

GEMDOCK 於虛擬藥物資料庫篩選之功能增進及套膜蛋白與醃亞胺水解酵素之實際應用

學生：沈再威

指導教授：楊進木

國立交通大學生物資訊所碩士班

摘 要

針對已知結構的蛋白質，應用虛擬藥物篩選的工具在化合物資料庫中尋找潛在的抑制劑，是目前電腦輔助藥物設計廣為使用的方法之一。其目的在於有效地縮短尋找潛在抑制劑的時間，並減低實驗所需的成本。在本篇論文中，除了希望將 GEMDOCK 全程自動化之外，還希望藉著計分程式的修正以增強 GEMDOCK 在虛擬藥物篩選上的功能。GEMDOCK 的演算法同時具有區域及全域搜尋最佳解的優點，而且修正後的計分程式不但提高了分子鉗合的準確度，並在虛擬資料庫篩選上有明顯的改進，確實減少了偽陽性 (false positive) 的數目。本研究先以 TK (thymidine kinase)、ER (estrogen receptor) 及 hDHFR (human dihydrofolate reductase) 為標的蛋白，從化合物資料庫 (ACD、MDDR) 中隨機挑選出 990 個化合物，再加上已知會與標的蛋白結合的配體各 10 個，做為測試組，以 GEMDOCK 預測這 1,000 個化合物分別與蛋白質鉗合的位置及能量，將結果依能量排序，以觀察 GEMDOCK 在篩選化合物資料庫上的表現。結果在 true hit% 達 100% 時，偽陽性的比例 (false positive rate) 分別為 9.7% (TK)、5.2% (ER antagonists)、21.2% (ER agonists) 及 8.6% (hDHFR)；若在蛋白質活性區域中重要的胺基酸上，根據 10 個已知配體的特性來加重計分，則偽陽性的比例分別減少為 2.9% (TK)、0.9% (ER antagonists)、1.9% (ER agonists) 及 2.0% (hDHFR)。在確認了 GEMDOCK 在篩選上的表現之後，我們將 GEMDOCK 應用於登革病毒套膜蛋白 (Envelope protein) 的抑制劑篩選上，登革熱是台灣夏季的流行性疾病，而登革病毒的套膜蛋白則為可能的藥物設計標的。GEMDOCK 也被應用在辨識 D-hydantoinase 的基質或抑制劑上，結果我們發現兩個新的基質，並與生物實驗的結果相符。

Enhancing GEMDOCK on Virtual Database Screening and Application to Envelope Protein and D-Hydantoinase

Student: Tsai-Wei Shen

Advisor: Jinn-Moon Yang

Institute of Bioinformatics
National Chiao Tung University

ABSTRACT

Virtual ligand screening is a method broadly used for computer-aided drug design. It will save much time and cost to find potential inhibitors for the target protein with aids of computers. In this thesis, we have developed an automatic tool with a novel scoring function for virtual screening by applying numerous enhancements and modifications to our original techniques, called GEMDOCK. 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. Our scoring function is indeed able to enhance the accuracy during the flexible docking and to improve the screening utility by reducing the number of false positives in the post-docking analysis. First we have verified our program with four screening sets for thymidine kinase (TK) substrates, estrogen receptor (ER) antagonists, estrogen receptor agonists, and human dihydrofolate reductase (hDHFR) ligands. These four screening sets were composed of ten known ligands for each target protein and 990 compounds randomly selected from the ACD or MDDR. The 1,000 compounds of the four screening sets were docked into each target protein and ranked according to their potentials. When the true hit rate was 100%, the false positive rates were 9.7% for TK, 5.2% for ER antagonists, 21.2% for ER agonists, and 8.6% for hDHFR. After adding pharmacological consensuses on important residues and ligand preferences according to the ten known ligands, the false positive rates were decreased to 2.9% for TK, 0.9% for ER antagonists, 1.9% for ER agonists, and 2.0% for hDHFR. After verifying the utility of GEMDOCK on virtual screening, we applied it to identifying potential inhibitors for the envelope protein of dengue virus. Dengue fever is an epidemic disease in Taiwan during the summer. The envelope protein (the PDB entry: 1oke) of dengue virus is a possible target for drug design. Finally, GEMDOCK has been also accessed on identifying new substrate/inhibitors of *Agrobacterium radiobacter* hydantoinase, which is an industrial enzyme. We have screened candidates for hydantoinase and identify two new substrates evaluated by wet experiments.

Acknowledgements

The most appreciation is for my advisor Dr. Jinn-Moon Yang. Because of his advice and instruction, I could have a start to learn about bioinformatics and finally finish the thesis. I am very grateful for his knowledge and suggestions that helped me to correct my mistakes. Without his guides and encouragement, it is impossible to finish the thesis during these two years. Thanks for Chun-Chen Chen, his biological knowledge is really helpful to me. Thanks for Yan-Fu Chen, his chemical sense corrects many wrong concepts of mine. Thanks for Yi-Yuan Chiu, he helps me a lot about the programming. Thanks for all members in my laboratory, this is a perfect team that we could discuss each other to solve someone's problem. Finally I would like to thank my parents and my family for supporting me through these two years.



CONTENTS

Abstract (in Chinese).....	I
Abstract	II
Acknowledgements	III
Contents.....	IV
List of Tables	VI
List of Figures	VIII
Chapter 1. Introduction	01
1.1 Motivations and Purposes	01
1.2 Related Works	02
1.3 Thesis Overview.....	03
Chapter 2. Materials and Methods	05
2.1 Preparation of Target Proteins and Ligand Databases	06
2.2 Molecular Docking.....	07
Chapter 3. Evaluation GEMDOCK on Virtual Database Screening.....	16
3.1 Parameters of GEMDOCK and GOLD	16
3.2 Thymidine Kinase	17
3.3 Estrogen Receptor	22
3.4 Dihydrofolate Reductase.....	25

Chapter 4. GEMDOCK on Practical Applications.....	30
4.1 Envelope Protein of Dengue Virus Type II	30
4.2 Molecular Docking of D-Hydantoinase	33
Chapter 5. Conclusions	35
5.1 Summary	35
5.2 Major Contributions and Future Perspectives.....	35
References	78

Appendix

A. Preparation of the Screening Set: Steps and Tools.....	A-1
B. Procedures of Running GEMDOCK.....	B-1
C. Compound Format Exchange: PDB MDL Mol and SYBYL Mol2	C-1
D. Source Codes of Several Self-developed Programs.....	D-1

List of Tables

Table 1. Atom Types of GEMDOCK.....	37
Table 2. Atom Formal Charge of GEMDOCK.....	37
Table 3. Parameters of GEMDOCK.....	38
Table 4. Interaction Preferences of Hot-spot Atoms of TK Evolved by Superimposing Known Active Ligands.....	38
Table 5. Ligand Preferences Evolved from Known Ligands Are Used to Screen the Lead Compounds for TK, ER, hDHFR, and E Protein.....	39
Table 6. Comparison GEMDOCK with GOLD, FlexX, and DOCK on Docking Ten Known Substrates of the TK with X-ray Structures into Their Native Proteins and the Reference Protein, 1kim.....	40
Table 7. Comparison of GEMDOCK with Four Methods by False Positive Rates (%) on Screening 990 Compounds and Ten Known Substrates of the TK.....	41
Table 8. Pharmacological Weights of Hot-spot Atoms of the ER α -antagonist Complex and ER α -agonist Complex Are Evolved by Overlapping Known Active Ligands.....	42
Table 9. Comparison GEMDOCK with GOLD on Docking Four Antagonists and Four Agonists with X-ray Structures into Their Native Proteins and the Reference Proteins That Are 3ert and 1gwr, Respectively.....	43
Table 10. Comparison of GEMDOCK with Four Methods by False Positive Rates (%) on Screening 990 Compounds and Ten Known Antagonists.....	44
Table 11. Interaction Preferences of Hot-spot Atoms of hDHFR Evolved by Overlapping Ten Known Active Ligands.....	45
Table 12. Comparison GEMDOCK with GOLD on Docking Ten Known Ligands of the hDHFR with X-ray Structures into Their Native Proteins and the Reference Protein, 1hfr.....	46

Table 13. Comparison of GEMDOCK with GOLD by False Positive Rates (%) on Screening 990 Compounds and Ten Known Ligands of the hDHFR.....	47
Table 14. The 35 Molecules of Intersection of Top 200 Scorer from the Screening Result of GEMDOCK and GOLD	48
Table 15. Specific Activity of D-Hydantoinase	52



Table of Figures

Figure 1. The main steps of GEMDOCK for virtual database screening.....	53
Figure 2. The linear energy function of the pair-wise atoms for the steric interactions, hydrogen bonds, and electrostatic potential in GEMDOCK.....	54
Figure 3. Ten known active ligands of HSV-1 thymidine kinase.....	55
Figure 4. Superimposing ten crystal structures of TK.....	56
Figure 5. The accuracy of GEMDOCK using different combinations of pharmacological preferences in screening the substrates of TK from a testing set.....	57
Figure 6. Ten antagonists of ER α	58
Figure 7. Ten agonists of ER α	59
Figure 8. (A) Superimposing ten docked poses of ER α antagonists (B) Superimposing ten docked poses of ER α agonists.....	60
Figure 9. The accuracy of GEMDOCK using different combinations of pharmacological preferences in screening ER α antagonists from a testing set.....	61
Figure 10. The accuracy of GEMDOCK using different combinations of pharmacological preferences in screening ER α agonists from a testing set.....	62
Figure 11. Ten known active ligands of hDHFR.....	63
Figure 12. Preparation of the testing set of hDHFR.....	64
Figure 13. Superimposing ten known ligands of hDHFR.....	65
Figure 14. The accuracy of GEMDOCK using different combinations of pharmacological preferences in screening ligands of hDHFR from a testing set.....	66
Figure 15. (A) The conformational rearrangement of the E protein to the reduced pH of an endosome (B) The detergent molecule of <i>n</i> -octyl- β -D-glucoside (β -OG) complex with the E protein in the crystal structure (C) Detergent binding marks the pocket	

as a potential site for small-molecule fusion inhibitors	67
Figure 16. Results of the biological experiment for dengue virus.....	68
Figure 17. (A) Overlapping docked poses of the nine compounds tested by the biological experiment. (B) The docked pose of MCMC00007079. (C) Hydrogen bonds (dashed lines) formed between MCMC00007079 and the binding site.....	69
Figure 18. The number of contacts between the 35 compounds and residues of the E protein	70
Figure 19. The pathway to produce D-amino acid	71
Figure 20. (A) Training set: 17 molecules were substrates of the D-hydantoinase (B) Ten molecules with biological activities in the testing set.....	72
Figure 21. K_m and k_{cat} values of substrates and their docked poses.....	73
Figure 22. Superimposing 17 known substrates of D-HYD.....	74
Figure 23. (A) The docked pose of allantoin (B) The spectrum with or without adding D-HYD.....	75
Figure 24. (A) The docked pose of parabanic acid (B)The spectrum with or without adding D-HYD.....	76
Figure 25. IC ₂₀ values of inhibitors and their docked poses	77

Chapter 1

Introduction

1.1 Motivations and Purposes

Discovery of novel lead compounds through structure-based virtual screening of chemical databases against protein structures is an emerging step in computer aided drug design. It has contributed to the introduction of ~50 compounds into clinical trials and to numerous drug approvals [1]. As the number of protein structures as pharmaceutical targets is persistently increasing, virtual screening will play a major role in rational drug design. It could be considered as a powerful computational filter for reducing the size of a chemical library that will be further experimentally tested.



GEMDOCK is a docking program that has a good performance on the prediction of the target-bound conformation and orientation of docked ligands. It can predict known protein-bound ligand poses with averaged deviations below 2.0 Å [2] and model the best possible poses of ligands in the target that has no crystal of the protein-ligand complex [3]. Based on the good performance of GEMDOCK on molecular docking, we enhanced it on virtual screening large database with a suitable scoring function to reduce the number of false positives when screening large database.

In Taiwan, dengue virus is always widespread during summer and can cause severe epidemics of diseases such as dengue fever and dengue hemorrhagic fever/dengue shock syndrome. To combat diseases caused by dengue viruses, it is urgent to develop novel

antiviral therapeutic agents. The envelope protein (E protein) of dengue virus was reported recently that some specific characteristics on the structure could serve as the target for drug design [4, 5]. Therefore we apply virtual screening to the E protein for discovering novel lead compounds against dengue virus. Besides, GEMDOCK has been also accessed on identifying new substrate/inhibitors of *Agrobacterium radiobacter* hydantoinase, which is an industrial enzyme for production of D-amino acid intermediate compounds through stereospecific hydrolysis of chemically synthesized cyclic hydantoins.

1.2 Related Works

Any virtual screening method has to face two critical issues: docking method and scoring function. Several docking programs are now available and consider the ligand as a flexible molecule because a flexible docking method clearly outperforms rigid body matches. A flexible docking is generally based on one of the following methods: incremental construction (FlexX [6], Hammerhand [7], DOCK 4.0 [8]), and genetic algorithm (GOLD [9], AutoDock3.0 [10]). For the virtual screening purpose, only methods able to dock within a reasonable time scale (20-200 seconds) per ligand are suited. After the ligand has been docked, it should be scored according to the interactions between the target and ligand. Several scoring functions have been described and based on either force-field methods (DOCK [11], GOLD [9]), empirical free energy scoring functions (Ludi [12], Chemscore [13], Score [14], Fresno [15], FlexX [6], Plp [16]), or knowledge-based potential of mean force (Pmf [17], Drugscore [18]). Each scoring function has its own form, depends on a distinct atom type, assigns atomic partial charge with different calculation methods, and has been trained on different data sets of protein-ligand complexes. All of them have been validated for various high-resolution protein-ligand X-ray structures and are generally able to predict absolute binding free energies within 7-10 kJ/mol [19]. However, no single scoring function could

have the same performance on every target protein. It has been proposed that combining multiple scoring functions (consensus scoring) improves the enrichment of true positives [19, 20]. An ideal scoring function must rank a correct pose of a molecule higher than an incorrect one and predict the rank order of binding affinities of ligands to the target protein. Therefore we hope to design a scoring function that could have the variables against different proteins according to the knowledge of known active ligands or characteristics of the binding site to improve the screening utility.

GEMDOCK is a tool with an evolutionary approach for flexible ligand docking and an empirical scoring function for rapid recognition of potential ligands [21, 22]. The evolutionary approach of GEMDOCK is more robust than the standard one with regard to several specific domains. The core idea of this evolutionary approach is to design multiple operators that cooperate using a family competition paradigm that is similar to a local search procedure. One operator is a differential evolution operator to reduce the disadvantages of Gaussian and Cauchy mutations; the other one is a new rotamer-based mutation operator to reduce the search space of ligand structure conformations. The scoring function of GEMDOCK consists of electrostatic, steric, and hydrogen bonding potentials. Steric and hydrogen bonding potentials use a linear model that is simple, has fewer local minima, and recognizes potential complexes rapidly. GEMDOCK may be run as a purely flexible or hybrid docking approach. It is an automatic system that generates all related docking variables, such as atom formal charge, atom type, and the ligand binding site of a protein. To evaluate the docking accuracy, GEMDOCK was tested on a diverse data set of 100 protein-ligand complexes from the Protein Data Bank (PDB). 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 [22].

1.3 Thesis Overview

We have enhanced GEMDOCK on virtual database screening and applied it to the envelope protein and D-hydantoinase. In chapter 2, we have prepared the four screening sets and target proteins. Each screening set contained ten known ligands and 990 randomly selected compounds from the MDDR or ACD. The target proteins were all derived from the PDB. After preparing screening sets and target proteins, we modified the scoring function according to ten known ligands and tested GEMDOCK with different combinations of pharmacological consensuses and ligand preferences to evaluate the performance and to observe the comparison of search behavior with other public docking tools.

In chapter 3, we evaluated the screening performance of GEMDOCK with four screening sets (TK, ER antagonists, ER agonists, and hDHFR) by the true hit, hit rate, goodness-of-hit (GH), and false positive rate. We used four combinations of none (E_{bind}), ligand preference (E_{bind} and E_{ligpre}), interaction preference (E_{bind} and E_{pharma}), and both (E_{tot}) to compare screening utilities with other public docking tools. The results showed that the scoring function considering both ligand preference and interaction preference was the most reliable.

In chapter 4, we applied GEMDOCK to the envelope protein of dengue virus and the microbial D-hydantoinase. For the envelope protein, we suggested 35 molecules from the intersection of GEMDOCK and GOLD top 200 scorers. Biological experiments to test their activities are still in process cooperated with Dr. Yun-Lian Yang. On the other hand, we identified two novel substrates for *Agrobacterium radiobacter* D-hydantoinase according to their docked poses and approved by wet experiments.

Chapter 5 presented some conclusions and future perspectives. The screening accuracy of

GEMDOCK was enhanced with the modified scoring function and it could effectively reduce the number of false positives. With the modified scoring function, we could design specific variables for virtual screening against different target proteins according to the knowledge of known active ligands or characteristics of the binding site.



Chapter 2

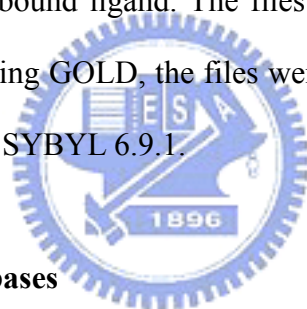
Materials and Methods

GEMDOCK, enhanced and modified from our original technique [22, 23], is nearly an automatic tool for virtual screening (Figure 1). GEMDOCK can sequentially be 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. The target protein is first prepared by specifying the atomic coordinates from the PDB, the ligand binding area, atom formal charge, and atom types. When active ligands of the target protein are available, GEMDOCK evolves a pharmacological consensus (e.g., hot spots) and ligand preferences from the target protein and these ligands by overlapping the docked ligand conformations or superimposing X-ray structures. The pharmacological consensuses and ligand preferences were incorporated into our scoring function to improve the screening accuracy. The ligand database was constructed from the public compound databases, e.g., the MDL Drug Data Report (MDDR) or the Available Chemical Directory (ACD), according to the characteristics of the target protein and ligand preferences mined from known active compounds. After the ligand database and the target protein are prepared and the pharmacological preferences are evolved, GEMDOCK sequentially predicts the binding conformation and estimates the binding affinity for each ligand in the compound database. Finally, GEMDOCK ranks these docked ligand conformations for use in the post-docking analysis.

2.1 Preparation of Target Proteins and Ligand Databases

A. Preparing for target proteins

We applied GEMDOCK to virtual screening against four targets, 1) herpes simplex virus types 1 thymidine kinase (TK, the PDB entry: 1kim) [24], 2) human estrogen receptor alpha ($ER\alpha$, the PDB entry: 3ert) [25], 3) human estrogen receptor alpha ($ER\alpha$, the PDB entry: 1gwr) [25], 4) human dihydrofolate reductase (hDHFR, the PDB entry: 1hfr) [26], 5) dengue virus envelope protein (E protein, the PDB entry: 1oke) [4], 6) *Thermus sp.* D-hydantoinase (the PDB entry: 1gkp) [27]. The coordinates were derived from the PDB. For each target protein, we defined that the binding site was the collection of amino acids enclosed within a 10 Å radius sphere centered on the bound ligand. The files in the PDB format were prepared for running GEMDOCK. For running GOLD, the files were added hydrogen atoms and stored in the SYBYL mol2 format using SYBYL 6.9.1.



B. Preparing for ligand databases

Two screening sets designed for virtual screening against TK and $ER\alpha$ antagonists were proposed by Bissantz et al. [19] in 2000 and retested by Jain in 2003 [28]. A TK library contained ten known ligands of TK and 990 randomly chosen molecules from the ACD; an $ER\alpha$ -antagonists library contained ten known antagonists of $ER\alpha$ and 990 randomly chosen molecules from the ACD. The two testing sets were downloaded from <http://jainlab.ucsf.edu/Downloads.html> proposed by Jain and the files in the SYBYL mol2 format were converted to the MDL mol format with Corina3.0 for running GEMDOCK. Besides, the screening set for $ER\alpha$ agonists included ten known agonists [29] and the same 990 molecules as the $ER\alpha$ -antagonists library.

The screening set designed for hDHFR included ten known ligands from the PDB. The other 990 molecules were randomly selected from the MDDR. Using MDL Integrated Scientific Information System (ISIS), the MDDR were first filtered with suitable molecular weights between 200 and 750. Then we removed analogues of 4-substituted 2-aminopyrimidine. Out of the remaining molecules, 990 were randomly chosen and downloaded from the MDDR in a multi-structure file in the MDL RDF format. We divided it into separated files and removed small fragments from multi-component records, such as counter-ion in salts, solvent molecules with Corina3.0. Finally we prepared the structure files of molecules in the MDL mol format for running GEMDOCK. Using Corina3.0, hydrogen atoms were added to the structures and stored in the SYBYL mol2 format for running GOLD.

The screening set for the E protein of dengue virus was derived from the drug database, Comprehensive Medicinal Chemistry (CMC). Using ISIS, the CMC were first filtered with molecular weights between 200 and 500 according to the molecular weight of the ligand (β -octylglucoside, molecular weight: 292) complex with the crystal in the PDB and then removed small fragments with Corina3.0. The structure files of molecules were stored in the MDL mol format for running GEMDOCK and in the SYBYL mol2 format for running GOLD.

2.2 Molecular Docking

A. Search method of flexible docking

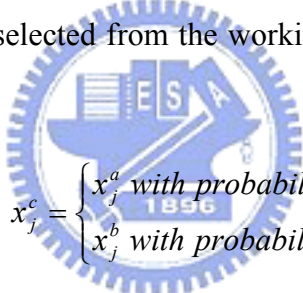
The search algorithm of GEMDOCK is a generic evolutionary method [22]. The core idea of our evolutionary approach was to design multiple operators that cooperate using the family competition model, which is similar to a local search procedure. The rotamer-based mutation operator, a discrete operator, is used to reduce the search space of ligand structure

conformations. The Gaussian and Cauchy mutations, continuous genetic operators, efficiently search the orientation and conformation of the ligand relating to the center of the target protein. GEMDOCK randomly generates a starting population with N solutions by initializing the orientation and conformation of the ligand relating to the center of the receptor. 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 represents the adjustable variables to be optimized in which x_1, x_2 , and x_3 are the 3-dimensional location of the ligand; x_4, x_5 , and x_6 are the rotational angles; 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 operated 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.

Recombination operator: 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 :

$$w'_j = w_j A(\cdot)$$

$$x'_j = x_j + w'_j D(\cdot)$$

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 $\exp[\tau'N(0, 1) + \tau N_j(0, 1)]$ 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

$$\sigma^c = \gamma \sigma^a$$

$$x_j^c = x_j^a + \sigma^c N_j(0,1)$$

B. 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:

$$E_{tot} = E_{bind} + E_{pharma} + E_{ligpre} \quad (1)$$

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

$$E_{bind} = E_{inter} + E_{intra} + E_{penal} \quad (2)$$

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

$$E_{inter} = \sum_{i=1}^{lig} \sum_{j=1}^{pro} \left[F(r_{ij}^{B_{ij}}) + 332.0 \frac{q_i q_j}{4r_{ij}^2} \right] \quad (3)$$

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, 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 terms lig and pro denote the number of heavy atoms in the ligand and receptor, respectively. $F(r_{ij}^{B_{ij}})$ is a simple atomic pair-wise potential function (Figure 2).

In this atomic pair-wise model, the interactive types include only hydrogen bonding and steric potentials having the same function form but different parameters, V_1, \dots, V_6 . The energy value of hydrogen bonding should be larger than that for steric potential. In this model, atoms are divided into four different atom types: 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, which is set to 5 or -5, respectively, if the electrostatic energy is more than 5 or less than -5.

The intramolecular energy of a ligand is

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

where $F(r_{ij}^{B_{ij}})$ is defined as Equation 3 except the value is set to 1000 when $r_{ij}^{B_{ij}} < 2.0 \text{ \AA}$ and *dihed* is the number of rotatable bonds of a ligand. We followed the work [16] to set the values of *A*, *m*, and θ_0 . For the sp^3 - sp^3 bond *A*, *m*, and θ_0 are set to 3.0, 3, and π ; for the sp^3 - sp^2 bond and *A* = 1.5, *m* = 6, and $\theta_0 = 0$.

C. Mining pharmacological consensuses

GEMDOCK evolves binding site pharmacological consensuses 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. For each known active ligand, GEMDOCK first yielded ten docked ligand conformations by docking the ligand into the target protein, and only the ligand with the lowest docked conformation 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 three different types, including hydrogen bonding, hydrogen-charged interactions, and hydrophobic 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

$$Q_j^k = \frac{f_j^k}{3N} \quad (5)$$

where f_j^k is the number of an atom j (in protein) interacting with ligands with the interaction type k and N is the number of known active ligands. In our present work, an atom j 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:

$$E_{pharma} = \sum_{i=1}^{lig} \sum_{j=1}^{hs} CW(B_{ij}) F(r_{ij}^{B_{ij}}) \quad (6)$$

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 Equation 3, lig is the number of the 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

$$CW(B_{ij}) = \begin{cases} 1.0 & \text{if } Q_j^k \leq 0.5 \text{ or } B_{ij} \neq k \\ 1.5 + 5(Q_j^k - 0.5) & \text{if } Q_j^k > 0.5 \text{ and } B_{ij} = k \end{cases} \quad (7)$$

Q_j^k is the atomic pharmacological-profile weight (Equation 5) and k is the interact type (e.g., hydrogen bonding, hydrogen-charged interactions, or hydrophobic interactions) 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 our purpose, the atomic pair-wise potential energies of the electrostatic, hydrogen bond, and steric potential were set to -5 , -2.5 , and -0.4 , respectively (Figure 2). If the binding site of a target protein is hydrophobic, the ligand preference (E_{ligpre}) is a penalty value for those screened ligands having many charged and polar atoms. The E_{ligpre} is given as

$$E_{ligpre} = LP_{elec} + LP_{hb} \quad (8)$$

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

$$LP_{elec} = \begin{cases} 10NA_{elec} & \text{if } NA_{elec} > UB_{elec} \\ 0 & \text{if } NA_{elec} \leq UB_{elec} \end{cases} \quad (9)$$

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

$$LP_{hb} = \begin{cases} 5NA_{hb} & \text{if } r_{hb} > Ur_{hb} \\ 0 & \text{if } r_{hb} \leq Ur_{hb} \end{cases} \quad (10)$$

where $r_{hb} = \frac{NA_{hb}}{NA_t}$ 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 formulated a normalization strategy defined as

$$E_{tot} = \frac{E_{tot}}{(NA_t)^K} \text{ where } K = \begin{cases} 0.5 & \text{if } \mu_{mw} \leq 15 \\ 0.5 - \frac{0.45(\mu_{mw} - 15)}{25} & \text{if } 15 < \mu_{mw} \leq 40 \\ 0.05 & \text{if } \mu_{mw} > 40 \end{cases} \quad (11)$$

where E_{tot} is the empirical binding energy (Equation 1), 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.



Chapter 3

Evaluation GEMDOCK on Virtual Database Screening

We have made four screening sets against different target proteins, thymidine kinase (TK), estrogen receptor with antagonists (ER-antagonists), estrogen receptor with agonists (ER-agonists) and dihydrofolate reductase (DHFR). Each screening set includes ten known active ligands and 990 randomly selected compounds. Four common metrics were used to evaluate the screening quality, including the true hit (the percentage of active ligands retrieved from the database), hit rate (the percentage of active ligands in the hit list), goodness-of-hit (GH), and false positive rate.



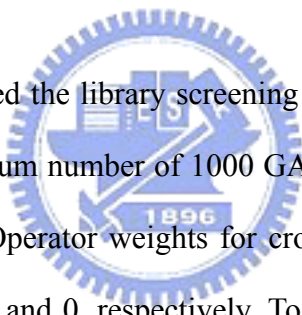
3.1 Parameters of GEMDOCK and GOLD

A. GEMDOCK parameters

Table 3 indicates the setting of GEMDOCK parameters in flexible search phase, including the initial step sizes, family competition length ($L=2$), population size ($N=200$), and recombination probability ($p_c = 0.3$) in this work. For each screening ligand, GEMDOCK optimization stops when either the convergence is below certain threshold value or the iterations exceed a maximal preset value that was set to 60. Therefore, GEMDOCK generated 800 solutions in one generation and terminated after it exhausted 48,000 solutions for each docking ligand. On average, GEMDOCK took 135 seconds for a docking run on a Pentium 1.4 GHz personal computer with a single processor.

B. GOLD 2.1 parameters

For molecular recognition, we use the standard default settings of GOLD for docking. For each of the ten independent genetic algorithm (GA) runs, a maximum number of 100,000 GA operations was performed on five islands and the population size of each island was 100. Operator weights for crossover, mutation, and migration in the entry box were 95, 95 and 10, respectively. To allow poor nonbonded contacts at the start of each GA run, the maximum distance between hydrogen donors and fitting points was set to 4.0 Å and nonbonded van der Waals energies were cut off at a value equal to $2.5 k_{ij}$ (well depth of the vander Waals energy for the atom pair i, j). The GA docking was stopped when the top three solutions were within 1.5 Å RMSD of each other.



For virtual screening, we used the library screening settings of GOLD. For each of the 10 independent GA runs, a maximum number of 1000 GA operations was performed on a single population of 50 individuals. Operator weights for crossover, mutation and migration in the entry box were set to 100, 100 and 0, respectively. To allow poor nonbonded contacts at the start of each GA run, the maximum distance between hydrogen donors and fitting points was set to 5.0 Å and nonbonded van der Waals energies were cut off at a value equal to $10.0 k_{ij}$ (well depth of the van der Waals energy for the atom pair i, j). The GA docking was stopped when the top three solutions were within 1.5 Å RMSD of each other.

3.2 Thymidine Kinase

A. Preparation of docking databases

Herpes simplex virus types 1 and 2 (HSV-1 and HSV-2) could cause painful epithelial ulcers near the mouth, on the cornea and genitals, as well as fatal encephalitis. HSV-1 TK is

the center of phosphorylation of nucleosides or nucleoside analogs such as acyclovir [30, 31]. Many antiviral drugs attack the replication of the viral genome with nucleoside analogs. These analogs are activated by phosphorylation with TK and prevent DNA synthesis by the introduction of a chain-terminating nucleoside at the 3' end of the growing DNA strand. Besides antiviral drugs, these analogs have been used in a virological study of TK mutations [32] and employed extensively in gene therapy for cancer [33, 34]. Therefore virtual screening for TK to exploit novel lead compounds would be of considerable value in many fields.

In order to evaluate GEMDOCK and to compare GEMDOCK with several widely used methods, we docked ten active substrates of HSV-1 TK into the complexes with experimentally X-ray structures from the PDB. Each ligand systematically named with four characters followed by three characters. For example, in the ligand "1kim.THM", "1kim" denotes the PDB code and "THM" is the ligand name in the PDB. When we evaluated the accuracy of GEMDOCK for molecular docking, the crystal coordinates of the ligand and protein atoms were taken from PDB, and were separated into different files. Our program then assigned the atom formal charge and atom type (i.e., donor, acceptor, both, or nonpolar) for each atom within the ligand and the protein. The bond type (sp^3-sp^3 , sp^3-sp^2 , or others) of a rotatable bond inside a ligand was also assigned.

To evaluate the virtual screening utility of GEMDOCK, we used HSV-1 TK as the target protein with a testing set proposed by Bissantz et al.. It included ten known active ligands (Figure 3) of TK and 990 randomly chosen non-active compounds from the ACD. When preparing the target protein, the atom coordinates for virtual screening were taken from the crystal structure of the TK complex with the ligand deoxythymidine (the PDB entry: 1kim). We thought that choosing the crystal coordinates of TK in complex with its nature substrate

(deoxythymidine) was a reasonable choice since the active site is open enough to accommodate a broad variety of ligands. The atom coordinates of each ligand were sequentially derived from the database. Our program automatically assigned the formal charge and atom type of each atom. The ligand characteristics (i.e., the number of electrostatic atoms, hydrogen donor, and hydrogen acceptor) and the bond types of single bonds inside a ligand were also calculated. These variables were used in Equation 3 to calculate the scoring value of a docked conformation. Finally GEMDOCK re-ranked and sorted all docked ligand conformations for the post-analysis.

Figure 4 shows the pharmacological consensus of the binding site and ligand preferences that were identified by superimposing ten crystal structures of TK shown in Figure 2. Four important residues of the pharmacological consensus were identified and marked. The dashed lines indicate the hydrogen binding. The phenolic ring of Y172 formed stack force with the ligand. According to the pharmacological consensus of ten protein-ligand complexes, we added pharmacological weights ($CW(B_{ij})$) listed in Table 4. These weights were used in Equation 6 for calculating the value of E_{pharma} . For ligand preferences, parameters for calculating E_{ligpre} were listed in Table 5.

B. Molecular docking results on ten TK complexes

First we evaluated the docking accuracy of GEMDOCK by docking ten TK ligands (Figure 3) back into their respective complex. GEMDOCK executed three independent runs for each complex. The solution with the lowest score was then compared with the observed ligand crystal structure. We based the results on root mean square deviation (RMSD) error in ligand heavy atoms between the docked conformation and the crystal structure. The RMSD values of all ten docked conformations are less than 1.5 Å (Table 6). Second we docked all ten TK

ligands into the reference protein (1kim) and the results were shown in Table 6. During flexible docking GEMDOCK obtained similar results whether the pharmacophore preferences (i.e., E_{pharma} and E_{ligpre}) were considered or not. The docked conformations with RMSD values less than 1.6 Å for seven pyrimidine-based ligands. On the other hand, three purine-based ligands (i.e., 1ki2.GA2, 1ki3.PE2, and 2ki5.AC2) could not be successfully docked into the X-ray poses because the side-chain conformation of Q125 in the reference protein 1kim differs from the ones of these purine-based complexes, i.e. 1ki2, 1ki3, and 2ki5. GEMDOCK was the best among these four competing methods (GEMDOCK, GOLD, FlexX, and DOCK) on this testing set.

C. Virtual screening of TK substrates

Figure 5 shows the overall accuracy of GEMDOCK using different combinations of pharmacological preferences in screening the substrates of HSV-1 thymidine kinase (TK) from a testing set with 1000 compounds. This testing set, including ten active and 990 random ligands proposed by Bissantz et al., was used to evaluate the performance of three docking tools (DOCK, FlexX, and GOLD) with different combinations of seven scoring functions. The results of the comparison are also shown in Table 7.

Four common metrics were used to evaluate the screening quality. The GH score is defined as

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

where A_h is the number of active ligands in the hit list, T_h is the total number of compounds in the hit list, A is total number of active ligands in the database, and T is the total number of compounds in the database. The yield (hit rate), false positive (FP) rates and true positive (TP)

rate can be given as

$$\begin{aligned} \text{yield} &= 100 \frac{A_h}{T_h} \% \\ \text{FP} &= 100 \frac{T_h - A_h}{T} \% \\ \text{TP} &= 100 \frac{A_h}{A} \% \end{aligned}$$

In the case of TK A and T are 10 and 1000, respectively.

The main objective of this study was to evaluate whether the new scoring function was applicable to both molecular docking and ligand scoring in virtual screening. Figure 5 shows the results of GEMDOCK using different combinations of ligand preferences (E_{ligpre}) and interaction preferences (E_{pharma}). We tested GEMDOCK with different combinations to evaluate the performance and to observe the search behavior of different parameters. GEMDOCK generally improves the screening quality by considering both ligand preferences and the pharmacological consensus. The ligand preference seemed more important than the interaction preference in the case of TK. As shown in Figure 5A, the hit rates of GEMDOCK for different combinations are 25.6% (both), 15.9% (ligand preferences), 13.5% (interaction preferences) and 9.4% (none) when the TP rate is 100%. When GEMDOCK applied both interaction preferences (E_{pharma}) and ligand preferences (E_{ligpre}) and the TP rate is 100%, the GH score is 0.43 (Figure 5B) and the FP rate is 2.93% (Figure 5C).

Table 7 compares GEMDOCK with four docking methods (Surflex [28], DOCK, FlexX, GOLD) on the same target protein and screening database at true positive rates ranging from 80% to 100%. For GEMDOCK on the target TK, the ranks of the ten active ligands were 5-7, 9-11, 13-14, 22 and 39. For the true positive rate of 100%, the FP rate for GEMDOCK is 2.9%. In contrast, the FP rates for competing methods are 3.2% (Surflex), 27% (DOCK), 19.4% (FlexX), and 9.3% (GOLD) [28]. The performance of DOCK is the worst and of

GEMDOCK is the best among these five approaches.

3.3 Estrogen Receptor

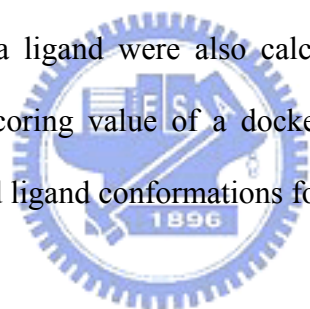
A. Preparation of docking databases

Estrogens such as 17β -estradiol are steroid hormones as key mediators of female reproductive glands and they also exert their actions on other systems. For example, estrogens contribute to the maintenance of bone tissue through a process involving bone resorption and bone formation [35]. Hormone replacement therapies have been used for the treatment of vasomotor symptoms related to the menopause and for prevention of osteoporosis [36, 37]. Compounds mimic estrogen in some tissues while antagonizing its action in others are named selective estrogen receptor modulators (SERMs) [38]. Many SERMs such as tamoxifen and raloxifene, are currently on the market for the treatment of hormone-dependent breast cancer [39] and prevention and treatment of osteoporosis [40], respectively. But there are often several intolerable side effects such as benign and malignant lesions of the uterus when patients take the treatment with SERMs for a long term. Therefore, the search for proper SERMs among both existing and new drugs has been a challenging task in recent years [41, 42].

We evaluated the docking accuracy of GEMDOCK and compared it with several widely used methods on docking four $ER\alpha$ -antagonists and four $ER\alpha$ -agonists back into their complexes with experimentally X-ray structures from the PDB. Each ligand is systematically named in the same way as TK. When we evaluated the accuracy of GEMDOCK for molecular docking, the crystal coordinates of the ligand and protein atoms were taken from the PDB, and were separated into different files. Our program then assigned the atom formal charge and atom type for each atom of both the ligand and protein. The bond type of a rotatable bond

inside a ligand was also assigned.

We have applied GEMDOCK to virtual screening against ER α with a testing set proposed by Bissantz et al. It was composed of ten known antagonists of ER α (Figure 6), ten known agonists of ER α (Figure 7) [29] and 990 randomly selected compounds from ACD (Available Chemicals Directory). When preparing the target proteins, the atom coordinates for virtual screening were taken from the crystal structure of ER α complex with the ligand 4-hydroxytamoxifen (the PDB entry: 3ert) for screening antagonists and with the ligand 17 β -estradiol (the PDB entry: 1gwr) for screening agonists. The atom coordinates of each ligand were sequentially taken from the database. Our program automatically decided the formal charge and atom type of each ligand atom. The ligand characteristics and the bond types of single bonds inside a ligand were also calculated. These variables were used in Equation 3 to calculate the scoring value of a docked conformation. Finally GEMDOCK re-ranked and sorted all docked ligand conformations for the post-analysis.



According to our methods, we have docked ten known active antagonists and agonists of ER α into the target protein for 20 times. Figure 8 shows the pharmacological consensus of the binding site and ligand preferences that were identified by superimposing ten docked poses of ER α antagonists (Figure 8A) and agonists (Figure 8B) with the lowest energy. The important residues of the pharmacological consensus and interactions were marked. The dashed lines indicated hydrogen bonds formed between ligands and important residues. According to these docked conformations and orientations, we defined following hot-spots: D351-OD1, E353-OE2, R394-NH2, G420-O and H524-ND1 for antagonists and E353-OE2, R394-NH2 and H524-ND1 for agonists. Table 8 showed weight values ($CW(B_{ij})$) that we set for hot-spots according to equation 5 and 7. These weights were used in Equation 6 for calculating the value of E_{pharma} for interaction preferences. For ligand preferences, parameters for calculating

E_{ligpre} were listed in Table 5.

B. Molecular docking results on four ER complexes

First we docked four antagonists and four agonists back into their native complexes, respectively. We based the results on root mean square deviation (RMSD) error in ligand heavy atoms between the docked conformation and the crystal structure. Second we docked four antagonists and four agonists into their reference proteins, 3ert and 1gwr, respectively. During flexible docking GEMDOCK obtained similar results whether interaction preferences and ligand preferences (i.e., E_{pharma} and E_{ligpre}) were considered or not. The RMSD values of docked conformations were less than 1.6 Å when docking into their native binding site (Table 9). When docking into the binding site of each reference protein, two ligands could not be successfully docked similar to the X-ray poses. In 1hj1.AOE and 1qkm.GEN, the values of RMSD were larger than 2.0 Å because the native proteins were crystal structures of ER β -ligand complexes. When docking 1hj1.AOE and 1qkm.GEN into the reference protein of ER α , the docked poses would be different from the crystal structures complex with ER β .

C. Virtual screening of ER α antagonists and agonists

Figure 9 shows the overall screening utility of GEMDOCK whether we used the interaction preferences and ligand preferences for antagonists and agonists of ER α with two testing sets of 1,000 compounds. One testing set included ten active antagonists and 990 random selected molecules proposed by Bissantz et al. They were used to evaluate the performance of three docking tools (DOCK, FlexX, and GOLD). The results of the comparison for antagonists were shown in Table 10. Another screening set included ten active agonists and 990 molecules that were the same as the former.

The main objective of this study was to evaluate whether the new scoring function was applicable to both molecular docking and ligand scoring in virtual screening. Figure 9 shows the screening results for antagonists using different combinations of ligand preferences (E_{ligpre}) and interaction preferences (E_{pharma}) and Figure 10 shows the results of agonists. We tested GEMDOCK on different combinations to evaluate the performance and to observe the search behavior of our program. GEMDOCK generally improves the screening utility by considering both ligand preferences and interaction preferences. The ligand preferences seem more important than interaction preferences in the case of ER α whether for antagonists or agonists. As shown in Figure 9A, the hit rates for ER α antagonists of GEMDOCK with different combinations are 52.6% (both), 31.3% (ligand preferences), 22.2% (interaction preferences) and 16.4% (none) when the TP rate is 100%. When GEMDOCK applied both interaction and ligand preferences and the TP rate is 100%, the GH score is 0.64 (Figure 9B) and the FP rate is 0.91% (Figure 9C). For ER α agonists (Figure 10), the hit rates with different combinations are 34.5% (both), 10.1% (ligand preferences), 8.6% (interaction preferences) and 4.5% (none) when the TP rate is 100%. When GEMDOCK applied both interaction and ligand preferences and the TP rate is 100%, the GH score is 0.50 (Figure 10B) and the FP rate is 1.9% (Figure 10C).

3.4 Dihydrofolate Reductase

A. Preparation of docking databases

Dihydrofolate reductase (DHFR) catalyzes the reduction of 7,8-dihydrofolate or folate to 5,6,7,8-tetrahydrofolate (THF) in an NADPH-dependent pathway. THF is an essential cofactor for other enzymes involving one-carbon-transfer reactions necessary for the biosynthesis of numerous amino acids and purines. THF also acts as a cofactor for thymidylate synthase, which is responsible for the methylation of deoxyuridylate to

thymidylate, a key component for synthesis of DNA. DHFR is found in cells of all living organisms, where it maintains the intracellular level of THF. Therefore, the inhibition of DHFR activity reduces the intracellular pool of THF resulting in inhibition of DNA synthesis and leading to cell death. Based on this mechanism, human DHFR (hDHFR) has become a major drug target in anticancer therapy. It is also a target for inhibition of bacterial, fungal, and protozoal DHFRs to treat human infectious diseases by many implicated microorganisms [43, 44]. With the wide use of these antifolate drugs, the resistance of DHFRs in human or other microorganisms is widespread. Therefore, it is urgent to search for new targets or new effective inhibitors to deal with the problem [45, 46].

We evaluated the docking accuracy of GEMDOCK and compared it with GOLD on docking ten known ligands (Figure 11) of hDHFR back into the complexes with experimentally X-ray structures from PDB. Each ligand is systematically presented in the same way as TK and ER. When we evaluated the accuracy of GEMDOCK for molecular docking, the crystal coordinates of the ligand and protein atoms were taken from PDB, and were separated into different files. Our program then assigned the atom formal charge and atom type for each atom of both the ligand and protein. The bond type of a rotatable bond inside a ligand was also assigned.

We used hDHFR as the target protein for virtual screening with a testing set prepared by ourselves. The testing set was composed of ten known active ligands of hDHFR and 990 randomly selected compounds from the MDL Drug Data Report (MDDR). The drug database of MDDR included 132,726 compounds until May, 2004. First we filtered the MDDR with molecular weights between 200 and 750 (119,106 compounds) and removed analogues of 4-substituted-2-aminopyrimidine (removing 2,197 compounds). After removing small fragments from multi-component records, we randomly selected 990 compounds from the

remainder (Figure 12). When preparing the target protein, the atom coordinates for virtual screening were taken from the crystal structure of the DHFR complex (the PDB entry: 1hfr). The atom coordinates of each ligand were sequentially taken from the database. Our program automatically decided the formal charge and atom type of each ligand atom. The ligand characteristics and the bond types of single bonds inside a ligand were also calculated. These variables were used in Equation 3 to calculate the scoring value of a docked conformation. Finally GEMDOCK re-ranked and sorted all docked ligand conformations for the post-analysis.

Figure 13 shows the pharmacological consensus of the binding site and ligand preferences that were identified by superimposing ten crystal structures of hDHFR shown in Figure 11. Three important residues of pharmacological consensus were marked. The dashed lines indicate the hydrogen binding. According to the pharmacological consensus of ten protein-ligand complexes, we added pharmacological weights ($CW(B_{ij})$) shown in Table 11. These weights were used in Equation 6 for calculating the value of E_{pharma} . For ligand preferences, parameters of hDHFR ligands for calculating E_{ligpre} were shown in Table 5.

B. Molecular docking results on ten hDHFR complexes

For molecular recognition, we docked ten known ligands back into their complexes, respectively. We based the results on root mean square deviation (RMSD) error in ligand heavy atoms between the docked conformation and the crystal structure. Second we docked these ten ligands into the reference protein (1hfr) and the results were shown in Table 12. During flexible docking GEMDOCK obtained similar results whether interaction preferences and ligand preferences (i.e., E_{pharma} and E_{ligpre}) were considered or not. The RMSD values of docked conformations were less than 1.5 Å when docking into their native binding site. On

the other hand, when docking into the binding site of the reference protein, all ligands could be successfully docked into the correspondent X-ray poses. The performance of GEMDOCK was superior to that of GOLD on this testing set.

C. Virtual screening of hDHFR ligands

We have applied GEMDOCK to virtual screening for hDHFR with a testing set designed by ourselves. The testing set composed of ten known ligands and 990 compounds randomly selected from the MDDR was used to evaluate the screening utility of GEMDOCK and compare the performance with GOLD.

Figure 14 shows the screening results for hDHFR ligands using different combinations of ligand preferences (E_{ligpre}) and interaction preferences (E_{pharma}). We tested GEMDOCK on different combinations to evaluate the performance and to observe the search behavior of our program. GEMDOCK generally improves the screening utility by considering both ligand preferences and interaction preferences. As shown in Figure 14A, the hit rates of GEMDOCK for different combinations are 33.3% (both), 29.4% (ligand preferences), 30.3% (interaction preferences), and 10.53% (none) when the TP rate is 100%. When GEMDOCK applied both pharmacophore and ligand preferences and the TP rate is 100%, the GH score is 0.49 (Figure 14B) and the FP rate is 2.0% (Figure 14C). Comparison of GEMDOCK with GOLD by false positive rates (%) on screening 990 compounds and ten known ligands of the hDHFR is shown in Table 13. False positive rates of GEMDOCK for different combinations are lower than the ones of GOLD when true positives rates are 80%, 90%, and 100%.

Chapter 4

GEMDOCK on Practical Applications

4.1 Envelope Protein of Dengue Virus Type II

A. Preparation of docking databases

Dengue virus, a member of the flavivirus family, is an emerging global health threat. There is no specific treatment for infection, and control of dengue virus by vaccination has proved elusive [4]. Besides, several other flaviviruses are important human pathogens, including yellow fever virus, West Nile virus, Tick-borne encephalitis virus (TBE), and Japanese encephalitis viruses (JE). Therefore related research about dengue virus has been an important target in epidemiology and virology. In Taiwan, it is always widespread in summer and can cause severe epidemics of diseases such as dengue fever and dengue hemorrhagic fever/dengue shock syndrome [47]. To combat diseases caused by dengue viruses, it is emergent to develop novel antiviral therapeutic agents.

Membrane fusion is the central molecular event during the entry of enveloped viruses into cells [5]. Dengue virus enters a host cell when the viral envelope glycoprotein, E, binds to a receptor and responds by conformational rearrangement to the reduced pH of an endosome. After conformational rearrangement, the dimeric prefusion form of the E protein changes to its trimeric postfusion state. The trimer of the E protein inserts into the host-cell membrane with three fusion loops at one end and then induces fusion of viral and host-cell membranes. A ligand-binding pocket in the E protein of dengue virus was found and there is a

protein-ligand complex in the PDB (the PDB entry: 1oke). The key difference between the two structures is a local rearrangement of the “kl” β -hairpin (residues 268-280) and the concomitant opening up of a hydrophobic pocket, occupied by a detergent molecule of *n*-octyl- β -D-glucoside (β -OG). Mutations affecting the pH threshold for fusion map to the hydrophobic pocket [48, 49], which we propose is a hinge point in the fusion-activating conformational change. Detergent binding marks the pocket as a potential site for small-molecule fusion inhibitors (Figure 15) [4, 5].

We defined that the binding site of the E protein was the collection of amino acids enclosed within a 10 Å radius sphere centered on the bound ligand, β -OG. The coordinates of atoms were derived from the PDB and stored in the PDB format for running GEMDOCK. Using SYBYL 6.9.1, we converted the structure file in the SYBYL mol2 format for running GOLD.

We prepared the screening set from the drug database, CMC. The CMC contained 7,937 compounds until May, 2004. It was first filtered with molecular weights between 200 and 600. In the remaining 5,961 compounds, 630 records with multi-component were removed and finally we had 5,331 molecules in the screening set.

B. Molecular docking results on the E protein

To evaluate the docking accuracy of GEMDOCK for the E protein of dengue virus, we docked the ligand (β -OG) into its native binding site of the complex. GEMDOCK executed 3 independent runs for each complex. The solution with the lowest scoring function was then compared with the observed ligand crystal structure. We based the results on root mean square deviation (RMSD) error in ligand heavy atoms between the docked conformation and the crystal structure. During flexible docking GEMDOCK obtained different results in this case

when the interaction preference (i.e. E_{pharma}) was considered. The RMSD value of the docked conformation was 1.63 Å when docking without considering the interaction preference. The RMSD value of the docked conformation predicted by GOLD was 1.55 Å. According to the protein-ligand complex, the ligand forms hydrogen bonds with E49-OE2 and Q271-OE1. Because there is no other known ligand, we set the $CW(B_{ij})$ (in Equation 6) was 3.0 for the two oxygen atoms. After considering the interaction preference, the RMSD value of the docked conformation was 1.20 Å. The interaction preference would be considered when we screened the screening set to find novel lead compounds.

C. Virtual screening for the E protein

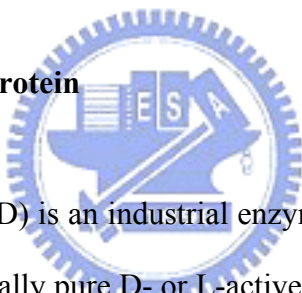
Based on the screening utility of GEMDOCK described above, we applied GEMDOCK to virtual screening for the E protein with a screening set including 5,331 molecules from the CMC. In the list of top 200 compound, we selected nine molecules to be tested by the biological experiment (cooperation with Dr. Yun-Lian Yang) and we found that MCMC00007079 could suppress the activity of dengue virus with concentrations of 1mM and 10mM (Figure 16). When we overlapped docked poses of the nine molecules (Figure 17A), MCMC00007079 was different from others because of the hydrogen bond forming with Q52. Q52 was mutated to affect the pH threshold of fusion in the previous experiment. Maybe it is a potential point to develop inhibitors for dengue virus.

According to the references and the protein-ligand complex, the binding site is a hydrophobic pocket. Therefore we set the UB_{elec} was zero and Ur_{hb} was 0.3 when we screened the 5,331 molecules with ligand preferences in the second virtual screening. Besides GEMDOCK, we also applied GOLD to virtual screening for the E protein with the same screening set. We got the intersection of top 200 scorers from the two results of virtual

screening by GEMDOCK and GOLD. Table 13 was the list of the intersection from GEMDOCK and GOLD. There were 35 compounds in the list and biological experiments to test their activities were still in process. Figure 18 showed the number of contacts between the 35 poses and residues of the E protein. Contacts mean that the docked ligand forms hydrogen bonds with specific residues. Among these residues, Q52 was mutated to affect the pH threshold of fusion in the previous experiment. If docked poses form more contacts with Q52, these compounds have more possibility to become potential inhibitors. According to our observation, they are MCMC00001935, MCMC00002025, MCMC00005147, MCMC00006126, MCMC00007533, and MCMC00009993.

4.2 Molecular Docking of D-Hydantoinase

A. Preparation of the target protein



Microbial hydantoinase (HYD) is an industrial enzyme. This enzyme has been used for the commercial production of optically pure D- or L-active amino acids in biocatalysis. It has also been used for the production of D-amino acid intermediate compounds through stereospecific hydrolysis of chemically synthesized cyclic hydantoins (Figure 19). These intermediates are widely used for semisynthetic antibiotics, peptide hormones, pyrethroids, and pesticides [27].

Because the crystal structure has no ligand complex with the protein, we defined that the binding site was the collection of amino acids enclosed within a 10 Å radius sphere centered on the residue KCX150 and two zinc ions [27]. The docked ligands proposed by Dr. Yuh-Shyong Yang were shown in Figure 20A. They were substrates that have been verified by biological experiments. Besides, they also provided us a testing set with 20 molecules (Figure 20B). We hope to recognize them according to their docked poses.

B. Molecular docking results of substrates

Figure 21 shows docked poses, K_m and k_{cat} of seven known substrates (proposed by Dr. Yuh-Shyong Yang) in the training set and Figure 22 shows the pharmacological consensus of the binding site and ligand preferences that were identified by overlapping 17 docked poses of substrates shown in Figure 20A. The important residues of the pharmacological consensus and interactions were marked. The dashed lines indicated hydrogen bonds formed between ligands and important residues. According to these docked conformations and orientations, we defined following hot spots: S288-N, S288-O, D315-OD1 and D315-OD2. Both substrates and inhibitors of D-HYD could enter the binding site and we could recognize them according to their docked poses among 20 compounds in the testing set.

On our research cooperating with Dr. Yuh-Shyong Yang, we have identified two new substrates for *Agrobacterium radiobacter* D-HYD by GEMDOCK. Figure 23A showed the docked conformation of allantoin in D-HYD. The docked pose formed hydrogen bonds with S288 and D315 in the meanwhile as known substrates. Therefore we deduced that allantoin should be a substrate for D-HYD and the experimental data also supported this hypothesis (Figure 23B). Besides allantoin, parabanic acid is also a substrate (Figure 24A, B) and the specific activity of the hydantoinase is shown in Table 14.

Figure 25 shows IC₂₀ values of inhibitors in Figure 20B and their docked poses are also shown. After molecular recognition, we could define important residues of the pharmacological consensus and it will be helpful when we virtually screen the molecular database later to find potential inhibitors.

Chapter 5

Conclusions

5.1 Summary

In summary, we have developed an automatic tool with a novel scoring function for virtual screening by applying numerous enhancements and modifications 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. Our new scoring function can be applied to both flexible docking and post-docking analysis for reducing the number of false positives. For different target proteins, our scoring function can consider the knowledge from known active ligands to improve the screening performance. Experiments verify that the proposed approach is robust and adaptable to virtual screening.

5.2 Major Contributions and Future Perspectives

To apply GEMDOCK to virtual screening, a well-designed screening set is essential. The screening set for hDHFR was prepared by ourselves to test the utility of GEMDOCK. We have set up a procedure to prepare a screening set according to the target protein (Appendix A). With a well-designed screening set, we could verify our scoring function for virtual screening by considering pharmacological preferences (E_{pharma} and E_{ligpre}) and the screening utility of GEMDOCK has been evaluated by four screening sets including two benchmark data sets (TK and ER antagonists) and two self-developed datasets (hDHFR and ER agonists). The performance of GEMDOCK is superior to those of other public docking tools.

The post-docking analysis will be the point to be developed in the future. We could cluster docked compounds according to their chemical characteristics and docked poses. We will try to find the consistence and make use of it to mind more potential inhibitors. Besides we could also apply consensus score to GEMDOCK to modify the screening utility.

After verifying the performance of GEMDOCK, we applied it to the envelope protein of dengue virus to screen potential inhibitors from the chemical database. The candidates we recommended will be tested by biological experiments. If we can find any lead compound to become potential inhibitors, it will be a good start to refine them by lead optimization. On the other hand, we also applied GEMDOCK to the D-hydantoinase to identify novel substrates or inhibitors. With the mutual proof of biological experiments, we will identify the common interactions between substrates/inhibitors and the binding site to modify our scoring function. The consistence will be used to train our scoring function by some approaches, such as genetic algorithm. After training, the scoring function will be more specific when we screen the chemical database against the D-hydantoinase to find more novel substrates or inhibitors by pharmacological preferences.

Table 1. Atom Types of GEMDOCK

Atom type	Heavy atom name
Donor	Primary and secondary amines, sulfur, and metal atoms
Acceptor	Oxygen and nitrogen with no bound hydrogen
Both	Structural water and hydroxyl groups
Nonpolar	Other atoms (such as carbon and phosphorus)

Table 2. Atom Formal Charge of GEMDOCK

Formal charge	Atom name
Receptor:	
0.5	N atom in His (ND1 & NE2) and Arg (NH1 & NH2)
-0.5	O atom in Asp (OD1 & OD2) and Glu (OE1 & OE2)
1.0	N atom in Lys (NZ)
2.0	metal ions (MG, MN, CA, ZN, FE, and CU)
0	other atoms
Ligand:	
0.5	N atom in $-C(NH_2)_2^+$
-0.5	O atom in $-COO^-$, $-PO_2^-$, $-PO_3^-$, $-SO_3^-$, and $-SO_4^-$
1.0	N atom in $-NH_3^+$ and $-N^+(CH_3)_3$
0	other atoms

Table 3. Parameters of GEMDOCK

Parameter	Value of parameters
Initial step sizes	$\sigma = 0.8, \nu = \psi = 0.2$ (in radius)
Family competition length	$L = 2$
Population size	$N = 200$
Recombination rate	$p_c = 0.3$
No. of the maximum generation	60

Table 4. Interaction Preferences of Hot-spot Atoms of TK Evolved by Superimposing Known Active Ligands

Residue Id ^a	Atom Id ^b	Hot-spots weight ($CW(B_{ij})$)		Interaction type
		TK-ligand complex		
Q125	OE1	4.00		H-bond (NH ↔ O) (NH group) [23]
Q125	NE2	3.50		H-bond (O ↔ NH) (carbonyl group) [23]
Y101	OH	2.00		H-bond (OH ↔ OH) (hydroxyl group) [23]
R163	NH1	1.50		H-bond (OH ↔ N) (hydroxyl group) [23]
	CG			
	CD1			
	CD2			
Y172	CE1	2.50		Van der Waal force (C ↔ C) [23]
	CE2			
	CZ			

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

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

Table 5. Ligand Preferences Evolved from Known Ligands Are Used to Screen the Lead Compounds for TK, ER, hDHFR, and E Protein

Ligand name	Electrostatic preferences (Equation 9)			Hydrophilic preferences (Equation 10)			Molecular weight (Equation 11)	
	θ_{elec}	σ_{elec}	UB_{elec}	θ_{hb}	σ_{hb}	Ur_{hb}	μ_{mw}	K
TK-substrate	0	0	0	0.50	0.05	0.55	17.10	0.46
ER-antagonist	2.00	0.63	2.63	0.15	0.02	0.17	34.00	0.16
ER-agonist	0	0	0	0.25	0.06	0.31	21.40	0.38
hDHFR-ligand	4.00	2.11	6.11	0.40	0.05	0.45	29.70	0.24
E protein-ligand	0	0	0	0.30	0	0.30	20.00	0.41



Table 6. Comparison GEMDOCK with GOLD, FlexX, and DOCK on Docking 10 Known Substrates of the TK with X-ray Structures into Their Native Proteins and the Reference Protein, 1kim

Lidand Id ^a	GEMDOCK				GOLD	FlexX	DOCK
	Native protein		Reference Protein				
	Pharma. ^b	Pharma.	Pharma.	Pharma.			
	consensus (yes)	consensus (no)	consensus (yes)	consensus (no)			
<i>1e2k.TMC</i>	1.08	0.99	0.75	0.79	1.19	1.11	7.56
<i>1e2m.HPT</i>	0.59	0.72	0.41	0.37	0.49	4.18	1.02
<i>1e2n.RCA</i>	0.42	0.29	1.54	1.41	2.33	13.30	9.62
<i>1e2p.CCV</i>	0.34	0.64	0.58	0.53	0.93	3.65	2.02
<i>1ki2.GA2</i>	0.55	0.38	3.56	2.15	3.11	6.07	3.01
<i>1ki3.PE2</i>	0.66	0.78	3.34	3.29	3.01	5.96	4.10
<i>1ki6.AHU</i>	0.57	0.65	0.43	0.39	0.63	0.88	1.16
<i>1ki7.ID2</i>	0.94	0.83	0.45	0.56	0.77	1.03	9.33
<i>1kim.THM</i>	0.39	0.53	0.47	0.48	0.72	0.78	0.82
<i>2ki5.AC2</i>	0.61	0.62	2.94	2.95	2.74	2.71	3.08

^a The four characters and three characters separated by a period denote the PDB code and the ligand name in PDB, respectively.

^b Pharmacological consensus

Table 7. Comparison of GEMDOCK with Four Methods by False Positive Rates (%) on Screening 990 Compounds and 10 Known Substrates of the TK

True positive%	GEMDOCK ^a	GEMDOCK ^b	Surflex ^c	DOCK ^c	FlexX ^c	GOLD ^c
80	4.7 ^d (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 GEMDOCK without pharmacophore and ligand preferences.

^b GEMDOCK with pharmacophore and ligand preferences.

^c These data were derived from reference 18 and 24.

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



Table 8. Pharmacological Weights of Hot-spot Atoms of the ER α -antagonist Complex and ER α -agonist Complex Are Evolved by Overlapping Known Active Ligands

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

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

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

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

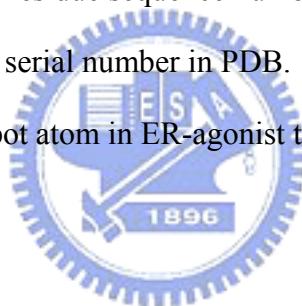


Table 9. Comparison GEMDOCK with GOLD on Docking Four Antagonists and Four Agonists with X-ray Structures into Their Native Proteins and the Reference Proteins That Are 3ert and 1gwr, Respectively

Lidand Id ^a	GEMDOCK				GOLD	
	Native protein		Reference Protein		Native protein	Reference Protein
	Pharma. consensus ^b	Pharma. consensus ^b	Pharma. consensus	Pharma. consensus		
	(yes)	(no)	(yes)	(no)		
EST01 (1err.RAL)	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
Tetrahydrochiolin (1uom.PTI)	0.80	0.43	0.89	0.85	0.56	1.56
ESA01 (1gwr.EST)	0.66	0.64	0.66	0.64	0.54	0.54
ESA02 (1l2i.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 The four characters and three characters separated by a period denote the PDB code and the ligand name in PDB, respectively.

^b Pharmacological consensus

Table 10. Comparison of GEMDOCK with Four Methods by False Positive Rates (%) on Screening 990 Compounds and 10 Known Antagonists

True positive%	GEMDOCK ^a	GEMDOCK ^b	Surflex ^c	DOCK ^c	FlexX ^c	GOLD ^c
80	1.5 ^d (15/990)	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)	1.9	18.9	- ^e	23.4

^a GEMDOCK without pharmacophore and ligand preferences.

^b GEMDOCK with pharmacophore and ligand preferences.

^c These data were derived from reference 18 and 24.

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

^e FlexX couldn't find the docked solution of RU-58668 (EST09).



Table 11. Interaction Preferences of Hot-spot Atoms of hDHFR Evolved by Overlapping 10 Known Active Ligands

Residue Id ^a	Atom Id ^b	Hot-spots weight ($CW(B_{ij})$)	
		hDHFR-ligand complex	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)

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

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

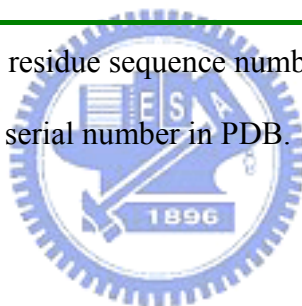


Table 12. Comparison GEMDOCK with GOLD on Docking 10 Known Ligands of the hDHFR with X-ray Structures into Their Native Proteins and the Reference Protein, 1hfr

Ligand Id ^a	GEMDOCK				GOLD	
	Native protein		Reference Protein		Native protein	Reference Protein
	Pharma. consensus ^b	Pharma. consensus	Pharma. consensus	Pharma. consensus		
	(yes)	(no)	(yes)	(no)		
<i>1boz.PRD</i>	1.20	1.13	1.00	0.95	2.03	2.43
<i>1dlr.MXA</i>	0.51	0.58	0.70	0.71	0.70	3.03
<i>1dls.MTX</i>	0.58	0.59	1.33	0.64	1.06	1.20
<i>1drf.FOL</i>	0.66	1.39	1.39	0.98	1.48	1.99
<i>1hfr.MOT</i>	0.62	0.79	0.81	0.87	0.72	1.21
<i>1kms.LIH</i>	1.10	0.68	1.16	1.17	0.46	0.65
<i>1kmv.LII</i>	0.35	0.35	0.94	0.90	0.50	2.68
<i>1mvs.DTM</i>	0.78	1.03	0.86	0.68	1.12	0.70
<i>1ohj.COP</i>	1.23	1.34	1.34	1.16	2.56	2.16
<i>2dhf.DZF</i>	0.86	0.53	1.09	1.45	1.17	2.09

^a The four characters and three characters separated by a period denote the PDB code and the ligand name in PDB, respectively.

^b Pharmacological consensus

Table 13. Comparison of GEMDOCK with GOLD by False Positive Rates (%) on Screening 990 Compounds and 10 Known Ligands of the hDHFR

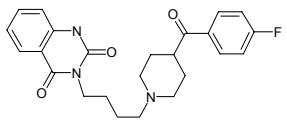
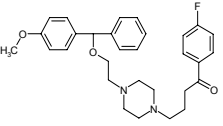
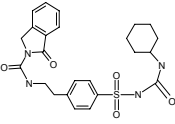
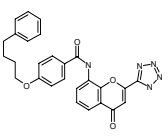
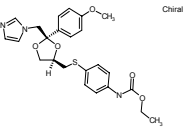
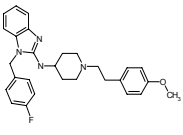
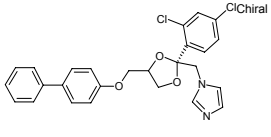
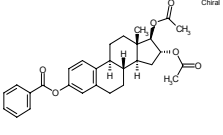
True positive %	GEMDOCK			GOLD	
	None ^a	Interaction preference ^b	Ligand preference ^c	Both ^d	
80	2.53	2.02	0.91	0.91	3.54
90	5.45	2.02	1.72	1.52	5.96
100	8.59	2.32	2.42	2.02	25.56

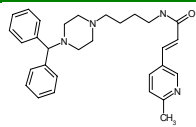
^{a,b,c,d} Use E_{bind} , $E_{bind} + E_{pharma}$, $E_{bind} + E_{ligpre}$, and E_{tot} as the scoring function. These energy

terms are defined in Equation 1.



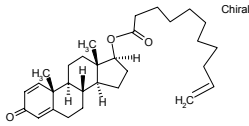
Table 14. The 35 Molecules of Intersection of Top 200 Scorer from the Screening Result of GEMDOCK and GOLD.

Structure	MDL number	Molecular weight	Generic name
	MCMC00005490	423.492	BUTANSERIN
	MCMC00004737	490.623	MOBENZOXAMINE
	MCMC00004946	484.579	GLISINDAMIDE
	MCMC00006316	481.515	PRANLUKAST
	MCMC00006126 ^a	469.559	ERBULOZOLE
	MCMC00004876	458.584	ASTEMIZOLE
	MCMC00004528	481.377	DOCONAZOLE
	MCMC00010202	476.566	Holin-Depot



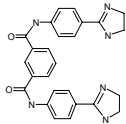
MCMC00007477

468.648 TAGORIZINE



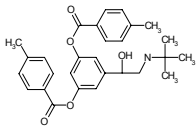
MCMC00002678

452.675 BOLDENONE
UNDECYLENATE



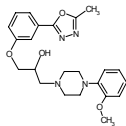
MCMC00002249

452.52 ISOTIC



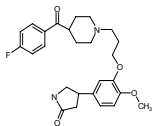
MCMC00005097

461.563 TOBUTEROL



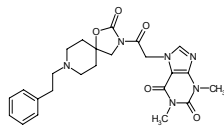
MCMC00005568

424.504 NESAPIDIL



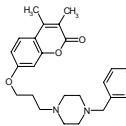
MCMC00005893

454.546 LIDANSERIN



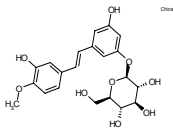
MCMC00005671

480.528 SPIROFYLLINE



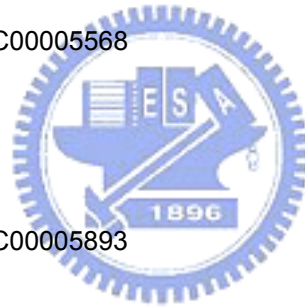
MCMC00003872

440.974 PICUMAST

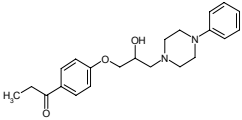
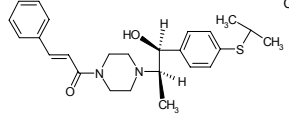
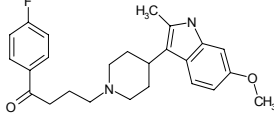
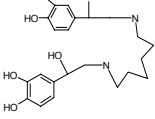
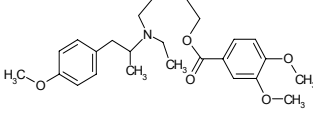
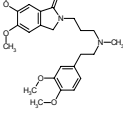
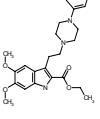
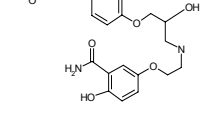
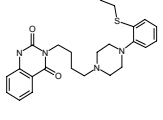


MCMC00009993 ^a

420.412 Ponticin



	MCMC00007533 ^a	384.436 TOBORINONE
	MCMC00010200	437.584 Diphenethindole
	MCMC00005777	441.51 PERBUFYLLINE
	MCMC00004752	391.558 FENRETINIDE
	MCMC00002264	449.592 DIFLUANINE
	MCMC00004923	380.491 MINDODIOLOL
	MCMC00000488	380.453 SCARLET RED
	MCMC00004342	483.616 ROPITOIN
	MCMC00006389	425.513 TULOPAFANT

	MCMC00010062	368.474 Centpropazine
	MCMC00005434	424.606 SUNAGREL
	MCMC00004095	408.521 MINDOPERONE
	MCMC00001935 ^a	420.51 HEXOPRENALINE
	MCMC00002025 ^a	429.561 MEBEVERINE
	MCMC00005183	428.533 FALIPAMIL
	MCMC00003427	437.544 ALPERTINE
	MCMC00005147 ^a	420.467 TRIGEVOLOL
	MCMC00004119	452.623 TIOPERIDONE

^a These molecules formed hydrogen bonds with Q52.

Table 15. Specific Activity of D-Hydantoinase

Substrate	Concentration		Specific activity ^{a,b} (μ mol/min/mg)
	(nm)	(mM)	
Allantoin	250	10	1.7 \pm 0.1 ^c
Parabanic acid	295	1	1.9 \pm 0.2

^a Enzymatic activity was measured in 25 and buffer (100 mM Tris-HCl at pH 7.5) were used.

^b Extinction coefficient of each substrate was determined experimentally by direct measurement with a spectrophotometer.

^c Each value is the average of at least three measurements, which differ by less than 10%.



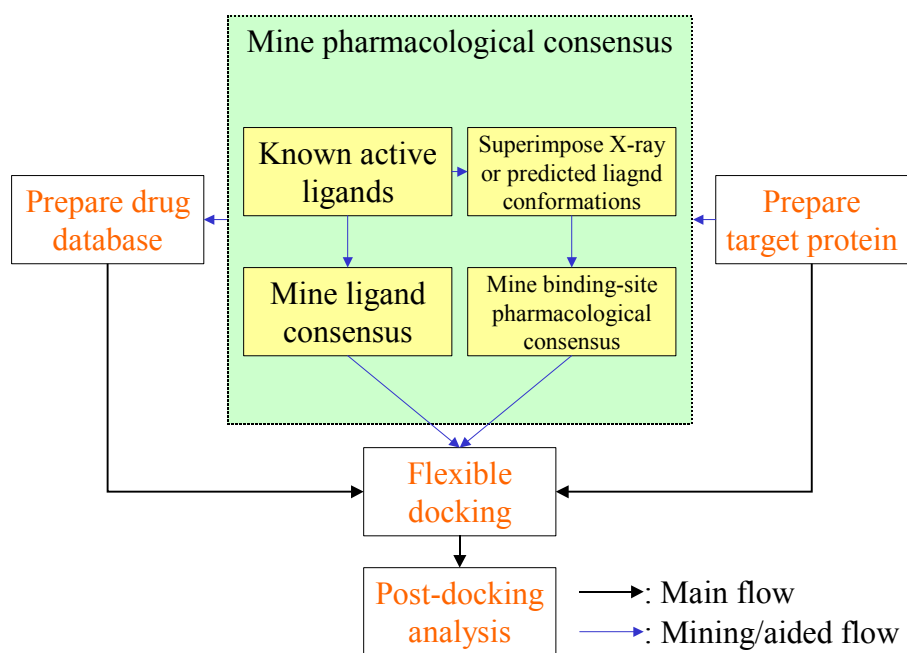
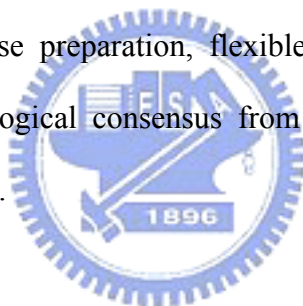


Figure 1. The main steps of GEMDOCK for virtual database screening, including the target protein and compound database preparation, flexible docking, and post-docking analysis. GEMDOCK yields pharmacological consensus from the target protein and known active ligands when they are available.



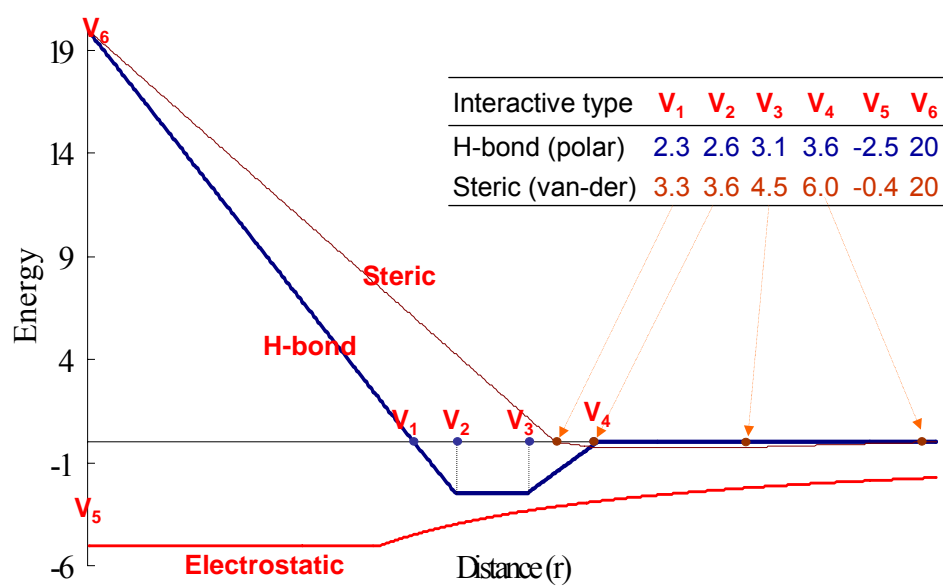
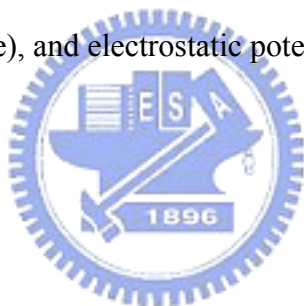


Figure 2. The linear energy function of the pair-wise atoms for the steric interactions (light line), hydrogen bonds (bold line), and electrostatic potential (middle line) in GEMDOCK.



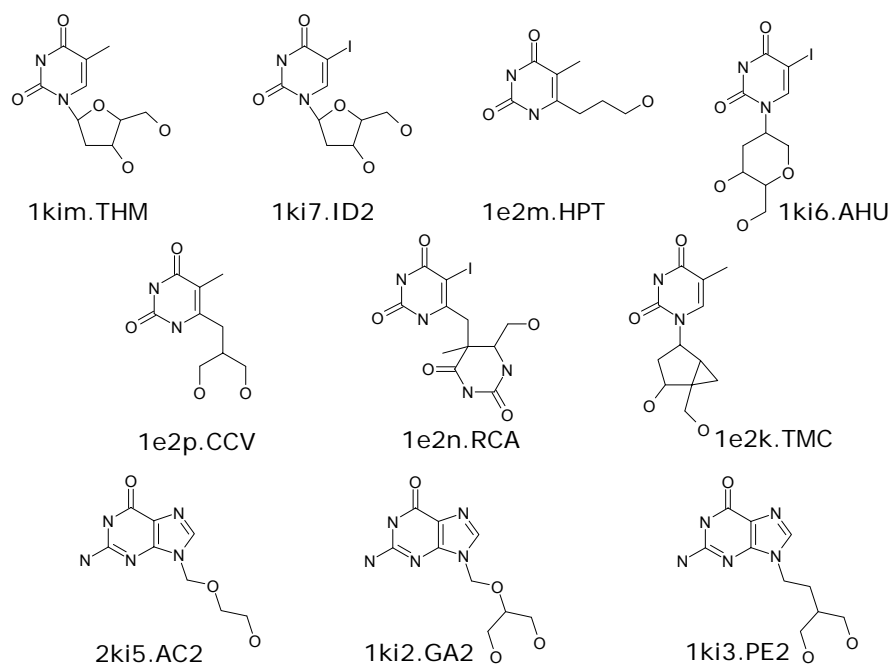


Figure 3. Ten known active ligands of HSV-1 thymidine kinase. The abbreviations are as follows: *1kim.THM*, deoxythymidine; *1ki7.ID2*, 5-iododeoxyuridine; *1e2m.HPT*, 6-(3-hydroxy-propyl-thymine); *1ki6.AHU*, 5-iodouracil anhydrohexitol nucleoside; *1e2p.CCV*, 6-(3-hydroxy-2-hydroxymethylpropyl)-5-methyl-1H-pyrimidine-2,4-dione; *1e2n.RCA*, 6-[6-hydroxy-methy-5-methyl-2,4-dioxohexahydropyrimidin-5-yl-methyl]-5-methyl-1H-pyrimidin-2,4-dione; *1e2k.TMC*, (North)methanocarbothymidine; *2ki5.AC2*, aciclovir; *1ki2.GA2*, ganciclovir; *1ki3.PE2*, penciclovir.

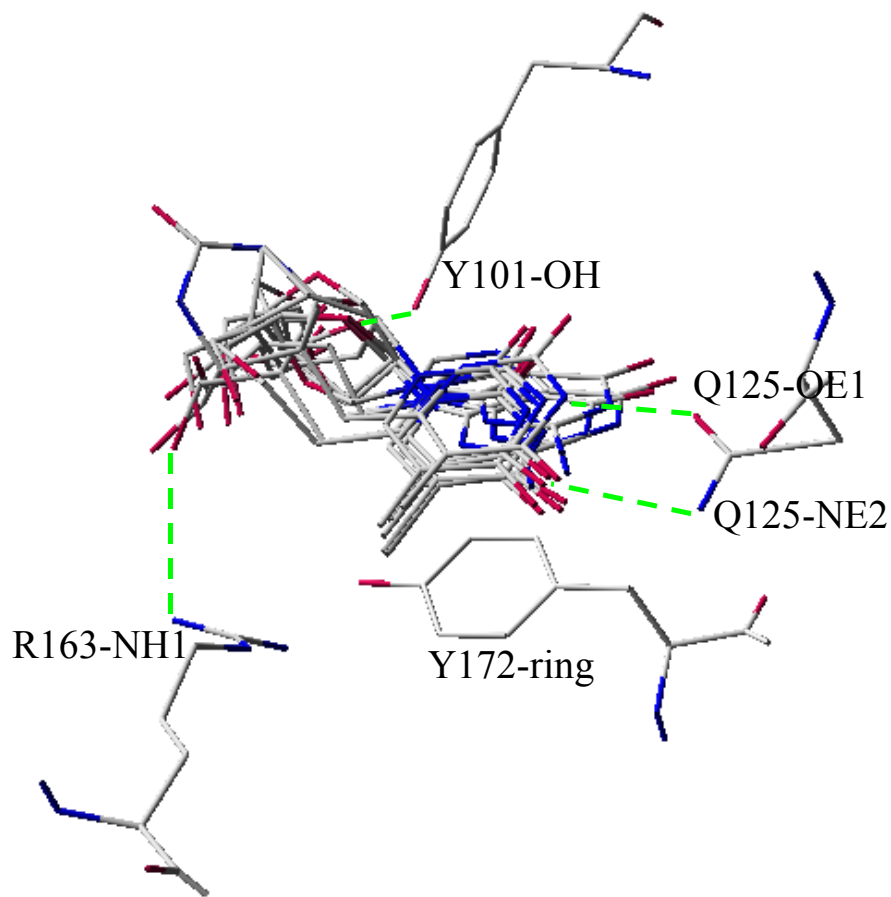
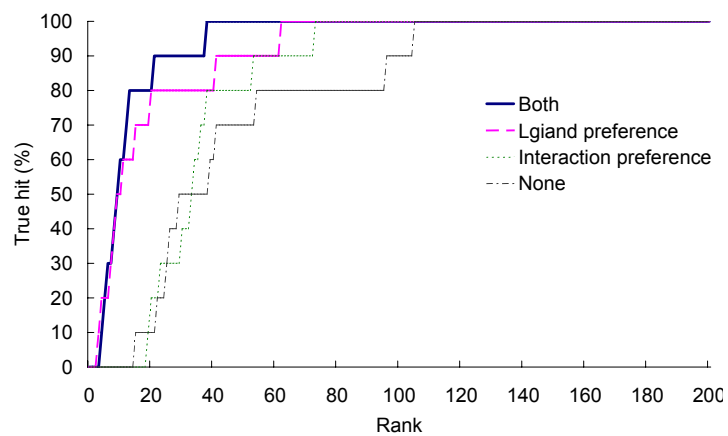
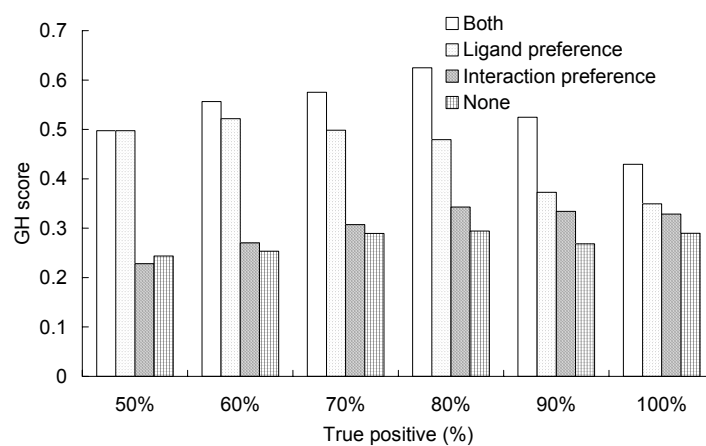


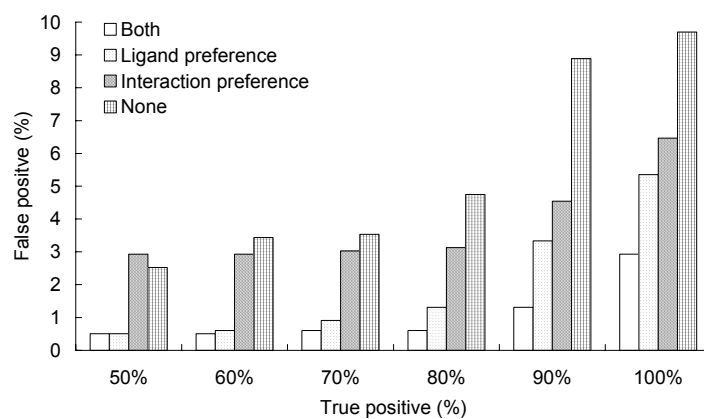
Figure 4. Superimposing ten crystal structures of TK. Four important residues of the pharmacological consensus were identified and marked. The dash lines indicate the hydrogen binding. The phenolic ring of Y172 formed stack force with the ligand.



A



B



C

Figure 5. The overall accuracy of GEMDOCK using different combinations of pharmacophore preferences in screening the substrates of TK from a testing set with 1000 compounds.

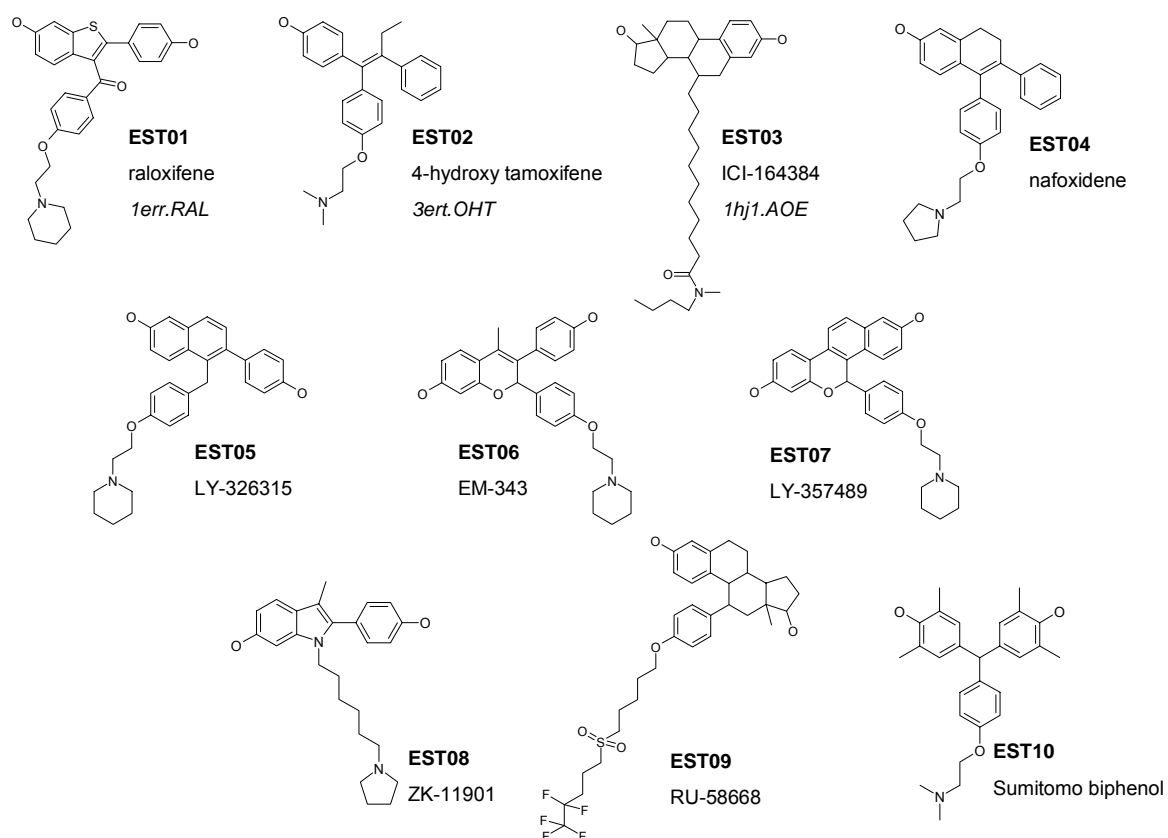


Figure 6 . Ten antagonists of ER α .



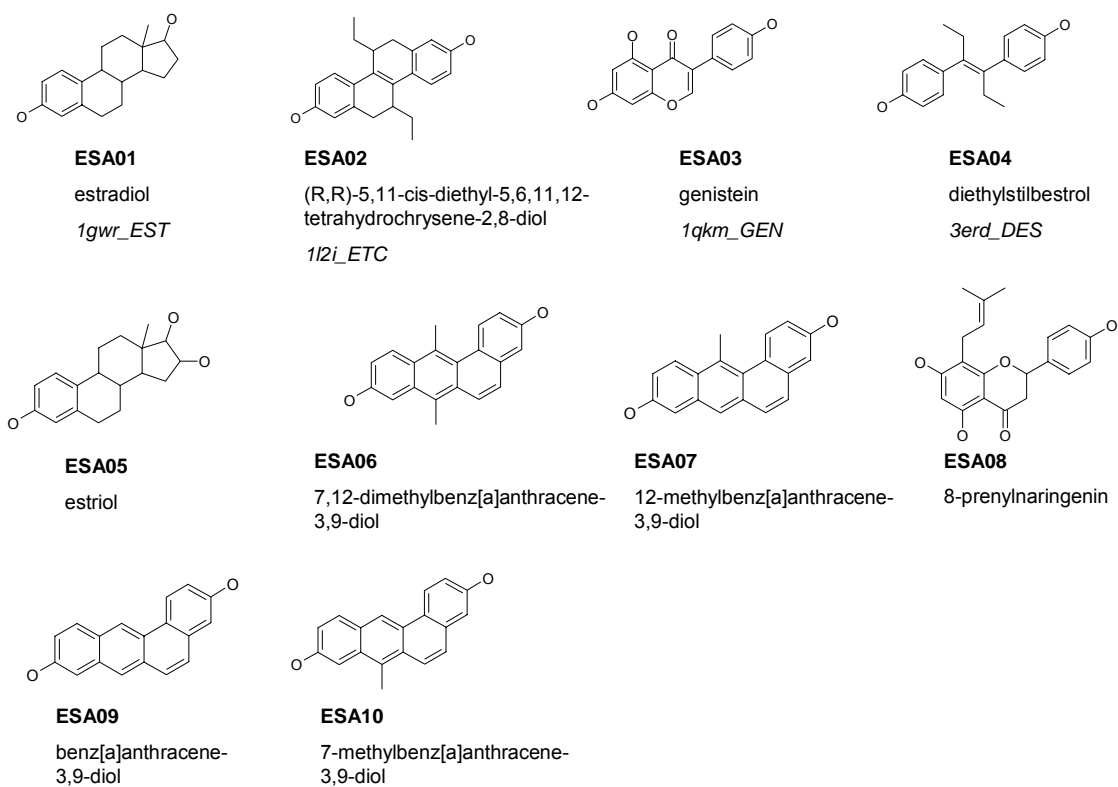


Figure 7. Ten agonists of ER α derived from the reference 35.



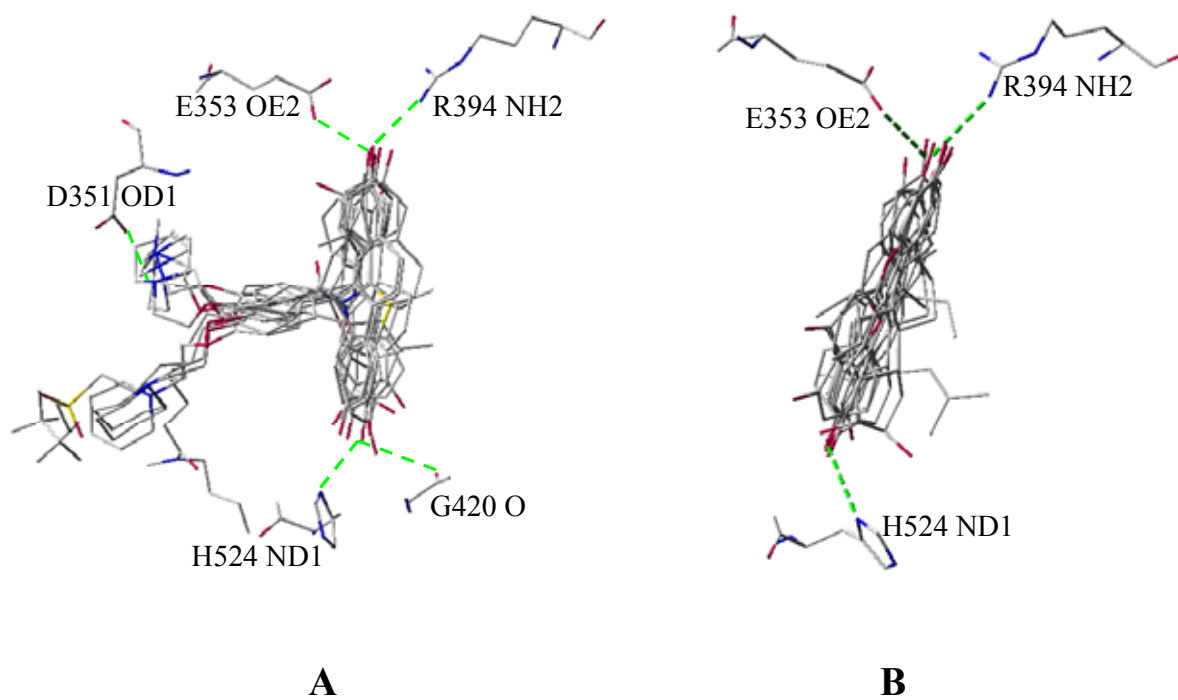
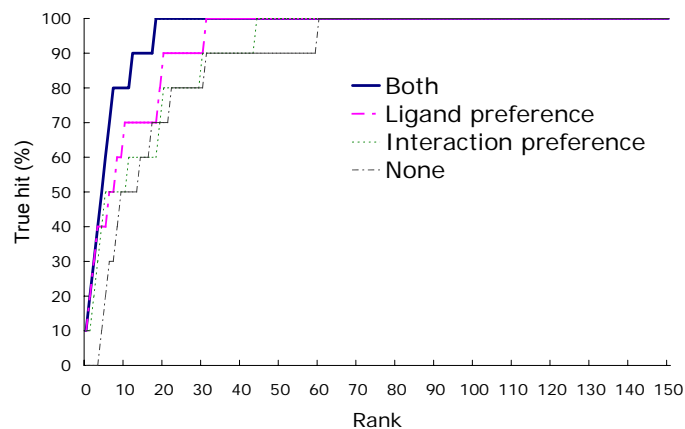
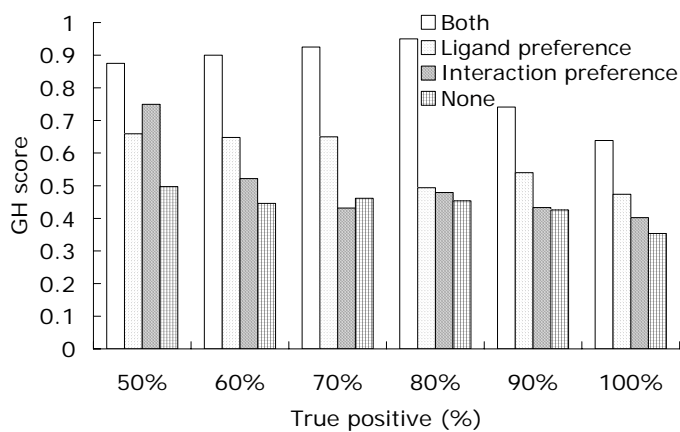


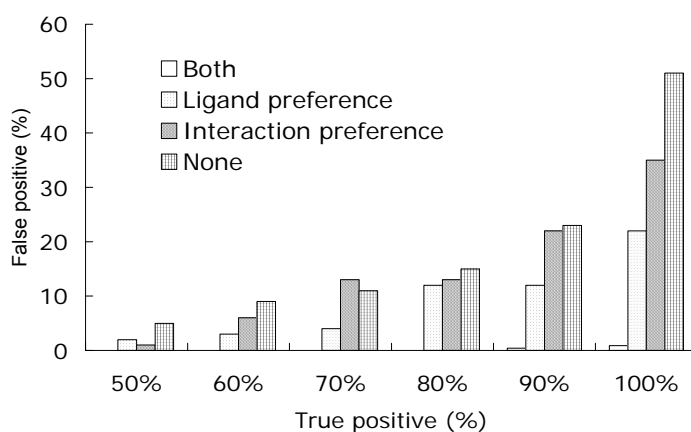
Figure 8. (A) Superimposing ten docked poses of ER α antagonists. Five important residues of the pharmacological consensus were identified and marked. (B) Superimposing ten docked poses of ER α agonists. Three important residues of the pharmacological consensus were identified and marker. The dash lines indicate the hydrogen binding.



A

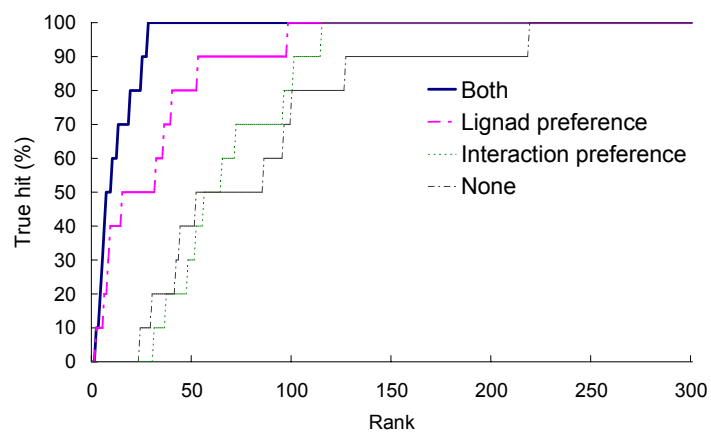


B

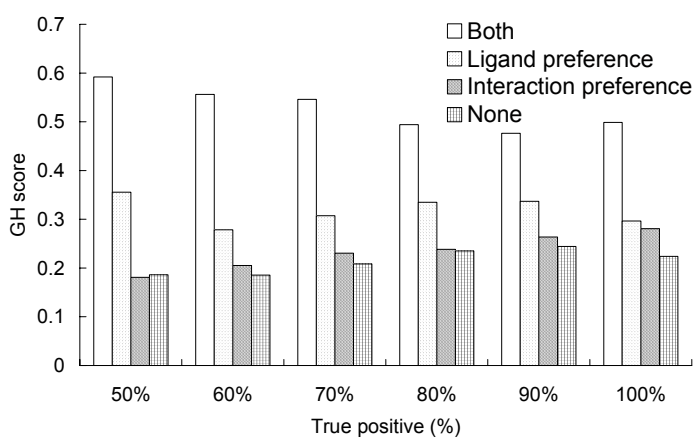


C

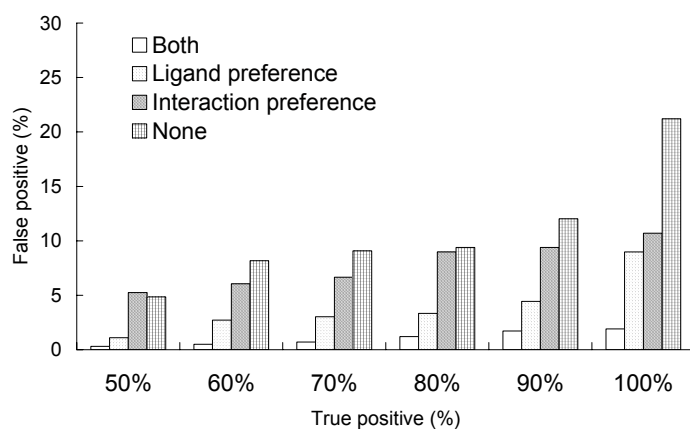
Figure 9. The overall accuracy of GEMDOCK using different combinations of pharmacophore preferences in screening ER α antagonists from a testing set with 1000 compounds.



A



B



C

Figure 10. The overall accuracy of GEMDOCK using different combinations of pharmacophore preferences in screening ER α agonists from a testing set with 1000 compounds.

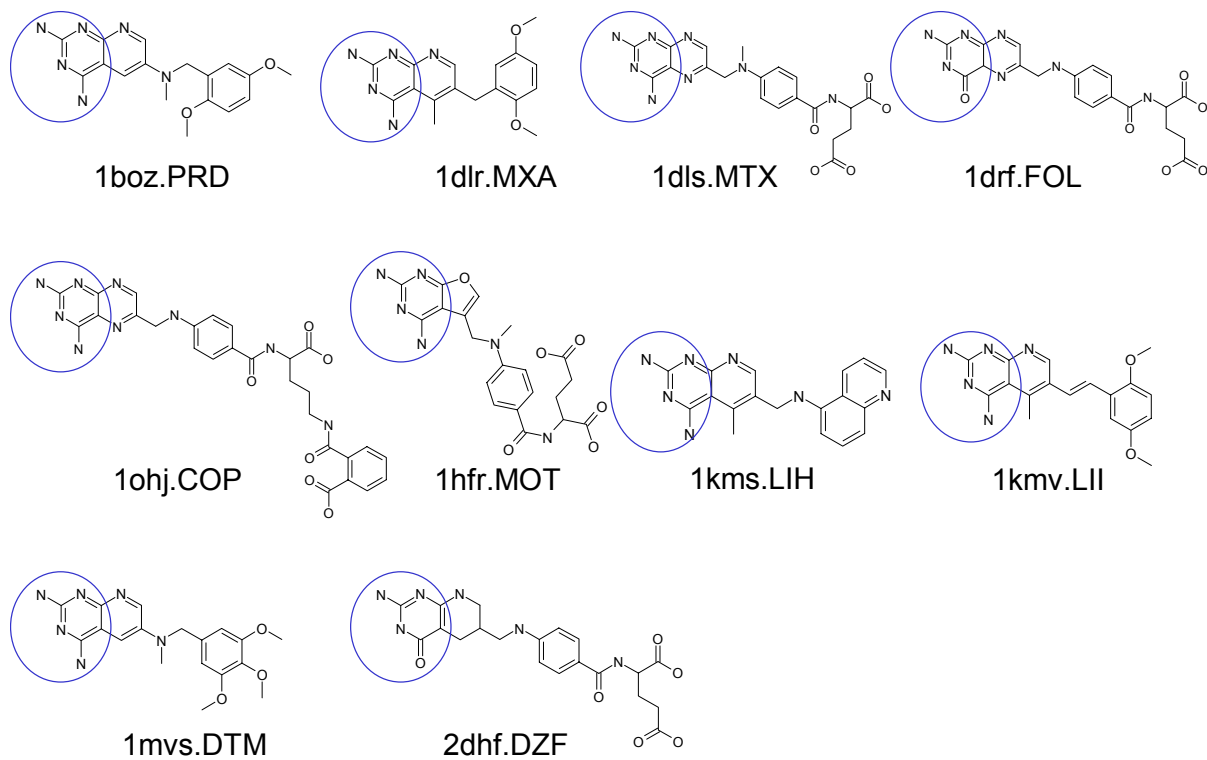


Figure 11. Ten known active ligands of hDHFR derived from the PDB.



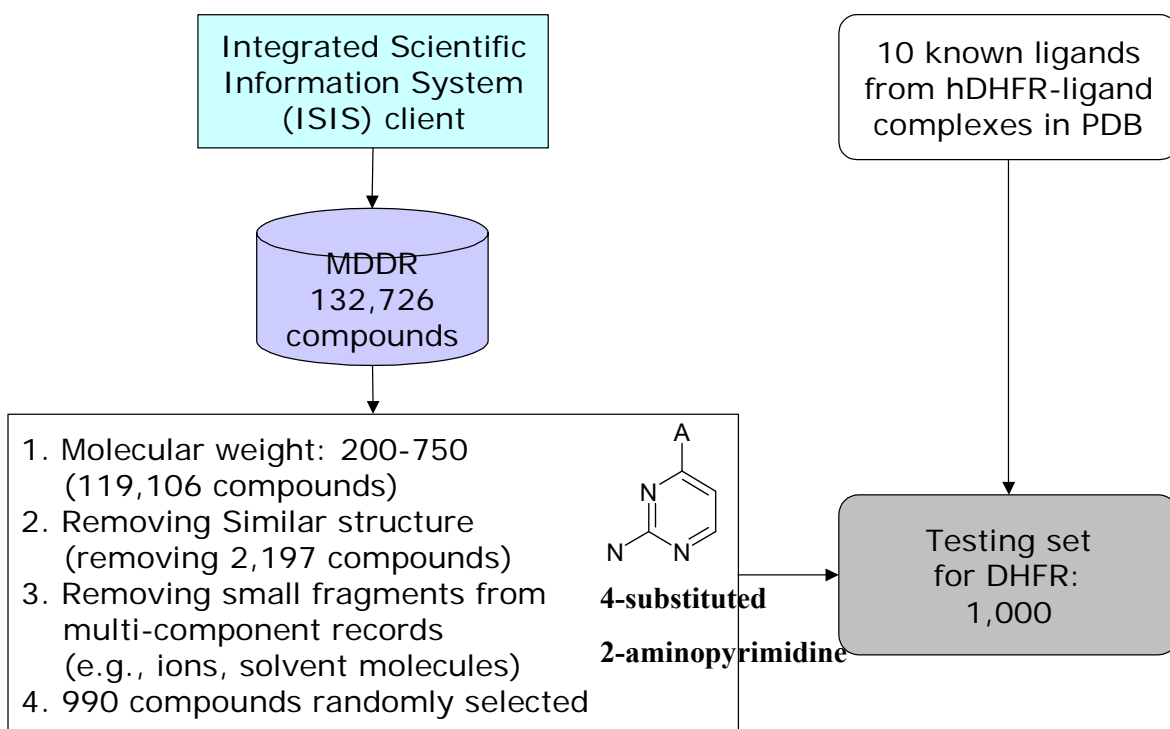


Figure 12. Preparation of the testing set of hDHFR. The testing set included ten known ligands from the PDB and 990 randomly chosen molecules from the MDDR. We have filtered the MDDR first with molecular weights. Then we removed similar structures of 4-substituted-2-aminopyrimidine and small fragments from multi-component records. Finally we randomly selected 990 compounds from the remainder to form the testing set.

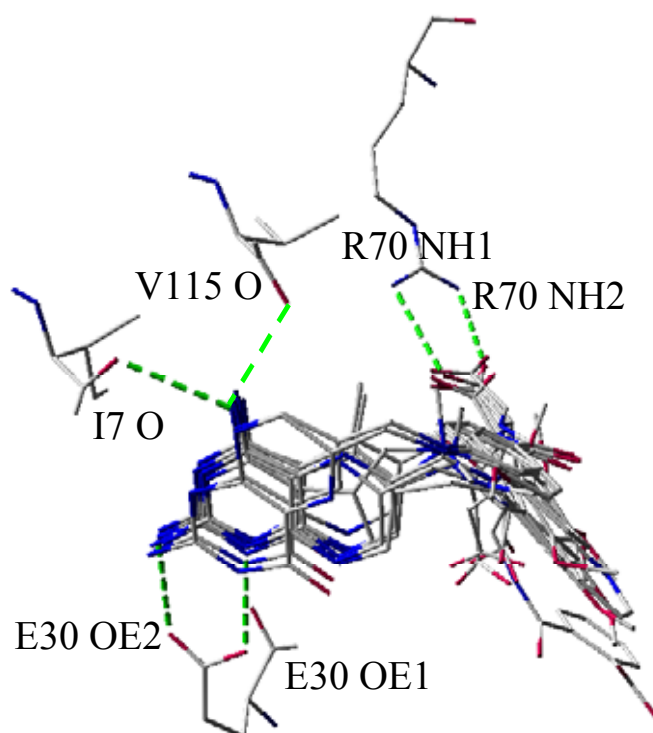
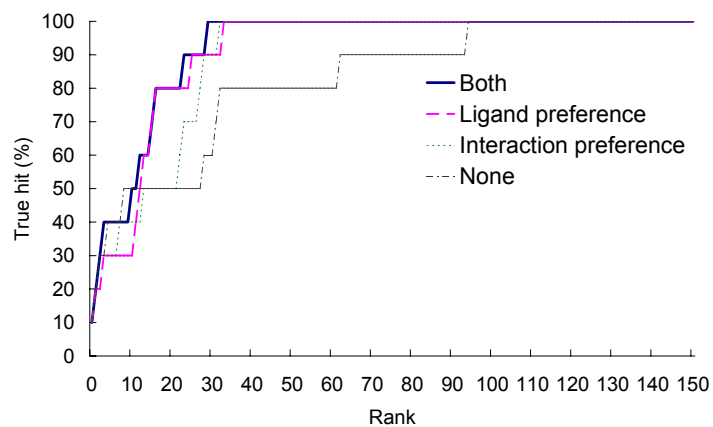
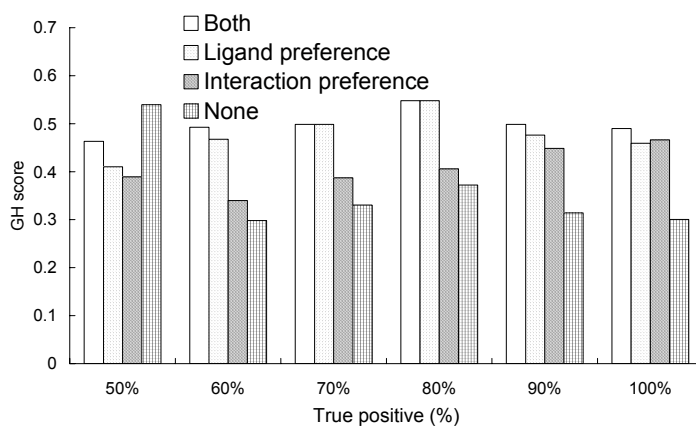


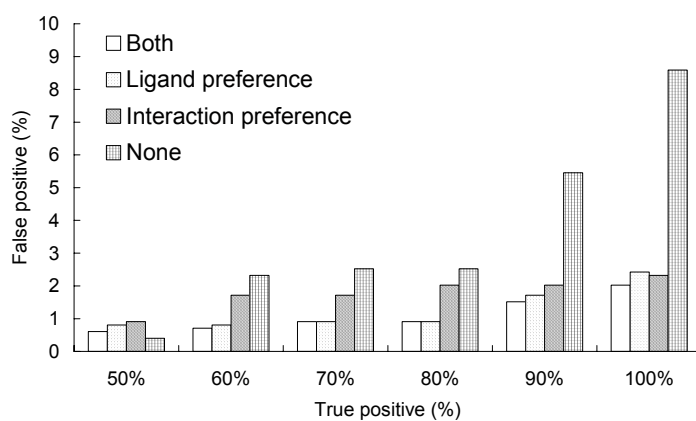
Figure 13. Superimposing ten known ligands of hDHFR. Four important residues of the pharmacological consensus were identified and marked. The dash lines indicate the hydrogen binding.



A



B



C

Figure 14. The overall accuracy of GEMDOCK using different combinations of pharmacological preferences in screening ligands of hDHFR from a testing set with 1000 compounds.

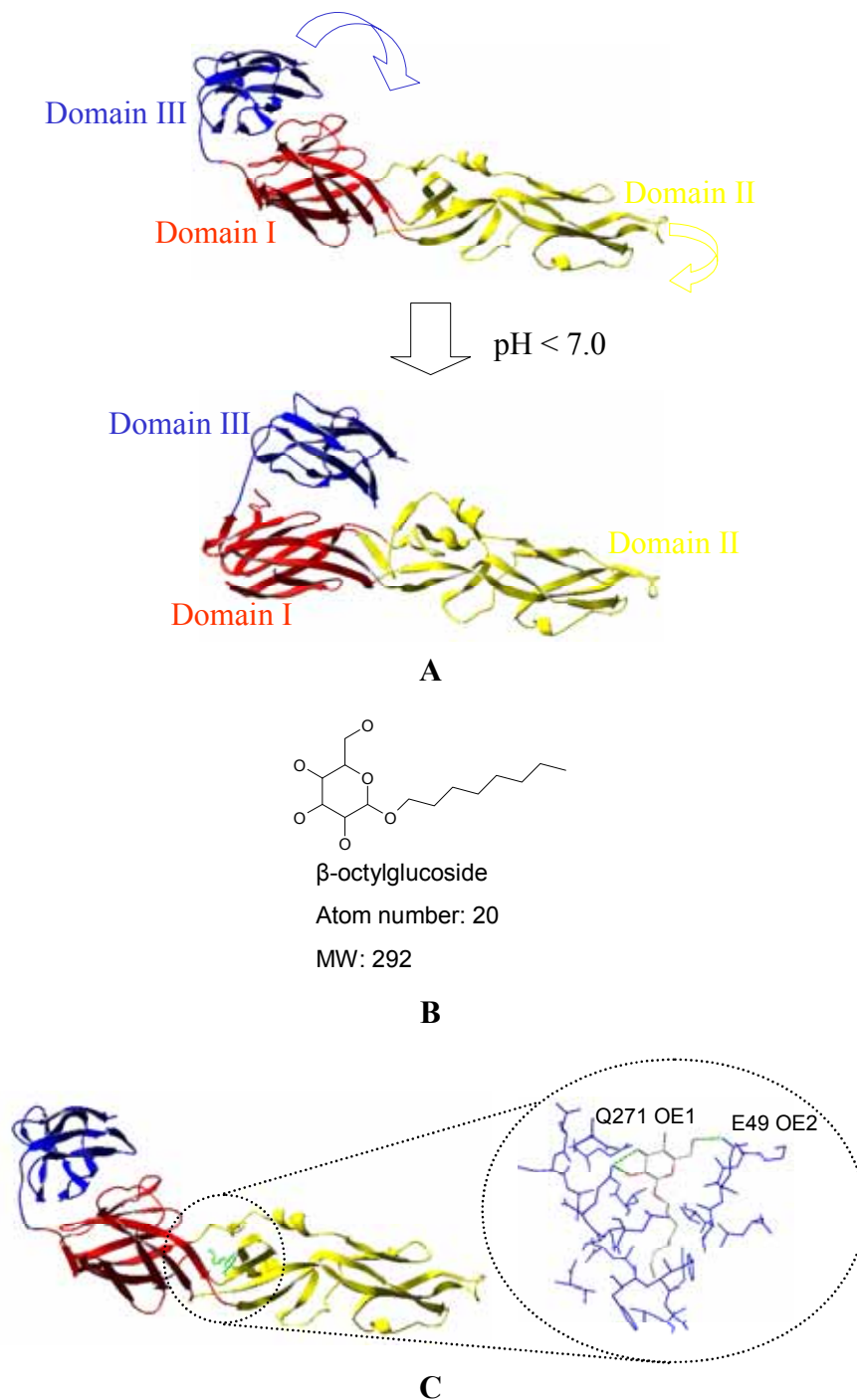


Figure 15. (A) The conformational rearrangement of the E protein to the reduced pH of an endosome. (B) The detergent molecule of *n*-octyl- β -D-glucoside (β -OG) complex with the E protein in the crystal structure (PDB entry: 1oke). (C) Mutations affecting the pH threshold for fusion map to the binding pocket that we propose is a hinge point in the fusion-activating conformational change. Detergent binding marks the pocket as a potential site for small-molecule fusion inhibitors.

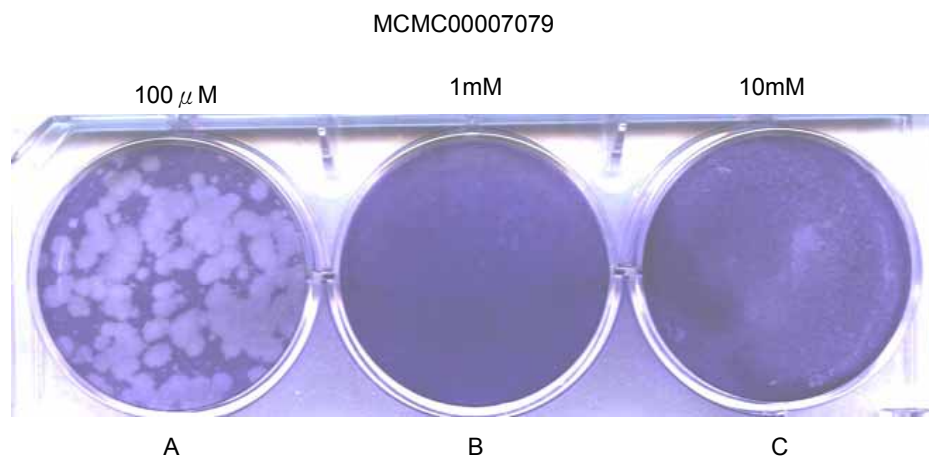
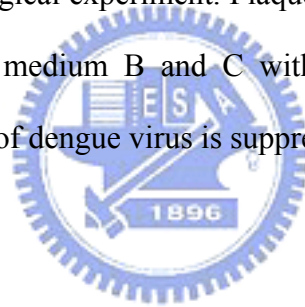


Figure 16. Results of the biological experiment. Plaques in the medium mean that the cell are infected by dengue virus. In medium B and C with concentrations 1mM and 10mM of MCMC00007079, the activity of dengue virus is suppressed and cells are alive.



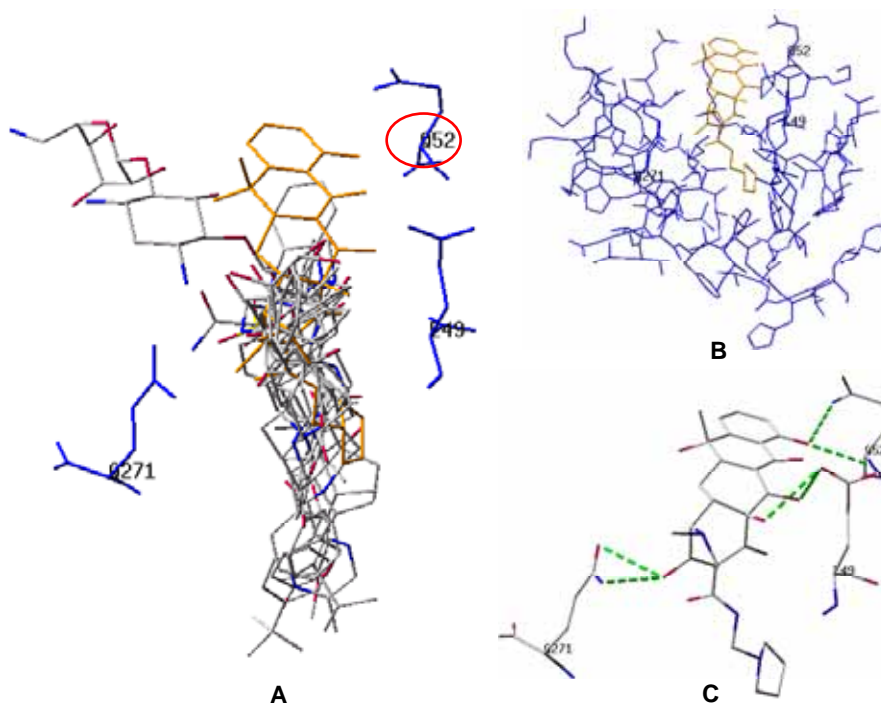
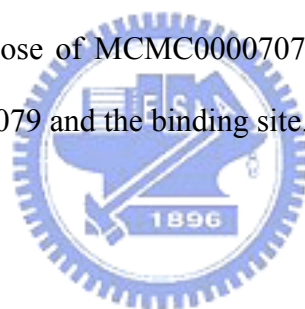


Figure 17. (A) Overlapping docked poses of the nine compounds tested by the biological experiment. (B) The docked pose of MCMC00007079. (C) Hydrogen bonds (dashed lines) formed between MCMC00007079 and the binding site.



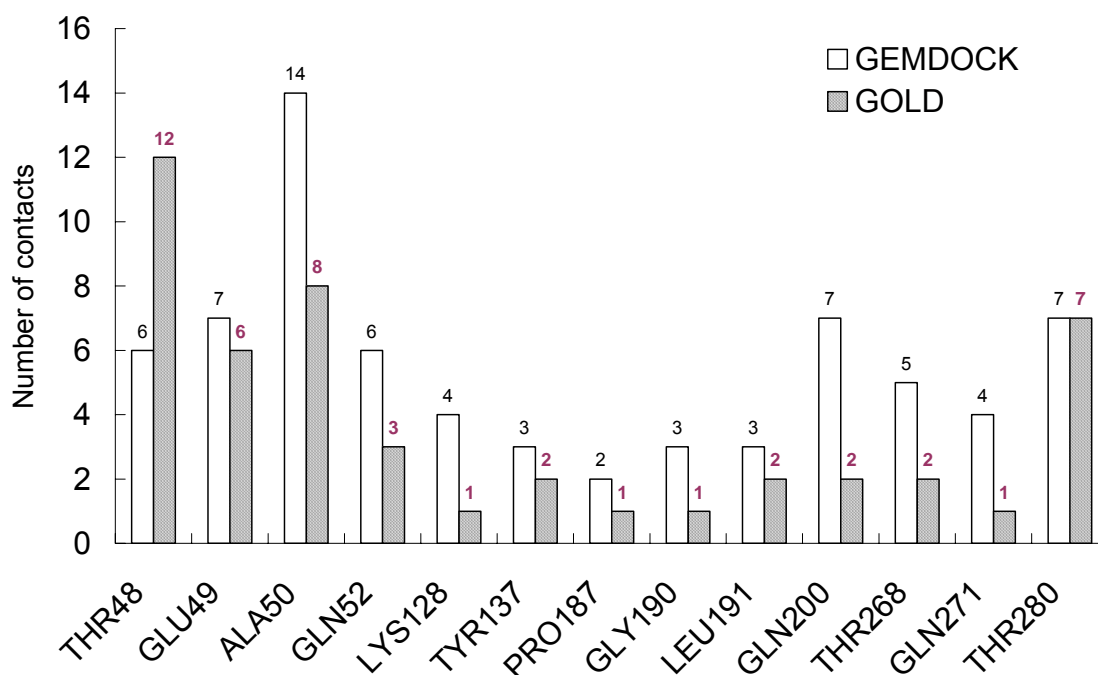


Figure 18. The number of contacts between the 35 compounds and residues of the E protein. Contacts mean that the docked ligand forms hydrogen bonds with specific residues. Among these residues, Q52 was mutated to affect the pH threshold of fusion in the previous experiment. If docked poses form more contacts with Q52, these compounds have more possibility to become potential inhibitors.

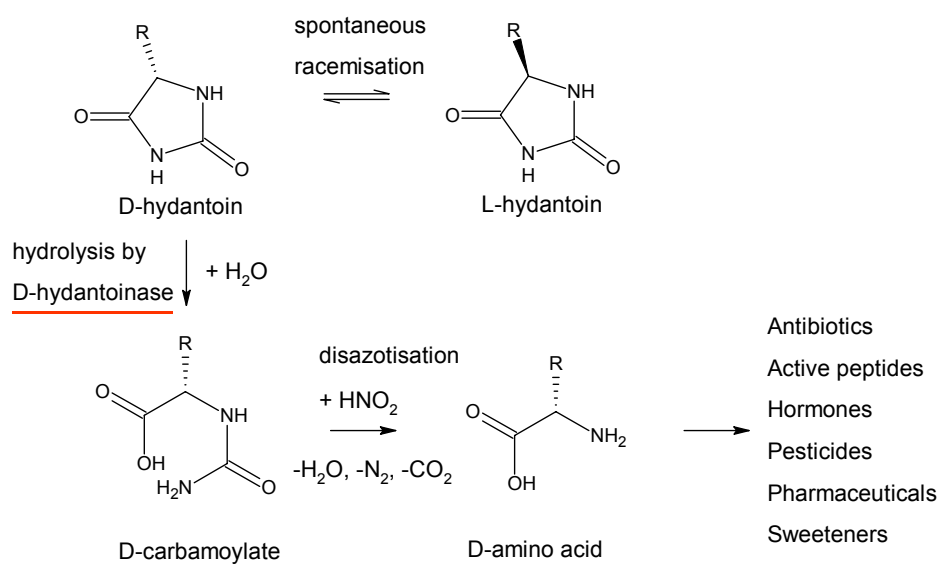
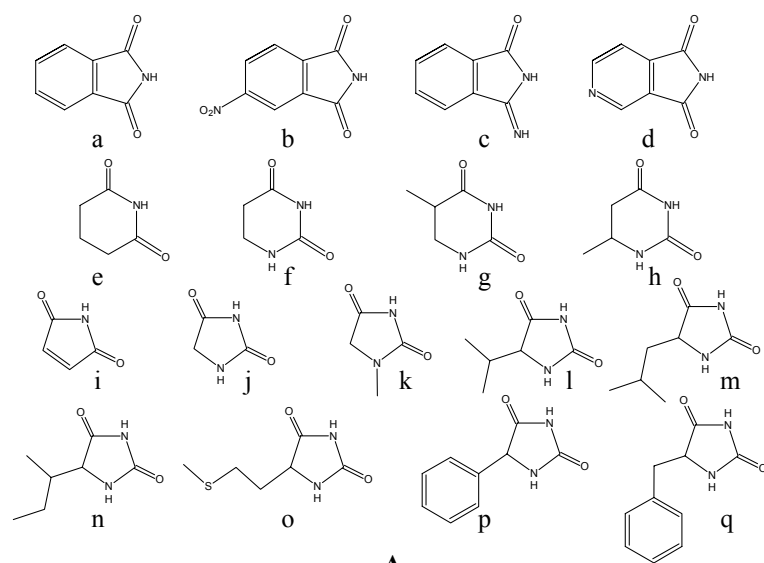
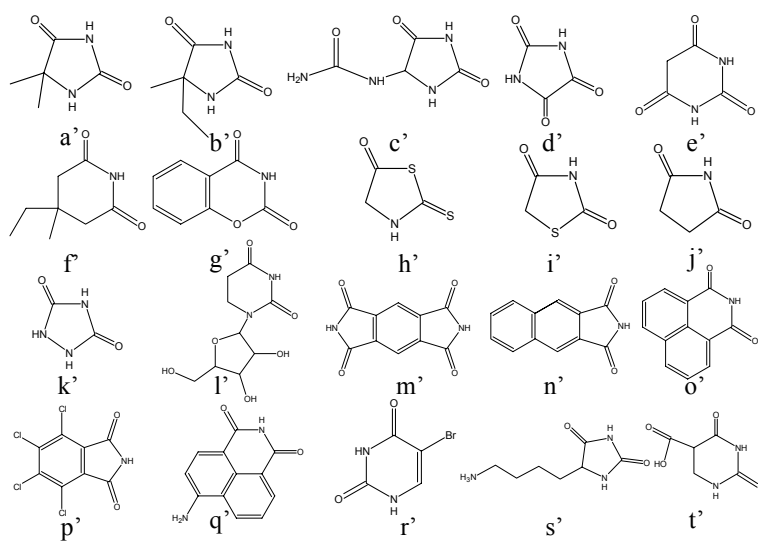


Figure 19. The pathway to produce D-amino acid. D-hydantoinase (red underline) serves as an important enzyme in the pathway.



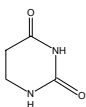
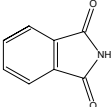
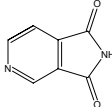
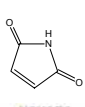
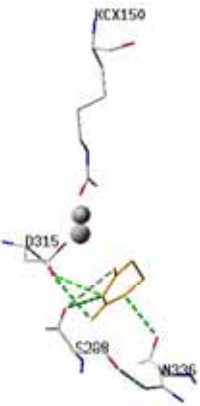
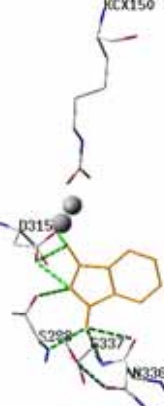
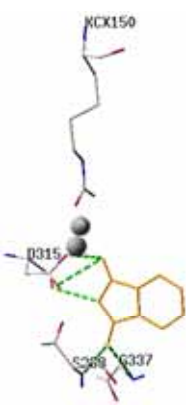
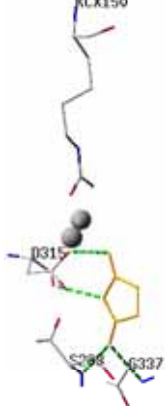


A



B

Figure 20. (A) Training set: 17 molecules were substrates of the D-hydantoinase. (B) Testing set: 20 molecules.

Name	Dihydrouracil	Phthalimide	3,4-pyridine dicarboximide	maleimide
Structure				
Docked pose				
K_m (mM)	50 ± 20	5.8 ± 1.9	40 ± 10	180 ± 30
k_{cat} (min^{-1})	860 ± 390	460 ± 120	3400 ± 850	9800 ± 1200

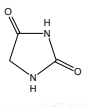
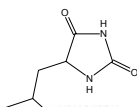
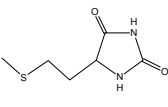
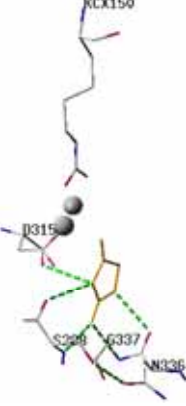
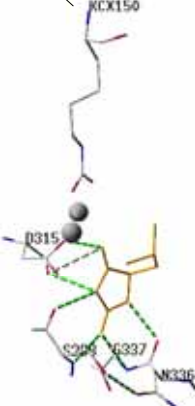
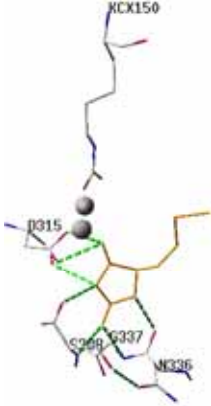
Name	Hydantoin	5-leuciny-hydantoin	5'-methionyl-hydantoin
Structure			
Docked pose			
K_m (mM)	850 ± 170	4.3 ± 0.2	8.6 ± 0.8
k_{cat} (min^{-1})	7700 ± 1100	3850 ± 60	40 ± 2

Figure 21. K_m and k_{cat} values of substrates and their docked poses. The dashed lines indicate the hydrogen binding formed between substrates and the protein.

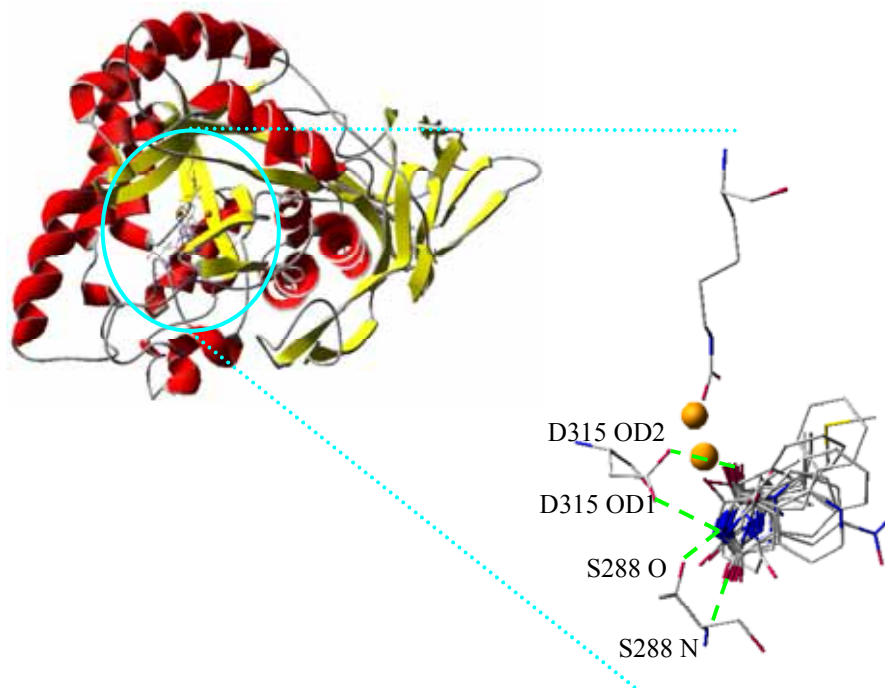


Figure 22. Superimposing 17 known substrates of D-HYD. Two important residues of the pharmacological consensus were identified and marked. The dash lines indicate the hydrogen binding.



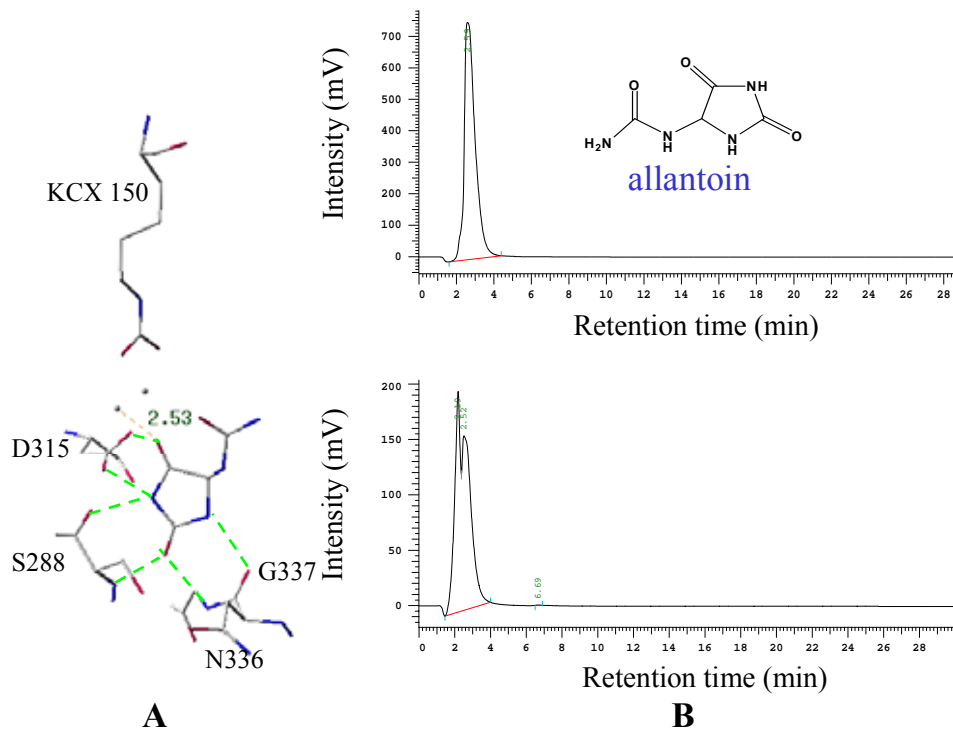
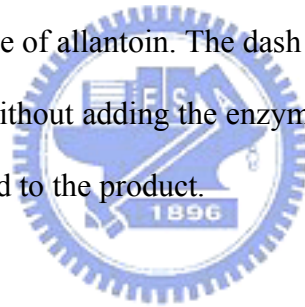


Figure 23. (A) The docked pose of allantoin. The dash lines indicate the hydrogen binding. (B) Upper figure is the spectrum without adding the enzyme. After adding D-hydantoinase (lower figure), allantoin is decomposed to the product.



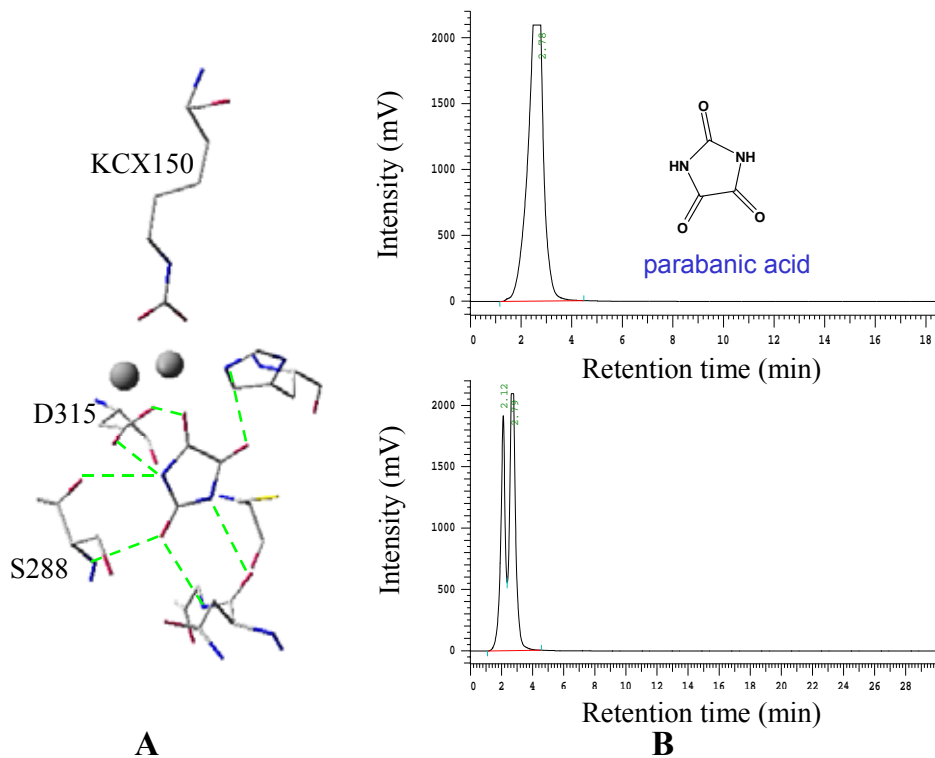


Figure 24. (A) The docked pose of parabanic acid. The dash lines indicate the hydrogen binding. (B) Upper figure is the spectrum without adding the enzyme. After adding D-hydantoinase (lower figure), parabanic acid is decomposed to the product.

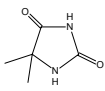
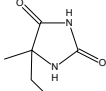
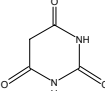
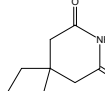
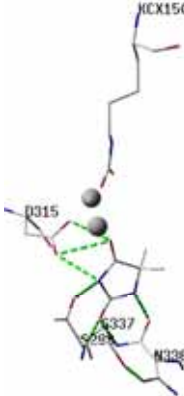
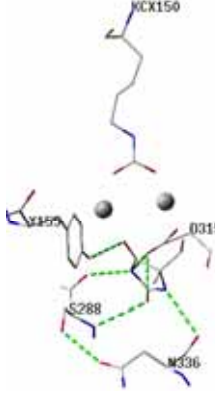
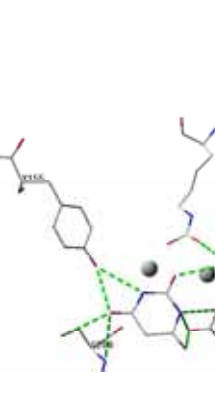
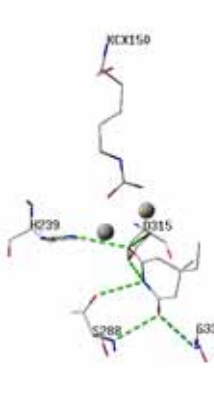
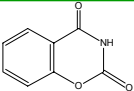



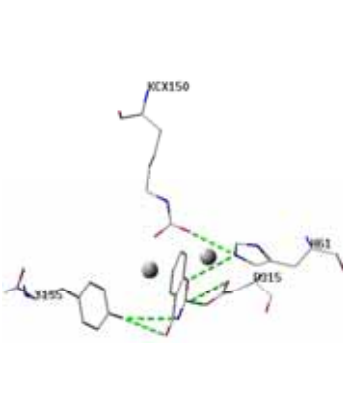
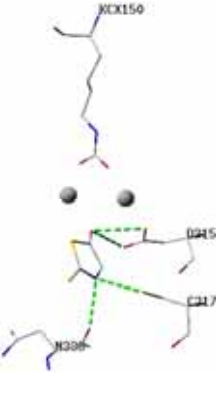
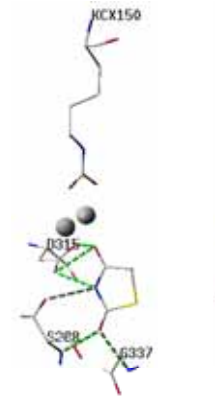
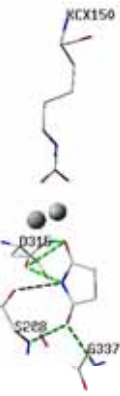
Name	5,5-dimethylhydantoin	5-ethyl-5-methylhydantoin	barbituric acid	bemegride
Structure				
Docked pose				
IC20	7.5mM	16.5mM	18.75mM	2.4mM
Name	2H-1,3-benzoxazine-2,4-dione	rhodanine	2,4-thiazoldinedione	succinimide
Structure				
Docked pose				
IC20	0.5mM	0.2mM	20mM	112.5mM

Figure 25. IC₂₀ values of inhibitors and their docked poses. The dash lines indicate the hydrogen binding formed between inhibitors and the protein. Reaction mixtures contained 100 mM of Bis-Tris propane (pH 7.0) buffer, 0.5 mM phthalimide, and docking compounds in a final volume of 1 mL.

References

- 1 W.L. Jorgensen, "The many roles of computation in drug discovery," *Science*, 2004, vol.303, pp. 1813-1818.
- 2 J.M. Yang and C.C. Chen, "GEMDOCK: a generic evolutionary method for molecular docking Development and evaluation of a generic evolutionary method for protein-ligand docking," *Proteins: Structure, Function, and Bioinformatics*, 2004, vol.55, pp. 288-304.
- 3 E.S. Lin, J.M. Yang, and Y.S. Yang, "Modeling the binding and inhibition mechanism of nucleotide and sulfotransferase using molecular docking," *Journal of the Chinese Chemical Society*, 2003, vol.50, pp. 655-663.
- 4 Y. Modis, S. Ogata, D. Clements, and S.C. Harrison, "A ligand-binding pocket in the dengue virus envelope glycoprotein," *Proceedings of the National Academy of Sciences of the United States of America*, 2003, vol.100(12), pp. 6986-6991.
- 5 Y. Modis, S. Ogata, D. Clements, and S.C. Harrison, "Structure of the dengue virus envelope protein after membrane fusion," *Nature*, 2004, vol.427, pp. 313-319.
- 6 M. Rarey, B. Kramer, T. Lengauer, and G. Klebe, "A fast flexible docking method using an incremental construction algorithm," *Journal of Molecular Biology*, 1996, vol.261, pp. 470-489.
- 7 W. Welch, J. Ruppert, and A.N. Jain, "Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites," *Chemistry & Biology*, 1996, vol.3, pp. 449-462.
- 8 T.J. Ewing, S. Makino, A.G. Skillman, and I.D. Kuntz, "DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases," *Journal of Computer-Aided Molecular Design*, 2001, vol.15, pp. 411-428.
- 9 G. Jones, P. Willett, R.C. Glen, A.R. Leach, and R. Taylor, "Development and validation of a genetic algorithm for flexible docking," *Journal of Molecular Biology*, 1997, vol.267, pp. 727-748.
- 10 G.M. Morris, D.S. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew, and A.J. Olson, "Automated docking using a lamarckian genetic algorithm and empirical binding free energy function," *Journal of Computational Chemistry*, 1998, vol.19, pp. 1639-1662.
- 11 I.D. Kuntz, J.M. Blaney, S.J. Oatley, R. Langridge, and T.E. Ferrin, "A geometric approach to macromolecule-ligand interactions," *Journal of Molecular Biology*, 1982, vol.161, pp. 269-288.
- 12 H.J. Bohm, "The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure," *Journal of Computer-Aided Molecular Design*, 1994, vol.8, pp. 243-256.
- 13 M.D. Eldridge, C.W. Murray, T.R. Auton, G.V. Paolini, and R.P. Mee, "Empirical

- scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes," *Journal of Computer-Aided Molecular Design*, 1997, vol.11, pp. 425-445.
- 14 R. Wang, L. Liu, L. Lai, and Y. Tang, "SCORE: a new empirical method for estimating the binding affinity of a protein-ligand complex," *Journal of Molecular Modeling*, 1998, vol.4, pp. 379-384.
- 15 D. Rognan, S.L. Lauemoller, A. Holm, S. Buus, and V. Tschinke, "Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins," *Journal of Medicinal Chemistry*, 1999, vol.42, pp. 4650-4658.
- 16 D.K. Gehlhaar, G.M. Verkhivker, P.A. Rejto, C.J. Sherman, D.B. Fogel, L.J. Fogel, and S.T. Freer, "Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming," *Chemistry & Biology*, 1995, vol.2, pp. 317-324.
- 17 I. Muegge, Y.C. Martin, P.J. Hajduk, and S.W. Fesik, "Evaluation of PMF scoring in docking weak ligands to the FK506 binding protein," *Journal of Medicinal Chemistry*, 1999, vol.42, pp. 2498-2503.
- 18 H. Gohlke, M. Hendlich, and G. Klebe, "Knowledge-based scoring function to predict protein-ligand interactions," *Journal of Molecular Biology*, 2000, vol.295, pp. 337-356.
- 19 C. Bissantz, G. Folkers, and D. Rognan, "Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations," *Journal of Medicinal Chemistry*, 2000, vol.43, pp. 4759-4767.
- 20 M. Stahl and M. Rarey, "Detailed analysis of scoring functions for virtual screening," *Journal of Medicinal Chemistry*, 2001, vol.44, pp. 1035-1042.
- 21 J.M. Yang, "Development and evaluation of a generic evolutionary method for protein-ligand docking," *Journal of Computational Chemistry*, 2004, vol.25, pp. 843-857.
- 22 J.M. Yang and C.C. Chen, "GEMDOCK: a generic evolutionary method for molecular docking," *Proteins: Structure, Function, and Bioinformatics*, 2004, vol.55, pp. 288-304.
- 23 J.M. Yang and C.Y. Kao, "Flexible ligand docking using a robust evolutionary algorithm," *Journal of Computational Chemistry*, 2000, vol.21, pp. 988-998.
- 24 J.N. Champness, M.S. Bennett, F. Wien, R. Visse, W.C. Summers, P. Herdewijn, E. de Clerq, T. Ostrowski, R.L. Jarvest, and M.R. Sanderson, "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*, 1998, vol.32, pp. 350-361.
- 25 A.K. Shiau, D. Barstad, P.M. Loria, L. Cheng, P.J. Kushner, D.A. Agard, and G.L.

- Greene, "The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen," *Cell*, 1998, vol.95, pp. 927-937.
- 26 V. Cody, N. Galitsky, J.R. Luft, W. Pangborn, R.L. Blakley, and A. Gangjee, "Comparison of ternary crystal complexes of F31 variants of human dihydrofolate reductase with NADPH and a classical antitumor furopyrimidine," *Anticancer Drug Design*, 1998, vol.13, pp. 307-315.
- 27 J. Abendroth, K. Niefind, and D. Schomburg, "X-ray structure of a dihydropyrimidinase from *Thermus* sp. at 1.3 Å resolution," *Journal of Molecular Biology*, 2002, vol.320, pp. 143-156.
- 28 A.N. Jain, "Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine," *Journal of Medicinal Chemistry*, 2003, vol.46, pp. 499-511.
- 29 M.M. van Lipzig, A.M. ter Laak, A. Jongejan, N.P. Vermeulen, M. Wamelink, D. Geerke, and J.H. Meerman, "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*, 2004, vol.47, pp. 1018-1030.
- 30 P.D. Griffiths, "Progress in the clinical management of herpesvirus infections," *Antiviral Chemistry & Chemotherapy*, 1995, vol.6, pp. 191-209.
- 31 G.K. Darby, "In search of the perfect antiviral," *Antiviral Chemistry & Chemotherapy*, 1995, vol.6, pp. 54-63.
- 32 M.E. Black, T.G. Newcomb, H.M. Wilson, and L.A. Loeb, "Creation of drug-specific herpes simplex virus type 1 thymidine kinase mutants for gene therapy," *Proceedings of the National Academy of Sciences of the United States of America*, 1996, vol.93, pp. 3525-3529.
- 33 E. Borrelli, R. Heyman, M. Hsi, and R.M. Evans, "Targeting of an inducible toxic phenotype in animal cells," *Proceedings of the National Academy of Sciences of the United States of America*, 1988, vol.85, pp. 7572-7576.
- 34 D. Klazmann, J. Philippon, C.A. Valery, and G. Bensimon, "Clinical protocol: Gene therapy for glioblastoma in adult patients: Safety and efficacy evaluation of an in situ injection of recombinant retroviruses producing cells carrying the thymidine kinase gene of the herpes simplex type 1 virus, to be followed with the administration of ganciclovir," *Human Gene Therapy*, 1996, vol.7, pp. 109-126.
- 35 M. Sato, T.A. Grese, J.A. Dodge, H.U. Bryant, and C.H. Turner, "Emerging therapies for the prevention or treatment of postmenopausal osteoporosis," *Journal of Medicinal Chemistry*, 1999, vol.42, pp. 1-24.
- 36 D.J. Torgerson, "HRT and its impact on the menopause, osteoporosis and breast cancer," *Expert Opinion on Pharmacotherapy*, 2000, vol.1, pp. 1163-1169.
- 37 C.P. Miller, "SERMs: evolutionary chemistry, revolutionary biology," *Current*

- Pharmaceutical Design*, 2002, vol.8, pp. 2089-2111.
- 38 M. Dutertre and C.L. Smith, "Molecular mechanisms of selective estrogen receptor modulator (SERM) action," *The Journal of Pharmacology and Experimental Therapeutics*, 2000, vol.295, pp. 431-437.
- 39 M. Maricic and O. Gluck, "Review of raloxifene and its clinical applications in osteoporosis," *Expert Opinion on Pharmacotherapy*, 2002, vol.3, pp. 767-775.
- 40 J.I. MacGregor and V.C. Jordan, "Basic guide to the mechanisms of antiestrogen action," *Pharmacological Reviews*, 1998, vol.50, pp. 151-196.
- 41 R. Gust, R. Keilitz, and K. Schmidt, "Synthesis, structural evaluation, and estrogen receptor interaction of 2,3-diarylpiperazines," *Journal of Medicinal Chemistry*, 2002, vol.45, pp. 2325-2337.
- 42 J. Renaud, S.F. Bischoff, T. Buhl, P. Floersheim, B. Fournier, C. Halleux, J. Kallen, H. Keller, J.M. Schlaeppli, and W. Stark, "Estrogen receptor modulators: identification and structure-activity relationships of potent ERalpha-selective tetrahydroisoquinoline ligands," *Journal of Medicinal Chemistry*, 2003, vol.46, pp. 2945-2957.
- 43 G.B. Elion, "Nobel lecture in physiology or medicine--1988. The purine path to chemotherapy," *In Vitro Cellular & Developmental Biology*, 1989, vol.25, pp. 321-330.
- 44 G.H. Hitchings, Jr., "Nobel lecture in physiology or medicine--1988. Selective inhibitors of dihydrofolate reductase," *In Vitro Cellular & Developmental Biology*, 1989, vol.25, pp. 303-310.
- 45 P.C. Wyss, P. Gerber, P.G. Hartman, C. Hubschwerlen, H. Locher, H.P. Marty, and M. Stahl, "Novel dihydrofolate reductase inhibitors. Structure-based versus diversity-based library design and high-throughput synthesis and screening," *Journal of Medicinal Chemistry*, 2003, vol.46, pp. 2304-2312.
- 46 G. Rastelli, S. Pacchioni, W. Sirawaraporn, R. Sirawaraporn, M.D. Parenti, and A.M. Ferrari, "Docking and database screening reveal new classes of Plasmodium falciparum dihydrofolate reductase inhibitors," *Journal of Computational Chemistry*, 2003, vol.46, pp. 2834-2845.
- 47 I. Kautner, M.J. Robinson, and U. Kuhnle, "Dengue virus infection: epidemiology, pathogenesis, clinical presentation, diagnosis, and prevention," *The Journal of Pediatrics*, 1997, vol.131, pp. 516-524.
- 48 F.A. Rey, F.X. Heinz, C. Mandl, C. Kunz, and S.C. Harrison, "The envelope glycoprotein from tick-borne encephalitis virus at 2 A resolution," *Nature*, 1995, vol.375, pp. 291-298.
- 49 E. Lee, R.C. Weir, and L. Dalgarno, "Changes in the dengue virus major envelope protein on passaging and their localization on the three-dimensional structure of the protein," *Virology*, 1997, vol.232, pp. 281-290.