

# Chapter 1

## Introduction

### 1.1 Motivations and Purposes

With the recent development of high-throughput X-ray crystallography, the total number of structures will grow at an even greater speed[1]. And the enormous advances in genomics have resulted in a large increase in the number of potential therapeutic targets that are available for investigation. This growth in potential targets has increased the demand for reliable target validation, as well as technologies that can identify rapidly several quality lead candidates. Virtual screening methods are a primary source for the discovery of lead molecules for drug development, with high-throughput docking algorithms being among the most extensively used of these methods. The application of virtual high-throughput screening[2, 3], to the drug discovery process invariably produces a large number of potential lead candidates. And it is well known that current scoring functions used in virtual screening campaigns are often inadequate at predicting the true binding affinity of a ligand for a receptor[4]. These prospective ligands need to be filtered in order to reduce their number for more precise and labor-intensive studies.

The purpose for utilizing post-analysis is to minimize the number of false positives in the selection list and to propagate the true hits to the top of the list. One of the post-analysis methods such as clustering based upon structural similarity can nonetheless generally improve the performance of the scoring function[5, 6]. Clustering molecules based upon similarity requires some quantitative measure (descriptor) of the similarity between two molecules. There are many different approaches to generate descriptor, include 2D and 3D methods.

Most of the 2D methods have focused on representing a molecule based upon its own structural and chemical composition, like Atom-Pair. But it regardless the information from protein that is important in the field of structure-based drug designs.

Traditionally, similarity methods have focused on representing a molecule based upon its own structural and chemical composition. With the objective of developing a method for post-analysis, a detailed understanding of intermolecular interactions between proteins and their ligands is of critical importance to structure-based drug design.

## 1.2 Relate Work

Deng and co-workers[5] described a novel approach to representing the properties of a ligand. As opposed to calculating the properties of a ligand from the perspective of its own structural and chemical components, the Structural Interaction Fingerprint (SIFt) method represents a ligand by the interactions it forms in the binding site of a protein. Using seven bits per binding-site residue to represent seven different types of interaction, the SIFt method encoded a ligand-protein interaction into a binary string. The types of interaction that considered are hydrogen bond and physical contact. Recently another approach proposed by Amari et al,[7]have developed a clustering tool for visualized cluster analysis of protein-ligand interaction (VISCANA) that analyzes the pattern of the interaction of the receptor and ligand on the basis of quantum theory. They applied the ab initio fragment molecular orbital (FMO)[8] method for represent the interaction between protein and ligand, which used the ab initio electronic structure calculation of proteins and encoding each docked pose into real number string. But the FMO method needed to obtain more reliable descriptions of van der Waals interactions and hydrogen bonds that are important for receptor-ligand binding.

In order to handle the large amounts of result from virtual screening and consider more specific information for protein-ligand binding, we used the empirical energy function from

GEMDOCK[9] which is specifically optimized for virtual screening, and utilized hierarchical clustering to help us better analyze the binding interactions between proteins and ligands. We have developed a method that using empirical energy function while sharing the basic premise of the above-mentioned SIFt method, extends it to encode more interaction-specific information into the real number string, hydrogen bond, Van der Waal force, and electrostatic. By representing the interactions at the atomic-level as opposed to the residue level and including measures of the strength of the interactions, we were able to better describe a ligand-protein interaction and produced a more informative analysis of virtual screening. In order to select the representative compounds with diverse structure but similar binding interaction, we applied the 2D topology descriptor, atom-pair for grouping compounds with similar structure.

GEMDOCK is a docking program with a good performance on the prediction of the target-bound conformation and orientation of docked ligand. It can predict known protein-bound ligand poses with averaged RMSD below 2.0 Å[9] and model the best possible poses of ligands in the target with no crystal protein complex[10]. GEMDOCK was improved and modified for virtual screening on large database with a suitable scoring function to reduce the number of false positive[11].

### 1.3 Application

The presence of *Helicobacter pylori* on the human gastric mucosa induces gastritis and may further develop into peptic ulcer and gastric cancer. Several factors including differential bacterial virulence, host immunity, and environmental factors are considered to influence the development into various clinical sequelae[50]. *H. pylori* resistance to antibiotics is the main factor for therapy failure, while other features remain largely unknown[49].

The shikimate pathway is a seven-step biosynthetic route that involves aromatic

amino acid biosynthesis and many aromatic secondary metabolites, including tetrahydrofolate and ubiquinone. The shikimate pathway is an attractive target for the discovery of new antimicrobial agents, herbicides and antiparasitic agents because it is essential in bacteria, fungi, and plants, but not mammals[46, 47, 51, 52].

Our research was cooperated with Dr. W. C. Wang. We have applied our two stages cluster method on the result of virtual screening of shikimate kinase from *Helicobacter pylori* By GEMDOCK. And we identified a potential inhibitor, furosemide (MCMC00000106).



## Chapter 2

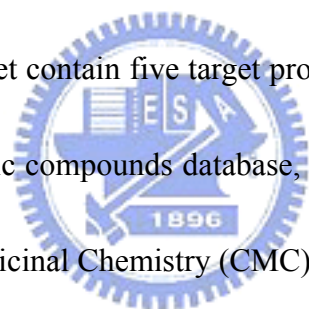
### Materials and Methods

The Main Step of our Cluster analysis methodology is show in Figure 1, the overall process was shown in Figure 2. Our clustering analysis was a two stages clustering, first stage was protein-ligand interaction based clustering and the next stage was Compound structure based clustering. In the section of protein-ligand interaction based cluster analysis, in order to visualize and analyze the large data from virtual screening, we converted every docked pose into one dimensional real number string by calculating atom-based protein-ligand interaction, implemented from our original technique by Yang et al[9, 12]. Because of the representation of the protein-ligand interaction, we were able to evaluate the distance of binding mode between two docked poses. Hierarchical clustering analysis[13] were carried out with MATLAB[14]. Compounds with similar binding mode were visualized and grouped. In the section of Compound structure based clustering, each compound structure was represented by an one dimension atom-pair descriptor, implemented from an approach proposed by Carhart et al[15, 16]. By evaluating the distance of structure between two compounds, similar structure compounds within a cluster by first stage clustering could be grouped together for selecting compounds with covering most chemotype in all clusters. Finally, our cluster analysis could select diversely compounds by protein-ligand interaction and compound

structure for use in bioassay. After analysis the distance between active and non-active compounds, a reference threshold was decided for demarcating similar compounds.

We generated two sets of structure-based virtual screening result and used them in our studies.

First set was designed to verify the protein-ligand interaction descriptor is suitable for identifying compounds with similar binding mode. The dataset consist of five classes known inhibitors and each inhibitor in a class has the same target protein. The second set was designed to evaluate the database enrichment potential and the property of compounds in the same cluster by docking a diverse set of compounds spiked with known inhibitors into the same target protein. This dataset contain five target proteins. Our testing compounds database was constructed from the public compounds database, e.g., the Available Chemical Directory (ACD) or Comprehensive Medicinal Chemistry (CMC).



## **2.1 Preparation of Target Protein and Compound Databases**

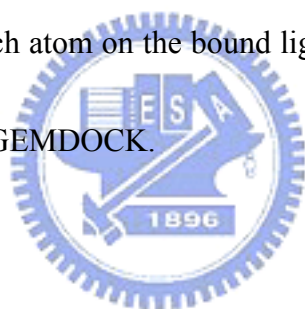
### **A. Preparing for target proteins**

We prepared virtual screening result against five target proteins

1. TK (herpes simplex virus type 1 thymidine kinase) PDB id : 1kim[17],
2. ER  $\alpha$  (human estrogen receptor alpha) PDB id : 3ert[18],

3. ER  $\alpha$  (human estrogen receptor alpha) PDB id : 1gwr[18],
4. hDHFR (human dihydrofolate reductase) PDB id : 1hfr[19],
5. NA (tern n9 influenza virus neuraminidase) PDB id : 1mwe[20],
6. SK (helicobacter pylori shikimate kinase) PDB id : 1zui[21]

Fifth and Sixth target proteins were also used on application study. Each structure file was first derived from the Protein Data Bank (PDB), then remove solvent and hydrogen by swiss pdb viewer. The Ligand binding site was defined as the collection of amine acids using a cutoff radius of 10Å from each atom on the bound ligand. The structure files were stored as PDB format for inputting into GEMDOCK.



## **B. Preparing for compound databases**

We constructed two compound sets for screening against each target protein.

The verifying dataset contained five class compounds and all compounds within a class were active compounds for a specific target protein. All structures excluded NA active compounds were derived from previous work[22]. The TK active compounds and ER  $\alpha$  antagonists were proposed by Bissantz et al,[23], contain ten TK active compounds and eleven ER  $\alpha$  antagonists. The ER  $\alpha$  agonist were proposed by Lipzig et al,[24], also contain

ten ER  $\alpha$  agonist. The NA active compounds was proposed by Birch et al,[25], contain twenty NA active compounds. The hDHFR active compounds were derived from Protein Data Bank (PDB), contain ten active compounds. All compounds were list as follow :

1. TK active compounds : 10[23],

2. ER  $\alpha$  antagonists : 11[23],

3. ER  $\alpha$  agonists : 10[24],

4. hDHFR active compounds : 10 ,

5. NA active compounds : 20[25],

Totally the verifying dataset contained 61 compounds.



Testing dataset contained 990 randomly selected compounds combined with known active compounds for each target protein. The set of 990 randomly selected compounds were derived from Bissantz et al,[23] . All compound structures were converted to mol format and removed hydrogen by CORINA3.0 for inputting into GEMDOCK.

The compound set for the helicobacter pylori shikimate kinase (SK) and neuraminidase of influenza virus (NA) was derived from the Comprehensive Medicinal Chemistry (CMC). All compounds in CMC were first filtered with molecular weight between 200 and 800, and then removed small fragments and hydrogen by CORINA3.0. The structure files were stored as



mol format for inputting into GEMDOCK.

## 2.2 Preparation of Virtual Screening Result for Cluster analysis

### A. Molecular docking

The GEMDOCK program was used to prepare the docked pose and predict binding affinity for each compound in dataset, which was enhanced and modified from our original technique[12] Two key component of the GEMDOCK is the searching algorithm and the scoring function. The searching algorithm of GEMDOCK is a generic evolutionary method, the detail of the algorithm was on our previous research by Yang et al.,[9]. Scoring function was based on an empirical energy function. 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  $E_{ligpre}$  (see Mining pharmacological consensus subsection) 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 energies, respectively, and  $E_{penal}$  is a large penalty value if the ligand is out of the range of the search box. For our present work,  $E_{penal}$  was set to 10,000. The intermolecular energy is defined as



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

Where  $r_{ij}$  is the distance between the atoms  $i$  and  $j$ ;  $q_i$  and  $q_j$  are the formal charges and 332.0 is a factor that converts the electrostatic energy into kilocalories per mole. The *lig* and *pro* denote the numbers of the heavy atoms in the ligand and receptor, respectively.  $F(r_{ij}^{B_{ij}})$  is a simple atomic pair-wise potential function (Figure 3), as defined in our previous study[9] 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. The energy value of a hydrogen bonding should be larger than that for steric potential. In this model, atoms are divided into four different atom types[9]: 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[9], 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

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

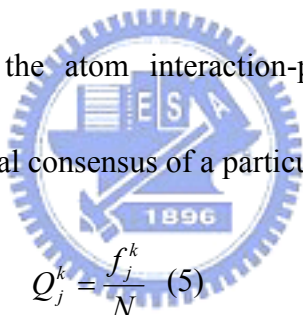
where  $F(r_{ij}^{B_{ij}})$  is defined as for Equation 3 except the value is set to 1000 when  $r_{ij}^{B_{ij}} < 2.0 \text{ \AA}$ , and *dihed* is the number of rotatable bonds in a ligand. We followed the work of Gehlhaar et al. [26] to set the values of *A*, *m*, and  $\theta_0$ . For the  $sp^3$ - $sp^3$  bond,  $A = 3.0$ ,  $m = 3$ , and  $\theta_0 = \pi$ ; for the  $sp^3$ - $sp^2$  bond,  $A = 1.5$ ,  $m = 6$ , and  $\theta_0 = 0$ .

## B. Mining pharmacophore consensus

From our previous research[11], GEMDOCK evolves the binding-site pharmacological consensus and ligand preferences from both known active ligands and the target protein to improve screening accuracy. We used the premise that previously acquired interactions (hot spots) between ligands and the target protein can be used to guide the selection of lead compounds for subsequent investigation and refinement. When known active ligands were available, GEMDOCK used a pharmacophore-based scoring function (Equation 1). On the

other hand,  $LP_{elec}$  and  $LP_{hb}$  were set to zero and GEMDOCK used a purely empirical-based scoring function (Equation 2) if known active compounds were not available.

For each known active compound, GEMDOCK first yielded 10 docked ligand conformations by docking the compound into the target protein, and only the docked ligand conformation with the lowest energy was retained for pharmacological consensus analysis. The protein-ligand interactions were extracted by overlapping these lowest-energy docked conformations, and the interactions were classified into two different types, including hydrogen bonding and hydrogen-charged interactions. After all of the protein-ligand interactions were calculated, the atom interaction-profile weight of the target protein representing the pharmacological consensus of a particular interaction was given as



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

Where  $N$  is the number of known active compounds and  $f_j^k$  is the total interaction number of an atom  $j$  (in a protein) interacting with an atom of known active ligands with the interaction type  $k$  (e.g., hydrogen bonding or hydrogen-charged interactions). An atom  $j$  in the reference protein was considered a hot-spot atom when  $Q_j^k$  was more than 0.5.

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

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

where  $CW(B_{ij})$  is a pharmacological-weight function of a hot-spot atom  $j$  with interaction type  $B_{ij}$ ,  $F(r_{ij}^{B_{ij}})$  is defined as in Equation 3,  $lig$  is the number of heavy atoms in a screened ligand, and  $hs$  is the number of hot-spot atoms in the protein. The  $CW(B_{ij})$  is given as

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

$Q_j^k$  is the atomic pharmacological-profile weight (Equation 5) and  $k$  is the interaction type of the hot-spot atom  $j$ .

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. The ligand preference ( $E_{ligpre}$ ) is a penalty value for those screened ligands that violate the electrostatic or hydrophilic constraints. The  $E_{ligpre}$  is given as

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

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.  $\theta_{elec}$  is the maximum number of charged atoms among known active compounds, and  $\sigma_{elec}$  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 compound and  $Ur_{hb}$  is the upper bound of the fraction of polar atoms calculated from known active compounds.  $NA_{hb}$  and  $NA_t$  are the number of polar atoms and the total number of the heavy atoms of a screened ligand, respectively.  $\theta_{hb}$  and  $\sigma_{hb}$  are the maximum ratio and the standard derivation of the ratios of polar atoms evolved from known compounds, respectively.

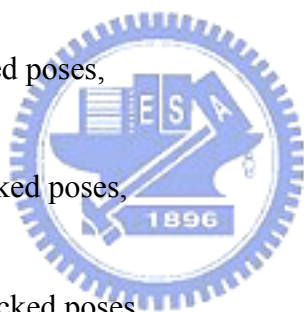
When the normalization strategy is applied, the energy function (Equation 1) is given as

$$E_{tot} = E_{bind}^{MW} + E_{pharma} + E_{ligpre} \quad (12)$$

### C. Verifying dataset

This dataset was designed to verify the protein-ligand interaction and atom-pair descriptor is suitable for identifying compounds with similar binding mode and similar structure. We constructed the verifying dataset by docking all 61 active compounds into each target protein. GEMDOCK yielded 10 docked conformations for each compound and filter out the conformation if the conformation didn't docked into the binding site. After filtering out the wrong conformation, only the conformation with lowest energy retained for generating the representative docked pose of each compound. Number of docked pose of each dataset was as follow.

1. TK (1kim) : 53 docked poses,
2. ER  $\alpha$  (3ert) : 61 docked poses,
3. ER  $\alpha$  (1gwr) : 52 docked poses,
4. hDHFR (1hfr) : 61 docked poses,
5. NA (1mwe) : 61 docked poses.



The parameter of GEMDOCK was listed on Table 4.

#### **D. Testing dataset**

The compound set for each target protein combined 990 randomly selected compounds and known active compounds of each target protein. GEMDOCK yielded 3 docked conformations

for each compound and only the conformation with lowest energy retained for generating the representative docked pose of each compound. Number of docked pose of each dataset was as follow.

1. TK (1kim) : 10 + 990 docked poses,
2. ER  $\alpha$  (3ert) : 11 + 990 docked poses,
3. ER  $\alpha$  (1gwr) : 10 + 990 docked poses,
4. hDHFR (1hfr) : 10 + 990 docked poses,
5. NA (1mwe) : 20 + 990 docked poses.

The parameter of GEMDOCK was listed on Table 5.



## 2.3 Generation of Descriptors

### A. protein-ligand interaction descriptors

We convert 3D docked pose into one dimension real number string by calculating the energy between each atom on protein and ligand. The protein-ligand interaction definition is show in Figure 4. We used the energy function of GEMDOCK for calculating the protein-ligand interaction. The protein-ligand interaction is divided into three types of interaction, steric, hydrogen bond, and electrostatic. The interaction energy of each atom  $j$  on



protein is defined as

$$E_j = \sum_{i=1}^{lig} \left[ F(r_{ij}^{B_{ij}}) + 332.0 \frac{q_i q_j}{4r_{ij}^2} \right] \quad (13)$$

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.

These two potentials are calculated by the same function form but different parameters,

$V_1, \dots, V_6$  given in (Figure 3).  $q_i$  and  $q_j$  are the formal charges and 332.0 is a factor that

converts the electrostatic energy into kilocalories per mole. The  $lig$  and denote the numbers of

the heavy atoms on the ligand.  $F(r_{ij}^{B_{ij}})$  is a simple atomic pair-wise potential function (Figure

3). In this atomic pair-wise model, atoms are divided into four different atom types[9]: 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[9], which is set to 5 or

-5, respectively, if the electrostatic energy is more than 5 or less than -5. By doing so, each

atom of protein was represented by a three type of interaction real number string.

In this study, we have test varies type of coding of protein-ligand interaction on our verifying

dataset. We used the atom-based, real number coding. Because the setting has the best

discrimination between active and non-active compounds. After the atom-pair descriptor was

generated, we removed the columns with values were all zero.

## B. Atom pair descriptors

Atom-pair descriptors are 2D topological descriptors, which count the distance between two atoms as the shortest path of bonds. The atom-pair definition is show in Figure 4. Atom pair descriptors was generated from the atom-pair generator program developed by our laboratory, the methodology was first proposed by Carhart et al.[15, 16]. Two major components for constructing a set of atom-pair descriptors include the definition of atom type and the number of distance bins between two atom types. An atom-pair is a simple type of substructure defined in term of the atom types and the shortest path graph distance between two atoms. The graph distance is defined as the smallest number of atoms along the path connecting two atoms in a molecular structure. The formula of an atom-pair is as follows :

$$\text{atom type } i \text{---(distance)---atom type } j$$

Where the (distance) is the graph distance between atom  $i$  and  $j$  in the case of a 2D atom-pair description. (The distance can also be defined as the physical distance between atom  $i$  and  $j$  in the case of a 3D atom-pair description.)

From previous study in our laboratory, SYBYL 23 atom types were clustered into 10 atom types (Table 1) in order to reduce the number of atom-pair descriptors and improve the accuracy.

In this study, we have test varies type of coding of atom-pair on our verifying dataset. We found that the setting of distance bins, coding, and similarity measurement which has the most discrimination in our dataset is the similar to the research by Hans Matter[27]

Procedure for preparing atom-pair descriptors :

1. Structure file in mol format
2. Remove hydrogen
3. Transform to mol2 format by utilizing CORINA3.0
4. Calculating atom-pair descriptor by atom-pair generator

Distance bins : 15

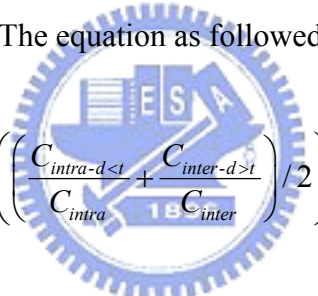


5. Store in binary coding form

The total number of pairwise combinations of all 10 atom types is 55. Furthermore, 15 distance bins were defined in the interval between graph distance zero (i.e., zero atoms separating an atom-pair) to 14. Thus, a total of 825 (55 x 15) atom-pair descriptors were generated for each molecular structure[28]. After the atom-pair descriptor was generated, we removed the columns with the values were all zero.

## **2.4 Reference Threshold for Protein-ligand Interaction and Atom-pair Descriptor**

A reference threshold of protein-ligand interaction should distinguish between similar binding mode and non-similar binding mode. The similar binding mode was defined as the binding modes of active compounds docked into its target protein, and the non-similar binding mode was defined as the binding mode of non-active compounds docked into the same target protein. The distance between similar binding modes it self should smaller then the distance between similar binding modes and non-similar binding modes. For designing a reference threshold of protein-ligand interaction, verifying dataset was used to deciding a reference threshold of distance by determining a maximum discrimination between similar and non-similar binding mode. The equation as followed



$$\max \left( \left( \frac{C_{intra-d < t} + C_{inter-d > t}}{C_{intra} + C_{inter}} \right) / 2 \right) \quad (14)$$

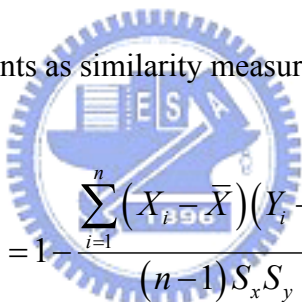
Where  $t$  is the reference threshold.  $C_{intra-d < t}$  is the number of intra active compound pairs with the distance  $<$  threshold.  $C_{inter}$  is the number of compound pairs between active and non-active compounds.

Designing a reference threshold of Atom-pair descriptors was similar, but the hDHFR and NA active compounds were divided into two classes because of the diverse compound structure (Figure 9) (Figure 10).

(Figure 17) shows the result of designing a reference threshold.

## 2.5 Method of Cluster Analysis

Hierarchical methods have the advantage of building up an interpretable relationship between the clusters. Hierarchical clustering analyses were carried out with MATLAB[14]. After translating the docked poses by protein-ligand interaction into real number string and compound structure into bit strings, we applied a simply two stages hierarchical clustering to visualize and analyze the protein-ligand interaction and atom-pair descriptor for each dataset by using an agglomerative hierarchical clustering approach[13]. At the first stage, we used protein-ligand interaction descriptor for clustering compounds with similar binding mode and applied the correlation coefficients as similarity measurements. Formula was as followed.


$$D_{xy}^{corr} = 1 - \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)S_x S_y} \quad (15)$$

Where  $D_{xy}^{corr}$  is the correlation distance between docked pose  $X$  and  $Y$ . The  $S_x$  is the standard deviation of  $X$ .  $X_i$  is the  $i$ th value of  $X$ .  $n$  is the number of descriptors. We applied the standard UPGMA clustering method for calculating the distance between two clusters while constructing the dendrogram. Formula was as followed.

$$D_{rs}^{clu} = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D_{risj}^{corr} \quad (16)$$

The reference threshold was calculated from verifying dataset by equation (14) for determining the number of clusters.

At the second stage, we applied atom-pair descriptor for clustering compounds within each cluster from first stage clustering and applied the tanimoto coefficients as similarity measurements. Formula was as followed.

$$D_{xy}^{\text{tani}} = \frac{|X \cap Y|}{|X \cup Y|} \quad (17)$$

Where  $D_{xy}^{\text{tani}}$  is the tanimoto distance between  $X$  and  $Y$ .  $|X \cap Y|$  is the number of ON bits common in both  $X$  and  $Y$ , and the  $|X \cup Y|$  is the number of ON bits present in either  $X$  or  $Y$ .

We also used the standard UPGMA clustering method for calculating the distance between two clusters while constructing the dendrogram. The equation is similar with equation (16), instead  $D_{xy}^{\text{corr}}$  by  $D_{xy}^{\text{tani}}$ . The reference threshold was calculated from verifying dataset by equation (14) for determining the number of clusters. Compounds with similar structure were grouped together within each cluster from first stage clustering, and the representative compounds were selected by the lowest energy compounds within each cluster. The dendrogram graph was plotted for visualizing the binding mode of multi docked poses by protein-ligand interaction.

## Chapter 3

### Results

We applied our cluster method to analyze two datasets of docking result. The target proteins were hDHFR (human dihydrofolate reductase) PDB id : 1hfr, ER $\alpha$  (human estrogen receptor alpha) PDB id : 3ert, ER $\alpha$  (human estrogen receptor alpha) PDB id : 1gwr, NA (tern n9 influenza virus neuraminidase) PDB id : 1mwe, and TK(herpes simplex virus type 1 thymidine kinase) PDB id : 1kim. The verifying dataset was constructed by cross-docking of all 61 known active compounds against 5 target proteins. And the testing dataset was constructed by docking known active compounds spiked into 990 randomly selected compounds against 5 target proteins. The docked pose and the Fitness (predicted binding affinity) were prepared by GEMDOCK for each compound.

First, we evaluated the accuracy of GEMDOCK for molecular docking. Second, we tested the significance of the distance generated by protein-ligand interaction and atom-pair using T-test on verifying dataset. Third, we decided a reference threshold of distance on protein-ligand interaction and atom-pair by determining a maximum discrimination between similar and non-similar binding mode and structure. Forth, we applied the cluster method to analyze the result of molecular docking on verifying and testing dataset. Fifth, we applied our

two-stage cluster method on virtual screening of NA and SK, and selected several potential candidates for use in bioassay.

### **3.1 Molecular Recognition and Setting of Pharmacophore Consensus**

From previous study[9, 11, 12], GEMDOCK evolved a pharmacophore consensus and a ligand preference from the target protein and known active compounds by overlapping the docked poses and quantified the consensus of protein-ligand interaction and the property of those known active compounds. After incorporating into scoring function while flexible docking, it was proved to reducing the number of false positives on screening large database.

In this work, we applied the pharmacophore consensus and the ligand preference for each set of target protein by superimposing known active compounds, respectively. The pharmacophore consensus was used on each molecular docking, while screening, we applied both the ligand preference and the pharmacophore consensus. The ligand preference and the pharmacophore consensus for each target protein were listed on Table 6 and Table 7.

The setting of GEMDOCK parameters applied on verifying dataset and testing dataset were listed on Table 4 and Table 5.

We evaluated the accuracy of GEMDOCK by performing molecular recognition on each pair of target protein and its known active compounds. After docking all known active compounds into the target protein, we based the result on root mean square deviation (RMSD) error



between docked pose and the original crystal ligand for verifying GEMDOCK. 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. The overall performance for each target protein was shown on Table 8.

#### **A. TK (Thymidine kinase)**

Virtual screening for exploiting diverse lead compounds of TK would be of considerable value in many fields. 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[29, 30]. 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.

The crystal coordinates of the ligand and protein atoms were taken from PDB, and were separated into different files. All active compound structures were shown on Figure 6. 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 opened enough to accommodate a broad variety of ligands. The average RMSD of all ten docked poses was 1.44 Å. The result was shown on Table 8.

Figure 11 shows the alignment of ligands from crystal structure and the pharmacological consensus of the binding site. The ligand preferences those were identified by superimposing 10 crystal structures of TK were listed on Table 6.

### **B. ER $\alpha$ (Estrogen receptor)**

The search for proper SERMs among both existing and new drugs has been a challenging task in recent years [31, 32]. Because 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. Estrogens such as 17 $\beta$ -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 [33]. Hormone replacement therapies have been used for the treatment of vasomotor symptoms related to the menopause and for prevention of osteoporosis [34, 35]. Compounds mimic estrogen in some tissues while antagonizing its action in others are named selective estrogen receptor modulators (SERMs) [36]. Many SERMs such as tamoxifen and raloxifene, are currently on the market for the treatment of hormone-dependent breast cancer [37] and prevention and treatment of osteoporosis [38], respectively.

The target protein structures of ER $\alpha$  (3ert) and ER $\alpha$  (1gwr) were derived from PDB, and the antagonists and agonists were derived from previous work[22]. We docked four antagonists into the target protein (3ert) and four agonists into another one (1gwr), and based

the results on root mean square deviation (RMSD) error in ligand heavy atoms between the docked pose and the crystal structure. The average RMSD of docked antagonists and agonists was 1.42 Å. In 1hj1.AOE and 1qkm.GEN, the values of RMSD were larger than 2.0 Å because the native proteins were crystal structures of Erβ-ligand complexes.

The pharmacophore consensus of ERα antagonist (3ert) and ERα agonist (1gwr) were obtained by docking antagonists and agonists into their cavity, respectively. We have docked 10 times for each antagonists and agonists, and retained the lowest energy conformation. The pharmacophore consensus was calculated by equation 5 and 7. Figure 12 shows the docked poses of 11 antagonists for calculating the pharmacophore consensus of the target protein (3ert). And Figure 13 shows the docked poses of 10 agonists for calculating the pharmacophore consensus of the target protein (1gwr). The ligand preferences those were obtained by superimposing known active compounds of ER (3ert) and ER (1gwr) were listed on Table 6.

### **C. hDHFR (human dihydrofolate reductase)**

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. 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 [39, 40]. 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 [41, 42].

The structures of the ligand and target protein were derived from the protein data bank (PDB). All structures of active compound were shown on Figure 9. For evaluating the accuracy of GEMDOCK for molecular docking, we docked 10 known active compounds into the target protein, and compare the RMSD between the docked pose and the bound ligand in crystal structure. The average RMSD of all ten docked poses was 1.03 Å. The RMSD of All active compounds were lower than 2 Å. It means that GEMDOCK was suitable for molecular docking on this target. The result was shown on Table 8.

Figure 14 shows the alignment of ligands from crystal structure and the pharmacological consensus of the binding site. The ligand preferences those were identified by superimposing 10 crystal structures of TK were listed on Table 6.

#### **D. NA (neuraminidase)**

Chemicals that inhibit neuraminidase (NA) can protect the host from viral infection[43].

Influenza is a major respiratory infection associated with significant morbidity in the general population and mortality in elderly and high-risk patients. It is an RNA virus that contains two major surface glycoproteins, neuraminidase (NA) and hemagglutinin (HA). These proteins are essential for infection. Neuraminidase (NA) has been found to be a potential target to control influenza virus[44]. Neuraminidase (NA) can cleave the a-ketosidic connections of sialic acid and nearby sugar residues. The removal of sialic acid lowers the viscosity of the virus, thus permitting the entry of the virus into epithelial cells. Neuraminidase (NA) also destroys hemagglutinin (HA) on the virus surface allowing the emergence of progeny virus units from infected cells.



The crystal coordinates of the known active ligand and target protein atoms were derived from PDB. All structures of active compounds were shown on Figure 10. For evaluating the accuracy of GEMDOCK for molecular docking, we docked 20 known active compounds into the target protein, and based the results on root mean square deviation (RMSD) error in ligand heavy atoms between the docked pose and the crystal structure. The average RMSD of all 20 docked poses was 0.95 Å. The result was shown on Table 8.

Figure 15 shows the alignment of ligands from crystal structure and the pharmacological consensus of the binding site. The ligand preferences those were identified by overlapping the 20 crystal structures of NA were listed on Table 6.

## 3.2 Significance of Descriptor on Verifying Dataset

### 3.2.1 Significance of Protein-ligand Interaction Descriptor

**Significance of known compounds in five classes:** First, we verified the protein-ligand interaction descriptor was suitable for identifying compounds with similar binding mode. The similar binding mode was defined as the binding modes of active compounds docked into its target protein, and the non-similar binding mode was defined as the binding mode of non-active compounds docked into the same target protein. To test whether the distance between similar binding mode which was represented by protein-ligand interaction are significantly lower than that for distance between non-similar binding mode, we calculated the P-value for the null hypothesis of no difference between the mean of distance between active compounds and the mean of distance between active and non-active compounds. The results are listed in Table 9. The T-scores are calculated as the standard two sample T-test statistics :

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

Where  $\mu$  is the mean of samples, and

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

The result of T-test shown that the translation of docked pose into protein-ligand interaction descriptor could significantly distinguish the distance between similar binding modes from distance between non-similar binding modes. We quantified the discrimination of protein-ligand interaction descriptor between similar and non-similar binding mode by equation 14, the maximum discrimination was determined, and the result was shown on Figure 17. The result shown that utilizing this method could correctly distinguish similar and non-similar binding mode by 88.89%.

**Significance of similar compounds:** Next, for the purpose of post-analysis, we test that the similar compounds have similar docked behavior (pose, interaction) on a protein receptor. In this test, each class of active compounds was defined as similar compounds. The docked behavior was defined as the protein-ligand interaction generated from each docked pose. There are five classes of similar compounds on each target protein. We tested whether the mean of distance between similar compounds which were represented by protein-ligand interaction are significantly lower than that for distance between non-similar compounds. The result of T-test was listed in Table 11. Almost all null hypotheses were denied, but there were four cases that passed. Two were hDHFR active compounds docked into ER (3ert) and ER (1gwr); others were EST antagonist docked into ER (1gwr) and TK (1kim). Because of the diverse structures of hDHFR active compounds and ER antagonists (Figure 9, 7).

### 3.2.2 Significance of Atom-pair Descriptor

We applied similar method to verify the atom-pair descriptor is suitable for identifying compounds with similar structure. The similar structure was defined as active compounds and the non-similar structure was defined as non-active compounds. The result of T-test shown on Table 10, and the translation of compound structure into atom-pair descriptor could significantly distinguish the distance between similar structures from distance between non-similar structures.

We also quantified the discrimination of the translation of structure into atom-pair descriptor by equation 14, but the active compounds of hDHFR and NA were divided into two classes because of the diverse compound structure (Figure 9) (Figure 10). Because the definition of atom-pair Table 1. There were two type of carbon on the core ring of the 20 known active compounds of NA. The type of carbons on the core ring of lig1ina\_ST6, lig1ing\_ST5, lig1ivc\_ST2, lig1ivd\_ST1, lig1ive\_ST3 was aromatic carbon, but the type of others was normally Non-aromatic. In the case of hDHFR known active compounds, because the old drugs contained two more carboxylic acid groups, it is different from the new drugs. The maximum discrimination between similar and non-similar structure was determined, and the result shown that utilizing this method could correctly distinguish between similar and non-similar structure by 91.45%.

### 3.3 Calculating Reference Threshold by Verifying Dataset

A reference threshold of protein-ligand interaction should distinguish between similar



binding mode and non-similar binding mode. For designing a reference threshold of protein-ligand interaction, verifying dataset was used to deciding a reference threshold of distance by determining a maximum discrimination between similar and non-similar binding mode. By applying the above-mentioned method, we calculated the distance threshold (Correlation coefficient: 0.3894) that had the maximum discrimination.

The reference threshold of atom-pair (Tanimoto coefficient: 0.5548) was calculated by the 7 classes of structures which were mentioned above. The reference threshold was shown on Figure 17. The property of complement between atom-pair descriptor and protein-ligand interaction descriptor was also show on Figure 17.

### **3.4 Cluster analysis of Molecular Docking Result on Verifying Dataset**

We applied our protein-ligand interaction and atom-pair descriptor cluster method individually on hDHFR and TK sets for analyzing the property of those two descriptors.

#### **A. Clustering by protein-ligand interaction:**

##### **3.4.1 Cluster analysis of Molecular Docking on hDHFR (human dihydrofolate reductase)**

In the case of hDHFR, there were no docked pose which was out of binding site, after removing the columns with value of all zero, a total number of 316 atoms involved were identified, contained 61 docked poses include 10 of known active compounds. Protein-ligand

interactions of all complexes were generated, each of which was composed of 316 real numbers. The result of hierarchical clustering was shown on Figure 18 represented as a heat map. By the reference threshold of protein-ligand interaction (Correlation coefficient: 0.3894), three major clusters can be identified in Figure 18, cluster c, d and e. First, all active compounds were group together into the cluster c, Figure 18(c), All hDHFR ligands in cluster c had hydrogen-bond (E30-OE1, E30-OE2, V115-O, I7-O), van der Waals force (I60-CAR, F31-RING), the binding interactions of each docked poses within the cluster c were similar. The cluster d contained 6 TK ligands and one NA ligand, and all docked poses of cluster e were NA ligands, by the Figure 18 (d), (e), all docked poses with in cluster d had hydrogen-bond (Y121-O, I7-O), and all docked poses with in cluster e had hydrogen-bond (E30-OE1, V8-N). The binding interactions of docked poses within those clusters were similar, and the binding interactions between those clusters were different, indicated that the clustering by protein-ligand interaction has separated the poses into different groups with distinct binding interactions.

There were two types of hDHFR active compounds (Figure 9) [11], the DHFR03, 04, 05, 09, 10, were old drugs and had two more carboxylic acid group. The old drugs had different binding affinity comparing with new drugs. Those two types of compounds can be split correctly by the dendrogram, and the detail binding interaction of these two types of drugs was shown on Figure 19. The old drugs contain additional hydrogen bonds (R70-NH1,

R70-NH<sub>2</sub>, and N64-ND<sub>2</sub>) comparing to new drugs, these were shown on the residue numbers in red of Figure 19(a). We could also identify important van der Waals force by the pointers in red on the heat map (I60-ven der Waals force, F31-stacking force, F34-stacking force, NAP-stacking force), those were the residues in yellow shown on (b), (c). Visual inspection confirms that our method could easily identify the difference of these two binding interactions by the blue frames on the heat map, Figure 19(a).

By comparing the binding interaction between cluster c, d, e, f, and g, our cluster method could separate docked poses into distinct clusters that reveal distinct binding interactions, and mining important interaction of protein ligand binding.

### 3.4.2 Cluster analysis of Molecular Docking on TK (Thymidine kinase)

We also applied our cluster method to analysis the result of molecular docking on TK (1kim). After filtering out the compounds which the docked pose was out of binding site, there were 53 docked poses include 10 docked poses of active compounds. A total number of 305 atoms involved were identified. Protein-ligand interactions were generated for all complexes, each of which was composed of 305 real numbers. The hierarchical clustering result of these fingerprints was shown in Figure 20 represented as a heat map. By the reference threshold (Correlation coefficient: 0.3894), two major clusters can be identified in Figure 20, cluster c and d. All docked poses of active compounds were grouped into cluster c. the major different of interaction between cluster c and d was shown in Figure 20 (b), (c), (d),

the cluster c contained the hydrogen bond with Q125-NE1 and NE2, the cluster d only interact with Q125-NE2. And the cluster d contained two positive van der Waals force on A167-C, but cluster c did not have the interaction. That was because the structures of cluster d were slightly larger than the volume of the cavity. Those observations could easily inspected by the visualized heat map, and it is useful for mining conserved interaction within a cluster.

### **3.4.3 Cluster analysis of Molecular Docking on NA, ER $\alpha$ (3ert), and ER $\alpha$ (1gwr)**

The process of cluster analysis was the same as described above, in the section of NA, the result was shown on Figure 21, all the known active compounds were grouped within a cluster (frame in red) and had hydrogen-bond with target protein (R152, E277, R292, and R371). In the section of ER $\alpha$  (3ert), the result was shown on Figure 22, the active compounds were divided into four clusters by the red frames on the heat map, two were singleton, one contains 4 inhibitors, and last cluster contained 5 inhibitors and 8 ER $\alpha$  agonists. We could inspect that the positive van der Waals force on (I424, M388, and L349) made EST11 and EST10 different from other inhibitors. In the section of ER $\alpha$  (1gwr), the result was shown on Figure 23, all active compounds except ESA08 were grouped into one cluster, and the ESA08 had additional interaction with target protein (T347 and L525).

#### **B. Clustering by atom-pair descriptor:**

### 3.4.4 Cluster analysis of Compound Structures on Verifying Dataset

The second stage cluster method was based on 2D topology: atom-pair descriptor. The descriptor was calculated by atom-pair representation, using tanimoto coefficient for measuring distance between two molecules. The dendrogram of hierarchical clustering of 61 known compound structures was shown on Figure 24(d). Under the reference threshold (tanimoto coefficient = 0.55), There were three major clusters, a, b, and c. In the cluster a, Figure 24(a), all 10 ER $\alpha$  agonists were grouped within the cluster. In the cluster b, Figure 24(b), all 11 ER $\alpha$  antagonists were also grouped within the cluster. In the cluster c, Figure 24(c), all 10 TK inhibitors and 14 NA were grouped together because the structures between TK and NA inhibitors were similar. By the observation on these three clusters, we could inspect that the atom-pair descriptor could group compounds with similar structures and divided compounds with different structures.

### 3.5 Cluster analysis of Virtual Screening Results on Testing Dataset

#### 3.5.1 First Stage Cluster analysis on hDHFR Dataset

We performed a virtual screen for a set of 10 hDHFR inhibitors spiked into 990 randomly selected compounds from ACD. After virtual screening, we analysis the result of virtual screening by protein-ligand interaction. The top 100 rankers predicted by GEMDOCK were selected for cluster analysis. After removing the columns with value of all zero, a total

number of 476 atoms involved were identified, contained 100 docked poses include 10 of known active compounds. Protein-ligand interactions of all complexes were generated, each of which was composed of 316 real numbers. The correlation coefficient was applied in hierarchical clustering for measuring the distance between two docked poses. The result of hierarchical clustering was shown on Figure 25 represented as a heat map. The dendrogram was on the left side of the heat map. All hDHFR inhibitors were grouped together into one cluster in the red frame. This cluster contain 45 compounds included 10 active compounds and 35 unknown compounds, the detail binding interactions of active and unknown compounds were shown on Figure 26(b), (c). In the Figure 26(b) and (c), all 10 active compounds had hydrogen-bond with target protein (I7-O, V115-O, E30-OE1, E30-OE2, N64-ND2) and ven der Waals force (F31-stacking force, F31-stacking force, I60-ven der Waals force, NAP-stacking force). All the 35 unknown compounds had similar hydrogen-bond with target protein (I7-O, V115-O, E30-OE1, E30-OE2, N64-ND2) and ven der Waals force (F31-stacking force, F31-stacking force, I60-ven der Waals force, NAP-stacking force). The binding interactions within the cluster were similar. By inspecting the compound structures in Figure 26 (b) and (c), the compounds within the cluster contained similar binding interaction but diverse compounds structures. By this observation, we utilized the 2D topology of compounds for selected the representative compounds within the cluster after clustering by protein-ligand interaction.

### 3.5.2 Second Stage Cluster analysis on hDHFR Dataset

Each cluster generated by first stage clustering was clustered by second stage clustering for selecting representative compounds within each sub-cluster. We performed the second cluster method to analysis the result of first stage clustering. In the case of hDHFR, we give the largest cluster from first stage clustering for example. The process and result was shown on Figure 27. Figure 27(a) was the binding interactions of the largest cluster generated from first stage clustering, the cluster contained 45 compounds include 10 active compounds and 35 unknown compounds. Each compound was represented by one dimension atom-pair binary string from 2D topology. We applied the tanimoto coefficient for measuring the distance between two compounds. After performing the hierarchical clustering, the result was shown on Figure 27(b), there were four major clusters identified by the dendrogram, Figure 27(c), (d), (e), (f). It was expected that the active compounds were spliced into two clusters, the old drugs Figure 27(d) and the new drugs Figure 27(e) because of the difference of the carboxylic acid group. The sub-structures within each cluster selected by the red circle in Figure 27(c) and (f) were similar, but the sub-structures were different between each cluster. After the second stage clustering, we selected the lowest energy compound within each cluster for representing all compounds within the cluster. Each representative compounds structures were shown on Figure 27(g), (h), (i), and (j).

Our method was able to identify and classify those sharing the same or similar binding

mode, and selecting the representative compounds within each cluster of similar binding mode for use in bioassay.





## Chapter 4

**Applications: Using two stages Cluster method for post-analysis on the results of virtual screening of Shikimate kinase of helicobacter pylori.**

### 4.1 Preparations of the Target Protein and Compound set

The shikimate pathway is a seven-step biosynthetic route that involves aromatic amino acid biosynthesis and many aromatic secondary metabolites, including tetrahydrofolate and ubiquinone. The shikimate pathway was shown on Figure 28. The shikimate pathway is an attractive target for the discovery of new antimicrobial agents, herbicides and antiparasitic agents because it is essential in bacteria, fungi, and plants, but not mammals[46, 47, 51, 52].

Until now, we know several successful drugs in shikimate pathway for the pathogenesis of a number of microorganisms. Shikimic acid analogs—(6*R*)-6-Fluoro-shikimate and (6*S*)-6-fluoro-shikimate exhibited different inhibitory effects on *Plasmodium falciparum* growth in vitro (4). 5-Enolpyruvylshikimate 3-phosphate (EPSP) synthase, which catalyzes the sixth step in the pathway, has been successfully targeted, with the development of glyphosate, one of the world's best-selling herbicides[53].

The shikimate kinase (SK; EC 2.7.1.71), the fifth enzyme of the pathway, Figure

28(c), catalyzes the specific phosphorylation of the 3-hydroxyl group of shikimic acid using ATP as a co-substrate. Cheng *et al.* have determined the crystallographic structure of shikimate kinase from *Helicobacter pylori* (HpSK; PDB code, 1ZHU), showing a three-layer  $\alpha/\beta$  fold consisting of a central sheet of five parallel  $\beta$ -strands flanked by eight  $\alpha$ -helices. They also determined a complex structure—HpSK-shikimate-PO<sub>4</sub> (PDB code, 1ZHI) and revealed induced-fit movement from an open to a closed form on substrate (shikimate) binding[45].

We used the closed form of HpSK (PDB code, 1ZHI) to be the target protein because the protein structure with substrate had the induced-fit movement would be more suitable for generating correct interaction between protein and ligand. We defined that the binding site of the HpSK was the collection of amino acids enclosed within a 8 Å radius sphere centered on the bound ligand, shikimate (SKM). The coordinates of atoms were derived from Dr. Cheng, W. C. and stored in the PDB format for inputting into GEMDOCK. After comparing the induced-fit movement of the lid structure from an open to a closed form by Figure 29 (b)(c), we removed the lid structure (residue number: 108~124) for capable docking larger potential compounds.

We prepared the compound set from the CMC drug database. It was first filtered with molecular weights between 200 and 800, then removed small fragment in record. Finally we had 6,443 molecules in the compound set.

## 4.2 Molecular Recognition and Setting of Pharmacophore Consensus on the Shikimate

## **Kinase**

To evaluate the docking accuracy of GEMDOCK for the shikimate kinase of dengue virus, we docked the substrate shikimate (SKM) into its binding site of the complex. GEMDOCK executed 10 independent runs for each docking. The solution with the lowest scoring function was then compared with the 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 result was shown on Figure 30. The RMSD value of the docked pose was 0.023 Å, indicated that after removing of the lid structure, the GEMDOCK still could dock the substrate back into the target protein correctly. According to the protein-ligand complex, the shikimate forms hydrogen bonds with D33-OD1, D33-OD2, R57-NH1, G80-N, R132-NH1, and R132-NH2. Because there is no other known ligand, we set the pharmacophore consensus by the hydrogen-bond of subtract listed in Table 7. According to the property of shikimate and the cavity was a hydrophilic pocket, the interaction preference was set the  $UB_{elec}$  was 2 and  $Ur_{hb}$  was 0.42 when we screened the CMC 6,443 compounds, listed in Table 6.

### **4.3 Virtual Screening for the Shikimate Kinase**

Based on the screening utility of GEMDOCK described above, we applied GEMDOCK on virtual screening for the shikimate kinase with a compound set including 6,443 compounds from the CMC. The pharmacophore consensus and ligand preference was used while screening.

### **4.4 Two Stage Cluster analysis of Result of Virtual Screening for Selecting Representative Compounds.**

The overall process was shown on Figure 33. First, we applied virtual screening on CMC 6443 compounds by GEMDOCK. Second, upon the rank of shikimate (rank: 268) in Figure 31, we selected top 300 compounds for two-stage cluster analysis. Third, compounds were clustered by protein-ligand interaction. The threshold was set manually for giving an appropriated number of clusters. The result of cluster analysis was shown on Figure 32. There were 8 major clusters and others compounds, totally 9 clusters. Forth, we applied the atom-pair descriptor for clustering compounds within each cluster and 23 representative compounds were selected for use in bioassay. We give more concern on the cluster that has shikimate and choice 5 compounds for use in bioassay. Figure 34 shown the structures of the 23 representative compounds. Fifth, 5 compounds were tested in vivo (compound structures with blue frames) and one had 36% inhibition on shikimate kinase (compound structure with red frame) at the concentration of 625  $\mu\text{m}$ .

Our research was cooperated with Dr. W. C. Wang, we have identified a potential inhibitor (MCMC00000106) for shikimate kinase from *Helicobacter pylori* by GEMDOCK.

## Chapter 5

### Conclusions

#### 5.1 Major Contributions and Future Works

We developed a two stages cluster method for analysis the results of virtual screening by protein-ligand interaction and compound structure. We validated this method on data sets having five classes: thymidine kinase inhibitors, dihydrofolate reductase inhibitors, estrogen receptor agonist, estrogen receptor antagonists and neuraminidase inhibitors. Our method on these pharmaceutical interest targets provided a suggestion of cluster threshold and helped to mining diversely representative structures from large number of virtual screening data. Our method also has been applied on the practical inhibitor screening analysis for *Helicobacter pylori* shikimate kinase (HpSK). After virtual screening in CMC database, we selected compounds from top 300 and selected 23 representative candidates. Five of 23 representative candidates were tested *in vivo*, and one of the five candidates, furosemide, was identified being able to inhibit HpSK by cooperated laboratory of Dr. Wen-Ching Wang.

An overall index for evaluating the accuracy of our two staged cluster method is useful for comparing with other methods.

## Tables

**Table 1.** Ten atom types used in atom-pair descriptors

	Description	Atom type	Mol2 format atom type
1	Aromatic carbons	C.ar	C.ar
2	Nonaromatic carbons	C.na	C.3 C.2 C.1 C.cat
3	Aromatic nitrogen	N.ar	N.ar
4	Nonaromatic nitrogen	N.na	N.3 N.2 N.1 N.am N.4 N.pl3
5	Oxygen atoms in the sp <sup>3</sup> hybridization state	O.3	O.3
6	Oxygen atoms in the sp <sup>2</sup> hybridization state	O.2	O.2
7	All sulfur atoms	S	S.3 S.2 S.O S.o2
8	Phosphorus atoms	P.3	P.3
9	Halogen atoms	X	F Cl Br I
10		Other atoms	Other atom types



**Table 2.** 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 3.** Atom formal charge of GEMDOCK.

Formal charge	Atom name
<b>Receptor:</b>	
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
<b>Ligand:</b>	
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 4.** Parameters used for docking on verify dataset

Parameter	Value of parameters
Population size	600
No. of the maximum generation	80
No. of runs	10

**Table 5.** Parameters used for docking on testing dataset

Parameter	Value of parameters
Population size	300
No. of the maximum generation	60
No. of runs	3





**Table 6.** The ligand preferences calculated from known active compounds used for virtual screening on TK, ER, hDHFR, NA, and HpSK

Ligand name	Electrostatic preferences (Equation 9)			Hydrophilic preferences (Equation 10)		
	$\theta_{elec}$	$\sigma_{elec}$	$UB_{elec}$	$\theta_{hb}$	$\sigma_{hb}$	$Ur_{hb}$
TK-substrate	0	0	0	0.50	0.05	0.55
ER-antagonist	2.00	0.56	2.56	0.15	0.03	0.18
ER-agonist	0	0	0	0.25	0.06	0.31
hDHFR-ligand	4.00	2.11	6.11	0.40	0.05	0.45
NA-ligand	4.00	0.75	4.75	0.50	0.05	0.55
SK-substrate	2.00	0	2.00	0.42	0	0.42



**Table 7.** The pharmacophore consensus calculated by superimposing known active compounds used for molecular docking on TK, ER, hDHFR, NA, and HpSK

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

Pharmacophore consensus			
weight ( $CW(B_{ij})$ )			
Residue	Atom	TK-ligand	Interaction type
Id <sup>a</sup>	Id <sup>b</sup>		
Q125	OE1	4.00	H-bond (NH ↔ O) (NH group)
Q125	NE2	3.50	H-bond (O ↔ NH) (carbonyl group)
Y101	OH	2.00	H-bond (OH ↔ OH) (hydroxyl group)
R163	NH1	1.50	H-bond (OH ↔ N) (hydroxyl group)
	CG		
	CD1		
	CD2		
Y172	CE1	2.50	van der Waals force (C ↔ C)
	CE2		
	CZ		
NA-ligand			
R371	NH1	2.0	H-bond (NH ↔ O) (NH group)
R371	NH2	2.0	H-bond (NH ↔ O) (NH group)
R292	NH1	1.5	H-bond (NH ↔ O) (NH group)
R292	NH2	1.5	H-bond (NH ↔ O) (NH group)
E276	OE2	1.5	H-bond (OH ↔ OH) (hydroxyl group)
R152	NH1	2.0	H-bond (O ↔ NH) (carbonyl group)

Pharmacophore consensus			
weight ( $CW(B_{ij})$ )			
SK-substrate			
D33	OD1	1.5	H-bond (OH ↔ OH) (hydroxyl group)
D33	OD2	1.5	H-bond (OH ↔ OH) (hydroxyl group)
R57	NH1	1.5	H-bond (O ↔ NH) (carbonyl group)
G80	N	1.5	H-bond (NH ↔ O) (NH group)
R132	NH1	1.5	H-bond (NH ↔ O) (NH group)
R132	NH2	1.5	H-bond (NH ↔ O) (NH group)

<sup>a</sup> One-code amino acid with the residue sequence number in PDB.

<sup>b</sup> The atom name with the atom serial number in PDB.



**Table 8.** The RMSD between docked poses and crystal ligands

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



<sup>a</sup> Four antagonists dock into target protein (3ert)

<sup>b</sup> Four agonists dock into target protein (1gwr)

**Table 9.** T-test of distance between similar and non-similar binding mode generated by converting the docked pose into protein-ligand interaction profile ( $\alpha=0.01$ ).

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

<sup>a</sup> Standard Deviation



**Table 10.** T-test of distance between similar and non-similar structure generated by atom-pair representation ( $\alpha=0.01$ ).

Target protein	H <sub>0</sub>	Similar : Average Distance	Non-similar : Average Distance	P-value	Similar : Std <sup>a</sup> of Distance	Non-similar : Std <sup>a</sup> of Distance
DHFR	Reject	0.42	0.63	5.84E-23	0.15	0.12
ESA	Reject	0.24	0.66	4.60E-65	0.11	0.14
EST	Reject	0.27	0.63	2.85E-56	0.14	0.14
NA	Reject	0.32	0.65	1.75E-131	0.18	0.17
TK	Reject	0.22	0.63	2.11E-93	0.09	0.19

<sup>a</sup> Standard Deviation



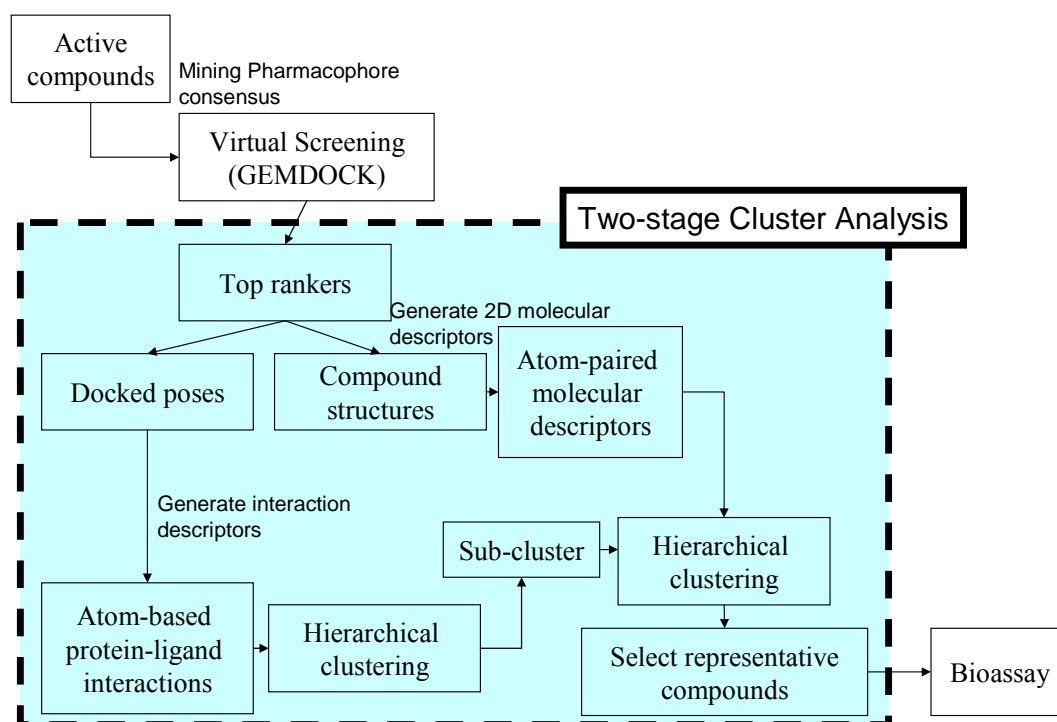
**Table 11.** T-test of distance between similar and non-similar compounds on each target protein. Descriptor was generated by converting the docked pose into protein-ligand interaction profile ( $\alpha=0.01$ ).

Target protein	Compound class	H <sub>0</sub>	Similar : Average Distance	Non-similar : Average Distance	P-value	Similar : Std <sup>a</sup> of Distance	Non-similar : Std <sup>a</sup> of Distance
DHFR	DHFR	Reject	0.21	0.50	1.71E-58	0.09	0.13
	ESA	Reject	0.52	0.58	2.73E-03	0.18	0.12
	EST	Reject	0.52	0.63	7.51E-07	0.21	0.13
	NA	Reject	0.46	0.55	5.34E-23	0.13	0.14
	TK	Reject	0.38	0.51	8.03E-11	0.16	0.13
ESA	DHFR	Pass	0.55	0.62	0.10111	0.28	0.16
	ESA	Reject	0.23	0.48	2.29E-31	0.14	0.14
	EST	Pass	0.67	0.76	0.23105	0.25	0.14
	NA	Reject	0.33	0.59	1.51E-58	0.24	0.20
	TK	Reject	0.46	0.57	0.000121	0.25	0.20
EST	DHFR	Pass	0.55	0.57	4.01E-01	0.21	0.14
	ESA	Reject	0.25	0.42	7.04E-20	0.13	0.12
	EST	Reject	0.31	0.48	7.94E-39	0.09	0.12
	NA	Reject	0.40	0.46	1.46E-09	0.15	0.15
	TK	Reject	0.28	0.43	2.17E-29	0.09	0.15
NA	DHFR	Reject	0.35	0.68	3.46E-25	0.22	0.25
	ESA	Reject	0.59	0.71	2.91E-04	0.28	0.24
	EST	Reject	0.56	0.66	2.46E-04	0.25	0.24
	NA	Reject	0.17	0.73	0.00E+00	0.07	0.20
	TK	Reject	0.48	0.60	3.46E-07	0.18	0.23
TK	DHFR	Reject	0.42	0.62	9.80E-12	0.13	0.10
	ESA	Reject	0.16	0.52	9.99E-62	0.07	0.13
	EST	Pass	0.58	0.65	6.28E-02	0.18	0.14
	NA	Reject	0.40	0.53	2.92E-53	0.11	0.15
	TK	Reject	0.19	0.47	3.89E-64	0.08	0.15

<sup>a</sup> Standard Deviation

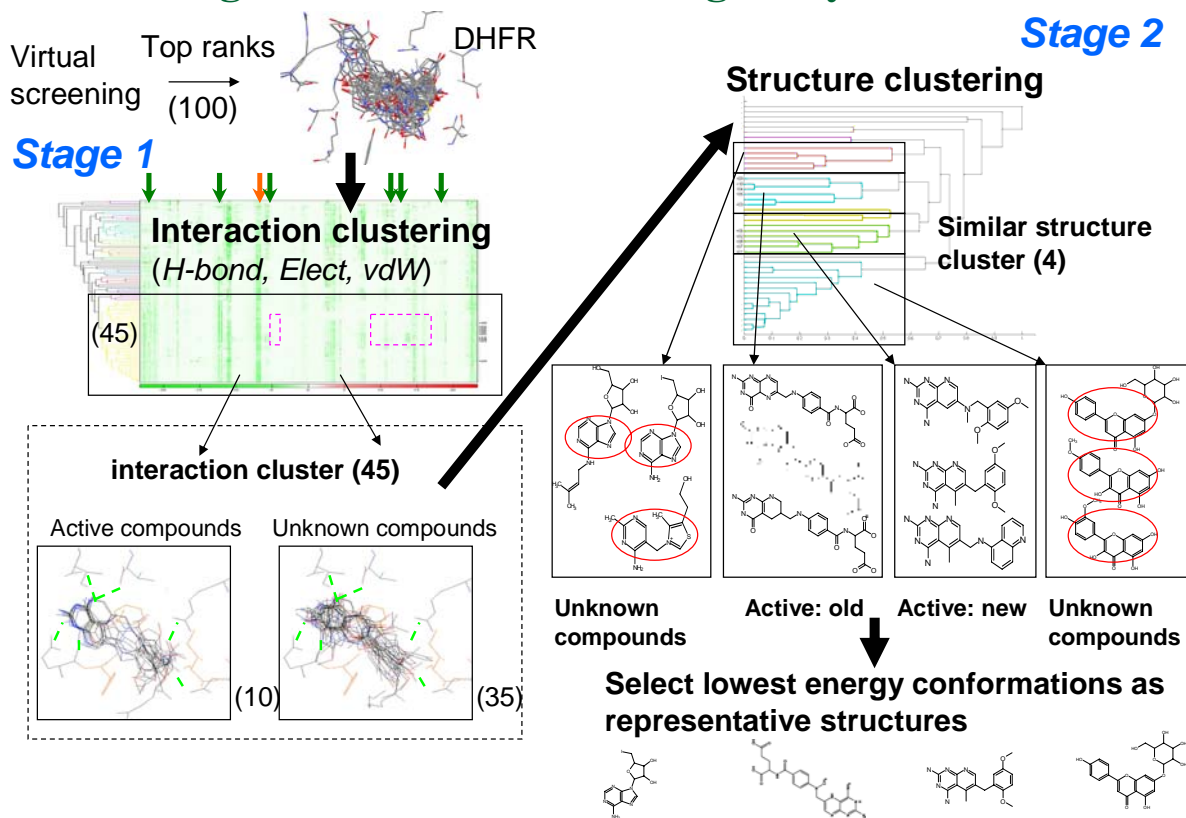


## Figures

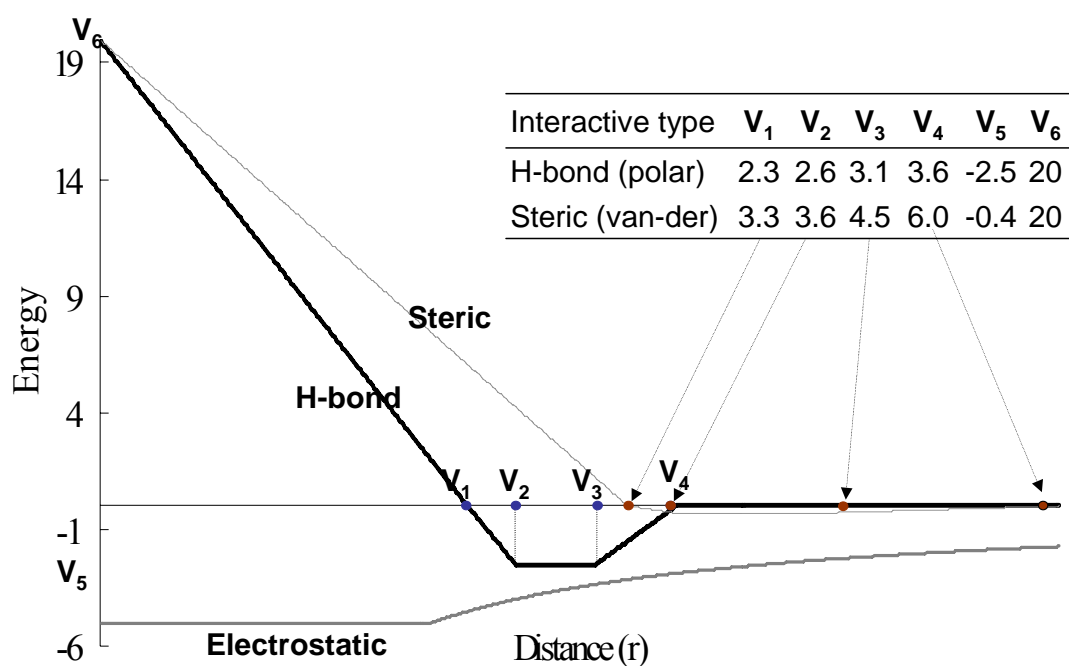


**Figure 1.** Main steps of our two-stage cluster method for analyzing the result of virtual screening. First we evaluate pharmacophore consensus from known active compounds and apply on virtual screening. Second, the GEMDOCK program was used to predict the docked conformation and rank a series of candidates by using flexible docking. Third, we translate the docked poses and structures of top rankers into one dimension real number string and binary string, respectively. Fourth, we utilize the protein-ligand interaction to cluster docked poses with similar binding mode, within each cluster, we use the atom-pair for cluster compounds with similar structure. Fifth, select representative compounds on each cluster manually for use in bioassay.

## Two stage hierarchical clustering analysis

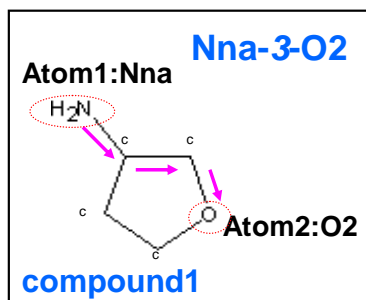


**Figure 2.** Overall process of the two stages hierarchical clustering analysis.



**Figure 3.** The linear energy functions of the pair-wise atoms for the steric interactions and hydrogen bonds in GEMDOCK (bold line) with a standard Lennard-Jones potential (light line).

atom type  $i$  - (distance) - atom type  $j$



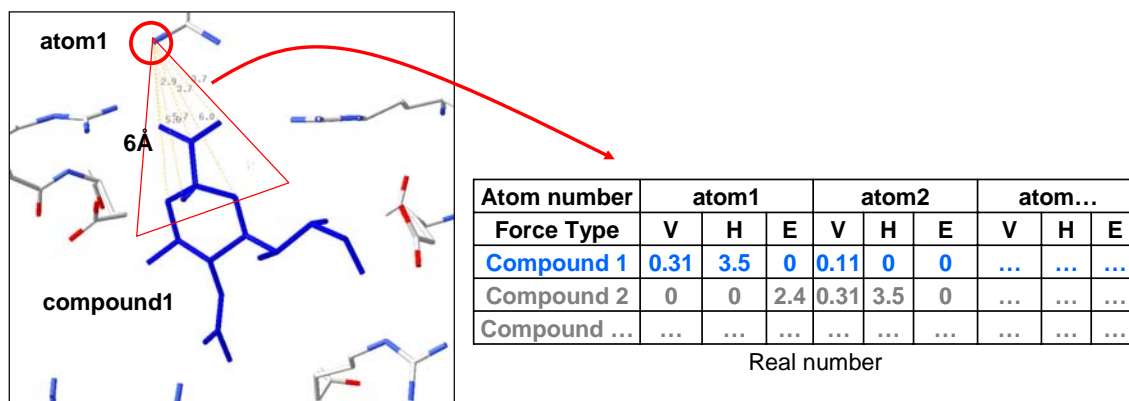
825 descriptors

Atom Pair descriptor	Car-1-Car	Can-1-O2	...	Nna-3-O2	...
Compound 1	0	1	...	1	...
Compound 2	0	0	...	1	...

Binary string

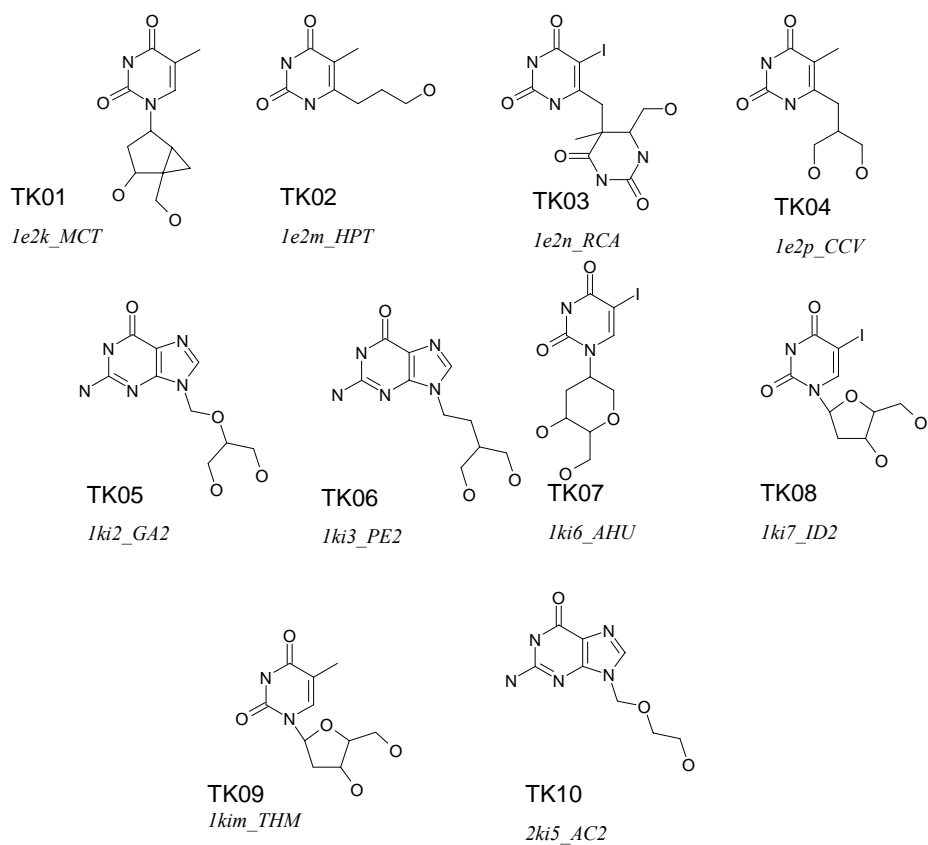
Figure 4. Definition of atom-pair representation.



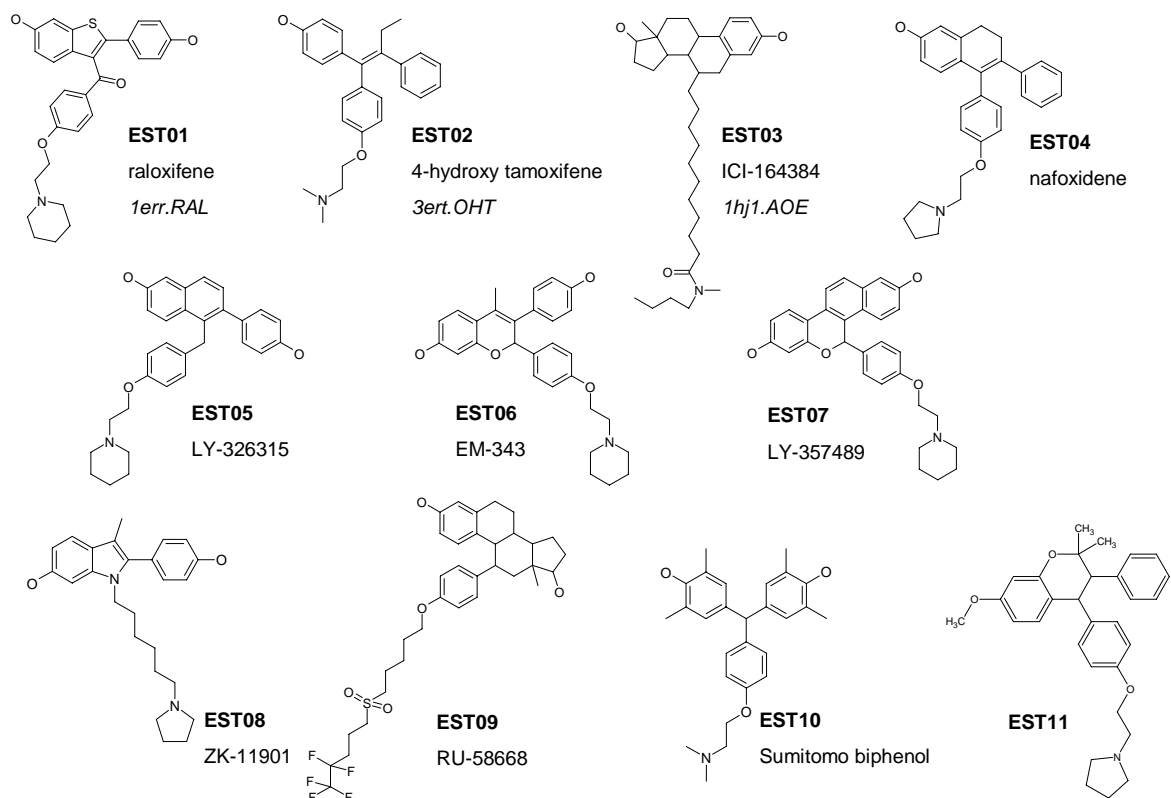


**Figure 5.** Definition of protein-ligand interaction.

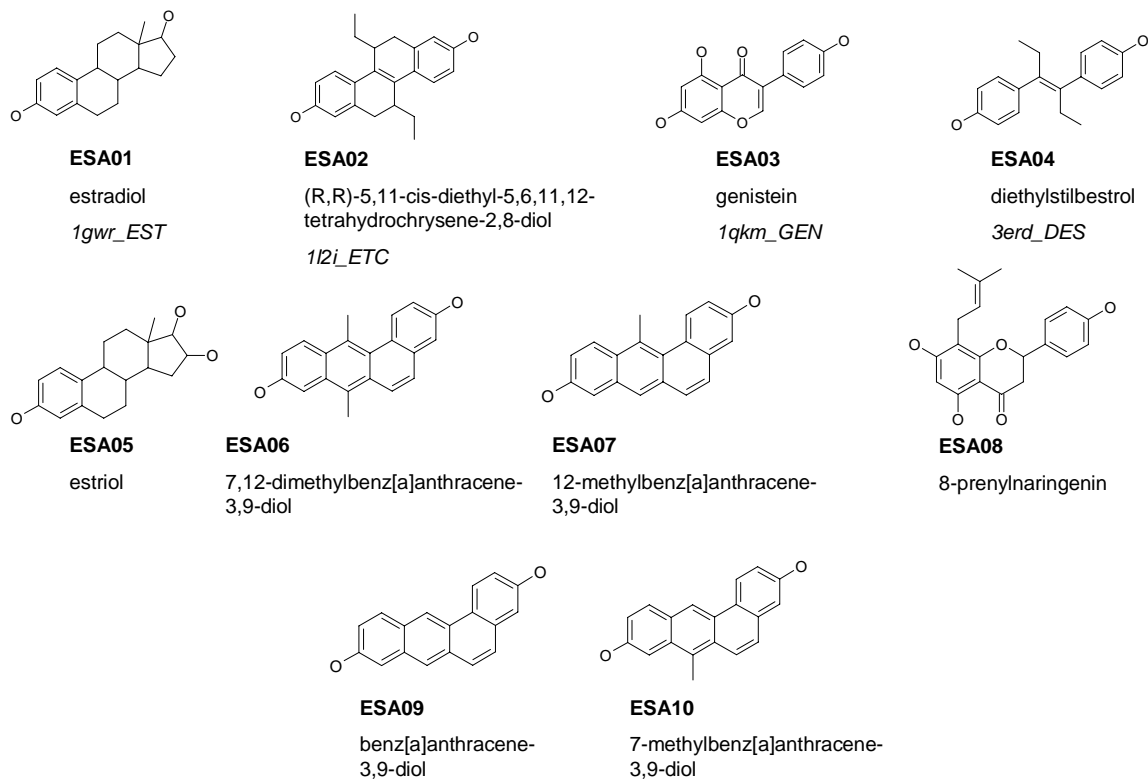




**Figure 6.** Ten TK (thymidine kinase) active compound structures.

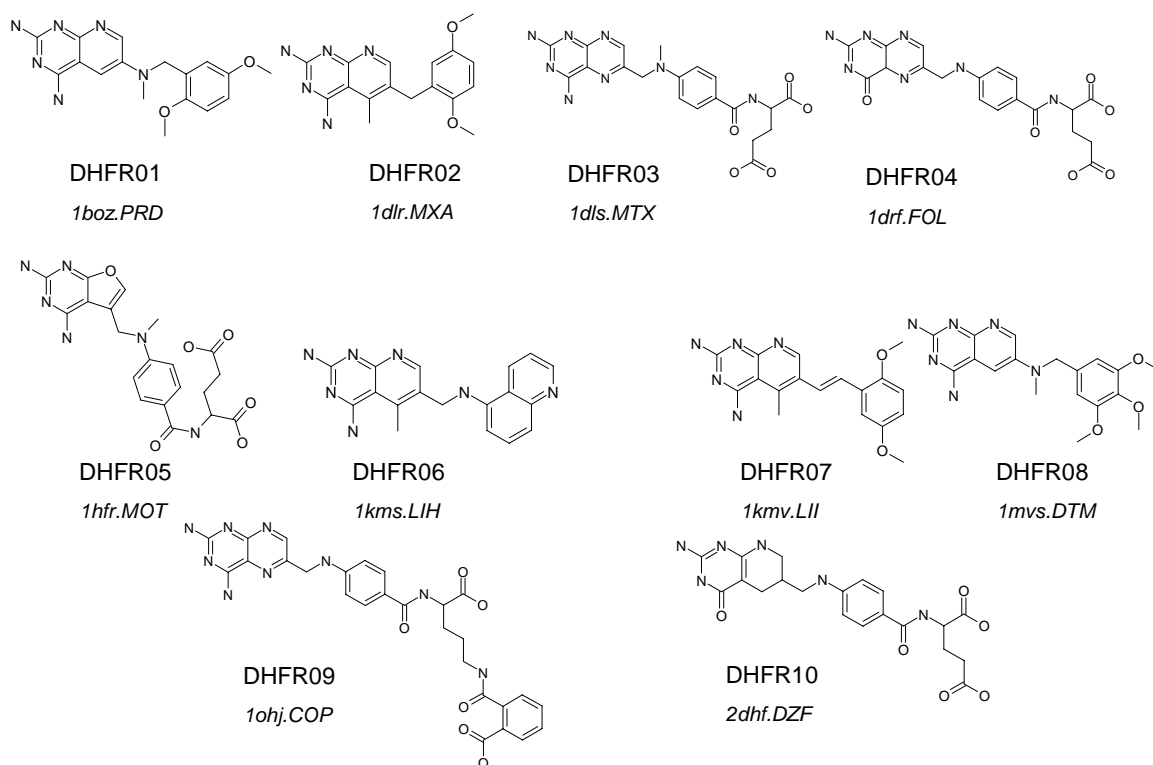


**Figure 7.** Eleven ER $\alpha$  (estrogen receptor) antagonist structures.

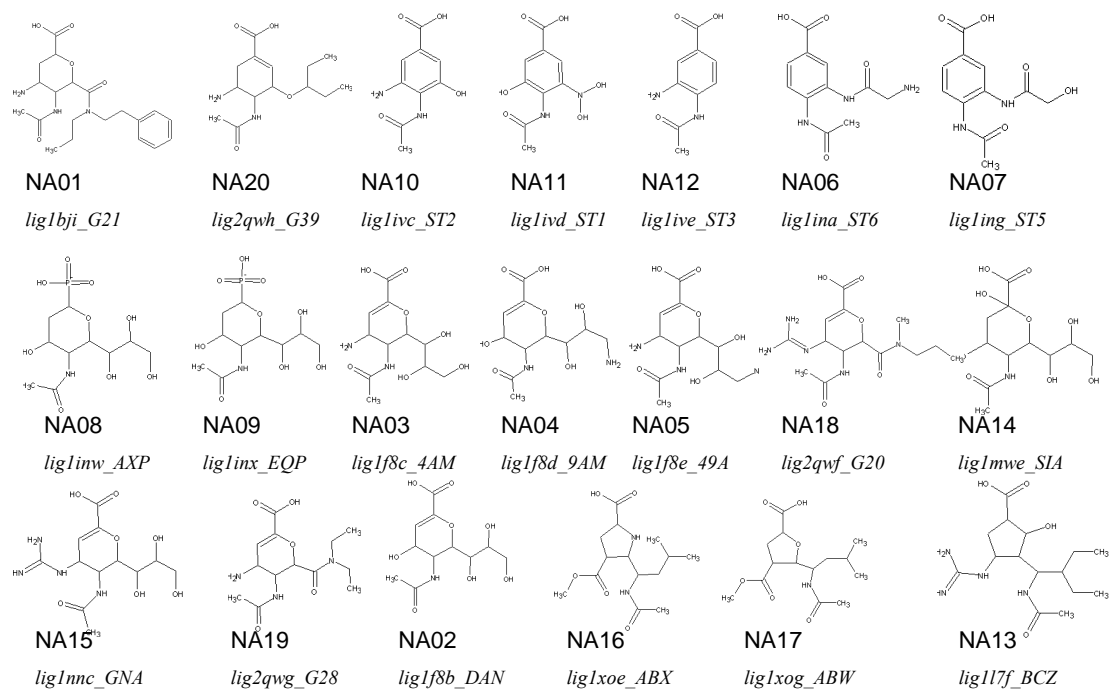


**Figure 8.** Ten ER $\alpha$  (estrogen receptor) agonist structures.



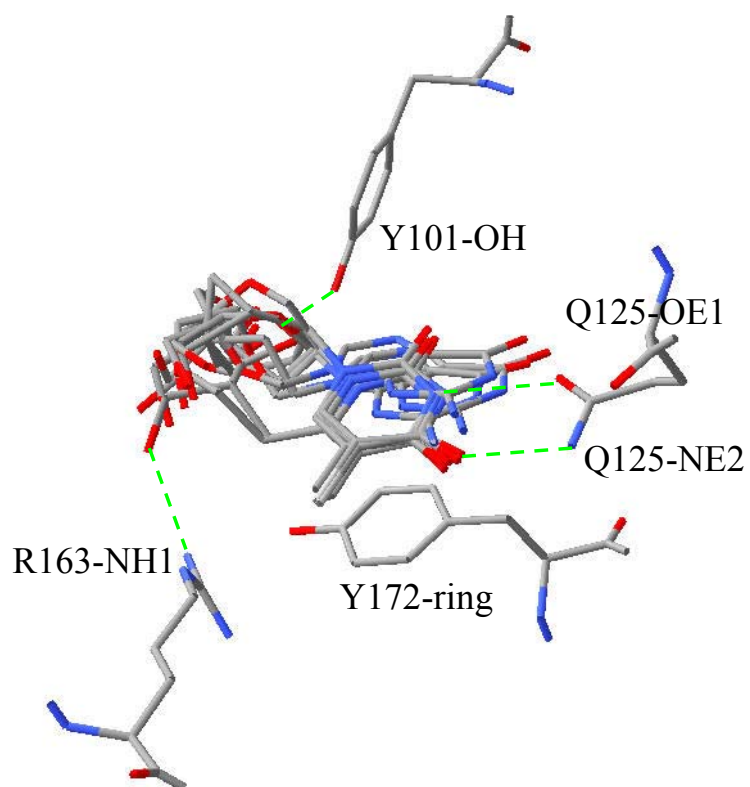


**Figure 9.** Ten hDHFR (human dihydrofolate reductase) active compound structures.

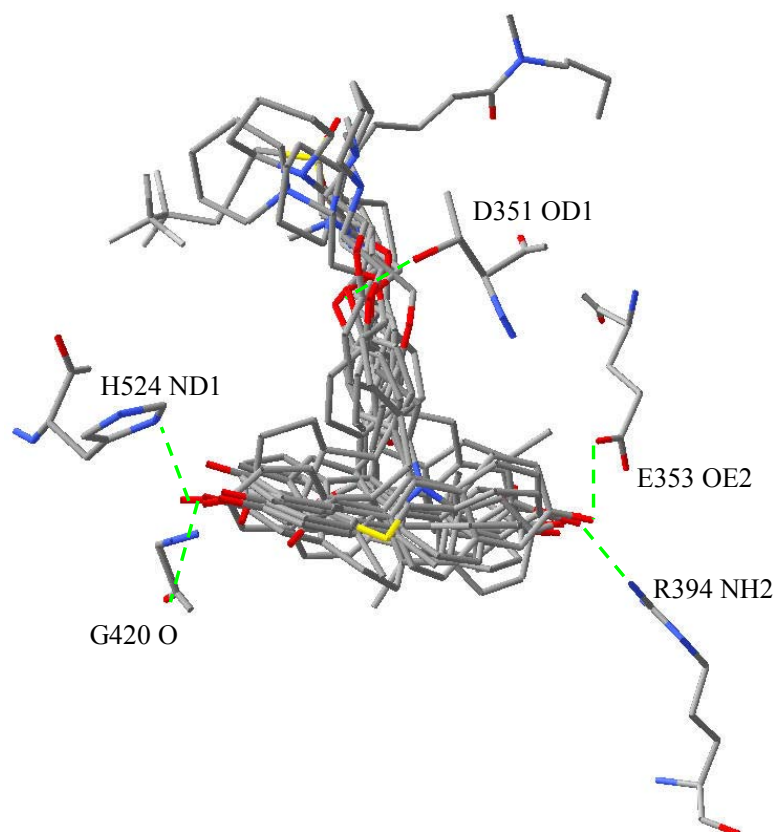


**Figure 10.** Twenty NA (neuraminidase) active compound structures.

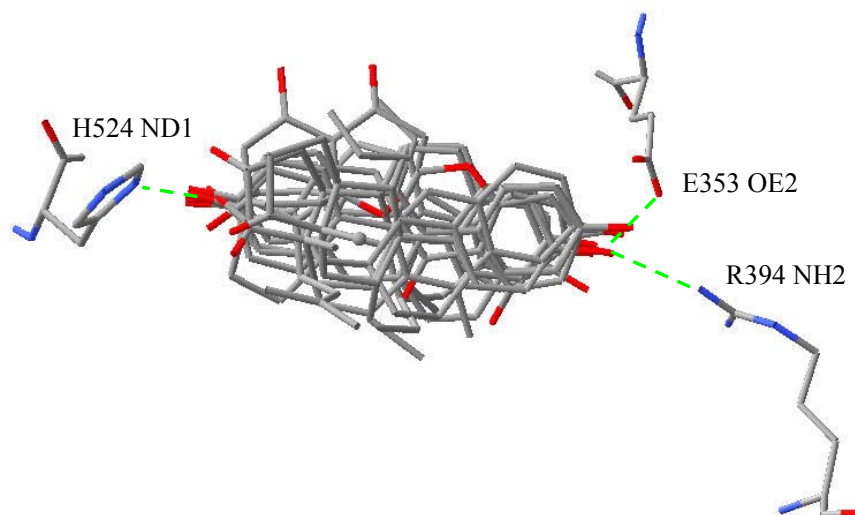




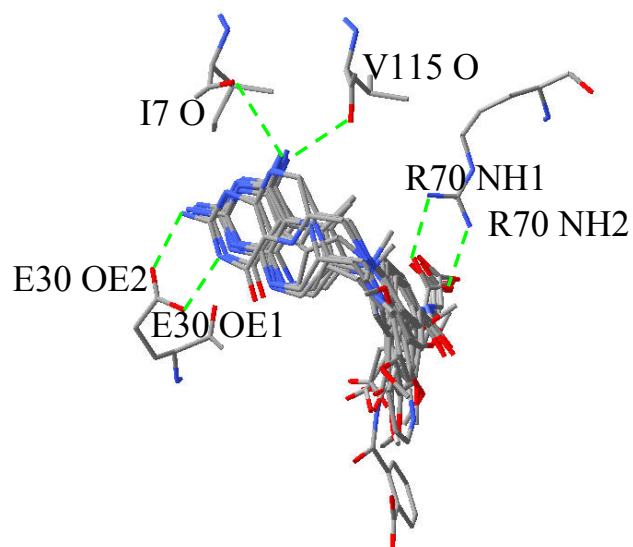
**Figure 11.** Overlapping 10 X-ray ligand structures on TK (1kim). Four important residues of the pharmacological consensus were identified and marked. The dash lines indicated the hydrogen binding. The phenolic ring of Y172 formed  $\pi - \pi$  stacking with the ligands.



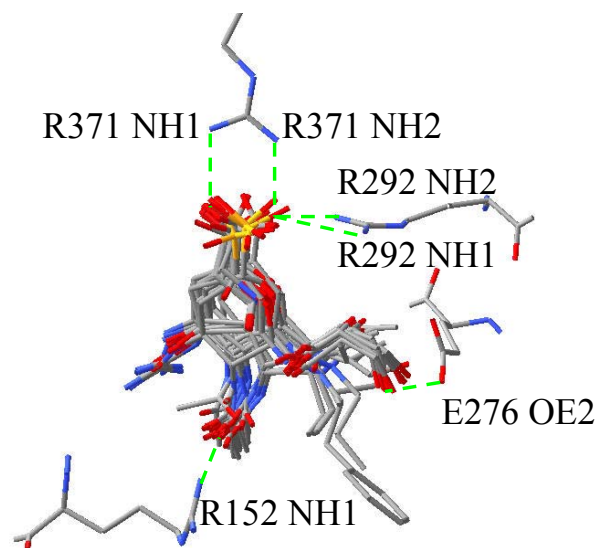
**Figure 12.** Overlapping 11 docked poses of ER $\alpha$  antagonists on ER $\alpha$  (3ert). Five important residues of the pharmacological consensus were identified and marked. The dash lines indicated the hydrogen binding.



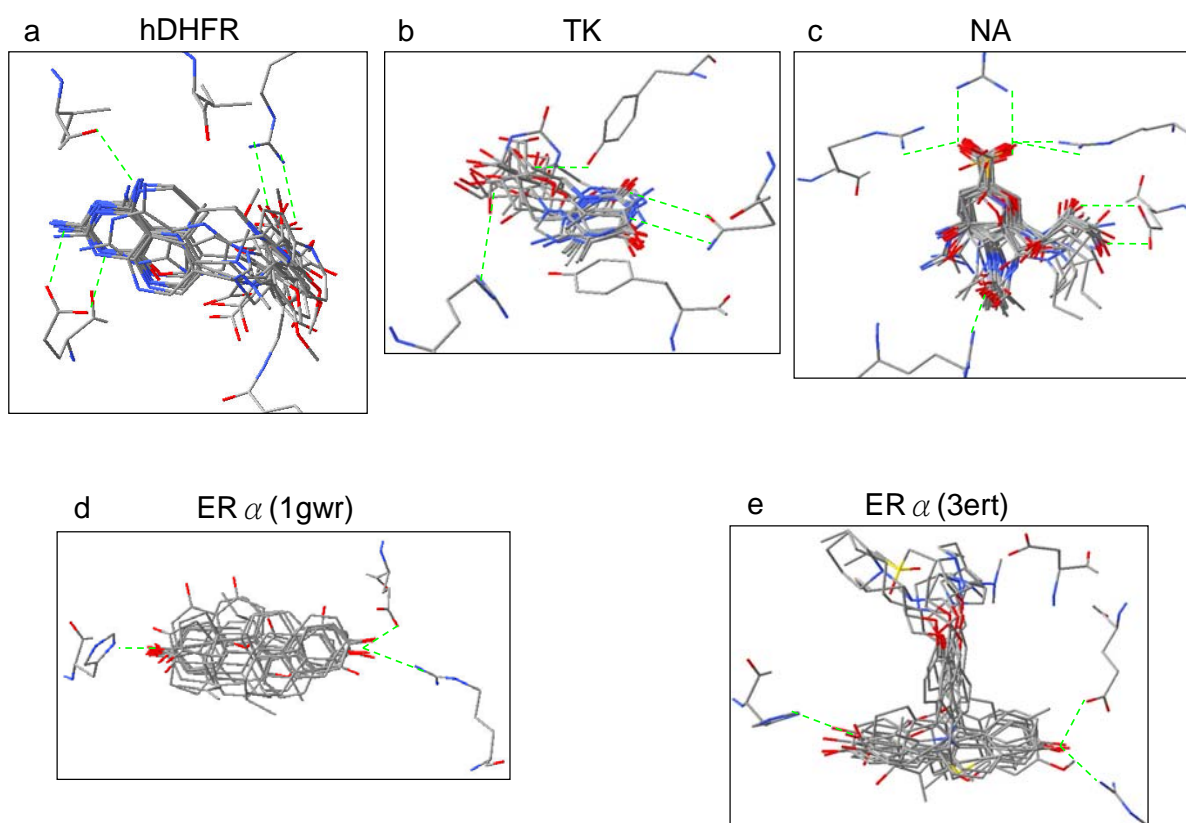
**Figure 13.** Overlapping 10 docked poses of ER $\alpha$  agonists on ER $\alpha$  (1gwr). Three important residues of the pharmacological consensus were identified and marked. The dash lines indicated the hydrogen binding.



**Figure 14.** Overlapping 10 X-ray structures of ligands of hDHFR (1hfr). Four important residues of the pharmacological consensus were identified and marked. The dash lines indicate the hydrogen binding.

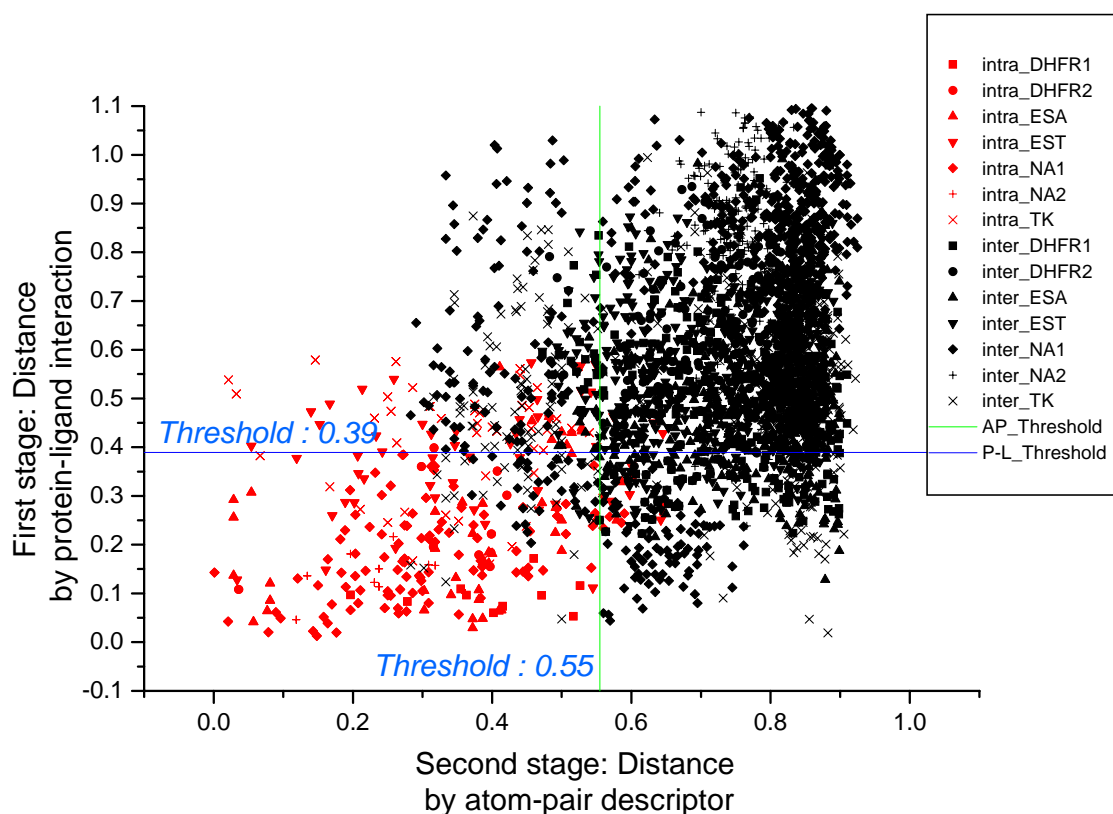


**Figure 15.** Overlapping 20 crystal structures of ligands of NA (1mwe). Four important residues of the pharmacological consensus were identified and marked. The dash lines indicated the hydrogen binding.

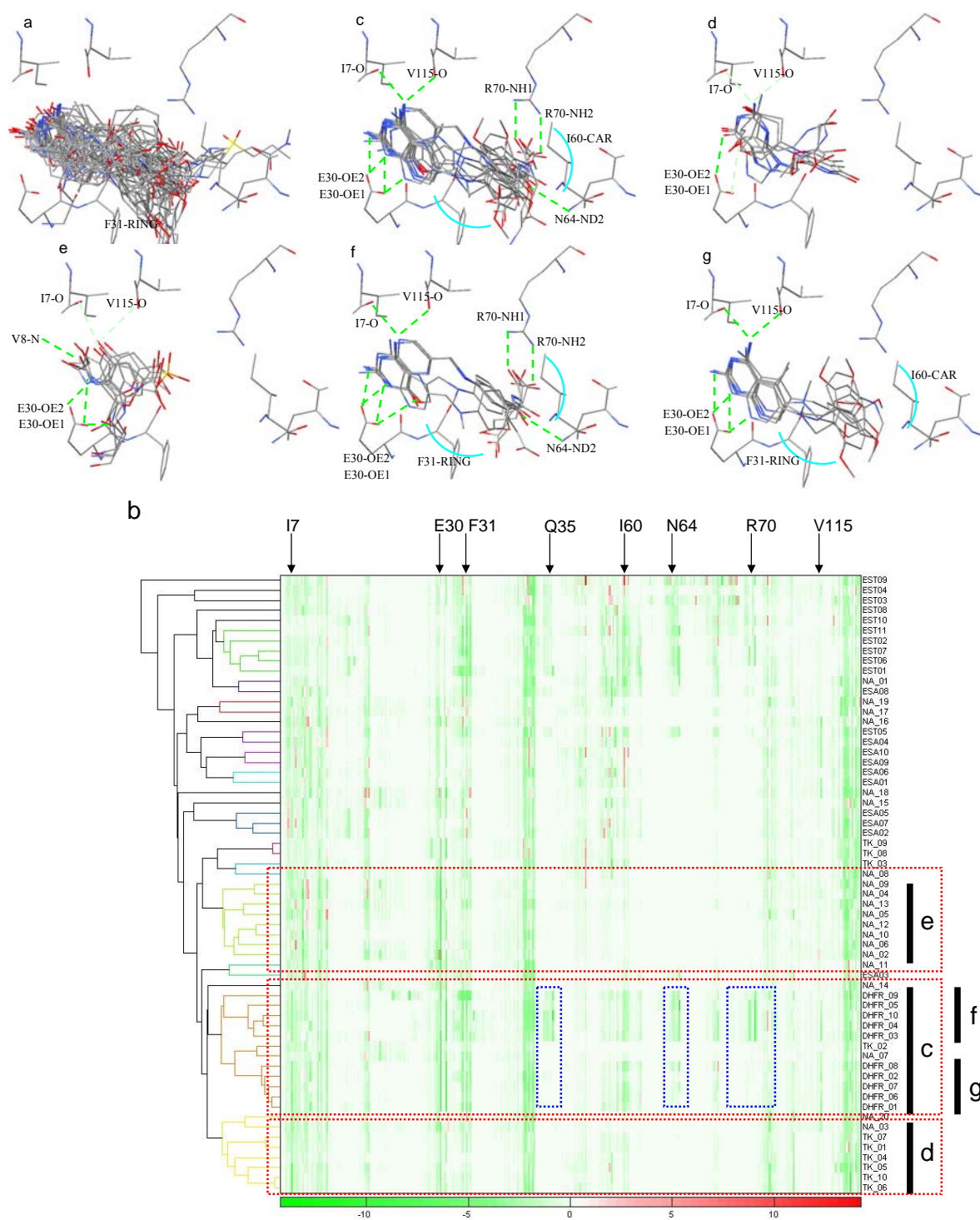


**Figure 16.** The result of molecular recognition on (a)hDHFR, (b)TK, (c)NA, (d)ER  $\alpha$  (1gwr), (e) ER  $\alpha$  (3ert)





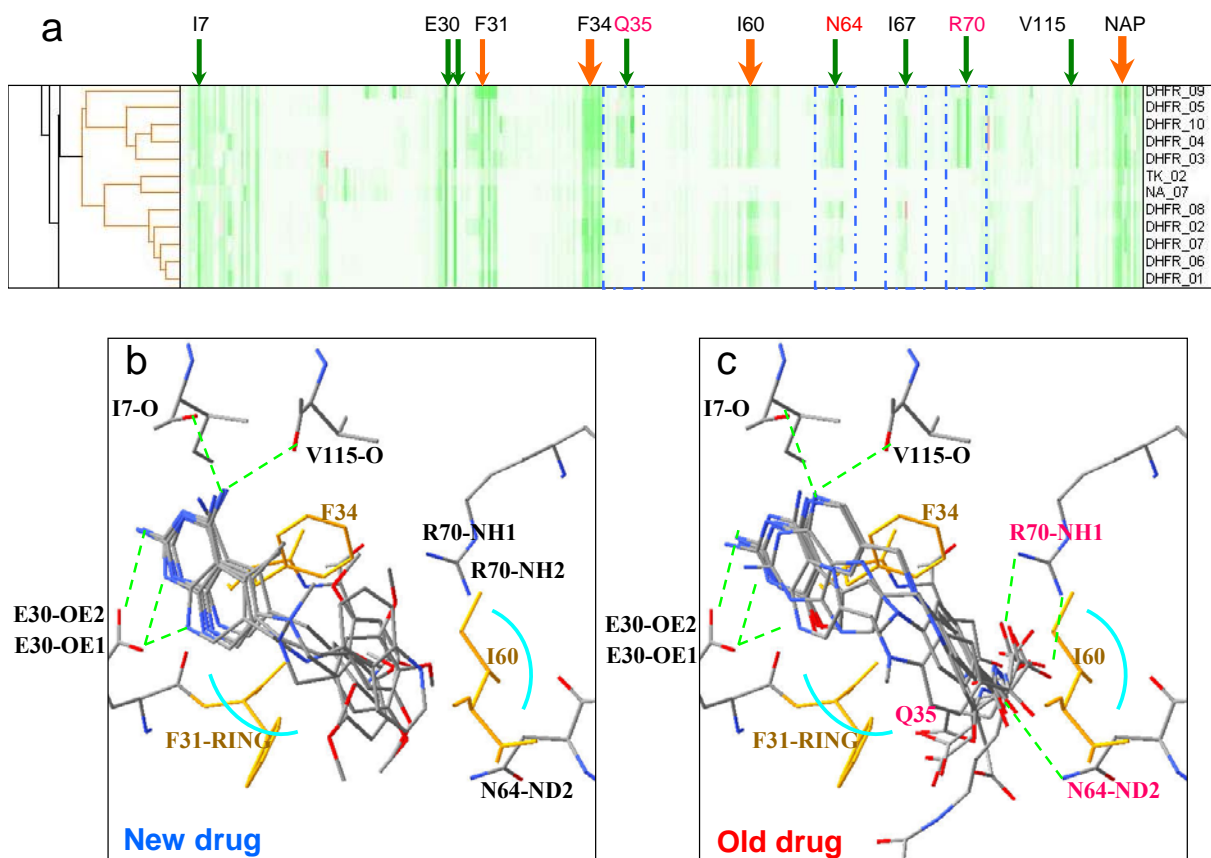
**Figure 17.** The result of designing a reference threshold of protein-ligand interaction and atom-pair descriptors. The property of complement between atom-pair descriptor and protein-ligand interaction descriptor was also show on this figure. The threshold of distance of atom-pair descriptor was 0.55 (tanimoto coefficients). The threshold of distance of protein-ligand interaction descriptor was 0.39 (correlation coefficients).



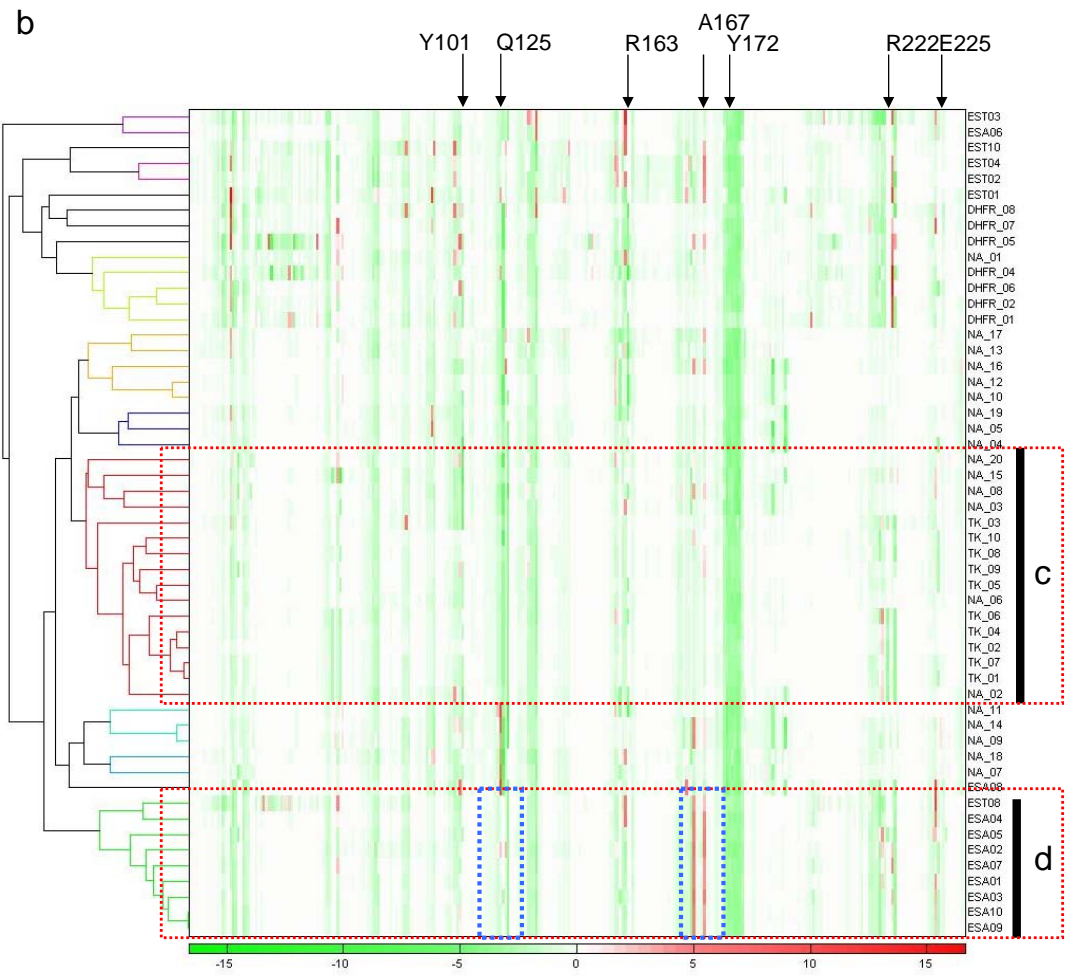
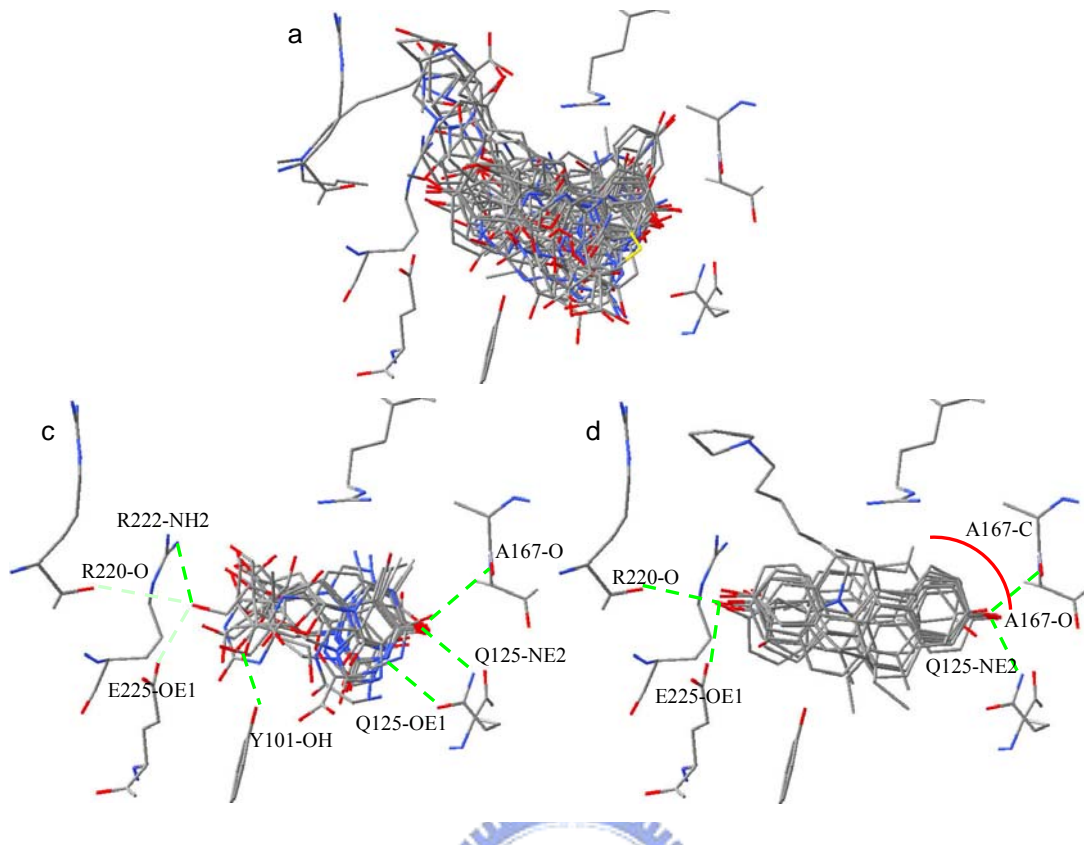
**Figure 18.** (a) Overlay of all 61 docked poses of known active compounds in the vicinity of the target protein hDHFR (PDB id: 1hfr). (b) Hierarchical clustering of protein-ligand interaction of 61 docked poses on hDHFR (PDB id: 1hfr). Each docked pose is represented as one line in the heat map in the middle of the figure, and the green being the lowest protein-ligand interaction energy and the red being the highest energy. The left side of the heat

map shows the hierarchical clustering results on the hDHFR, including the dendrogram. Docked poses in the heat map are rearranged according to the order given by hierarchical clustering marked by the black bar 'c' in the right side of the heat map. The amino acids identified for description were also shown in the top side of the heat map. (c) Overlay of docked poses of the cluster with most number of known active compounds, and shown the important interaction between protein and ligand. (d)(e) Overlay of docked poses of the cluster with most number of unknown compounds, and shown the important interaction between protein and ligand. (f)(g) Overlay of docked poses of the sub-cluster within hDHFR active compounds, and shown the important interaction between protein and ligand. The blue frames in the heat map were the major interaction that different between cluster f and g.



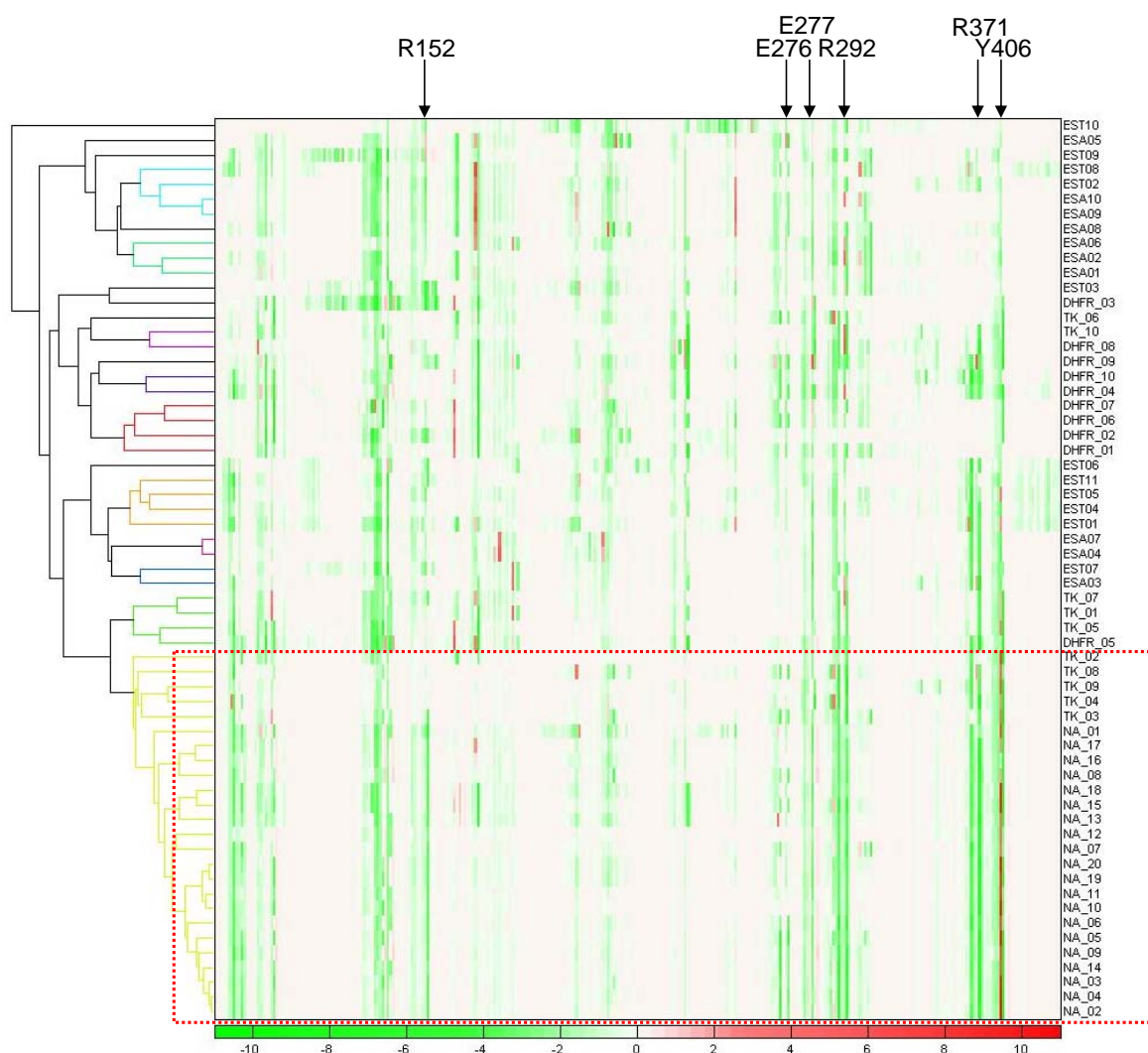


**Figure 19.** The detail of difference of binding interactions between new drugs and old drugs of hDHFR on verifying dataset. The cluster contained 12 compounds include 5 new drugs (DHFR01, 02, 06, 07, 08) and 5 old drugs (DHFR03, 04, 05, 09, 10). The binding interactions of new drugs were shown on (b), and of old drugs were shown on (c). In (a) The residue numbers in red were the major differences of interactions (Q35, N64, and R70). We could also identify important van der Waals force by the pointers in red on the heat map (I60-ven der Waals force, F31-stacking force, F34-stacking force, NAP-stacking force), those were the residues in yellow shown on (b), (c). (c) The old drugs contained additional hydrogen-bound with target protein (Q35, N64, and R70).

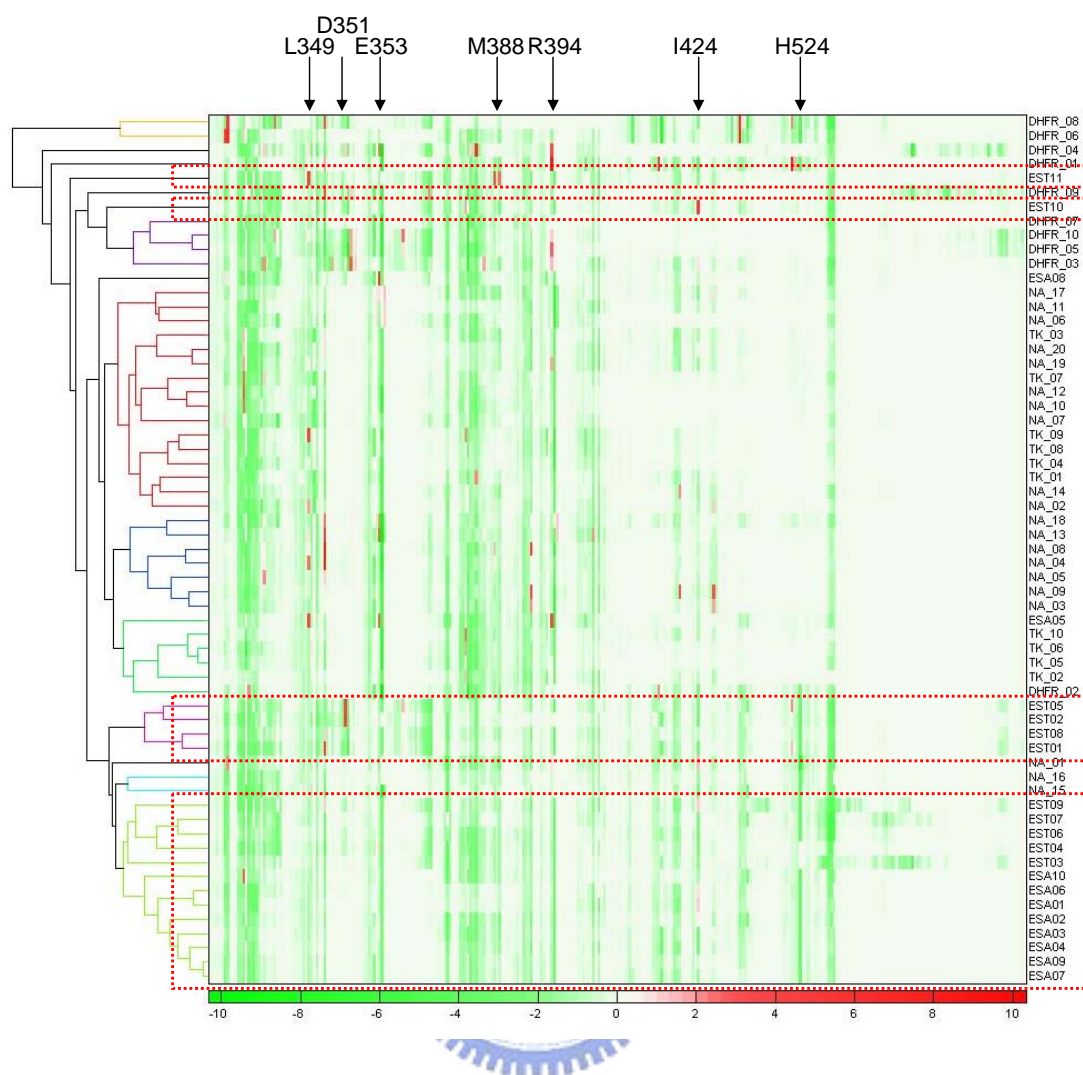


**Figure 20.** (a) Overlay of all 53 docked poses of known active compounds in the vicinity of the target protein TK (PDB id: 1kim). (b) Hierarchical clustering of protein-ligand interaction of 53 docked poses on TK (PDB id: 1kim). Each docked pose is represented as one line in the heat map in the middle of the figure, and the red being the lowest protein-ligand interaction energy and the green being the highest energy. The left side of the heat map shows the hierarchical clustering results on the TK, including the dendrogram. Docked poses in the heat map are rearranged according to the order given by hierarchical clustering marked by the black bar 'c' in the right side of the heat map. The hot spots identified from overlapping known active compounds were also shown in the top side of the heat map. (c) Overlay of docked poses of the cluster with most number of known active compounds, and shown the important hydrogen bonds between protein and ligand. (d) Overlay of docked poses of the cluster with most number of unknown compounds, and shown the important hydrogen bonds between protein and ligand. The blue frames in the heat map were the major interaction that different between cluster c and d.





**Figure 21.** Hierarchical clustering of protein-ligand interaction of 61 docked poses on NA (PDB id: 1mwe). Each docked pose is represented as one line in the heat map in the middle of the figure, and the red being the lowest protein-ligand interaction energy and the green being the highest energy. The left side of the heat map shows the hierarchical clustering results on the NA, including the dendrogram. Docked poses in the heat map are rearranged according to the order given by hierarchical clustering. The hot spots identified from overlapping known active compounds were also shown in the top side of the heat map. All the known active compounds were grouped within a cluster (frame in red) and had hydrogen-bond with target protein (R152, E277, R292, and R371).

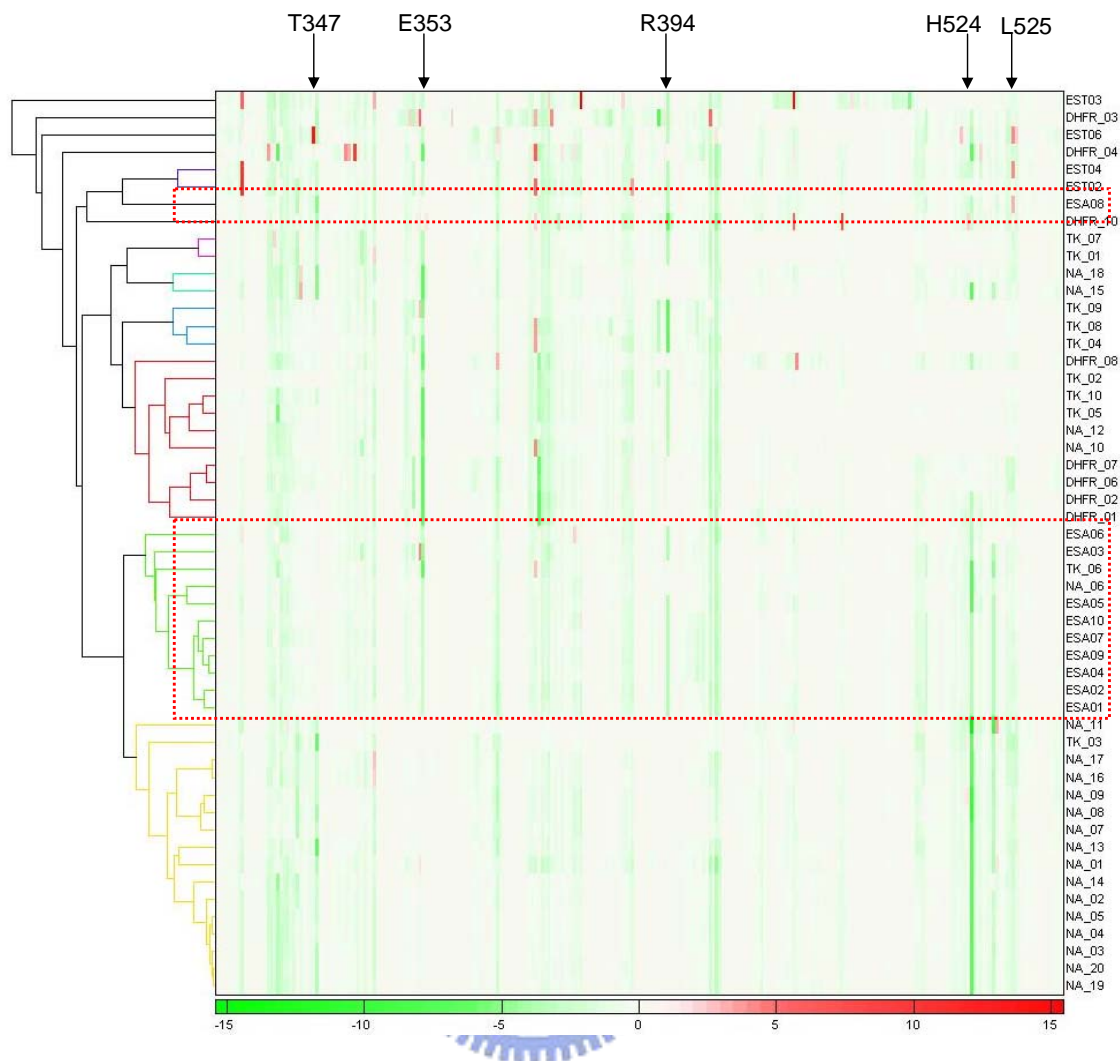


**Figure 22.** Hierarchical clustering of protein-ligand interaction of 61 docked poses on ER  $\alpha$  (PDB id: 3ert). Each docked pose is represented as one line in the heat map in the middle of the figure, and the red being the lowest protein-ligand interaction energy and the green being the highest energy. The left side of the heat map shows the hierarchical clustering results on the ER  $\alpha$ , including the dendrogram. Docked poses in the heat map are rearranged according to the order given by hierarchical clustering. The hot spots identified from overlapping known active compounds were also shown in the top side of the heat map. The active compounds were divided into four clusters by the red frames on the heat map, two were singleton, one contains 4 inhibitors, and last cluster contained 5 inhibitors and 8 ER  $\alpha$  agonists. We could inspect that the positive van der Waals force on (I424, M388, and L349) made EST11 and

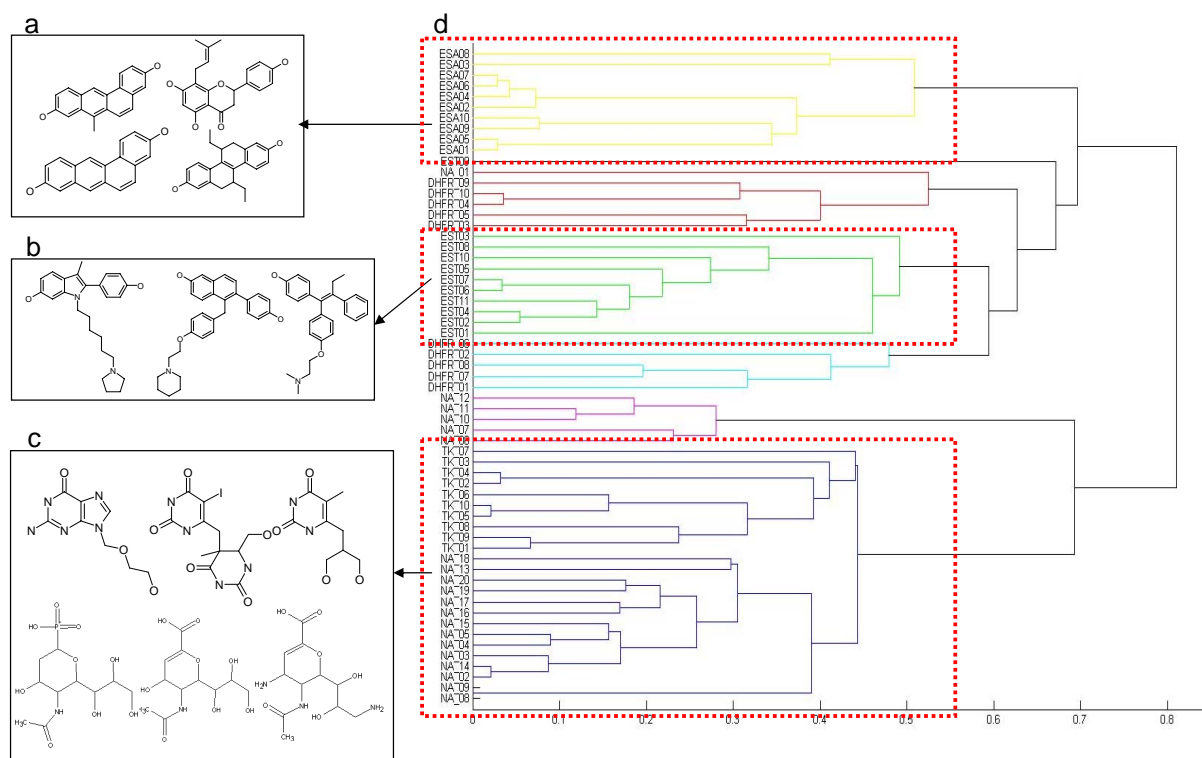


EST10 different from other inhibitors.

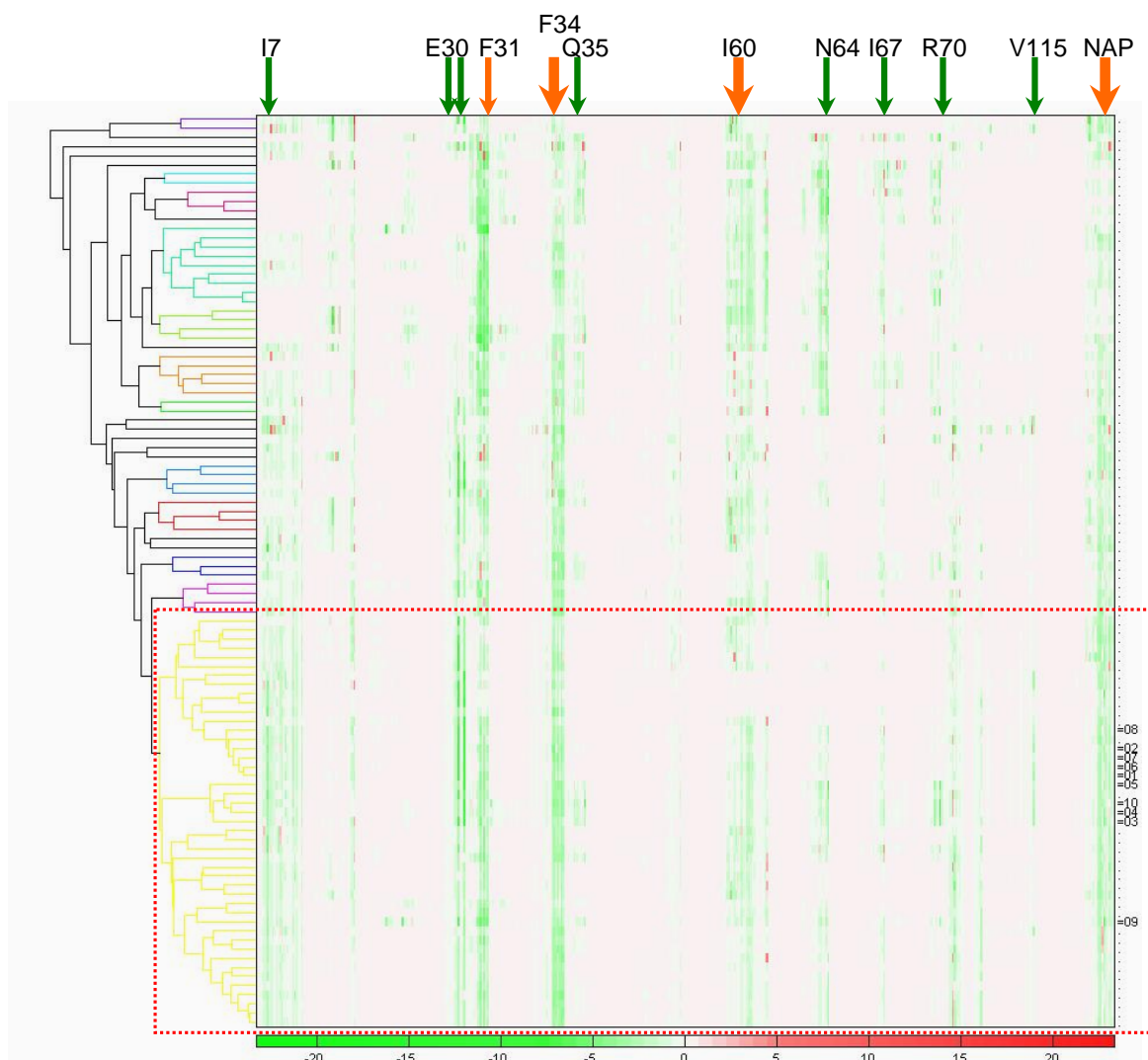




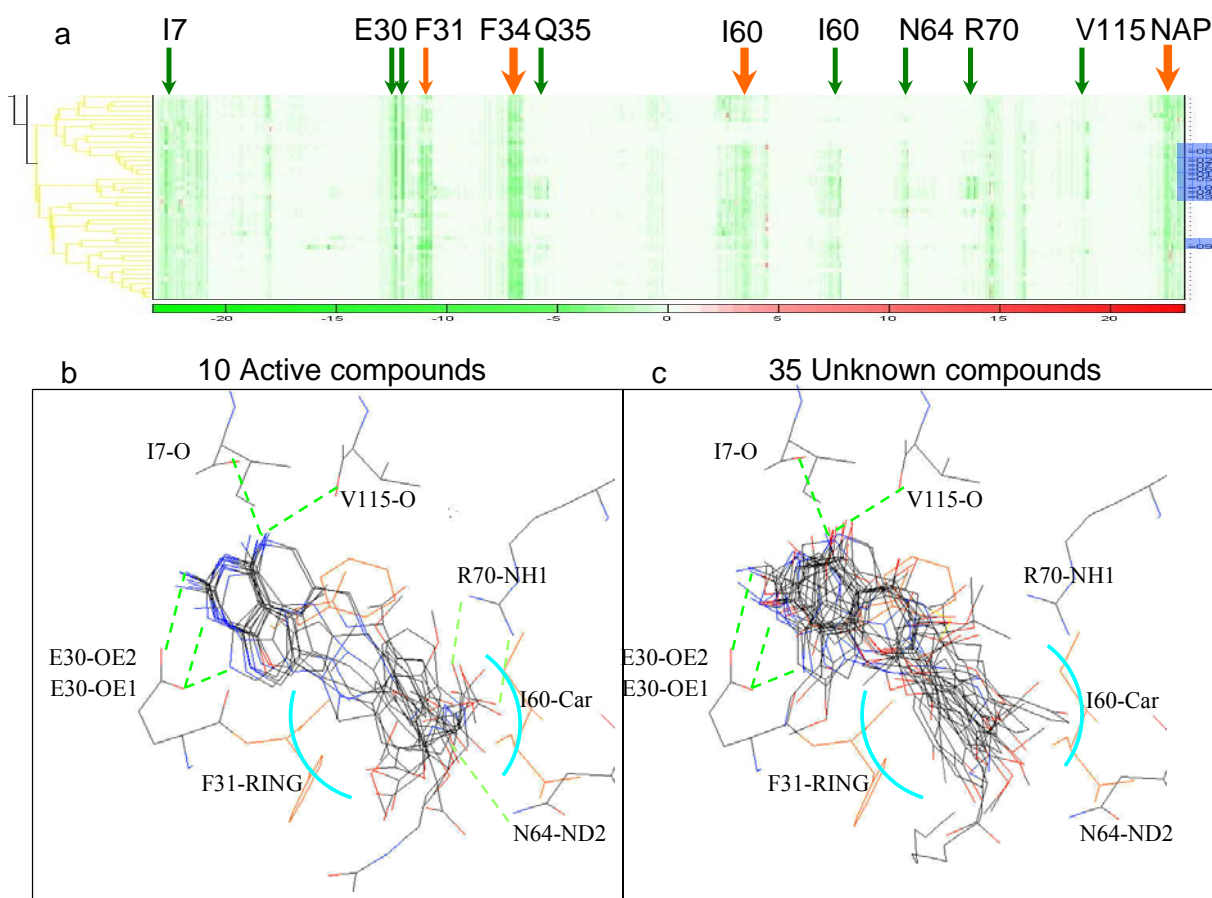
**Figure 23.** Hierarchical clustering of protein-ligand interaction of 61 docked poses on ER  $\alpha$  (PDB id: 1gwr). Each docked pose is represented as one line in the heat map in the middle of the figure, and the red being the lowest protein-ligand interaction energy and the green being the highest energy. The left side of the heat map shows the hierarchical clustering results on the ER  $\alpha$ , including the dendrogram. Docked poses in the heat map are rearranged according to the order given by hierarchical clustering. The hot spots identified from overlapping known active compounds were also shown in the top side of the heat map. All active compounds except ESA08 were grouped into one cluster, the ESA08 had additional interaction with target protein (T347 and L525).



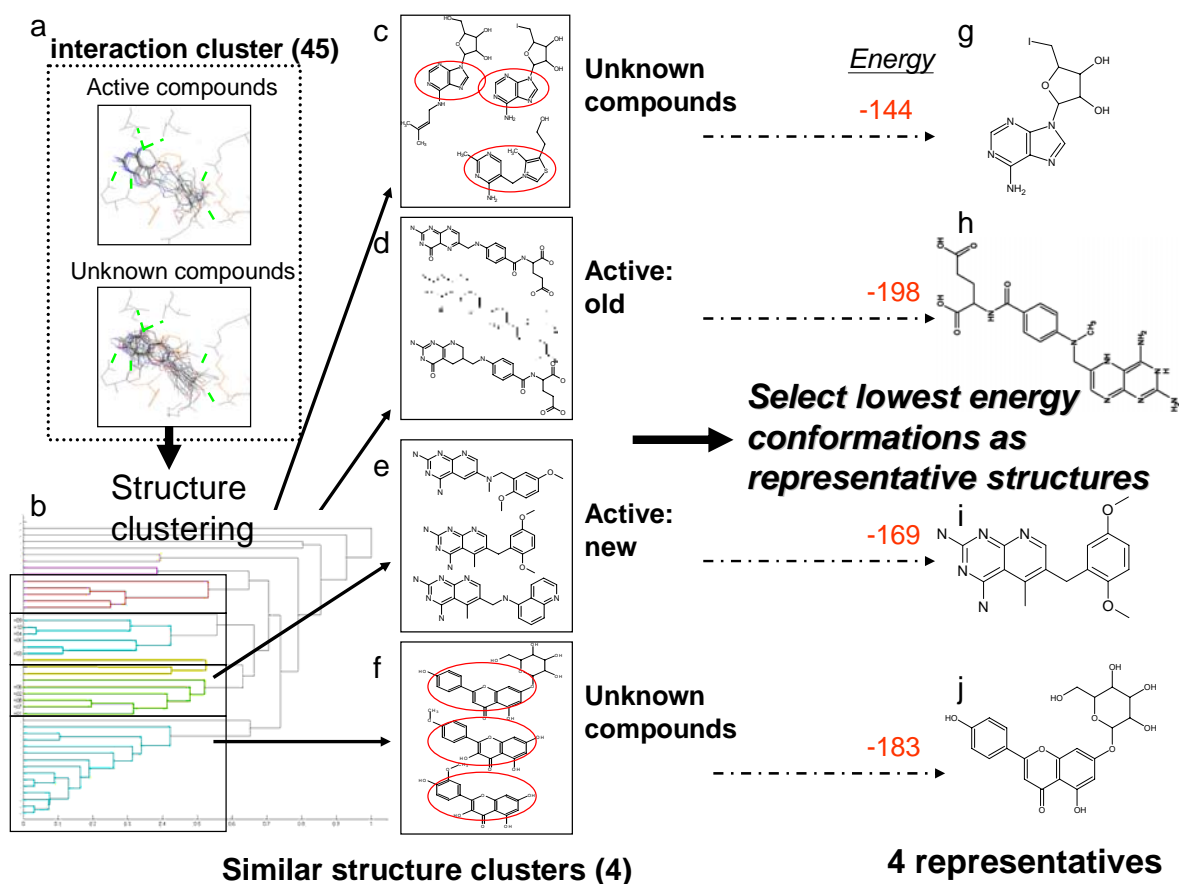
**Figure 24.** (d) The dendrogram of Hierarchical clustering of 61 known compound structures. The descriptor was calculated by atom-pair representation, using tanimoto coefficient for measuring distance between two molecules. Under the reference threshold (tanimoto coefficient = 0.55), There were three major clusters, a, b, and c. (a) In the cluster a, all 10 ER $\alpha$  agonists were grouped within the cluster. (b) In the cluster b, all 11 ER $\alpha$  antagonists were also grouped within the cluster. (c) In the cluster c, all 10 TK inhibitors and 14 NA were grouped together because the structures between TK and NA inhibitors were similar. By the observation on these three clusters, we could inspect that the atom-pair descriptor could group compounds with similar structures and divided compounds with different structures.



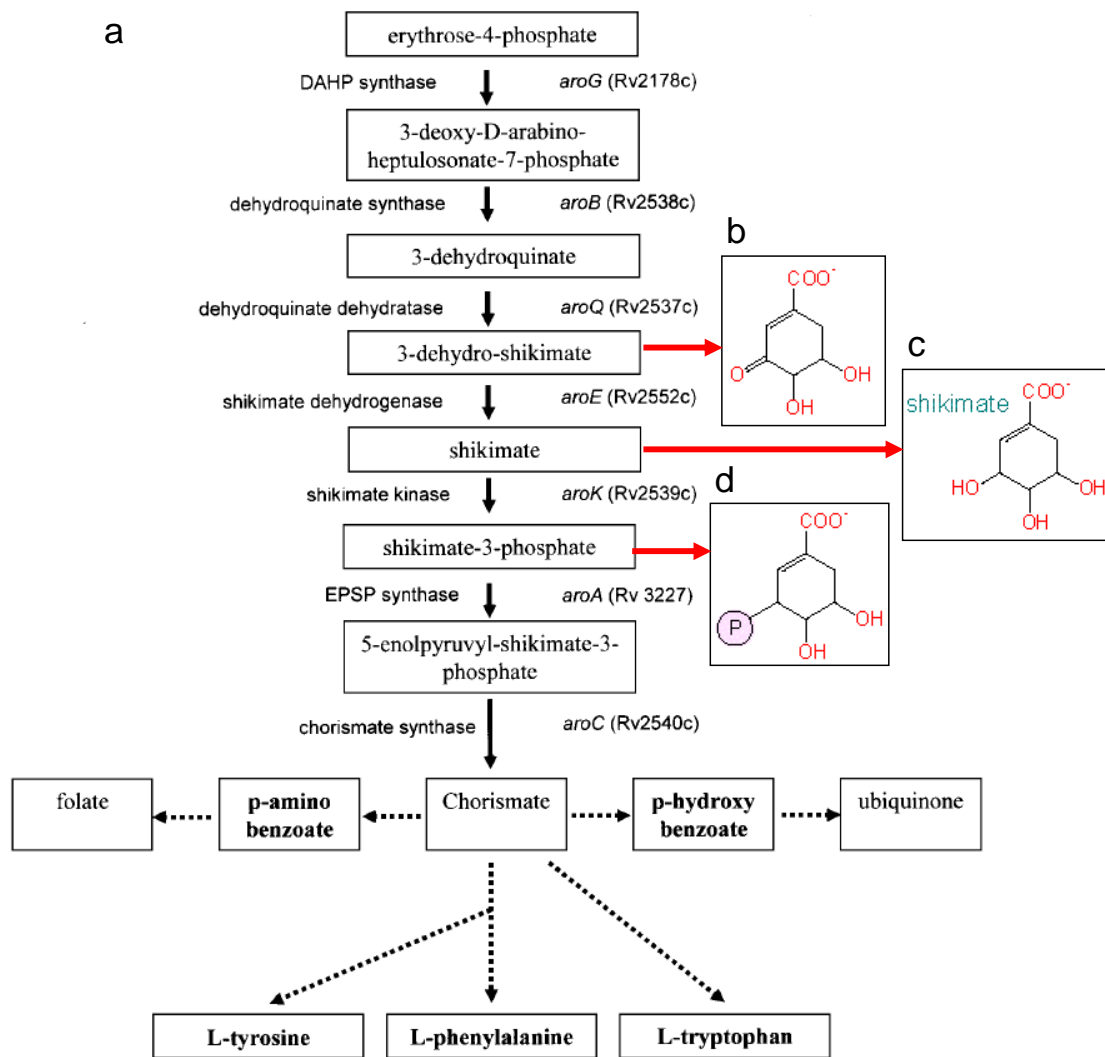
**Figure 25.** The first stage cluster analysis of hDHFR dataset. The compound set was combined 990 random selected compounds from ACD and 10 hDHFR inhibitors. The top 100 rankers of GEMDOCK were selected for clustering analysis. After the correlation coefficients was applied for measuring distance between docked poses, we performed hierarchical clustering of protein-ligand interaction on top 100 docked poses of hDHFR (PDB id: 1hfr). Each docked pose is represented as one line in the heat map in the middle of the figure, and the red being the lowest protein-ligand interaction energy and the green being the highest energy. The left side of the heat map shows the hierarchical clustering results on the hDHFR, including the dendrogram. Docked poses in the heat map are rearranged according to the order given by hierarchical clustering. Each active compounds of hDHFR was sires marked by the number '=01' in the right side of the heat map. The amino acids identified for description were also shown in the top side of the heat map. The pointers in red were ven der Waals forces. All active compounds were grouped into one cluster.



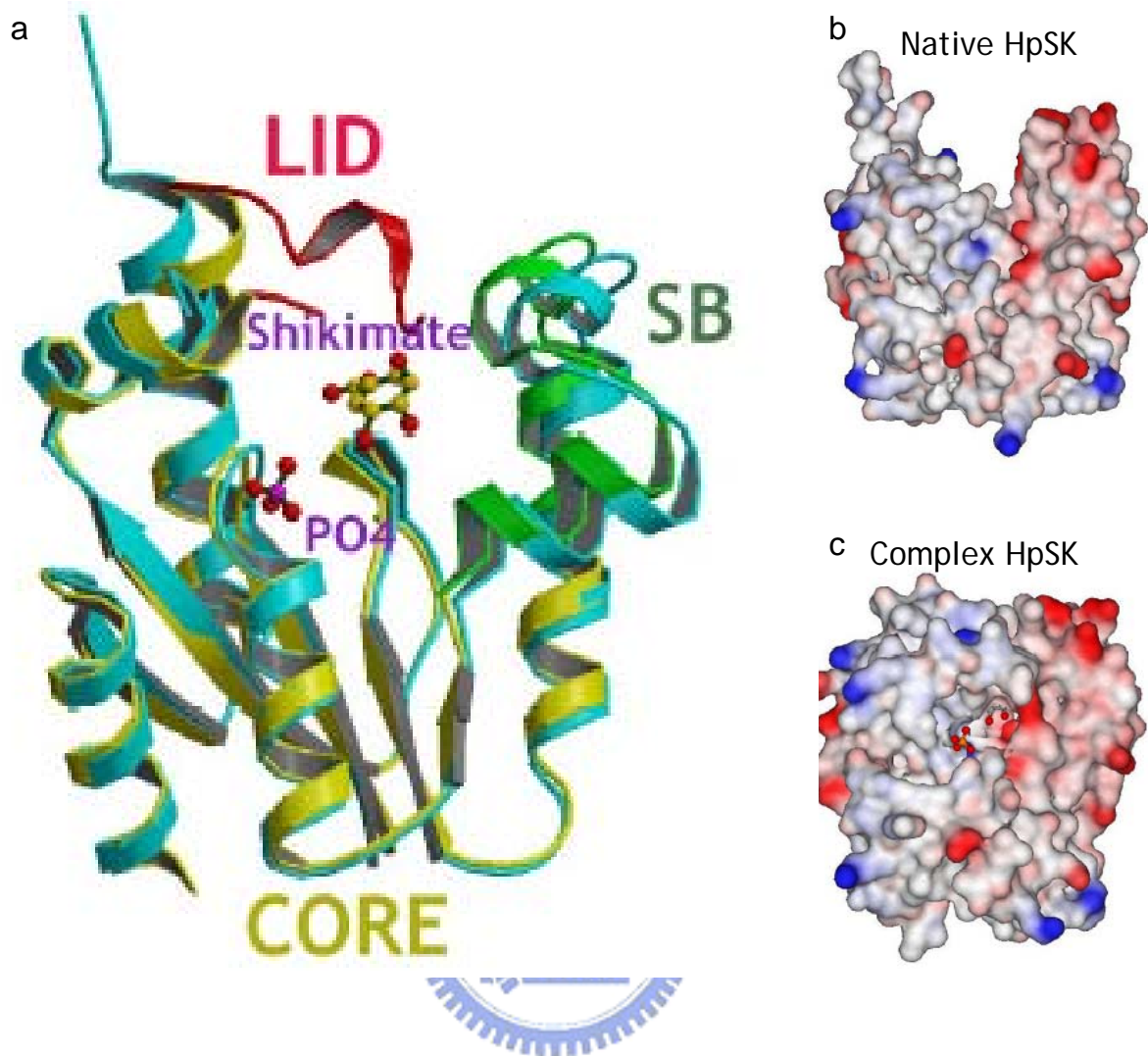
**Figure 26.** The detail of binding interactions within the largest cluster on hDHFR testing dataset. (a) The cluster contained 45 compounds include 10 active compounds and 35 unknown compounds. The pointers in red were the major ven der Waals force interactions. Each active compounds of hDHFR was sires marked by the number ‘=01’ in the right side of the heat map. The amino acids identified for description were also shown in the top side of the heat map. (b)(c)The detail binding interactions of active and unknown compounds within the cluster. The residues in yellow were major contribution of ven der Waals force. The binding interactions between active compounds and unknown compounds were similar, but the compound structures within the clusters were diverse.



**Figure 27.** The process and result of second stage cluster analysis on hDHFR testing dataset. The largest cluster generated by first stage clustering was clustered by second stage clustering to selecting representative compounds within each cluster. (a) The binding interactions of the largest cluster generated from first stage clustering, the cluster contained 45 compounds include 10 active compounds and 35 unknown compounds. (b) The result of hierarchical clustering, there were four major clusters identified by the dendrogram, (c), (d), (e), (f). The active compounds were spliced into two clusters, the old drugs (d) and the new drugs (e) because of the difference of the carboxylic acid group. The sub-structures within each cluster selected by the red circle in (c) and (f) were similar, but the sub-structures were different between each cluster. (g), (h), (i), (j), the lowest energy compound within each cluster for representing all compounds within the cluster.

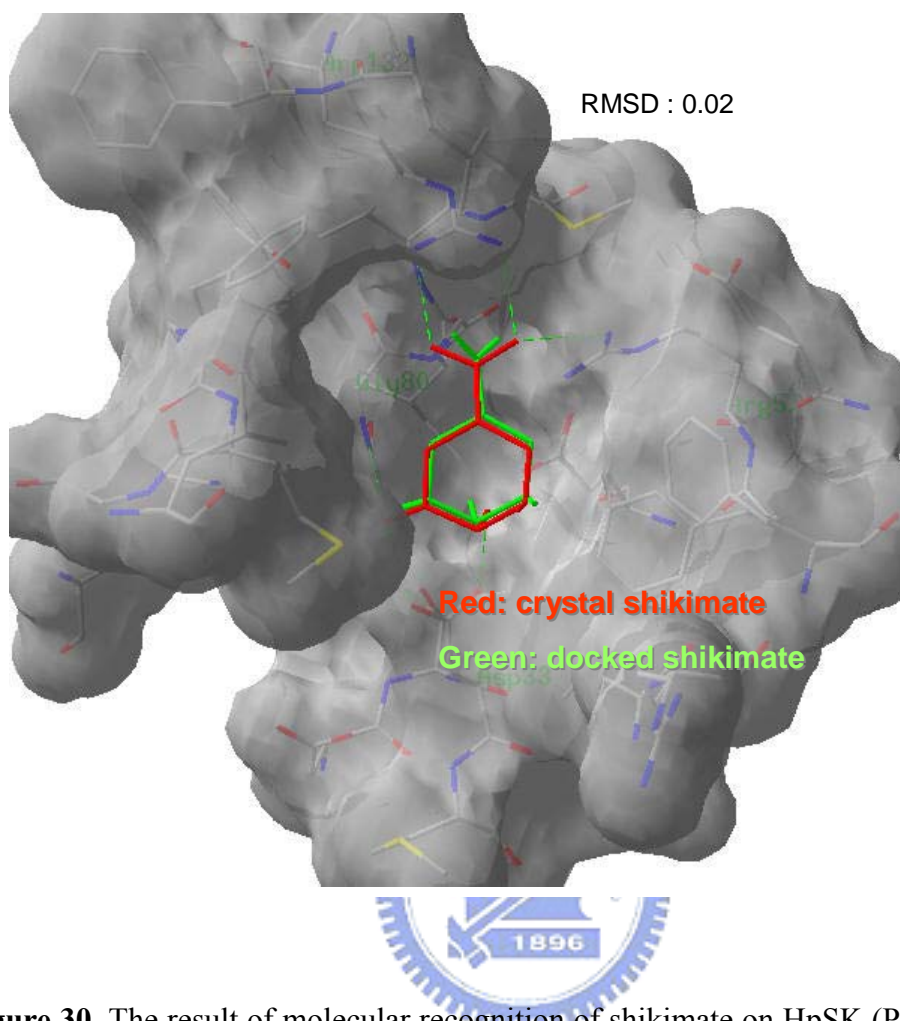


**Figure 28.** (a) The shikimate pathway. Our target protein was shikimate kinase. (b) The structure of 3-dehydro-shikimate. (c) The structure of shikimate, substrate of shikimate kinase. (d) The shikimate-3-phosphate, product of shikimate kinase.

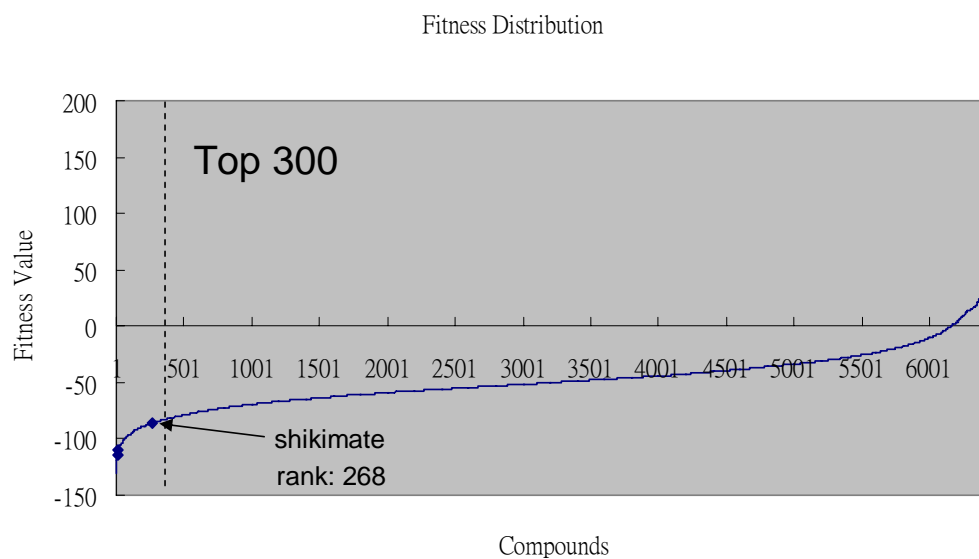


**Figure 29.** Comparing open and close form of shikimate kinase (open form PDB id: 1ZHU, close form PDB id: 1ZHI). (a) The induced-fit movement of the lid structure from an open to a closed form. (b)(c) The electrostatic surface of open and close form of shikimate kinase.



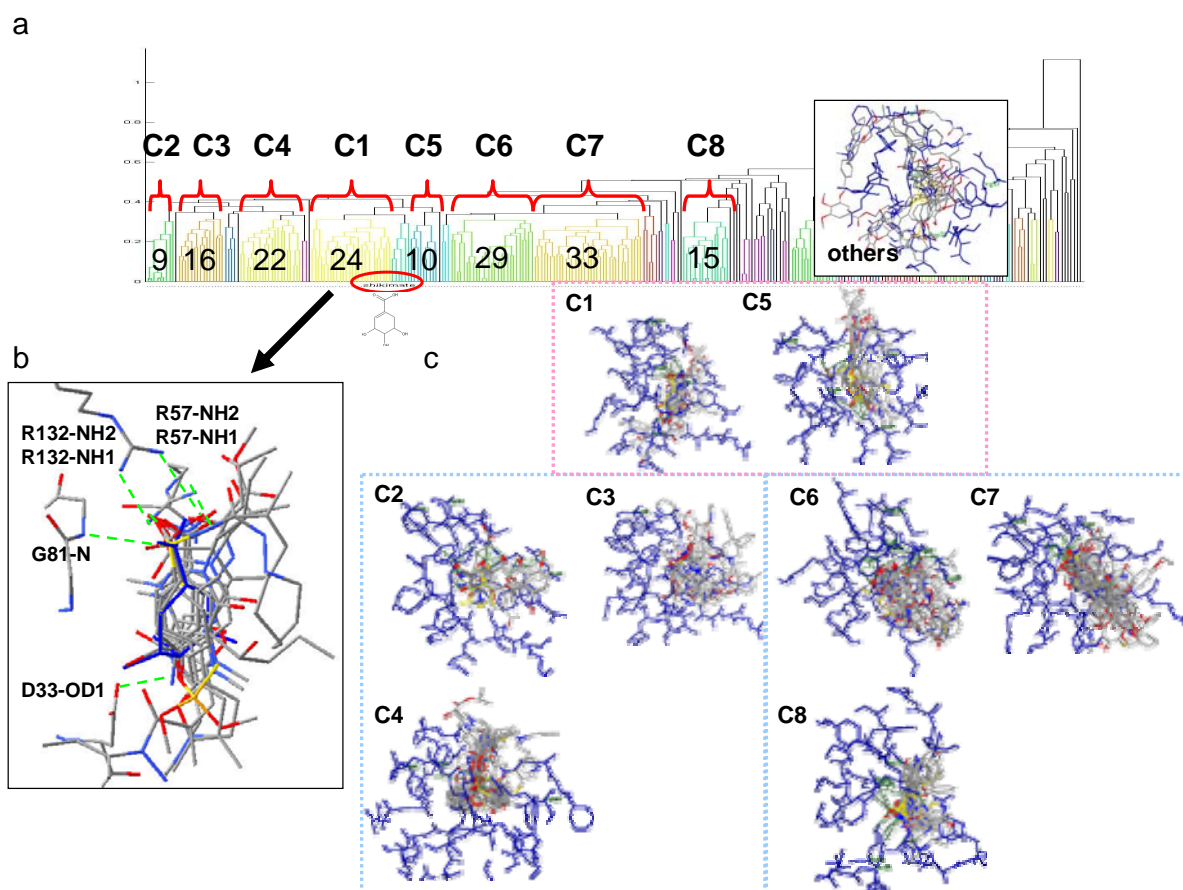


**Figure 30.** The result of molecular recognition of shikimate on HpSK (PDB id: 1zui). The lid structure (residue id: 108~124) was removed. The RMSD between crystal shikimate and docked shikimate was 0.02, indicated that after removing of the lid structure, the GEMDOCK still could dock the substrate back into the target protein correctly.



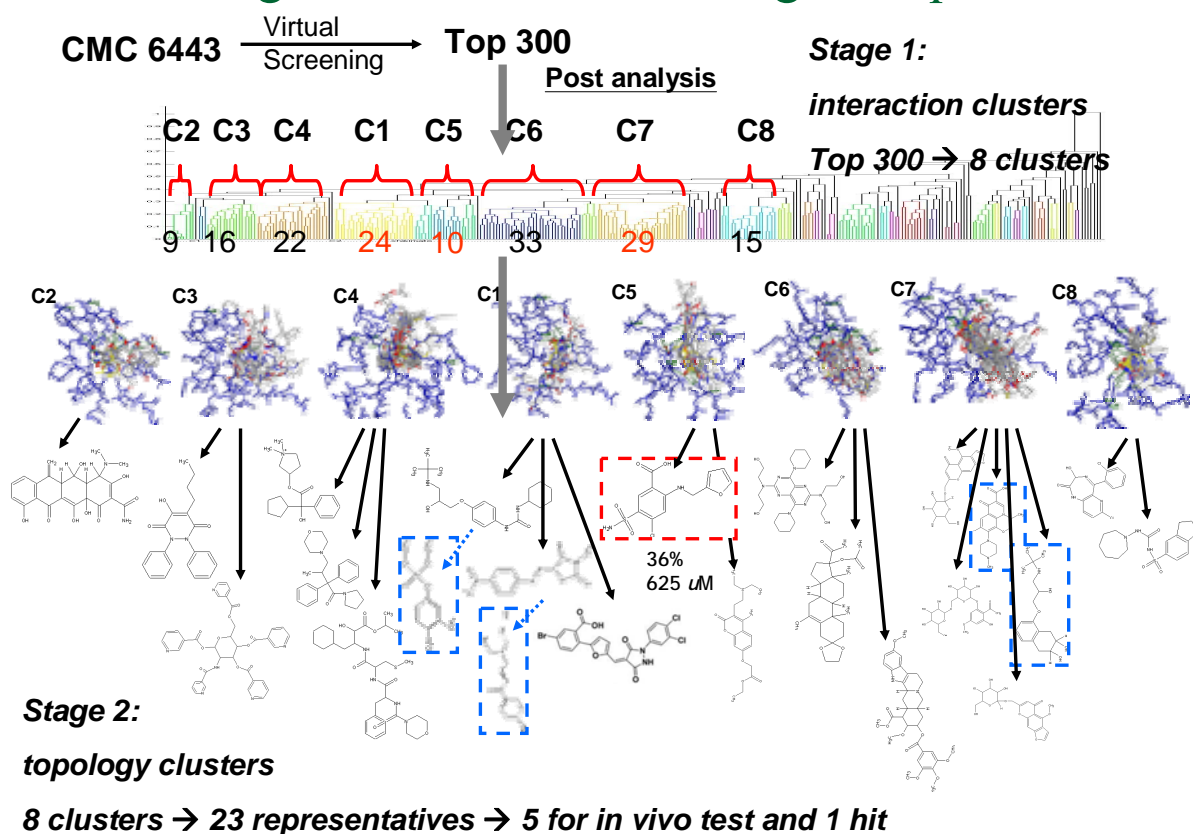
**Figure 31.** The distribution of fitness of compounds while screening on shikimate kinase. we selected top 300 rankers for performing cluster analysis, because the rank of substrate shikimate was 268.



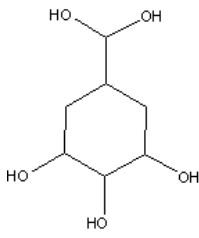
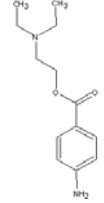
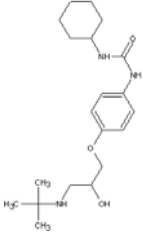
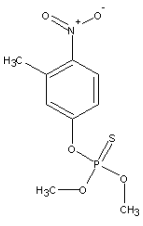
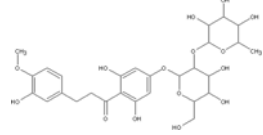
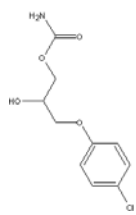
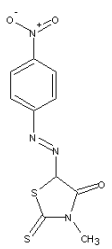
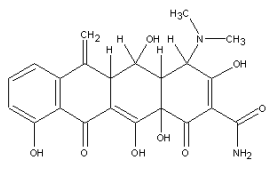


**Figure 32.** Result of clustering top 300 rankers by first stage clustering on shikimate kinase. (a) There are clusters C1-C8 and others, totally 9 clusters. (b) The binding interactions of compounds within the cluster, which were similar to the substrate. All compounds within the cluster had the hydrogen-bonds with target protein (R132-NH1, R132-NH2, R57-NH1, R57-NH2, G81-N, and D33-OD1). (c) The alignment of docked poses of each cluster. Each cluster had different distribution of conformation.

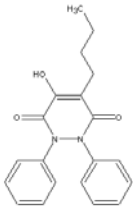
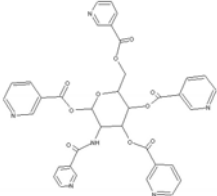
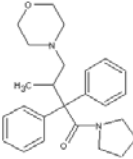
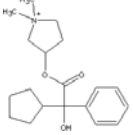
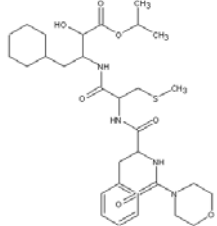
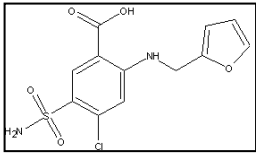
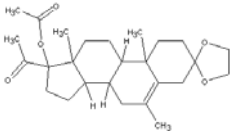
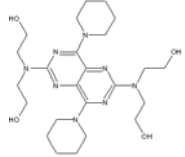
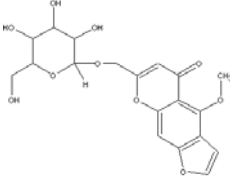
## Two-stage hierarchical clustering for HpSK



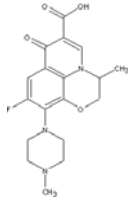
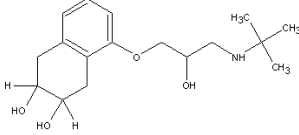
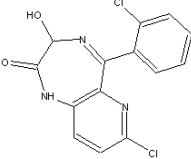
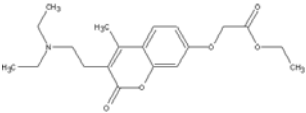
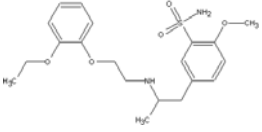
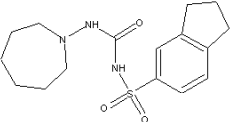
**Figure 33.** Overall process and result of two-stage clustering on shikimate kinase. First, virtual screening on CMC 6443 compounds by GEMDOCK. Second, we selected top 300 compounds for two-stage cluster analysis. Third, after clustering by protein-ligand interaction, there were 8 major clusters and others. Forth, we applied the atom-pair descriptor for clustering compounds within each cluster and 23 representative compounds were selected for use in bioassay. Fifth, 5 compounds were tested in vivo (compound structures with blue frames) and one had inhibition on shikimate kinase (compound structure with red frame).

CMC number	Rank	Fitness	Structures
Shikimate	268	-85.6673	
MCMC00000215	220	-87.8138	
MCMC00004411	24	-106.441	
MCMC00005807	103	-95.6473	
MCMC00010190	139	-92.5776	
MCMC00001504	288	-85.0355	
MCMC00001523	30	-105.279	
MCMC00001512	2	-126.82	



MCMC00003958	115	-94.7342	
MCMC00005287	10	-113.875	
MCMC00005814	50	-101.49	
MCMC00001378	55	-100.448	
MCMC00006402	159	-91.1382	
<b>MCMC00000106</b>	<b>58</b>	<b>-99.8947</b>	
mMCMC00001499	176	-90.3963	
mMCMC00001476	54	-100.682	
mMCMC00000188	1	-130.827	



mMCMC00002991	15	-111.331	
mMCMC00005346	42	-103.464	
mMCMC00003948	28	-105.808	
mMCMC00003970	78	-97.586	
mMCMC00001470	83	-97.4571	
mMCMC00001916	8	-114.836	

**Figure 34.** The structures of the 23 representative candidates on the HpSK. Five of 23 representative candidates were tested *in vivo*, and one of the five candidates, furosemide (MCMC00000106), was identified being able to inhibit shikimate kinase.

## Reference

1. Blundell, T.L., H. Jhoti, and C. Abell, *High-throughput crystallography for lead discovery in drug design*. Nature Reviews Drug Discovery, 2002. **1**(1): p. 45-54.
2. Lyne, P.D., *Structure-based virtual screening: an overview*. Drug Discovery Today, 2002. **7**(20): p. 1047-55.
3. Stahl, M. and T. Schulz-Gasch, *Practical database screening with docking tools*. Ernst Schering Res Found Workshop, 2003(42): p. 127-151.
4. Pearlman, D.A. and P.S. Charifson, *Are free energy calculations useful in practice? A comparison with rapid scoring functions for the p38 Matom-pair kinase protein system*. Journal of Medicinal Chemistry, 2001. **44**(21): p. 3417-23.
5. Deng, Z., C. Chuaqui, and J. Singh, *Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions*. Journal of Medicinal Chemistry, 2004. **47**(2): p. 337-44.
6. Kallblad, P., R.L. Mancera, and N.P. Todorov, *Assessment of multiple binding modes in ligand-protein docking*. Journal of Medicinal Chemistry, 2004. **47**(13): p. 3334-7.
7. Amari, S., et al., *VISCANA: visualized cluster analysis of protein-ligand interaction based on the ab initio fragment molecular orbital method for virtual ligand screening*. Journal of Chemical Information and Modeling, 2006. **46**(1): p. 221-30.
8. Nakano, T., et al., *Fragment molecular orbital method: use of approximate electrostatic potential*. The Journal of Chemical Physics, 2002. **351**: p. 475-480.
9. Yang, J.M. and C.C. Chen, *GEMDOCK: a generic evolutionary method for molecular docking*. Proteins, 2004. **55**(2): p. 288-304.
10. Lin, E.S. and J.M. Yang, *Modeling the binding and inhibition mechanism of nucleotide and sulfotransferase using molecular docking*. Journal of the Chinese Chemical Society, 2003. **50**: p. 655-663.
11. Yang, J.M. and T.W. Shen, *A pharmacophore-based evolutionary approach for screening selective estrogen receptor modulators*. Proteins, 2005. **59**(2): p. 205-20.
12. Yang, J.M., *Development and evaluation of a generic evolutionary method for protein-ligand docking*. Journal of Computational Chemistry, 2004. **25**(6): p. 843-57.
13. Dubes, R. and A.K. Jain, *Clustering methodologies in exploratory data analysis*. Adv. Comput., 1980(19): p. 113-228.
14. The MathWorks, I. and M. Natick, *MATLAB, version 7.0*. 2006.
15. CARHART, R.E., D.H. SMITH, and R. VENKATARAGHAVAN, *Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications*. Journal of chemical information and computer sciences, 1985(25): p. 64-13.
16. Jain, A.N., *Ligand-based structural hypotheses for virtual screening*. Journal of Medicinal Chemistry, 2004. **47**(4): p. 947-961.



17. Champness, J.N., et al., *Exploring the active site of herpes simplex virus type-1 thymidine kinase by X-ray crystallography of complexes with aciclovir and other ligands*. Proteins, 1998. **32**(3): p. 350-61.
18. Shiau, A.K., et al., *The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen*. Cell, 1998. **95**(7): p. 927-37.
19. Cody, V., et al., *Comparison of ternary crystal complexes of F31 variants of human dihydrofolate reductase with NADPH and a classical antitumor furopyrimidine*. Anti-cancer drug design., 1998. **13**(4): p. 307-15.
20. Varghese, J.N., et al., *Structural evidence for a second sialic acid binding site in avian influenza virus neuraminidases*. Proceedings of the National Academy of Sciences of the United States of America, 1997. **94**(22): p. 11808-12.
21. Cheng, W.C., Y.N. Chang, and W.C. Wang, *Structural basis for shikimate-binding specificity of Helicobacter pylori shikimate kinase*. Journal of Bacteriology, 2005. **187**(23): p. 8156-8163.
22. Yang, J.M., et al., *Consensus scoring criteria for improving enrichment in virtual screening*. Journal of Chemical Information and Modeling, 2005. **45**(4): p. 1134-46.
23. Bissantz, C., G. Folkers, and D. Rognan, *Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations*. Journal of Medicinal Chemistry, 2000. **43**(25): p. 4759-67.
24. van Lipzig, M.M., et al., *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. **47**(4): p. 1018-30.
25. Birch, L., et al., *Sensitivity of molecular docking to induced fit effects in influenza virus neuraminidase*. Journal of Computer-Aided Molecular Design, 2002. **16**(12): p. 855-69.
26. Gehlhaar, D.K., et al., *Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming*. Chemistry & Biology, 1995. **2**: p. 317-324.
27. Matter, H., *Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors*. Journal of Medicinal Chemistry, 1997. **40**(8): p. 1219-29.
28. Zheng, W. and A. Tropsha, *Novel variable selection quantitative structure--property relationship approach based on the k-nearest-neighbor principle*. Journal of chemical information and computer sciences, 2000. **40**(1): p. 185-94.
29. Griffiths, P.D., *Progress in the clinical management of herpesvirus infections*. Antiviral Chemistry & Chemotherapy, 1995. **6**: p. 191-209.
30. Darby, G.K., *In search of the perfect antiviral*. Antiviral chemistry & chemotherapy, 1995. **6**: p. 54-63.

31. Gust, R., R. Keilitz, and K. Schmidt, *Synthesis, structural evaluation, and estrogen receptor interaction of 2,3-diarylpiperazines*. Journal of Medicinal Chemistry, 2002. **45**: p. 2325-2337.
32. Renaud, J., et al., *Estrogen receptor modulators: identification and structure-activity relationships of potent ERalpha-selective tetrahydroisoquinoline ligands*. Journal of Medicinal Chemistry, 2003. **46**: p. 2945-2957.
33. Sato, M., et al., *Emerging therapies for the prevention or treatment of postmenopausal osteoporosis*. Journal of Medicinal Chemistry, 1999. **42**: p. 1-24.
34. Torgerson, D.J., *HRT and its impact on the menopause, osteoporosis and breast cancer*. Expert Opinion on Pharmacotherapy, 2000. **1**: p. 1163-1169.
35. Miller, C.P., *SERMs: evolutionary chemistry, revolutionary biology*. Current Pharmaceutical Design, 2002. **8**: p. 2089-2111.
36. Dutertre, M. and C.L. Smith, *Molecular mechanisms of selective estrogen receptor modulator (SERM) action*. The Journal of Pharmacology and Experimental Therapeutics, 2000. **295**: p. 431-437.
37. Maricic, M. and O. Gluck, *Review of raloxifene and its clinical applications in osteoporosis*. Expert Opinion on Pharmacotherapy, 2002. **3**: p. 767-775.
38. MacGregor, J.I. and V.C. Jordan, *Basic guide to the mechanisms of antiestrogen action*. Pharmacological Reviews, 1998. **50**: p. 151-196.
39. Elion, G.B., *Nobel lecture in physiology or medicine--1988. The purine path to chemotherapy*. In Vitro Cellular & Developmental Biology, 1989. **25**: p. 321-330.
40. Hitchings, G.H., Jr., *Nobel lecture in physiology or medicine--1988. Selective inhibitors of dihydrofolate reductase*. In Vitro Cellular & Developmental Biology, 1989. **25**: p. 303-310.
41. Wyss, P.C., et al., *Novel dihydrofolate reductase inhibitors. Structure-based versus diversity-based library design and high-throughput synthesis and screening*. Journal of Medicinal Chemistry, 2003. **46**: p. 2304-2312.
42. Rastelli, G., et al., *Docking and database screening reveal new classes of Plasmodium falciparum dihydrofolate reductase inhibitors*. Journal of Computational Chemistry, 2003. **46**: p. 2834-2845.
43. Palese, P., K. Tobita, and M. Ueda, *Virology*. 1974. **61**: p. 397.
44. Verma, R.P. and C. Hansch, *A QSAR study on influenza neuraminidase inhibitors*. Bioorganic & Medicinal Chemistry, 2006. **14**(4): p. 982-96.
  
45. Cheng, W. C., Y. N. Chang, and W. C. Wang, *Structural Basis for shikimate-binding specificity of Helicobacter pylori shikimate kinase*. J. Bacteriol, 2005. **187**:p. 8156-8163.
46. Coggins, J. R., C. Abell, L. B. Evans, M. Frederickson, D. A. Robinson, A. W. Roszak, and A. P. Laphorn. *Experiences with the shikimate-pathway enzymes as targets for*

- rational drug design*. Biochem. Soc. Trans. 2003. **31**:p. 548-52.
47. Herrmann, K. M., and L. M. Weaver. *The Shikimate Pathway*. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 1999. **50**:p. 473-503.
  48. McConkey, G. A. Targeting the shikimate pathway in the malaria parasite *Plasmodium falciparum*. *Antimicrob. Agents Chemother.* 1999. **43**:p. 175-177.
  49. Megraud, F., and H. Lamouliatte. *Review article: the treatment of refractory Helicobacter pylori infection*. *Aliment. Pharmacol. Ther.* 2003. **17**:p. 1333-1343.
  50. Parsonnet, J. *Helicobacter pylori and gastric cancer*. *Gastroenterol. Clin. North Am.* 1993. **22**:p. 89-104.
  51. Ridley, R. G. *Planting new targets for antiparasitic drugs*. *Nat. Med.* 1998. **4**:p. 894-895.
  52. Roberts, F., C. W. Roberts, J. J. Johnson, D. E. Kyle, T. Krell, J. R. Coggins, G. H. Coombs, W. K. Milhous, S. Tzipori, D. J. Ferguson, D. Chakrabarti, and R. McLeod. *Evidence for the shikimate pathway in apicomplexan parasites*. *Nature* 1998. **393**:p. 801-805.
  53. Steinrucken, H. C., and N. Amrhein. *The herbicide glyphosate is a potent inhibitor of 5-enolpyruvyl-shikimic acid-3-phosphate synthase*. *Biochem. Biophys. Res. Commun.* 1980. **94**:p. 1207-1212.

