An Evolutionary Approach with Pharmacophore-Based Scoring Functions for Virtual Database Screening

Jinn-Moon Yang, Tsai-Wei Shen, Yen-Fu Chen, and Yi-Yuan Chiu

Department of Biological Science and Technology & Institute of Bioinformatics
National Chiao Tung University, Hsinchu, 30050, Taiwan
moon@cc.nctu.edu.tw

Abstract. We have developed a new tool for virtual database screening. This tool, referred to as the Generic Evolutionary Method for molecular DOCKing (GEMDOCK), combines an evolutionary approach and a new pharmacophore-based scoring function. The former integrates discrete and continuous global search strategies with local search strategies to speed up convergence. The latter simultaneously serves as the scoring function of both molecular docking and post-docking analysis to improve the number of the true positives. We accessed the accuracy of our approach on HSV-1 thymidine kinase using a ligand database on which competing tools were evaluated. The accuracies of our predictions were 0.54 for the GH score and 1.62% for the false positive rate when the true positive rate was 100%. We found that our pharmacophore-based scoring function indeed is able to reduce the number of the false positives. These results suggest that GEMDOCK is robust and can be a useful tool for virtual database screening.

1 Introduction

Virtual screening of compound databases has emerged as one of the most powerful and inexpensive approach to discover novel rational lead compounds for drug development [1,2]. It is based on high-throughput molecular docking methods and the crystal structures of the target protein. Virtual screening is increasingly used for a number of drivers: explosions of high-resolution crystal protein structures, advent of the structural proteomics technologies, enriching the hit rate of high-throughput screening [2], and reducing cost of drug discover. Virtual screening encompasses four phases, including target protein modeling, compound database preparation, molecular docking, and post-docking analysis. In general, a computational method for virtual screening involves two basic critical elements that are molecular docking and a good scoring method.

A molecular docking method for virtual screening should be able to screen a large number of potential ligands with reasonable accuracy and speed. Many molecular docking approaches have been developed and can be roughly divided into rigid docking [3], flexible ligand docking [4,5], and protein flexible docking methods. Recently, the flexible docking tools were mostly used for virtual screening, such as incremental and fragment-based approaches (DOCK [6] and FlexX [5]) and genetic algorithms (GOLD [4], AutoDock [7], and GEMDOCK [8]).

Scoring methods for virtual screening should encompass two basic features: effectively discriminating between correct binding states and non-native docked conforma-

K. Deb et al. (Eds.): GECCO 2004, LNCS 3102, pp. 481–492, 2004.

tions during the molecular docking phase, and discriminating a small number of active compounds from hundreds of thousands of non-active compounds during the post-docking analysis phase. Various scoring functions have been developed for calculating binding free energy, including knowledge-based [9], physic-based [10], and empirical [11] scoring functions. In general the performance of these scoring functions is often inconsistent across different systems [12,13]. The inaccuracy, inadequately predicting the true binding affinity of a ligand for a receptor, of the scoring methods is probably major weakness for virtual screening. Combining multiple scoring functions, called consensus scoring, is a popular strategy and has been shown to improve the enrichment of true positive [12,13].

In this paper, we proposed a tool, GEMDOCK (Generic Evolutionary Method for DOCKing molecules) modified from our previous studies [8,14], for virtual screening. Our tool used a pharmacophore-based scoring function and an evolutionary approach. The former is able to simultaneously serve as the scoring function of both molecular docking and post-docking analysis. In order to balance exploration and exploitation, the core idea of our evolutionary approach, an efficient flexible docking tool, is to design multiple operators cooperating with each other by using the family competition which is similar to a local search procedure.

Our new pharmacophore-based scoring function is able to reduce the number of false positives for screening large database. This scoring function integrates a simple empirical scoring function and a pharmacophore-based scoring function. The former is used to quickly recognize potential ligands for the target receptor. It consists of electrostatic, steric, and hydrogen-bonding potentials with a linear model. The latter encompasses ligand preferences and the pharmacophore preferences that exploit knowledge from existing ligands to aid the docking process. The electrostatic and hydrophilic constrains were considered for ligand preferences. The pharmacophore-based preferences were assigned according to the binding-site preferences of protein-ligand interactions, such as hydrogen bonding and stacking force.

To evaluate the strengths and limitations of GEMDOCK and to compare with several widely used methods (e.g. DOCK, GOLD, and FlexX), we first tested our program on docking 10 active ligands, obtained from Protein Data Bank (PDB), back the respective complexes with experimentally x-ray structures. Second, we tested GEMDOCK on HSV-1 thymidine kinase, proposed by Bissabtz et al. [12], to evaluate GEMDOCK's screening utility. The docking accuracy of GEMDOCK was comparable with the best available methods and the screening performance of GEMDOCK was better than that of competing methods on these test cases.

2 Method

GEMDOCK is a nearly automatic tool, which was enhanced and modified from our original technique [8,15], for virtual screening. GEMDOCK consists of four computational phases, including target protein and ligand database preparation, molecular docking and post-docking analysis. First we specified the coordinates of target protein atoms from the PDB, the ligand binding area, atom formal charge, and atom types (Table 1). When we prepared the target protein and ligand database, GEMDOCK filters out some impossible

compounds and pharmacological preferences by exploiting knowledge from existing ligands to improve screening speed. After GEMDOCK prepares the ligand database and the target protein, GEMDOCK sequentially reads the atom coordinates of a ligand from the database and executes flexible docking for each ligand. Finally GEMDOCK re-ranks all docked ligand conformations for the post-docking analysis according to the scoring values of our pharmacophore-based scoring function.

Here, we briefly presented our approach for flexible docking. Please refer our previous studies [8,16] for the details. First our method randomly generates a starting population with N solutions by initializing the orientation and conformation of the ligand relating to the center of the receptor. Each solution is represented as a set of three n-dimensional vectors (x^i, σ^i, ψ^i), where n is the number of adjustable variables of a docking system and $i = 1, \ldots, N$ where N is the population size. The vector x represents the adjustable variables to be optimized in which x_1, x_2, and x_3 are the 3-dimensional location of the ligand; x_4, x_5, and x_6 are the rotational angles; and from x_7 to x_n are the twisting angles of the rotatable bonds inside the ligand. σ and ψ are the step-size vectors of decreasing-based Gaussian mutation and self-adaptive Cauchy mutation. In other words, each solution x is associated with some parameters for step-size control. The initial values of x_1, x_2, and x_3 are randomly chosen from the feasible box, and the others, from x_4 to x_n, are randomly chosen from 0 to 2π in radians. The initial step sizes σ is 0.8 and ψ is 0.2. After GEMDOCK initializes the solutions, it enters the main evolutionary loop which consists of two stages in every iteration: decreasing-based Gaussian mutation and self-adaptive Cauchy mutation. Each stage is realized by generating a new quasi-population (with N solutions) as the parent of the next stage. These stages apply a general procedure "FC_adaptive" with only different working population and the mutation operator.

The FC_adaptive procedure employs two parameters, namely, the working population (P, with N solutions) and mutation operator (M), to generate a new quasi-population. The main work of FC_adaptive is to produce offspring and then conduct the family competition. Each individual in the population sequentially becomes the "family father." With a probability p_c, this family father and another solution that is randomly chosen from the rest of the parent population are used as parents for a recombination operation. Then the new offspring or the family father (if the recombination is not conducted) is operated by the rotamer mutation or by differential evolution to generate a quasi offspring. Finally, the working mutation operates on the quasi offspring to generate a new offspring. For each family father, such a procedure is repeated L times called the family competition length. Among these L offspring and the family father, only the one with the lowest scoring function value survives. Since we create L children from one "family father" and perform a selection, this is a family competition strategy. This method avoids the population prematureness but also keeps the spirit of local searches. Finally, the FC_adaptive procedure generates N solutions because it forces each solution of the working population to have one final offspring.

2.1 Recombination Operators

GEMDOCK implemented modified discrete recombination and intermediate recombination [17]. A recombination operator selected the "family father (a)" and another

solution (b) randomly selected from the working population. The former generates a child as follows:

$$x_j^c = \begin{cases} x_j^a \text{ with probability } 0.8 \\ x_j^b \text{ with probability } 0.2. \end{cases} \tag{1}$$

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

$$w_j^c = w_j^a + \beta(w_j^b - w_j^a)/2, \tag{2}$$

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

2.2 Mutation Operators

After the recombination, a mutation operator, the main operator of GEMDOCK, is applied to mutate adjustable variables (x).

Gaussian and Cauchy Mutations: Gaussian and Cauchy Mutations are accomplished by first mutating the step size (w) and then mutating the adjustable variable x:

$$w_j' = w_j' A(\cdot), \tag{3}$$
$$x_j' = x_j + w_j' D(\cdot), \tag{4}$$

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

$$\psi_j^c = \psi_j^a \exp[\tau' N(0,1) + \tau N_j(0,1)], \tag{5}$$
$$x_j^c = x_j^a + \psi_j^c C_j(t). \tag{6}$$

We set τ and τ' to $(\sqrt{2n})^{-1}$ and $(\sqrt{2\sqrt{n}})^{-1}$, respectively, according to the suggestion of evolution strategies [17]. A random variable is said to have the Cauchy distribution ($C(t)$) if it has the density function: $f(y;t) = \frac{t/\pi}{t^2+y^2}$, $-\infty < y < \infty$. In this paper t is set to 1. Our decreasing-based Gaussian mutation uses the step-size vector σ with a fixed decreasing rate $\gamma = 0.95$ and works as

$$\sigma^c = \gamma \sigma^a, \tag{7}$$
$$x_j^c = x_j^a + \sigma^c N_j(0,1). \tag{8}$$

Table 1. Atom types of GEMDOCK

Atom type	Heavy atom name
Donor	primary and secondary amines, sulfur, and metal atoms
Acceptor	oxygen and nitrogen with no bound hydrogen
Both	structural water and hydroxy1 groups
Nonpolar	other atoms (such as carbon and phosphorus)

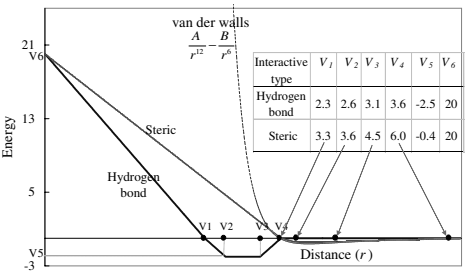


Fig. 1. The linear energy function of the pair-wise atoms for the steric interactions and hydrogen bonds in GEMDOCK (bold line) with a standard Lennard-Jones potential (light line).

2.3 Scoring Function

In this work, we have developed a new scoring function which was able to simultaneously serve as the scoring function of both molecular docking and post-docking analysis. It consisted of a simple empirical scoring function and a pharmacophore-based scoring function to reduce the number of false positives. The energy function can be dissected into the following terms:

$$E_{tot} = E_{bind} + E_{phama} + E_{ligpre}, \tag{9}$$

where E_{bind} is the empirical binding energy used during the molecular docking ; E_{phama} is the energy of binding-site pharmacophores; E_{ligpre} is a penalty value if the ligand unsatisfied the ligand preferences. E_{phama} and E_{ligpre} were used to improve the number of true positives by discriminating active compounds from hundreds of thousands of non-active compounds. The empirical binding energy (E_{bing}) is given as

$$E_{bind} = E_{inter} + E_{intra} + E_{penal}, \tag{10}$$

where E_{inter} and E_{intra} are the intermolecular and intramolecular energy, respectively, and E_{penal} is a large penalty value if the ligand is out of range of the search box. In this paper, E_{penal} is set to 10000. The intermolecular energy is defined as

$$E_{inter} = \sum_{i=1}^{lig} \sum_{j=1}^{pro} \left[F(r_{ij}^{B_{ij}}) + 332.0 \frac{q_i q_j}{4 r_{ij}} \right], \tag{11}$$

$r_{ij}^{B_{ij}}$ is the distance between the atoms i and j with the interaction type B_{ij} forming by the pair-wise heavy atoms between ligands and proteins; B_{ij} is either a hydrogen bond or a steric state; q_i and q_j are the formal charges and 332.0 is a factor that converts the electrostatic energy into kilocalories per mole. The 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 1) modified from previous works [8,11]. In this atomic pair-wise model, the interactive types are only hydrogen binding and steric potential which have the same function form but with different parameters, $V_1, \ldots, V_6$ (defined in Figure 1). The energy value of hydrogen binding should be larger than the one of steric potential. In this model, the atom is divided into four different atom types (Table 1) : donor, acceptor, both, and nonplar. The hydrogen binding can be formed by the following pair atom types: donor-acceptor, donor-both, acceptor-both, and both-both. Other pair-atom combinations form the steric state.

The intramolecular energy of a ligand is

$$E_{intra} = \sum_{i=1}^{lig} \sum_{j=i+2}^{lig} F(r_{ij}^{B_{ij}}) + \sum_{k=1}^{dihed} A[1 - \cos(m\theta_k - \theta_0)], \tag{12}$$

where $F(r_{ij}^{B_{ij}})$ is defined as Equation 11 except the value is set to 1000 when $r_{ij}^{B_{ij}} <$ 2.0 Å and $dihed$ is the number of rotatable bonds. We followed the work of Gehlhaar et al. (1995) to set the values of A, m, and θ_0. For the $sp^3 - sp^3$ bond A, m, and θ_0 are set to 3.0, 3, and π; and $A = 1.5$, $m = 6$, and $\theta_0 = 0$ for the $sp^3 - sp^2$ bond.

The pharmacophore-based interaction (E_{phama}) between the ligand and the protein is calculated by summing up all hot-spot atoms:

$$E_{phama} = \sum_{i=1}^{lig} \sum_{j=1}^{hs} f(w_j, B_{ij}) F(r_{ij}^{B_{ij}}), \tag{13}$$

where w_j is the pharmacophore weight of the hot-spot atom j, $F(r_{ij}^{B_{ij}})$ is defined as Equation 11, lig is number of the heavy atoms in the ligand, and hs is the number of hot-spot atoms in the receptor. The value of $f(w_j, B_{ij})$ is w_j or 0. $f(w_j, B_{ij})$ is w_j if the interaction type (B_{ij}) equals to the type of hot spots found between the target receptor and ligands.

In this paper the ligand preferences include electrostatic (i.e., the number of electrostatic atoms) and hydrophilic characteristic (i.e., the atom numbers of hydrogen donor and acceptor). The E_{ligpre} is a penalty value for a ligand which is unable to satisfy the ligand preferences and is defined as

$$E_{ligpre} = WP_{elec} + WP_{hb} \tag{14}$$

where WP_{elec} and WP_{hb} are the penalties for the electrostatic and hydrophilic preferences, respectively. In this paper WP_{elec} and WP_{hb} are set to 20.

3 Results

3.1 Parameters of GEMDOCK

Table 2 indicates the setting of GEMDOCK parameters, such as initial step sizes, family competition length ($L = 2$), population size ($N = 200$), and recombination probability ($p_c = 0.3$) in this work. The GEMDOCK optimization stops when either the convergence is below certain threshold value or the iterations exceed a maximal preset value which was set to 60. Therefore, GEMDOCK generated 800 solutions in one generation and terminated after it exhausted 48000 solutions in the worse case. These parameters were decided after experiments conducted to recognize complexes of test docking systems with various values. On average, GEMDOCK took 135 seconds for a docking run on a Pentium 1.4 GHz personal computer with a single processor.

Table 2. Parameters of GEMDOCK

Parameter	Value of parameters
Initial step sizes	$\sigma = 0.8$, $v = \psi = 0.2$ (in radius)
Family competition length	L = 2
Population size	$N = 200$
Recombination rate	p_c =0.3
# of the maximum generation	60

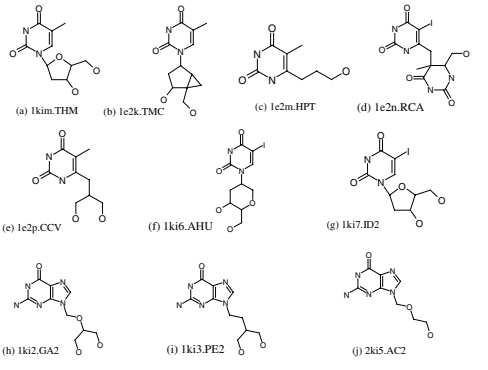


Fig. 2. Ten HSV-1 thymidine kinase ligands used as active compounds in evaluating docking accuracy and in screening performance. Each ligand systematically using four characters followed by three characters. For example, in the ligand "1kim.THM", "1kim" denotes the PDB code and "THM" is the ligand name in the PDB.

3.2 Target and Database Preparations

In order to evaluate GEMDOCK and to compare GEMDOCK with several widely used methods, we tested GEMDOCK on docking 10 active ligands (Figure 2) of HSV-1 thymidine kinase [18] back the complexes with experimentally x-ray structures from PDB. Each ligand systematically using four characters followed by three characters. For example, in the ligand "1kim.THM", "1kim" denotes the PDB code and "THM" is the ligand name in the PDB. When we evaluated the accuracy of GEMDOCK for molecular docking, the crystal coordinates of the ligand and protein atoms were taken from PDB, and were separated into different files. Our program then assigned the atom formal charge and atom type (i.e., donor, acceptor, both, or nonpolar) for each atom of both the ligand and protein. The bond type ($sp^3 - sp^3$, $sp^3 - sp^2$, or others) of a rotatable bond inside a ligand was also assigned.

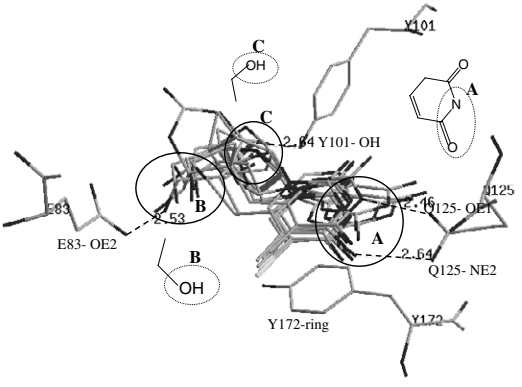


Fig. 3. Binding-site pharmacophores identified by superimposing ten crystal structures of HSV-1 thymidine kinase shown in Figure 2. Three pharmacological preferences and interactions are identified and circled as A (an amide binding site), B (a hydroxyl binding site), and C (a hydroxyl binding site). A stack force binding area (Y172-ring) is also indicated. The dash lines indicate the hydrogen binding.

To evaluate GEMDOCK's screening utility, we used HSV-1 thymidine kinase (TK), proposed by Bissabtz et al. [12] as the target protein with a ligand database, including 10 known active ligands (Figure 2) of TK and 990 randomly chosen non-active compounds from the ACD. When preparing the target protein, the atom coordinates for virtual screening were taken from the crystal structure of the TK complex (PDB entry 1kim). The atom coordinates of each ligand were sequentially taken from the database. Our program automatically decided the formal charge and atom type of each ligand atom. The ligand characteristics (i.e., the numbers of electrostatic atoms, hydrogen donor, and hydrogen acceptor) and the bond types of single bonds inside a ligand were also calculated. These variables were used in Equation 9 to calculate the scoring value of a docked conformation. Finally GEMDOCK re-ranked all docked ligand conformations for the post-analysis.

Figure 3 shows the binding-site pharmacophores and ligand preferences that were identified by superimposing ten crystal structures of TK shown in Figure 2. Three binding-site pharmacological preferences and interactions were identified and circled as A, B, and C. A stack force binding area (Y172-ring) was also indicated. The dash lines indicate the hydrogen binding. According to these observations, we added following pharmacological weights: Q125-OE1 and Q125-NE2 are hydrogen bonds with weighted value 4.0; Y101-OH and E83-OE2 are hydrogen bonds with weighted value 2.5; and six C atoms of Y172-ring form stacking force with weighted value 1.5. These weights were used in Equation 13 for calculating the value E_{phama}. For TK ligand preferences, the number of electrostatic atoms was set to 2 because all active ligands (Figure 2) have no charged atoms. The hydrophilic preference was not assigned in this target. In Equation 14, therefore, WP_{hb} is zero and WP_{elec} is 20 if the number of charged atoms inside a ligand was more than 2.

Table 3. Comparison GEMDOCK with three docking methods on docking 10 thymidine kinase ligands into the binding site of the target protein 1kim

	No. of	No. of	GEMDOCK				
	polar	hydrogen	pharmacophore	pharmacophore			
Ligand[a]	atoms[b]	bonds[c]	weight (yes)	weight (no)	GOLD[d]	FlexX[d]	DOCK[d]
1e2k.TMC	6	5	0.75	0.79	1.19	7.56	1.11
1e2m.HPT	5	6	0.41	0.37	0.49	1.02	4.18
1e2n.RCA	9	6	1.54	1.41	2.33	9.62	13.3
1e2p.CCV	6	8	0.58	0.53	0.93	2.02	3.65
1ki2.GA2	9	4	3.56	2.15	3.11	3.01	6.07
1ki3.PE2	8	5	3.34	3.29	3.01	4.10	5.96
1ki6.AHU	7	6	0.43	0.39	0.63	1.16	0.88
1ki7.ID2	7	6	0.45	0.56	0.77	9.33	1.03
1kim.THM	7	4	0.47	0.48	0.72	0.82	0.78
2ki5.AC2	8	5	2.94	2.95	2.74	3.08	2.71

[a] The four characters and three characters separated by a period denote the PDB code and the ligand name in the Protein Data Bank, respectively.
[b] The number of the atoms that may form a hydrogen bond; i.e., the atom type is either both, donor, or acceptor.
[c] The number of hydrogen bonds formed between the ligand and the protein was derived from the native crystal conformations based on our scoring function (Equation 10).
[d] These results were directly taken from [12].

3.3 Molecular Docking Results on Ten TK Complexes

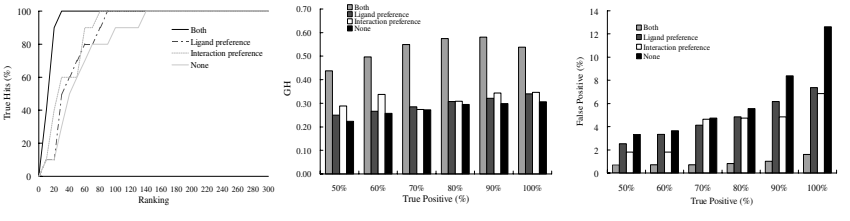
GEMDOCK executed 3 independent runs for each complex. The solution with lowest scoring function was then compared with the observed ligand crystal structure. First GEMDOCK docked each ligand of 10 TK ligands (Figure 2) back into its respective complex. We based the results on root mean square deviation (RMSD) error in ligand heavy atoms between the docked conformation and the crystal structure. The RMSD

values of all ten docked conformations are less than 1.0 Å. Second we docked all ten TK ligands into the reference protein (1kim) and the results were shown in Table 3. During flexible docking GEMDOCK obtained similar results whether the pharmacophore preferences (i.e., E_{phama} and E_{ligpre}) were considered or not. The docked conformations with RMSD values less than 1.5 Å for seven pyrimidine-based ligands. On the other hand, three purine-based ligands (i.e., 1ki2.GA2, 1ki3.PE2, and 2ki5.AC2) could not be successfully docked into the reference protein because the side-chain conformation of GLN125 in the reference protein 1kim differs from the ones of these purine-based complexes, i.e., 1ki2, 1ki3, and 2ki5. GEMDOCK was the best among these four competing methods (GEMDOCK, GOLD, FlexX, and DOCK) on this test set.

Table 4. Comparison of GEMDOCK with four methods on screening 1000 compounds with false positive rates

True Positive(%)	GEMDOCK	Surflex[a]	DOCK[a]	FlexX[a]	GOLD[a]
80	0.8[b] (8/990)	0.9	23.4	8.8	8.3
90	1.0 (10/990)	2.8	25.5	13.3	9.1
100	1.6 (16/990)	3.2	27.0	19.4	9.3

[a] These results were directly taken from [12] and [19].
[b] the false positive rate from 990 random ligands (%).



(a) true hit (b) GH-score (c) false positive

Fig. 4. GEMDOCK results for (a) true hit, (b) GH-score, and (c) the false positive rate for different true positive rates. GEMDOCK yielded good performance when it used both ligand and receptor pharmacological preferences.

3.4 Virtual Screening of TK Substrates

Figure 4 shows the overall accuracy of GEMDOCK using different combinations of pharmacophore preferences in screening the substrates of HSV-1 thymidine kinase (TK) from a data set with 1000 compounds. This data set, including 10 active and 990 random ligands proposed by Bissantz [12], was used to evaluate the performance of three

docking tools (DOCK, FlexX, and GOLD) with different combinations of seven scoring functions[12]. The results of the comparison are also shown in Table 4.

Four common metrics were used to evaluate the screening quality, including true hit (the percentage of active ligands retrieved from database), yield (the percentage of active ligands in the hit list), goodness-of-hit (GH), and false positive rate. The GH score is defined as

$$GH = (\frac{A_h(3A + T_h)}{4T_hA})/(1 - \frac{T_h - A_h}{T - A}), \tag{15}$$

where A_h is the number of active ligands in the hit list, T_h is the total number of compounds in the hit list, A is total number of active ligands in the database, and T is the total number of compounds in the database. The yield (hit rate) can be given as $100\frac{A_h}{T_h}\%$. The false positive (FP) rate is given as $100\frac{T_h - A_h}{T}\%$. In the TK case A and T are 10 and 1000, respectively.

The main objective of this study was to evaluate whether the new scoring function was applicable to both molecular docking and ligand scoring in virtual screening. Figure 4 shows these results of GEMDOCK using different combinations of pharmacophore preferences that are ligand preferences (E_{ligpre}) and binding-site pharmacophore (E_{phama}). GEMDOCK generally improves the screening quality by considering both ligand preferences and binding-site pharmacophore weights although we did not attempt to refine any parameters of these combinations. The binding-site pharmacophores seem more important than ligand preferences. As shown in Figure 4(a), the hit rates of GEMDOCK for different combinations are 38% (both), 12% (ligand preferences), 13% (binding-site pharmacophore) and 7% (none) when the TP rate is 100%. If GEMDOCK applied binding-site and ligand preferences, the GH score is 0.54 (Figure 4(b)) and the FP rate is 1.62% (Figure 4(c)) when the TP rate is 100%.

Table 4 compares GEMDOCK with four docking methods (Surflex, DOCK, FlexX, and GOLD) on the same target protein and screening database at true positive rates ranging from 80% to 100%. For GEMDOCK on the target TK, the ranks of the ten active ligands were 3, 7-9, 12-14, 16, 19, and 26. For the true positive rate of 100%, the FP rate for GEMDOCK is 1.6%. In contrast, the FP rates for competing methods are 3.2% (Surflex), 27% (DOCK), 19.4% (FlexX), and GOLD (9.3%). DOCK is the worst and GEMDOCK is the best among these five approaches on this data set.

4 Conclusions

In summary, we have developed an automatic tool with a novel scoring function for virtual screening by applying numerous enhancements and modifications to our original techniques. By integrating a number of genetic operators, each having a unique search mechanism, GEMDOCK seamlessly blends the local and global searches so that they work cooperatively. Our new scoring function is able to be applied to both flexible docking and post-docking analysis for reducing the number of false positives. Experiments verify that the proposed approach is robust and adaptable to virtual screening.

References

1. P. D. Lyne. Structure-based virtual screening: an overview. *Drug Discovery Today*, 7:1047–1055, 2002.
2. B. K. Shoichet, S. L. McGovern, B. Wei, and J. Irwin. Lead discovery using molecular docking. *Current Opinion in Chemical Biology*, 6:439–446, 2002.
3. I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin. A geometric approach to macromolecular-ligand interactions. *Journal of Molecular Biology*, 161:269–288, 1982.
4. G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor. Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology*, 267:727–748, 1997.
5. B. Kramer, M. Rarey, and T. Lengauer. Evaluation of the flexX incremental construction algorithm for protein-ligand docking. *Proteins: Structure, Function, and Genetics*, 37:228–241, 1999.
6. T. J. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz. Dock 4.0: search strategies for automated molecular docking of flexible molecule databases. *Journal of Computer-Aided Molecular Design*, 15:411–428, 2001.
7. G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson. Automated docking using a lamarckian genetic algorithm and empirical binding free energy function. *Journal of Computational Chemistry*, 19:1639–1662, 1998.
8. J.-M. Yang. Development and evaluation of a generic evolutionary method for protein-ligand docking. *Journal of Computational Chemistry*, 25:843–857, 2004.
9. H. Gohlke, M. Hendlich, and G. Klebe. Knowledge-based scoring function to predict protein-ligand interactions. *Journal of Molecular Biology*, 295:337–356, 2000.
10. S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, Jr., and P. Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society*, 106:765–784, 1984.
11. D. K. Gehlhaar, G. M. Verkhivker, P. Rejto, C. J. Sherman, D. B. Fogel, L. J. Fogel, and S. T. Freer. Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chemistry and Biology*, 2(5):317–324, 1995.
12. C. Bissantz, G. Folkers, and D. Rognan. Protein-based virtual screening of chemical databases. 1. evaluation of different docking/scoring combinations. *Journal of Medicinal Chemistry*, 43:4759–4767, 2000.
13. M. Stahl and M. Rarey. Detailed analysis of scoring functions for virtual screening. *Journal of Medicinal Chemistry*, 44:1035¡V1042, 2001.
14. J.-M. Yang and C.-Y. Kao. A robust evolutionary algorithm for training neural networks. *Neural Computing and Application*, 10(3):214–230, 2001.
15. J.-M. Yang and C.-C. Chen. Gemdock: A generic evolutionary method for molecular docking. *Proteins: Structure, Function, and Bioinformatics*, 55:288–304, 2004.
16. J.-M. Yang, C.-H. Tsai, M.-J. Hwang, H.-K. Tsai, J.-K. Hwang, and C.-Y. Kao. GEM: A gaussian evolutionary method for predicting protein side-chain conformations. *Protein Science*, 11:1897–1907, 2002.
17. T. Bäck. *Evolutionary Algorithms in Theory and Practice*. Oxford University Press, New York, USA, 1996.
18. J. N. Champness, M. S. Bennett, F. Wien, R. Visse, C. W. Summers, P. Herdewijn, E. de Clerq, T. Ostrowski, R. L. Jarvest, and M. R. Sanderson. Exploring the active site of herpes simplex virus type-1 thymidine kinase by x-ray crystallography of complexes with aciclovir and other ligands. *Proteins*, 32:350–361, 1998.
19. A. N. Jain. Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *Journal of Medicinal Chemistry*, 46:499–511, 2003.