₂	2.08
	2,4-Dihydroxy-phenylethylamine	-CH ₂ NH ₂	1.32
		-C ₂ H ₄ NH ₂	1.64
		-C ₃ H ₆ NH ₂	1.82
		-C ₄ H ₈ NH ₂	2.37
	Dopamine	-CH ₂ NH ₂	1.29
		-C ₂ H ₄ NH ₂	1.55
		-C ₃ H ₆ NH ₂	1.87
		-C ₄ H ₈ NH ₂	2.02
	Tyramine	-CH ₂ NH ₂	1.17
		-C ₂ H ₄ NH ₂	1.76
		-C ₃ H ₆ NH ₂	2.00
		-C ₄ H ₈ NH ₂	1.99

^a The predicted $\log(1/K_m)$ values from the final model generated by GEMQSAR.

The AGHO QSAR analysis selected 14 interaction features for describing the affinities of AGHO substrates (detail shown in Table 4.2.12). The importance of residues in AGAO has been studied in previous researches¹⁶⁶ but there are few researches focused on AGHO. The sequence identity between the AGHO and AGAO is 61%, suggesting a high structural and functional homology between them. Most selected residues in AGHO QSAR analysis are conserved on the AGHO and AGAO. We used the homologous residues of AGAO to understand the functions of selected residues in AGHO. Residues Ala155, Pro156, Tyr315, Tyr321, and Phe426 correspond to the

residues of the hydrophobic pocket in binding site of AGAO and they could form the van der Waals contact interactions in ligand binding.

Table 4.2.12. List of important atoms in the AGHO QSAR analysis

Selected Residues	Atom Type	Description[]
PHE125	CD2	- ^a
ALA155	O	The residue in the hydrophobic pocket
PRO156	N	The residue in the hydrophobic pocket
TYR315	OH	The residue in the hydrophobic pocket
ASP317	OD1	The general base in the active site
TYR321	CE1 、 CZ	The residue in the hydrophobic pocket
VAL398	C 、 CB	-
ASN400	CA 、 CG 、 OD1	The conserved residue in CuAOs
PHE426	CD2 、 CE2	The residue in the hydrophobic pocket

^a not available.

4.2.4 Conclusions

We employed GEMDOCK to generate the atom-based protein-ligand interactions as features, which are used by GEMPLS and GEMkNN to construct the QSAR models. The QSAR analysis method has been verified on huAChE QSAR analysis with $q^2=0.82$ and $r^2=0.72$ values therefore showing improvement over previous QSAR studies that apply the consensus features for modeling. The comprehensive results show good performances for QSAR analyses of huAChE and AGHO and in the process of QSAR model construction, generating the consensus feature set improves the quality and stability of QSAR analysis. On the application of AGHO, the new substrate structure of phenylmethanamine was identified by QSAR analysis and validated by further enzyme assay. These verifications and applications show that GEMQSAR method is adaptable to QSAR analysis and useful to mine the important interactions related to activities.

References

1. Bissantz, C., Folkers, G. & Rognan, D. (2000). Protein-based virtual screening of chemical databases. 1.evaluation of different docking/scoring combinations. *Journal of Medicinal Chemistry* **43**, 4759-4767.
2. Yang, J.-M. & Chen, C.-C. (2004). GEMDOCK: a generic evolutionary method for molecular docking. *Proteins: Structure, Function, and Bioinformatics* **55**, 288-304.
3. Kramer, B., Rarey, M. & Lengauer, T. (1999). Evaluation of the flexX incremental construction algorithm for protein-ligand docking. *Proteins: Structure, Function, and Bioinformatics* **37**, 228-241.
4. Morris, G. M., Goodsell, D. S., Huey, R. & Olson, A. J. (1996). Distributed automated docking of flexible ligands to proteins: parallel applications of autodock 2.4. *Journal of Computer-Aided Molecular Design* **10**, 293-304.
5. Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology* **267**, 727-748.
6. Bender, A., Mussa, H., Glen, R. & Reiling, S. (2004). Molecular similarity searching using atom environments, information-based feature selection, and a naïve Bayesian classifier. *Journal of Chemical Information and Computer Sciences* **44**, 170-178.
7. Amari, S., Aizawa, M., Zhang, J., Fukuzawa, K., Mochizuki, Y., Iwasawa, Y., Nakata, K., Chuman, H. & Nakano, T. (2006). 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**, 221-30.
8. Deng, Z., Chuaqui, C. & Singh, J. (2004). Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *Journal of Medicinal Chemistry* **47**, 337-44.
9. Yang, J.-M. & Shen, T.-W. (2005). A pharmacophore-based evolutionary approach for screening selective estrogen receptor modulators. *Proteins: Structure, Function, and Bioinformatics* **59**, 205-220.
10. Yang, J. M., Chen, Y. F., Shen, T. W., Kristal, B. S. & Hsu, D. F. (2005). Consensus scoring criteria for improving enrichment in virtual screening. *Journal of Chemical Information and Modeling* **45**, 1134-46.
11. Yang, J. M., Chen, Y. F., Tu, Y. Y., Yen, K. R. & Yang, Y. L. (2007). Combinatorial computational approaches to identify tetracycline derivatives as flavivirus inhibitors. *PLoS ONE*, e428.
12. Lin, E. S., Yang, J. M. & Yang, Y. S. (2003). Modeling the binding and inhibition mechanism of nucleotide and sulfotransferase using molecular docking. *Journal of the Chinese Chemical Society* **50**, 655-663.

13. Sayle, R. A. & Milner-White, E. J. (1995). RASMOL: biomolecular graphics for all. *Trends in Biochemical Sciences* **20**, 374.
14. Champness, J. N., Bennett, M. S., Wien, F., Visse, R., Summers, W. C., Herdewijn, P., Clercq, E. d., Ostrowski, T., Jarvest, R. L. & Sanderson, M. R. (1998). Exploring the active site of Herpes simplex virus type-1 thymidine kinase by X-ray crystallography of complexes with aciclovir and other ligands. *Proteins: Structure, Function, and Bioinformatics* **32**, 350-361.
15. Deshpande, N., Address, K. J., Bluhm, W. F., Merino-Ott, J. C., Townsend-Merino, W., Zhang, Q., Knezevich, C., Xie, L., Chen, L., Feng, Z., Green, R. K., Flippen-Anderson, J. L., Westbrook, J., Berman, H. M. & Bourne, P. E. (2005). The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Research* **33**, D233-D237.
16. Nissink, J. W., Murray, C., Hartshorn, M., Verdonk, M. L., Cole, J. C. & Taylor, R. (2002). A new test set for validating predictions of protein-ligand interaction. *Proteins: Structure, Function and Bioinformatics* **49**, 457-471.
17. Bissantz, C., Folkers, G. & Rognan, D. (2000). Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *Journal of Medicinal Chemistry* **43**, 4759-67.
18. Jain, A. J. (2003). Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *Journal of Medicinal Chemistry* **46**, 499-511.
19. Leyssen, P., Clercq, E. D. & Neyts, J. (2000). Perspectives for the treatment of infections with Flaviviridae. *Clinical Microbiology Reviews* **13**, 67-82.
20. Modis, Y., Ogata, S., Clements, D. & Harrison, S. C. (2003). A ligand-binding pocket in the dengue virus envelope glycoprotein. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 6986-6991.
21. Mukhyapathyay, S., Kuhn, R. J. & Rossmann, M. G. (2005). A structural perspective of flavivirus life cycle. *Nature Reviews Microbiology* **3**, 13-22.
22. Modis, Y., Ogata, S., Clements, D. & Harrison, S. C. (2004). Structure of the dengue virus envelope protein after membrane fusion. *Nature* **427**, 313-319.
23. Rey, F. A., Heinz, F. X., Mandl, C., Kunz, C. & Harrison, S. C. (1995). The envelope glycoprotein from tick-borne encephalitis virus at 2 Å resolution. *Nature* **375**, 291-298.
24. Lee, E., Weir, R. C. & Dalgarno, L. (1997). Changes in the dengue virus major envelope protein on passaging and their localization on the three-dimensional structure of the protein. *Virology* **232**, 281-290.
25. Lyne, P. D. (2002). Structure-based virtual screening: an overview. *Drug Discovery Today* **7**, 1047-1055.
26. Shoichet, B. K., McGovern, S. L., Wei, B. & Irwin, J. (2002). Lead discovery using molecular docking. *Current Opinion in Chemical Biology* **6**, 439-446.

27. Shoichet, B. K. (2004). Virtual screening of chemical libraries. *Nature* **432**, 862-865.
28. Yang, J.-M., Shen, T.-W., Chen, Y.-F. & Chiu, Y.-Y. (2004). An evolutionary approach with pharmacophore-based scoring functions for virtual database screening. *Lecture Notes in Computer Science* **3102**, 481-492.
29. Chiu, M.-W. & Yang, Y.-L. (2003). Blocking the dengue virus 2 infections on BHK-21 cells with purified recombinant dengue virus 2 E protein expressed in *Escherichia coli*. *Biochemical and Biophysical Research Communications* **309**, 672-678.
30. Shen, M., LeTiran, A., Xiao, Y., Golbraikh, A., Kohn, H. & Tropsha, A. (2002). Quantitative structure-activity relationship analysis of functionalized amino acid anticonvulsant agents using k nearest neighbor and simulated annealing PLS methods. *Journal of Medicinal Chemistry* **45**, 2811-2823.
31. Beasley, D. W. C. & Aaskov, J. G. (2001). Epitopes on the Dengue 1 Virus Envelope Protein Recognized by Neutralizing IgM Monoclonal Antibodies. *Virology* **279**, 447-458.
32. Serafin, I. L. & Aaskov, J. G. (2001). Identification of epitopes on the envelope (E) protein of dengue 2 and dengue 3 viruses using monoclonal antibodies. *Archives of Virology* **146**, 2469-2479.
33. Monath, T. P., Arroyo, J., Levenbook, I., Zhang, Z. X., Catalan, J., Draper, K. & Guirakhoo, F. (2002). Single mutation in the flavivirus envelope protein hinge region increases neurovirulence for mice and monkeys but decreases viscerotropism for monkeys: relevance to development and safety testing of live, attenuated vaccines. *Journal of Virology* **76**, 1932-1943.
34. Hurrelbrink, R. J. & McMinn, P. C. (2001). Attenuation of Murray Valley encephalitis virus by site-directed mutagenesis of the hinge and putative receptor-binding regions of the envelope protein. *Journal of Virology* **75**, 7692-7702.
35. Guirakhoo, F., Zhang, Z., Myers, G., Johnson, B. W., Pugachev, K., Nichols, R., Brown, N., Levenbook, I., Draper, K., Cyrek, S., Lang, J., Fournier, C., Barrere, B., Delagrave, S. & Monath, T. P. (2004). A single amino acid substitution in the envelope protein of chimeric yellow fever-dengue 1 vaccine virus reduces neurovirulence for suckling mice and viremia/viscerotropism for monkeys. *Journal of Virology* **78**, 9998-10008.
36. Chopra, I. & Roberts, M. (2001). Tetracycline antibiotics: mode of action, applications, molecular biology, and epidemiology of bacterial resistance. *Microbiology and Molecular Biology Reviews* **65**, 232-260.
37. Brodersen, D. E., Clemons, J. W. M., Carter, A. P., Morgan-Warren, R. J., Wimberly, B. T. & Ramakrishnan, V. (2000). The Structural Basis for the Action of the Antibiotics Tetracycline, Pactamycin, and Hygromycin B on the 30S Ribosomal Subunit. *Cell* **103**, 1143-1154.

38. Connell, S. R., Tracz, D. M., Nierhaus, K. H. & Taylor, D. E. (2003). Ribosomal protection proteins and their mechanism of tetracycline resistance. *Antimicrobial Agents and Chemotherapy* **47**, 3675-3681.
39. Orth, P., Schnappinger, D., Hillen, W., Saenger, W. & Hinrichs, W. (2000). Structural basis of gene regulation by the tetracycline inducible Tet repressor–operator system. *Nature* **7**, 215-219.
40. Laurie, A. T. & Jackson, R. M. (2005). Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* **21**, 1908-1916.
41. Burley, S. K., Almo, S. C., Bonanno, J. B., Capel, M., Chance, M. R., Gaasterland, T., Lin, D., Sali, A., Studier, F. W. & Swaminathan, S. (1999). Structural genomics: beyond the human genome project. *Nature Genetics* **23**, 151-157.
42. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389-3402.
43. Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C. & Murzin, A. G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Research* **32**, D226-D229.
44. Kolodny, R., Koehl, P., Guibas, L. & Levitt, M. (2002). Small libraries of protein fragments model native protein structures accurately. *Journal of Molecular Biology* **323**, 297-307.
45. Banner, D. W., Bloomer, A. C., Petsko, G. A., Phillips, D. C. & Wilson, I. A. (1976). Atomic coordinates for triose phosphate isomerase from chicken muscle. *Biochemical and Biophysical Research Communications* **72**, 146-155.
46. Kumar, S. & Bansal, M. (1998). Geometrical and sequence characteristics of α -helices in globular proteins. *Biophysical Journal* **75**, 1935-1944.
47. Barlow, D. J. & Thornton, J. M. (1988). Helix geometry in proteins. *Journal of Molecular Biology* **201**, 601-619.
48. Takano, K., Yamagata, Y. & Yutani, K. (2000). Role of amino acid residues at turns in the conformational stability and folding of human lysozyme. *Biochemistry* **39**, 8655-8665.
49. Milner-White, E. J. (1988). Recurring loop motif in proteins that occurs in righthanded and left-handed forms. Its relationship with α -helices and b-bulge loops. *Journal of Molecular Biology* **199**, 503-511.
50. Hutchinson, E. G. & Thornton, J. M. (1996). PROMOTIF—a program to identify and analyze structural motifs in proteins. *Protein Science* **5**, 212-220.
51. Aung, Z. & Tan, K. L. (2004). Rapid 3D protein structure database searching using information retrieval techniques. *Bioinformatics* **20**, 1045-1052.
52. Doman, T. N., McGovern, S. L., Witherbee, B. J., Kasten, T. P., Kurumbail, R., Stallings, W. C., Connolly, D. T. & Shoichet, B. K. (2002). Molecular docking and high-throughput

- screening for novel inhibitors of protein tyrosine phosphatase-1B. *Journal of Medicinal Chemistry* **45**, 2213-2221.
53. Kubinyi, H. (1997). QSAR and 3-D QSAR in drug design. 1. Methodology. *Drug Discovery Today* **2**, 457-467.
54. Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, T. E. (1982). A geometric approach to macromolecular-ligand interactions. *Journal of Molecular Biology* **161**, 269-288.
55. Ewing, T. J., Makino, S., Skillman, A. G. & Kuntz, I. D. (2001). DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *Journal of Computer-Aided Molecular Design* **15**, 411-428.
56. Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K. & Olson, A. J. (1998). Automated docking using a lamarckian genetic algorithm and empirical binding free energy function. *Journal of Computational Chemistry* **19**, 1639-1662.
57. Gohlke, H., Hendlich, M. & Klebe, G. (2000). Knowledge-based scoring function to predict protein-ligand interactions. *Journal of Molecular Biology* **295**, 337-356.
58. Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S. J. & Weiner, P. (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society* **106**, 765-784.
59. Gehlhaar, D. K., Verkhivker, G. M., Rejto, P., Sherman, C. J., Fogel, D. B., Fogel, L. J. & Freer, S. T. (1995). Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chemistry & Biology* **2**, 317-324.
60. Stahl, M. & Rarey, M. (2001). Detailed analysis of scoring functions for virtual screening. *Journal of Medicinal Chemistry* **44**, 1035-1042.
61. Langer, T. & Krovat, E. M. (2003). Chemical feature-based pharmacophores and virtual library screening for discovery of new leads. *Current Opinion in Drug Discovery & Development* **6**, 370-376.
62. Fradera, X., Knegtel, R. M. A. & Mestres, J. (2000). Similarity-driven flexible ligand docking. *Proteins: Structure, Function, and Bioinformatics* **40**, 623-637.
63. Hindle, S. A., Rarey, M., Buning, C. & Lengauer, T. (2002). Flexible docking under pharmacophore type constraints. *Journal of Computer-Aided Molecular Design* **16**, 129-149.
64. Pegg, S. C.-H., Haresco, J. J. & Kuntz, I. D. (2001). A genetic algorithm for structure-based de novo design. *Journal of Computer-Aided Molecular Design* **15**, 911-933.
65. Muegge, I., Martin, Y. C., Hajduk, P. J. & Fesik, S. W. (1999). Evaluation of PMF scoring in docking weak ligands to the FK506 binding protein. *Journal of Medicinal Chemistry* **42**, 2498-2503.
66. Yang, J.-M. (2004). Development and evaluation of a generic evolutionary method for protein-ligand docking. *Journal of Computational Chemistry* **25**, 843-857.

67. Good, A. C., Cheney, D. L., Sitkoff, D. F., Tokarski, J. S., Stouch, T. R., Bassolino, D. A., Krystek, S. R., Li, Y., Mason, J. S. & Perkins, T. D. (2003). Analysis and optimization of structure-based virtual screening protocols. 2. Examination of docked ligand orientation sampling methodology: mapping a pharmacophore for success. *Journal of Molecular Graphics and Modelling* **22**, 31-40.
68. Joseph-McCarthy, D., Thomas, B. E. I., Belmarsh, M., Moustakas, D. & Alvarez, J. C. (2003). Pharmacophore-based molecular docking to account for ligand flexibility. *Proteins: Structure, Function, and Bioinformatics* **51**, 172-188.
69. Miller, C. P. (2002). SERMs: evolutionary chemistry, revolutionary biology. *Current Pharmaceutical Design* **8**, 2089-2111.
70. Dutertre, M. & Smith, C. L. (2000). Molecular mechanisms of selective estrogen receptor modulator (SERM) action. *The Journal of Pharmacology and Experimental Therapeutics* **295**, 431-437.
71. Shiau, A. K., Barstad, D., Loria, P. M., Cheng, L., Kushner, P. J., Agard, D. A. & Greene, G. L. (1998). The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell* **95**, 927-937.
72. van Lipzig, M. M., ter Laak, A. M., Jongejan, A., Vermeulen, N. P., Wamelink, M., Geerke, D. & Meerman, J. H. (2004). Prediction of ligand binding affinity and orientation of xenoestrogens to the estrogen receptor by molecular dynamics simulations and the linear interaction energy method. *Journal of Medicinal Chemistry* **47**, 1018-1030.
73. Warnmark, A., Treuter, E., Gustafsson, J. A., Hubbard, R. E., Brzozowski, A. M. & Pike, A. C. (2002). Interaction of transcriptional intermediary factor 2 nuclear receptor box peptides with the coactivator binding site of estrogen receptor alpha. *The Journal of Biological Chemistry* **277**, 21862-21868.
74. Brzozowski, A. M., Pike, A. C., Dauter, Z., Hubbard, R. E., Bonn, T., Engstroem, O., Oehman, L., Greene, G. L., Gustafsson, J. A. & Carlquist, M. (1997). Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature* **389**, 753-758.
75. Renaud, J., Bischoff, S. F., Buhl, T., Floersheim, P., Fournier, B., Halleux, C., Kallen, J., Keller, H., Schlaeppli, J. M. & Stark, W. (2003). Estrogen receptor modulators: identification and structure-activity relationships of potent ER-alpha-selective tetrahydroisoquinoline ligands. *Journal of Medicinal Chemistry* **46**, 2945-2957.
76. Gust, R., Keilitz, R. & Schmidt, K. (2002). Synthesis, structural evaluation, and estrogen receptor interaction of 2,3-diarylpiperazines. *Journal of Medicinal Chemistry* **45**, 2325-2337.
77. Garg, R., Kapur, S. & Hansch, C. (2001). Radical toxicity of phenols: a reference point for obtaining perspective in the formulation of QSAR. *Medicinal Research Reviews* **21**, 73-82.
78. Fisher, L. S. & Guner, O. F. (2002). Seeking novel leads through structure-based pharmacophore design. *Journal of The Brazilian Chemical Society* **13**, 777-787.

79. Bissantz, C., Folkers, G. & Rognan, D. (2000). Protein-Based Virtual Screening of Chemical Databases. 1.Evaluation of Different Docking/Scoring Combinations. *J Med Chem* **43**, 4759-4767.
80. Sadowski, J., Gasteiger, J. & Klebe, G. (1994). Comparison of automatic three-dimensional model builders using 639 x-ray structures. *Journal of Chemical Information and Computer Sciences* **34**, 1000-1008.
81. Blundell, T. L., Jhoti, H. & Abell, C. (2002). High-throughput crystallography for lead discovery in drug design. *Nature Reviews Drug Discovery* **1**, 45-54.
82. Lyne, P. D. (2002). Structure-based virtual screening: an overview. *Drug Discovery Today* **7**, 1047-55.
83. Stahl, M. & Schulz-Gasch, T. (2003). Practical database screening with docking tools. *Ernst Schering Res Found Workshop*, 127-151.
84. Pearlman, D. A. & Charifson, P. S. (2001). Are free energy calculations useful in practice? A comparison with rapid scoring functions for the p38 MAP kinase protein system. *Journal of Medicinal Chemistry* **44**, 3417-23.
85. Kallblad, P., Mancera, R. L. & Todorov, N. P. (2004). Assessment of multiple binding modes in ligand-protein docking. *Journal of Medicinal Chemistry* **47**, 3334-7.
86. Nakano, T., Kaminuma, T., Sato, T., Fukuzawa, K. & Akiyama, Y. (2002). Fragment molecular orbital method: use of approximate electrostatic potential. *The Journal of Chemical Physics* **351**, 475-480.
87. CARHART, R. E., SMITH, D. H. & VENKATARAGHAVAN, R. (1985). Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences*, 64-73.
88. Jain, A. N. (2004). Ligand-based structural hypotheses for virtual screening. *Journal of Medicinal Chemistry* **47**, 947-961.
89. CARHART, R. E., SMITH, D. H. & VENKATARAGHAVAN, R. (1985). Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *Journal of chemical information and computer sciences*, 64-13.
90. Dubes, R. & Jain, A. K. (1980). Clustering methodologies in exploratory data analysis. *Advances in Computers*, 113-228.
91. The MathWorks, I. & Natick, M. (2006). MATLAB, version 7.0.
92. van Lipzig, M. M., ter Laak, A. M., Jongejan, A., Vermeulen, N. P., Wamelink, M., Geerke, D. & Meerman, J. H. (2004). Prediction of ligand binding affinity and orientation of xenoestrogens to the estrogen receptor by molecular dynamics simulations and the linear interaction energy method. *Journal of Medicinal Chemistry* **47**, 1018-30.
93. Birch, L., Murray, C. W., Hartshorn, M. J., Tickle, I. J. & Verdonk, M. L. (2002). Sensitivity of molecular docking to induced fit effects in influenza virus neuraminidase. *Journal of Computer-Aided Molecular Design* **16**, 855-69.

94. Shiau, A. K., Barstad, D., Loria, P. M., Cheng, L., Kushner, P. J., Agard, D. A. & Greene, G. L. (1998). The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell* **95**, 927-37.
95. Cody, V., Galitsky, N., Luft, J. R., Pangborn, W., Blakley, R. L. & Gangjee, A. (1998). Comparison of ternary crystal complexes of F31 variants of human dihydrofolate reductase with NADPH and a classical antitumor furopyrimidine. *Anti-cancer Drug Design* **13**, 307-15.
96. Varghese, J. N., Colman, P. M., van Donkelaar, A., Blick, T. J., Sahasrabudhe, A. & McKimm-Breschkin, J. L. (1997). Structural evidence for a second sialic acid binding site in avian influenza virus neuraminidases. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 11808-12.
97. Cheng, W. C., Chang, Y. N. & Wang, W. C. (2005). Structural basis for shikimate-binding specificity of *Helicobacter pylori* shikimate kinase. *Journal of Bacteriology* **187**, 8156-8163.
98. Matter, H. (1997). Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors. *Journal of Medicinal Chemistry* **40**, 1219-29.
99. Zheng, W. & Tropsha, A. (2000). Novel variable selection quantitative structure--property relationship approach based on the k-nearest-neighbor principle. *Journal of Chemical Information and Computer Sciences* **40**, 185-194.
100. DiMasi, J. A., Hansen, R. W. & Grabowski, H. G. (2003). The price of innovation: new estimates of drug development costs. *Journal of Health Economics* **22**, 151-185.
101. Charifson, P. S., Corkery, J. J., Murcko, M. A. & Walters, W. P. (1999). 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.
102. Ng, K. B. & Kantor, P. B. (2000). Predicting the effectiveness of naive data fusion on the basis of system characteristics. *Journal of the American Society of Information Science* **51**, 1177-1189.
103. Belkin, N. J., Kantor, P. B., Fox, E. A. & Shaw, J. A. (1995). Combining evidence of multiple query representation for information retrieval. *Information Processing and Management* **31**, 431-448.
104. Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E. & Schuffenhauer, A. (2004). Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *Journal of Chemical Information and Computer Sciences* **44**, 1177-85.
105. Salim, N., Holliday, J. & Willett, P. (2003). Combination of fingerprint-based similarity coefficients using data fusion. *Journal of Chemical Information and Computer Sciences* **43**, 435-42.
106. Hsu, D. F., Shapiro, J. & Taksa, I. (2002). Methods of data fusion in information retrieval: rank vs. score combination. *DIMACS Technical Report* **58**, 1-47.

107. Hsu, D. F. & Taksa, I. (2005). Comparing rank and score combination methods for data fusion in information retrieval. *Information Retrieval*, (In press).
108. Verdonk, M. L., Berdini, V., Hartshorn, M. J., Mooij, W. T., Murray, C. W., Taylor, R. D. & Watson, P. (2004). Virtual screening using protein-ligand docking: avoiding artificial enrichment. *Journal of Chemical Information and Computer Sciences* **44**, 793-806.
109. Wang, R. & Wang, S. (2001). How does consensus scoring work for virtual library screening? An idealized computer experiment. *Journal of Chemical Information and Computer Sciences* **41**, 1422-6.
110. Chuang, H. Y., Liu, H. F., Chen, F. A., Kao, C.-Y. & Hsu, D. F. (2004). *Proceedings of I-SPAN'04*.
111. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research* **28**, 235-242.
112. Yang, J.-M. & Shen, T.-W. (2004). A pharmacophore-based evolutionary approach for screening estrogen receptor antagonists. *Congress of Evolutionary Computation (CEC 2004)*, 1028-1035.
113. Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V. & Mee RP, R. P. (1997). Empirical scoring functions: 1. the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *Journal of Computer-Aided Molecular Design* **11**, 425-445.
114. Raymond, J. W., Jalaie, M. & Bradley, M. P. (2004). Conditional probability: a new fusion method for merging disparate virtual screening results. *Journal of Chemical Information and Computer Sciences* **44**, 601-9.
115. Hsu, D. F. & Palumbo, A. (2004). *Proceedings of the 7th International Symposium on Parallel Architectures, Algorithms, and Networks (I-SPAN)*.
116. Wiseman, T., Williston, S., Brandts, J. F. & Lin, L. N. (1989). Rapid measurement of binding constants and heats of binding using a new titration calorimeter. *Analytical Biochemistry* **179**, 131-137.
117. Raffa, R. B. & Porreca, F. (1989). Thermodynamic analysis of the drug-receptor interaction. *Life Sciences* **44**, 245-258.
118. Porstmann, T. & Kiessig, S. T. (1992). Enzyme immunoassay techniques. An overview. *Journal of Immunological Methods* **150**, 5-21.
119. Villar, H. O., Yan, J. & Hansen, M. R. (2004). Using NMR for ligand discovery and optimization. *Current Opinion in Chemical Biology* **8**, 387-391.
120. Lofas, S. (2004). Optimizing the hit-to-lead process using SPR analysis. *Assay and Drug Development Technologies* **2**, 407-415.
121. Beydon, M. H., Fournier, A., Drugeault, L. & Becquart, J. (2000). Microbiological high throughput screening: an opportunity for the lead discovery process. *Journal of Biomolecular Screening* **5**, 13-22.

122. Wang, R., Lai, L. & Wang, S. (2002). Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *Journal of Computer-Aided Molecular Design* **16**, 11-26.
123. Puvanendrapillai, D. & Mitchell, J. B. (2003). L/D Protein Ligand Database (PLD): additional understanding of the nature and specificity of protein-ligand complexes. *Bioinformatics* **19**, 1856-1857.
124. Wang, R., Fang, X., Lu, Y. & Wang, S. (2004). The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *Journal of Medicinal Chemistry* **47**, 2977-2980.
125. Warnmark, A., Treuter, E., Gustafsson, J. A., Hubbard, R. E., Brzozowski, A. M. & Pike, A. C. (2002). Interaction of transcriptional intermediary factor 2 nuclear receptor box peptides with the coactivator binding site of estrogen receptor alpha. *Journal of Biological Chemistry* **277**, 21862-21868.
126. Smith, G. M., Alexander, R. S., Christianson, D. W., McKeever, B. M., Ponticello, G. S., Springer, J. P., Randall, W. C., Baldwin, J. J. & Habecker, C. N. (1994). Positions of His-64 and a bound water in human carbonic anhydrase II upon binding three structurally related inhibitors. *Protein Science* **3**, 118-125.
127. Bissantz, C., Folkers, G. & Rognan, D. (2000). Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *Journal of Medicinal Chemistry* **43**, 4759-4767.
128. Grzybowski, B. A., Ishchenko, A. V., Kim, C. Y., Topalov, G., Chapman, R., Christianson, D. W., Whitesides, G. M. & Shakhnovich, E. I. (2002). Combinatorial computational method gives new picomolar ligands for a known enzyme. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 1270-1273.
129. Stams, T., Chen, Y., Boriack-Sjodin, P. A., Hurt, J. D., Liao, J., May, J. A., Dean, T., Laipis, P., Silverman, D. N. & Christianson, D. W. (1998). Structures of murine carbonic anhydrase IV and human carbonic anhydrase II complexed with brinzolamide: molecular basis of isozyme-drug discrimination. *Protein Science* **7**, 556-563.
130. Hakansson, K. & Liljas, A. (1994). The structure of a complex between carbonic anhydrase II and a new inhibitor, trifluoromethane sulphonamide. *FEBS Letters* **350**, 319-322.
131. Kim, C.-Y., Chang, J. S., Doyon, J. B., Baird, T. T., Fierke, C. A., Jain, A. & Christianson, D. W. (2000). Contribution of Fluorine to Protein-Ligand Affinity in the Binding of Fluoroaromatic Inhibitors to Carbonic Anhydrase II. *Journal of the American Chemical Society* **122**, 12125-12134.
132. Gutteridge, A. & Thornton, J. M. (2005). Understanding nature's catalytic toolkit. *Trends in Biochemical Sciences* **30**, 622-629.
133. Chen, Y.-C., Lo, Y.-S., Hsu, W.-C. & Yang, J.-M. (2007). 3D-partner: a web server to infer interacting partners and binding models. *Nucleic Acids Research*, W561-W567.

134. Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N. & Yeh, L. S. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research* **32**, D115-D119.
135. Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 10915-10919.
136. Kubinyi, H. (1997). QSAR and 3-D QSAR in drug design. 2. Applications and problems. *Drug Discovery Today* **2**, 538-546.
137. Osterberg, F., Morris, G. M., Sanner, M. F., Olson, A. J. & Goodsell, D. S. (2002). Automated docking to multiple target structures: Incorporation of protein mobility and structural water heterogeneity in autodock. *Proteins: Structure, Function, and Bioinformatics* **46**, 34-40.
138. Draper, N. & Smith, H. (1981). *Applied regression analysis*. 2nd edit, John Wiley and Sons.
139. Rarey, M., Kramer, B., Lengauer, T. & Klebe, G. (1996). A fast flexible docking method using an incremental construction algorithm. *Journal of Molecular Biology* **261**, 470-489.
140. Xiang, S., Short, S. A., Wolfenden, R. & Carter, C. W., Jr. (1995). Transition-state selectivity for a single hydroxyl group during catalysis by cytidine deaminase. *Biochemistry* **34**, 4516-4523.
141. Vincent, F., Gloster, T. M., Macdonald, J., Morland, C., Stick, R. V., Dias, F. M., Prates, J. A., Fontes, C. M., Gilbert, H. J. & Davies, G. J. (2004). Common inhibition of both beta-glucosidases and beta-mannosidases by isofagomine lactam reflects different conformational itineraries for pyranoside hydrolysis. *ChemBiochem* **5**, 596-1599
142. Sleigh, S. H., Seavers, P. R., Wilkinson, A. J., Ladbury, J. E. & Tame, J. R. (1999). Crystallographic and calorimetric analysis of peptide binding to OppA protein. *Journal of Molecular Biology* **291**, 393-415.
143. Davies, T. G., Hubbard, R. E. & Tame, J. R. (1999). Relating structure to thermodynamics: the crystal structures and binding affinity of eight OppA-peptide complexes. *Protein Science* **8**, 1432-1444
144. Krammer, A., Kirchhoff, P. D., Jiang, X., Venkatachalam, C. M. & Waldman, M. (2005). LigScore: a novel scoring function for predicting binding affinities. *Journal of Molecular Graphics & Modelling* **23**, 395-407.
145. Bohm, H. J. (1992). The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *Journal of Computer-Aided Molecular Design* **6**, 61-78.
146. Cozzini, P., Fornabaio, M., Marabotti, A., Abraham, D. J., Kellogg, G. E. & Mozzarelli, A. (2002). Simple, intuitive calculations of free energy of binding for protein-ligand complexes. 1. Models without explicit constrained water. *Journal of Medicinal Chemistry* **45**, 2469-2483.

147. Conklin, D. J., Langford, S. D. & Boor, P. J. (1998). Contribution of serum and cellular semicarbazide-sensitive amineoxidase to amine metabolism and cardiovascular toxicity. *Toxicological Sciences* **46**, 386-392.
148. Perez, C., Pastor, M., Ortiz, A. R. & Gago, F. (1998). Comparative binding energy analysis of HIV-1 protease inhibitors: incorporation of solvent effects and validation as a powerful tool in receptor-based drug design. *Journal of Medicinal Chemistry* **41**, 836-852.
149. Ortiz, A. R., Pisabarro, M. T., Gago, F. & Wade, R. C. (1995). Prediction of drug binding affinities by comparative binding energy analysis. *Journal of Medicinal Chemistry* **38**, 2681-2691.
150. Green, S. M. & Marshall, G. R. (1995). 3D-QSAR: a current perspective. *Trends in Pharmacological Sciences* **16**, 285-291.
151. Chen, Y. C., Yang, J. M., Tsai, C. H. & Kao, C. Y. (2005). GEMPLS: A new QSAR method combining generic evolutionary method and partial least squares. *Lecture Notes in Computer Science* **3449**, 125-135.
152. Guo, J., Hurley, M. M., Wright, J. B. & Lushington, G. H. (2004). A docking score function for estimating ligand-protein interactions: application to acetylcholinesterase inhibition. *Journal of Medicinal Chemistry* **47**, 5492-5500.
153. Sippl, W., Contreras, J. M., Parrot, I., Rival, Y. M. & Wermuth, C. G. (2001). Structure-based 3D QSAR and design of novel acetylcholinesterase inhibitors. *Journal of Computer-Aided Molecular Design* **15**, 395-410.
154. Sippl, W. (2002). Development of biologically active compounds by combining 3D QSAR and structure-based design methods. *Journal of Computer-Aided Molecular Design* **16**, 825-830.
155. Kua, J., Zhang, Y. & McCammon, J. A. (2002). Studying enzyme binding specificity in acetylcholinesterase using a combined molecular dynamics and multiple docking approach. *Journal of the American Chemical Society* **124**, 8260-8267.
156. Bernard, P., Kireev, D. B., Chretien, J. R., Fortier, P. L. & Coppet, L. (1999). Automated docking of 82 N-benzylpiperidine derivatives to mouse acetylcholinesterase and comparative molecular field analysis with natural alignment. *Journal of Computer-Aided Molecular Design* **13**, 355-371.
157. Cho, S. J., Garsia, M. L., Bier, J. & Tropsha, A. (1996). Structure-based alignment and comparative molecular field analysis of acetylcholinesterase inhibitors. *Journal of Medicinal Chemistry* **39**, 5064-5071.
158. Kryger, G., Harel, M., Giles, K., Toker, L., Velan, B., Lazar, A., Kronman, C., Barak, D., Ariel, N., Shafferman, A., Silman, I. & Sussman, J. L. (2000). Structures of recombinant native and E202Q mutant human acetylcholinesterase complexed with the snake-venom toxin fasciculin-II. *Acta Crystallographica. Section D, Biological crystallography* **56**, 1385-1394.

159. Kryger, G., Silman, I. & Sussman, J. L. (1999). Structure of acetylcholinesterase complexed with E2020 (Aricept): implications for the design of new anti-Alzheimer drugs. *Structure with Folding and Design* **7**, 297-307.
160. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410.
161. Kishishita, S., Okajima, T., Kim, M., Yamaguchi, H., Hirota, S., Suzuki, S., Kuroda, S., Tanizawa, K. & Mure, M. (2003). Role of copper ion in bacterial copper amine oxidase: spectroscopic and crystallographic studies of metal-substituted enzymes. *Journal of the American Chemical Society* **125**, 1041-1055.
162. Wilmot, C. M., Hajdu, J., McPherson, M. J., Knowles, P. F. & Phillips, S. E. (1999). Visualization of dioxygen bound to copper during enzyme catalysis. *Science* **286**, 1724-1728.
163. Schwede, T., Kopp, J., Guex, N. & Peitsch, M. C. (2003). SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Research* **31**, 3381-3385.
164. Boomsma, F., Derkx, F. H., van den Meiracker, A. H., Man in 't Veld, A. J. & Schalekamp, M. A. (1995). Plasma semicarbazide-sensitive amine oxidase activity is elevated in diabetes mellitus and correlates with glycosylated haemoglobin. *Clinical Science* **88**, 675-679.
165. O'Sullivan, J., Unzeta, M., Healy, J., O'Sullivan, M. I., Davey, G. & Tipton, K. F. (2004). Semicarbazide-sensitive amine oxidases: enzymes with quite a lot to do. *Neurotoxicology* **25**, 303-315.
166. O'Connell, K. M., Langley, D. B., Shepard, E. M., Duff, A. P., Jeon, H. B., Sun, G., Freeman, H. C., Guss, J. M., Sayre, L. M. & Dooley, D. M. (2004). Differential inhibition of six copper amine oxidases by a family of 4-(aryloxy)-2-butynamines: evidence for a new mode of inactivation. *Biochemistry* **43**, 10965-10978.

計畫成果自評(Self-evaluation of The Project Achievements)

We have published eight journal papers, one conference paper, five posters and won the 2007 national innovation award. Eight masters were supported by this research project. The details are described as bellow.

Awards:

1. 2007 National Innovation Award, Yen-Fu Chen, Yu-Ju Chen, and Jinn-Moon Yang, "GEMDOCK: An Integrated Environment for Computer-aided Drug Design and Its Applications", Taiwan

Journal papers:

1. Y.-Y. Chiu, J.-K. Hwang, J.-M. Yang*, "Soft energy function and generic evolutionary method for discriminating native from non-native protein conformations," *Journal of Computational Chemistry*, vol. 29, pp. 1364-1373, 2008 (SCI, IF: 4.89)
2. M.-C. Yang, H.-H. Guan, M.-Y. Liu, Y.-H. Lin, J.-M. Yang, W.-L. Chen, C.-J. Chen, and Simon J. T. Mao*, "Crystal structure of a secondary vitamin D3 binding site of milk β -lactoglobulin," *Proteins: Structure, Function, and Bioinformatics*, vol. 71, pp. 1197-1210, 2008. (SCI, IF: 3.73)
3. Y.Y. Yao, K.L. Shrestha, Y.J. Wu, H.J. Tasi, C.C. Chen, J.-M. Yang, A. Ando, C.Y. Cheng, Y.K. Li*, "Structural simulation and protein engineering to convert an endo-chitosanase to an exo-chitosanase," *Protein Engineering, Design & Selection*, 2008, in press. (SCI, IF: 3.0)
4. C.-H. Tung, J.-W. Huang and J.-M. Yang*, "Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for fast protein structure database search," *Genome Biology*, vol. 8, pp. R31.1~R31.16, 2007. (SCI, IF: 7.17)
5. C.-H. Tung and J.-M. Yang*, "fastSCOP: a fast web server for recognizing protein structural domains and SCOP superfamilies," *Nucleic Acids Research*, pp. W438-W443, 2007. (SCI, IF: 6.31)
6. J.-M. Yang, Y.-F. Chen, Y.-Y. Tu, K.-R. Yen, and Y.-L. Yang*, "Combinatorial computation approaches identifying tetracycline derivatives as flaviviruses inhibitors," *PLoS ONE*, pp. e428.1-e428.12, 2007.
7. J.-M. Yang* and T.-W. Shen, "A pharmacophore-based evolutionary approach for screening selective estrogen receptor modulators," *Proteins: Structure, Function, and Bioinformatics*, vol. 59, pp. 205-220, 2005. (SCI, IF: 3.73) (Times Cited: 13)
8. 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. (SCI, IF: 3.2) (Times Cited: 25)

Conferences Papers:

1. K-C Hsu, Y-F Chen, and J-M Yang*, "Binding affinity analysis of protein-ligand complexes," 2nd International Conference on Bioinformatics and Biomedical Engineering, pp. 167-171, 2008.

Posters

1. Y.-F. Chen, L.-J. Chang, J.-M. Yang*, "Integrating GEMDOCK with GEM-PLS and GEM-kNN for QSAR modeling of huAChE and AGHO," in 15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 6th European Conference on Computational Biology (ECCB), Vienna, Austria, 2007.
2. C.-H. Tung, T.-K. Yang, and J.-M. Yang*, "Structural Binding Pocket Clustering and Protein-Ligand Interaction Analysis for ATP-binding Proteins," in 15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 6th European Conference on Computational Biology (ECCB), Vienna, Austria, 2007.
3. J.-M. Yang, Y.-F. Chen, C.-Y. Chen and Y.-L. Yang, "Identifying Two Tetracycline-Derivates as Effective Novel Inhibitors on the Propagation of Dengue Virus Type 2 Using Virtual Screening against the Envelope Protein", in Annual Conference on Biotechnology, Hsinchu, Taiwan, 2006 (Excellent work)
4. C.-N. Ko, Y.-F. Chen, Y.-J. Chen and J.-M. Yang, "Cluster analysis of Structure-based Virtual Screening by Using Protein-ligand Interactions and Compound Structures", in Annual Conference on Biotechnology, Hsinchu, Taiwan, 2007 (Award)
5. Y.-T. Chen and J.-M. Yang, "A New Profile Method for Predicting Protein-ligand Binding Site", in 2008 Annual Conference on Biotechnology, Hsinchu, Taiwan, 2008 (Award)

Databases and web-based services

GEMDOCK: <http://gemdock.life.nctu.edu.tw/dock/>

Binding site analysis: http://gemdock.life.nctu.edu.tw/cavity_web/

3D-BLAST: <http://3d-blast.life.nctu.edu.tw/>

Awards in the past three years

Table 1. The awards of principal investigators during 2005-2008

Name of PI	Date	Prize
J.-M. Yang	2006	獲得國立交通大學 2006 年傑出人士榮譽獎勵
J.-M. Yang	2007	國家新創獎
J.-M. Yang	2007~	生物資訊協會理事
J.-M. Yang	2005	指導研究生獲資訊學會碩博士論文獎佳作獎

Table 2. The awards of graduate students joined in this project

Student	Professor	Date	Prize
陳佑德	J.-M. Yang	2008	交通大學生物科技學院 2008 生物科技學術壁報競賽優等
陳彥甫 陳右儒	J.-M. Yang	2007	國家新創獎第三名
陳彥甫	J.-M. Yang	2007	2007 年生物科技學術研討會暨壁報比賽 (優等)
董其樺	J.-M. Yang	2007	2007 年生物科技學術研討會暨壁報比賽 (優等)
董其樺	J.-M. Yang	2006	2006 年生物科技學術研討會暨壁報比賽 (優等)
陳彥甫	J.-M. Yang	2006	2006 年生物科技學術研討會暨壁報比賽 (佳作)
董其樺	J.-M. Yang	2005	資訊學會最佳碩博士論文

Table 3. Summary of conferences that our students have joined during 2005-2008

Student	Professor	Date	Conference
許凱程	J.-M. Yang	2008/05	The 2 nd International Conference on Bioinformatics and Biomedical Engineering (iCBBE2008)
董其樺	J.-M. Yang	2007/08	The 15 th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB)
陳彥甫	J.-M. Yang	2007/08	The 15 th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB)

In summary, we believe that we have achieved fruitful results in this project. This project covers virtual screening, pharmacophore identification, post-analysis of virtual screening, and prediction of binding affinity and QSAR analysis. These four parts construct an efficient and fast platform for drug discovery. We consider that the achievements in this project will be advantageous and valuable to researchers to study computer-aided drug design.