PROTEINS: Structure, Function, and Bioinformatics 55:288–304 (2004)

# GEMDOCK: A Generic Evolutionary Method for Molecular Docking

**Jinn-Moon Yang*** **and Chun-Chen Chen**
*Department of Biological Science and Technology, National Chiao Tung University, Hsinchu, Taiwan*

***ABSTRACT*** **We have developed an evolutionary approach for flexible ligand docking. This approval, GEMDOCK, uses a Generic Evolutionary Method for molecular DOCKing and an empirical scoring function. The former combines both discrete and continuous global search strategies with local search strategies to speed up convergence, whereas the latter results in rapid recognition of potential ligands. GEMDOCK was tested on a diverse data set of 100 protein–ligand complexes from the Protein Data Bank. In 79% of these complexes, the docked lowest energy ligand structures had root-mean-square derivations (RMSDs) below 2.0 Å with respect to the corresponding crystal structures. The success rate increased to 85% if the structure water molecules were retained. We evaluated GEMDOCK on two cross-docking experiments in which each ligand of a protein ensemble was docked into each protein of the ensemble. Seventy-six percent of the docked structures had RMSDs below 2.0 Å when the ligands were docked into foreign structures. We analyzed and validated GEMDOCK with respect to various search spaces and scoring functions, and found that if the scoring function was perfect, then the predicted accuracy was also essentially perfect. This study suggests that GEMDOCK is a useful tool for molecular recognition and may be used to systematically evaluate and thus improve scoring functions. Proteins 2004; 55:288–304.** © **2004 Wiley-Liss, Inc.**

**Key words: cross-docking; evolutionary algorithm; molecular recognition; protein–ligand docking; hybrid docking; structure-based drug design**

## INTRODUCTION

The protein–ligand docking problem is the prediction of a ligand conformation and orientation relative to the active site of a target protein. A computer-aided docking process, identifying the lead compounds by minimizing the energy of intermolecular interactions, is an important approach for structure-based drug designs.[1] Using a computer method to find a solution to a protein–ligand docking problem involves two critical elements: a good scoring function and an efficient algorithm for searching conformation and orientation spaces.[2,3]

A good scoring function should be able to screen a large number of potential solutions rapidly and simply, while effectively discriminating between correct binding states and non-native docked conformations. Various scoring functions have been developed for calculating the free energy of binding, including knowledge-based,[4,5] empirical,[6,7] physics-based,[8,9] and solvent-based scoring functions.[10] In general, the binding energy landscapes of these scoring functions are often complex and exhibit a rugged funnel shape.[11] Therefore, an efficient search algorithm is required to find a global solution for various scoring functions.

Many automated docking approaches have been developed and can be roughly divided into rigid docking, flexible ligand-docking, and flexible protein-docking methods. The rigid-docking methods, such as the DOCK program,[12] treat both the ligand and protein as rigid. In contrast, the ligand is considered flexible and the protein rigid for flexible ligand-docking methods, including evolutionary algorithms,[6,8,13–15] simulated annealing,[16] the fragment-based approach,[17] and other algorithms.[18–20] For reasonably fast addressing protein flexibility problems, in which both ligands and proteins are flexible, these docking methods often allowed a limited model of protein variations, such as the side-chain flexibility or small motions of loops in the binding site.[21–24] Most of these previous docking methods were evaluated using small test sets (<20 protein–ligand complexes). In contrast, GOLD[13] and FlexX[17] were evaluated using a test set of over 100 such complexes.

Despite the diversity of the scoring functions and search algorithms used in these methods, they are either flexible or rigid docking methods. It is not clear to what extent the nature of hybrid docking methods (e.g., involving both rigid and flexible docked conformations simultaneously) has influenced the accuracy of a search method in docking problems. A new docking method should be capable of determining which factors (e.g., search algorithms, scoring functions, the role of water, or protein and ligand flexibility) are primarily responsible for ligand docking errors.[2,3]

To address the above questions, we developed a molecular docking approach termed GEMDOCK (Generic Evolu-

Grant sponsor: National Science Council of Taiwan; Grant number: NSC-91-2320-B-009-001. Grant sponsor: Department of Health of Taiwan; Grant number: DOH92-TD-1132.

*Correspondence to: Jinn-Moon Yang, Department of Biological Science and Technology, National Chiao Tung University, Hsinchu 30050, Taiwan. E-mail: moon@cc.nctu.edu.tw

Received 18 July 2003; Accepted 1 October 2003

Published online 27 February 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20035

**TABLE I. Parameters of GEMDOCK**

| Parameter | Value of parameters |
|---|---|
| Initial step sizes | $\sigma = 0.8, \psi = 0.2$ (in radius) |
| Family competition length | $L = 2$ |
| Population size | $N = 300$ |
| Recombination rate | $p_c = 0.3$ |
| No. of the maximum generation | 70 |

**TABLE II. Atom Formal Charge of GEMDOCK**

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

tionary Method for molecular DOCKing). The GEMDOCK software is available on the Web at http://gemdock.life.nctu.edu.tw. This program uses an empirical scoring function and an evolutionary approach that is more robust than standard evolutionary approaches[25–27] with regard to several specific domains.[28–31] The GEMDOCK energy function consists of electrostatic, steric, and hydrogen-bonding potentials. Steric and hydrogen bonding potentials use a linear model that is simple and recognizes potential complexes rapidly. The core idea of this evolutionary approach is to design multiple operators that cooperate using a family competition paradigm that is similar to a local search procedure.

Numerous enhancements and modifications were applied to the original technique,[29] thereby improving the reliability and applicability of the method. There are four main differences in methodology between the present work and our previous studies. First, we developed an empirical scoring function having fewer local minima to replace the relatively complicated AMBER-based energy function. Second, we added a differential evolution operator[32] to reduce the disadvantages of Gaussian and Cauchy mutations, and a new rotamer-based mutation operator to reduce the search space of ligand structure conformations. Third, GEMDOCK may be run as either a purely flexible or hybrid docking approach. Finally, GEMDOCK is an automatic system that generates all related docking variables, such as atom formal charge, atom type, and the ligand binding site of a protein.

## MATERIALS AND METHODS

### GEMDOCK Parameters

Table I indicates the setting of GEMDOCK parameters, such as initial step sizes, family competition length ($L$ = 2), population size ($N$ = 300), and recombination probability ($pc$ = 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 70. Therefore, GEMDOCK generated 1200 solutions in one generation and terminated after it exhausted 84,000 solutions in the worse case. These parameters were decided after experiments were conducted to recognize complexes of test docking systems with various values.

### Scoring Function

In this work, we used an empirical scoring function given as

$$E_{tot} = E_{inter} + E_{intra} + E_{penal}, \tag{1}$$

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. $E_{penal}$ is set to 10,000.

The intermolecular energy is defined as

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

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. The formal charge of a receptor and ligand atom is indicated in Table II. $F(r_{ij}^{B_{ij}})$ is a simple atomic pairwise potential function (Fig. 1) modified from previous works[6,33] and given as

$$F(r_{ij}^{B_{ij}}) = \begin{cases} V_6 - \dfrac{V_6 r_{ij}^{B_{ij}}}{V_1} & \text{if } r_{ij}^{B_{ij}} \le V_1 \\ \dfrac{V_5 (r_{ij}^{B_{ij}} - V_1)}{V_2 - V_1} & \text{if } V_1 < r_{ij}^{B_{ij}} \le V_2 \\ V_5 & \text{if } V_2 < r_{ij}^{B_{ij}} \le V_3 \\ V_5 - \dfrac{V_5 (r_{ij}^{B_{ij}} - V_3)}{V_4 - V_3} & \text{if } V_3 < r_{ij}^{B_{ij}} \le V_4 \\ 0 & \text{if } r_{ij}^{B_{ij}} > V_4. \end{cases} \tag{3}$$

$r_{ij}^{B_{ij}}$ is the distance between the atoms $i$ and $j$ with the interaction type $B_{ij}$ forming by the pairwise heavy atoms between ligands and proteins where $B_{ij}$ is either a hydrogen bond or a steric state. In this atomic pairwise model, these two potentials are calculated by the same function form but with different parameters, $V_1$,...,$V_6$ given in Figure 1. The energy value of a hydrogen bond should be larger than the one of the steric potential. In this model, the atom is divided into four different atom types (Table III): donor, acceptor, both, and nonpolar. A hydrogen bond can be formed by the following atom-pair types: donor–acceptor (or acceptor–donor), donor–both (or both–donor), acceptor–both (or both–acceptor), and both–both. Other atom-pair combinations are to used form the steric state.

The intramolecular energy of a ligand is

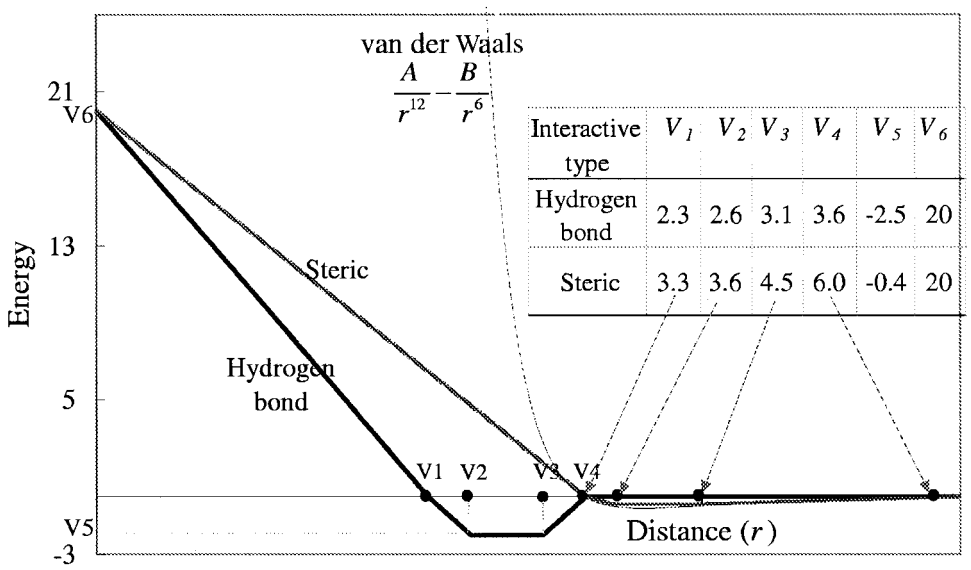


Fig. 1. The linear energy function of the pairwise atoms for the steric interactions and hydrogen bonds in GEMDOCK (bold line) with a standard Lennard–Jones potential (light line).

**TABLE III. Atom Types of GEMDOCK**

| Atom type | Heavy atom name |
|---|---|
| Donor | Primary and secondary amines, sulfur, and metal ions |
| Acceptor | Oxygen and nitrogen with no bound hydrogen |
| Both | Structural water and hydroxy 1 groups |
| Nonpolar | Other atoms (such as carbon and phosphorus) |

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

where $F(r_{ij}^{B_{ij}})$ is defined as Eq. (3) except that the value is 1000 to discard unreasonable conformations when $r_{ij}^{B_{ij}} <$ 2.0 Å and *dihed* is the number of rotatable bonds. We followed the work of Gehlhaar et al.[6] to set the values of $A$, $m$, and $\theta_0$. The $sp^3 - sp^3$ bond, $A$, $m$, and $\theta_0$ are set to 3.0, 3.0, and $\pi$, respectively; and $A = 1.5$, $m = 6$, and $\theta_0 = 0$ for the $sp^3 - sp^2$ bond.

## GEMDOCK Algorithm Details

In the following subsections, we present the details of our approach for molecular docking (see Appendix). The core idea of our evolutionary approach was to design multiple operators that cooperate using the family competition model, which is similar to a local search procedure. We designed a new rotamer-based mutation operator for reducing the search space of ligand structure conformations, and used a differential evolution operator[32] for reducing the disadvantages of Gaussian and Cauchy mutations. GEMDOCK is a nearly automatic docking tool for generating all experimental variables, and may serve as a flexible or hybrid docking program. First we specified the coordinates of ligand and protein atoms, the ligand binding area, atom formal charge (Table II), and atom types (Table III). Crystal coordinates of the ligand and protein atoms were taken from the Protein Data Bank (PDB) and separated into different files. GEMDOCK then automatically determined the center of the receptor and the search cube of a binding site according to the maximum and minimum of coordinates of these selected protein atoms.

After it prepares the ligand and protein, GEMDOCK works as follows: It 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 ($x^i$, $\sigma^i$, $\psi^i$), 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 three-dimensional (3D) 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. $\sigma$ and $\psi$ 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 $2\pi$ in radians. For the initial step sizes, $\sigma$ is 0.8 and $\psi$ 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. As shown in the Appendix, these stages apply a general procedure "FC_Adaptive," with only different working population and the mutation operator.

GEMDOCK can be a flexible docking method or a hybrid docking method that evolves simultaneously both flexible and rigid conformation solutions of a ligand. GEMDOCK is

a flexible docking tool if it evolves the conformation variables ($x_7$,…,$x_n$) of each solution in a population. On the other hand, GEMDOCK is a hybrid approach if the conformation variables of part of the solutions (e.g., $\eta N$ solutions) are fixed and set to the values of a native binding state. We set $\eta$ to 0.2 when GEMDOCK is a hybrid method that simultaneously evolves fixed and flexible ligand conformations by the recombination operators.

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

In the following, genetic operators are briefly described. We use $a = (x^a, \sigma^a, \psi^a)$ to represent the "family father" and $b = (x^b, \sigma^b, \psi^b)$ as another parent. The offspring of each operation is represented as $c = (x^c, \sigma^c, \psi^c)$. The symbol $x_j^s$ is used to denote the $j$th adjustable optimization variable of a solution $s$, $\forall j \in \{1,\ldots n\}$.

## Recombination Operators

GEMDOCK implements modified discrete recombination and intermediate recombination.[25] A recombination operator selected the "family father ($a$)" and another solution ($b$) randomly selects from the working population. The former generates a child as follows:

$$x_j^c = \begin{cases} x_j^a \text{ with probability } 0.8 \\ x_j^b \text{ with probability } 0.2. \end{cases} \quad (5)$$

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

$$w_j^c = w_j^a + (w_j^b - w_j^a)/2, \quad (6)$$

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

## Mutation Operators

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

### *Gaussian and Cauchy mutations*

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

$$w_j' = w_j A(\cdot), \quad (7)$$

$$x_j' = x_j + w_j' D(\cdot), \quad (8)$$

where $w_j$ and $x_j$ are the $i$th component of $w$ and $x$, respectively, and $w_j$ is the respective step size of the $x_j$ where $w$ is $\sigma$ or $\psi$. 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, and $N_j(0,1)$ 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

$$\psi_j^c = \psi_j^a \exp[\tau' N(0,1) + \tau N_j(0,1)], \quad (9)$$

$$x_j^c = x_j^a + \psi_j^c C_j(t). \quad (10)$$

We set $\tau$ and $\tau'$ to $(\sqrt{2n})^{-1}$ and $(\sqrt{2\sqrt{n}})^{-1}$, respectively, according to the suggestion of evolution strategies.[25] A random variable is said to have the Cauchy distribution [$C(t)$] if it has the density function: $f(y; t) = (t/\pi)/(t^2 + y^2)$, $-\infty < y < \infty$. In this article, $t$ is set to 1. Our decreasing-based Gaussian mutation uses the step-size vector $\sigma$ with a fixed decreasing rate $\gamma = 0.95$ and works as

$$\sigma^c = \gamma \sigma^a, \quad (11)$$

$$x_j^c = x_j^a + \sigma^c N_j(0,1). \quad (12)$$

### *Differential evolution*

An offspring of differential evolution is generated as

$$x_j^c = \begin{cases} u_j^m & \text{if } \mathrm{rand}(0,1) \le CR \\ x_j^a & \text{otherwise} \end{cases} \quad (13)$$

and

$$u_j^m = x_j^a + K(x_j^b - x_j^c), \quad (14)$$

where $a$ is the "family father"; $b$ and $c$ are two solutions randomly selected from the working population subjected to $a \neq b \neq c$. In this work, $K$ and $CR$ are set to 0.5 and 0.9, respectively.

### *Rotamer mutation*

This operator is only used for $x_7$ to $x_n$ to find the conformations of the rotatable bonds inside the ligand. For each ligand, this operator mutates all of the rotatable angles according to the rotamer distribution and works as

**TABLE IV. The 100 Test Complexes**

1aaq[a]:PSI[b] 1abe:ARA 1acj:THA 1ack:EDR 1acl:DME 1acm:PAL 1aco:TRA 1aec:E64 1aha:ADE 1apt:IVA 1ase:NOP 1azm:AZM 1baf:NPP 1blh:FOS 1cbx:BZS 1coy:AND 1cps:CPM 1dbb:STR 1dbj:AE2 1did:DIG 1die:DNJ 1dr1:BIO 1dwd:MID 1eap:HEP 1eed:BOC 1epb:REA 1eta:T44 1etr:MQI 1fkg:SB3 1fki:SB1 1ghb:PPP(I)[c] 1glq:GTB 1hdc:CBO 1hdy:NAD 1hef:PPP(I) 1hri:S57 1hsl:PPP(D) 1hyt:BZS 1icn:OLA 1ida:QND 1igj:DGX 1ive:ST3 1ldm:NAD 1lic:HDS 1lst:PPP([d]) 1mcr:PPP(P) 1mdr:SAA 1mrk:FMC 1mup:TZL 1nis:NTC 1pbd:PAB 1pha:PFZ 1phd:PIM 1phg:MYT 1poc:GEL 1rds:GPC 1rne:C60 1rob:C2P 1slt:NAG 1srj:NAB 1stp:BTN 1tdb:UFP 1tka:N3T 1tmn:PPP(I) 1tpp:APA 1ulb:GUN 1xid:ASC 1xie:ASO 2ada:HPR 2ak3:AMP 2cgr:GAS 2cht:BAR 2ctc:LOF 2dbl:S5H 2mcp:PC 2mth:MPB 2phh:APR 2pk4:ACA 2plv:SPH 2r07:W33 2sim:DAN 2yhx:OTG 3aah:PQQ 3cla:CLM 3cpa:PPP(S) 3gch:CIN 3hvt:NEV 3ptb:BEN 3tpi:PPP(S) 4cts:OAA 4dfr:MTX 4est:PPP(I) 4fab:FDS 4phv:VAC 5p2p:DHG 6abp:ARA 6rnt:2AM 6rsa:UVC 7tim:PGH 8gch:PPP(C)

[a]A 4-character PDB code used in the Protein Data Bank.
[b]A 3-character ligand code used in the Protein Data Bank.
[c]"PPP" denotes a peptide ligand and the uppercase character in ( ) denotes the chain code of the peptide ligand.
[d]The chain code of the peptide ligand is empty.

$$x_j = r_{ki} \text{ with probability } p_{ki}, \quad (15)$$

where *rki* and $p_{ki}$ are the angle value and the probability, respectively, of *i*th rotamer of *k*th bond type including $sp^3 - sp^3$ and $sp^3 - sp^2$ bond. The values of *rki* and $p_{ki}$ are based on the energy distributions of these two bond types.

## RESULTS

### Test Data Set and Docking Protocols

To evaluate the strengths and limitations of GEMDOCK, we tested the program on a highly diverse data set of 100 protein–ligand complexes (Tables IV and V) proposed by Jones et al.[13] In addition, our program was evaluated using 2 cross-docking ensembles of protein structures, 8 complexes of the human immunodeficiency virus type 1 (HIV-1), protease[34] and 5 complexes of the Fab' fragment of monoclonal antibody DB3. Crystal coordinates of the ligand and protein atoms were taken from the PDB and were separated into different files. Our program then assigned the atom formal charge and atom type (i.e., donor, acceptor, both, or nonpolar) for each atom of both the ligand and protein. The bond type ($sp^3$ and $sp^3 - sp^3$, $sp^3 - sp^2$, or others) of a rotatable bond inside a ligand was also assigned. These variables were used in Eq. (1) to calculate the scoring value of a docked conformation (see Materials and Methods section).

When preparing the proteins, the size and location of the ligand-binding site was determined by considering the protein atoms located $< 10$ Å from each ligand atom. The metal atoms in the active site were also retained. We duplicated the Jones et al. work,[13] in that all structure water molecules were removed. In addition, we evaluated our method in the presence of the structure water molecules and compared the results. GEMDOCK then automatically decided the search cube of a binding site based on the maximum and minimum values of coordinates among these selected protein atoms. Among these 100 complexes tested, the minimum cube was 23 Å × 24 Å × 20 Å (2mcp) and the maximum cube was 41 Å × 39 Å × 42 Å (1rne).

Seven selected complexes, which are shown in Table VI to illustrate the docking variables, are used to compare GEMDOCK with other approaches. The ligand variables and the protein–ligand interactions were derived from the native crystal complexes. Ligand variables included the number of single bonds, the number of polar atoms (donor, acceptor, or both), and the number of charged atoms (when the formal charge was not zero). Protein–ligand interactions included the binding energy of the native state, the number of hydrogen bonds, and the number of electrostatic interactions.

The root-mean-square deviation (RMSD) of heavy atom positions between the docked conformation and the crystal structure was used to assess the accuracy of docking predictions. The successful percentage (the proportion of docking experiments that found a solution within 2.0 Å RMSD) was determined to evaluate the robustness of a docking method. The RMSD commonly used in previous studies[13,17] is defined as

$$\left\{ \sum_{i=1}^{M} [(X_i - x_i)^2 + (Y_i - y_i)^2 + (Z_i - z_i)^2]/M \right\}^{1/2}, \quad (16)$$

where *M* is the heavy atom number of a ligand; $(X_i, Y_i, Z_i)$ and $(x_i, y_i, z_i)$ are the coordinates of the *i*th atom of X-ray crystal and docked structures, respectively. We defined an acceptable docked result as a solution that had an RMSD value $< 2.0$ Å.[13,17]

GEMDOCK was initially evaluated on two docking systems (Table VII) with different binding search areas, namely, the whole protein and the selected binding site. The selected binding site is the part of the protein where the protein atoms are located $< 10$ Å from each ligand atom. One hundred independent docking runs were performed in each test case and the results are summarized in Table VII and Figure 2. GEMDOCK generated the best RMSD values of 0.79 Å (1etr) and 0.58 Å (4dfr), and the average RMSD values of 3.24 Å (1etr) and 1.23 Å (4dfr) when the selected-area atoms were used as the binding site. The successful percentages under these conditions were 65% and 93% for 1etr and 4dfr, respectively. When the whole protein was considered as the search binding area, the results generated by GEMDOCK were slightly different. Figure 2(A and B) show the docked conformations of methotrexate (MTX) into dihydrofolate reductase (DHFR) using the whole-protein and selected-area atoms as the binding search areas, respectively. The docked lowest energy conformation (gray) is identical with the crystal structure (dark) for most ligand groups. The predicted ligand conformations can be divided into two main clusters, one (gray) near the native binding state (87%) and the other (dark gray) in the incorrect position [13%; Fig. 2(a)].

**TABLE V. GEMDOCK Results for 100 Complexes**

| RMSD (Å) | A (no water[a]) First rank | A (no water[a]) Any rank | A (water[b]) First rank | A (water[b]) Any rank | B (first rank) No water | B (first rank) Water | PDB code with Method A (no water) and the solutions at the first rank |
|---|---|---|---|---|---|---|---|
| ≤0.5 | 23 | 25 | 24 | 37 | 35 | 47 | 1abe 1ack 1acm 1aco 1dr1 1eap 1epb 1fki 1hdy 1nis 1pbd 1pha 1phd 1srj 1stp 1tpp 2ada 2cht 6rsa 2mcp 2r07 3aah 6abp |
| >0.5, ≤1.0 | 36 | 40 | 42 | 42 | 34 | 26 | 1acj 1aha 1baf 1cbx 1dbb 1dbj 1die 1eed 1etr 1hri 1hsl 1hyt 1ida 1ldm 1lst 1mrk 1phg 1rds 1ghb 1rob 1tka 1ulb 1xid 2cgr 2ctc 2dbl 2pk4 1fkg 2sim 3cpa 3hvt 3tpi 4cts 4dfr 4phv 7tim |
| >1.0, ≤1.5 | 9 | 15 | 15 | 7 | 7 | 11 | 1aec 1azm 1coy 1glq 1mdr 1slt 1tmn 2ak3 4fab |
| >1.5, ≤2.0 | 11 | 5 | 4 | 3 | 6 | 3 | 1aaq 1ase 1cps 1dwd 1hdc 1lic 1poc 1tdb 3ptb 5p2p 8gch |
| >2.0, ≤2.5 | 3 | 3 | 2 | 2 | 5 | 2 | 1apt 1icn 1xie |
| >2.5, ≤3.0 | 5 | 3 | 4 | 5 | 4 | 1 | 1acl 1mcr 2yhx 3gch 4est |
| >3.0 | 13 | 9 | 9 | 4 | 9 | 10 | 1blh 1did 1eta 1hef 1igj 1ive 1mup 1rne 2mth 2phh 2plv 3cla 6rnt |

A, GEMDOCK works as a flexible docking method evolving $N$ flexible ligand conformation solutions, where $N$ is the population size.
B, GEMDOCK works as a hybrid docking method evolving both 0.2 $N$ rigid and 0.8 $N$ flexible ligand conformation solutions.
[a]Removing all structure water molecules from proteins.
[b]Retaining structure water molecules in the binding site.

**TABLE VI. Seven Test Systems Selected From 100 Complexes in Table IV**

| Protein–ligand complex (PDB code) | Search Cartesian volume (Å) | Ligand: No. of torsions | Ligand: No. of polar atoms[a] | Ligand: No. of charge atoms[b] | Interaction between ligand and receptor[c]: No. of hydrogen bonds | Interaction between ligand and receptor[c]: No. of electrostatic interactions | Interaction between ligand and receptor[c]: Energy of native binding |
|---|---|---|---|---|---|---|---|
| Alpha-thrombin/NAPAP (1dwd) | 31 Å × 51 Å × 46 Å | 9 | 8 | 2 | 7 | 4 | −137.86 |
| Streptavidin/biotin (1stp) | 32 Å × 31 Å × 30 Å | 5 | 5 | 2 | 7 | 0 | −101.90 |
| Thermolysin/CLT-LEU-TRP (1tmn) | 31 Å × 29 Å × 32 Å | 10 | 8 | 3 | 16 | 12 | −129.03 |
| Fab McPc-603/phosphocholine (2mcp) | 26 Å × 27 Å × 23 Å | 3 | 4 | 4 | 5 | 3 | −56.81 |
| Carboxypeptidase/TYR-GLY (3cpa) | 29 Å × 26 Å × 35 Å | 8 | 6 | 1 | 13 | 1 | −97.38 |
| Trypsinogen/VAL-ILE (3tpi) | 32 Å × 30 Å × 30 Å | 9 | 5 | 1 | 10 | 0 | −113.04 |
| HIV-1 protease/VAC (4phv) | 33 Å × 37 Å × 33 Å | 11 | 7 | 0 | 13 | 0 | −191.36 |

[a]The number of the atoms that may form a hydrogen bond (i.e., the atom type is either both, donor, or acceptor).
[b]The number of the atoms with nonzero formal charge.
[c]Statics were derived from the native crystal conformations based on our scoring function [Eq. (1)].

**TABLE VII. GEMDOCK Results on Two Docking Examples With Either the Whole Protein or the Selected Binding Site as the Search Binding Area**

| PDB code | Search Cartesian volume (Å) | Ligand: No. of torsion | Ligand: No. of polar atoms | Ligand: No. of charge atoms | Interaction[a]: No. of hydrogen bonds | Interaction[a]: No. of electrostatic interactions | Results: Best RMSD (Å) | Results: Average RMSD (Å) | Results: Successful percentage |
|---|---|---|---|---|---|---|---|---|---|
| Selected binding site: | | | | | | | | | |
| 1etr | 41 Å × 39 Å × 42 Å | 8 | 10 | 4 | 11 | 2 | 0.79 | 3.24 | 65 |
| 4dfr | 37 Å × 33 Å × 31 Å | 10 | 12 | 4 | 9 | 4 | 0.58 | 1.23 | 93 |
| Whole protein: | | | | | | | | | |
| 1etr | 57 Å × 55 Å × 57 Å | 8 | 10 | 4 | 11 | 2 | 0.85 | 4.19 | 56 |
| 4dfr | 46 Å × 39 Å × 46 Å | 10 | 12 | 4 | 9 | 4 | 0.54 | 1.84 | 87 |

[a]Protein–ligand interactions were derived from the native crystal structures.

### Overall Accuracy on 100 Complexes

The overall accuracy of GEMDOCK in predicting the docked ligand conformations of 100 test complexes is shown in Table V. The results generated by GEMDOCK are compared with those of other methods in Tables VIII and IX. All results are derived from 10 independent


Fig. 2. GEMDOCK results of docking methotrexate (MTX) into dihydrofolate reductase (4dfr), with the search binding areas defined as either (**A**) the whole protein or (**B**) the selected binding site. The docked ligand conformation is dark and the crystal ligand structure is gray. (**A**) Among 100 docking runs, GEMDOCK achieved 87 docked ligand conformations near the native structure and 13 conformations in wrong positions (dark gray).

**TABLE VIII. Comparing GEMDOCK With Other Methods on Test Cases Presented in Table VI**

| | GEMDOCK (this study) | | | GOLD[b] | FlexX[c] | MSA[d] | ConsDock[e] | AutoDock[f] |
|---|---|---|---|---|---|---|---|---|
| PDB code | Minimum energy | Best RMSD (Å) | Successful percentage[a] | Best RMSD (Å) | Best RMSD (Å) | Best RMSD (Å) | Best RMSD (Å) | Best RMSD (Å) |
| 1dwd | −146.99 | 1.57 | 30 | 1.71 | 2.12 | 2.01 | 4.55 | NA |
| 1stp | −107.51 | 0.41 | 90 | 0.69 | 0.81 | 1.02 | 0.51 | 0.89 |
| 1tmn | −135.29 | 0.82 | 60 | 1.68 | 0.87 | 1.93 | NA | NA |
| 2mcp | −62.98 | 0.61 | 70 | 4.37 | NA | 1.71 | NA | 0.96 |
| 3cpa | −108.50 | 0.87 | 100 | 1.58 | 3.08 | 0.62 | 2.26 | NA |
| 3tpi | −124.75 | 0.56 | 70 | 0.80 | 0.58 | 1.73 | 0.69 | NA |
| 4phv | −219.54 | 0.62 | 60 | 1.11 | 1.04 | 1.97 | 10.7 | NA |

GEMDOCK results were derived from 10 independent docking runs, and the docked lowest energy conformation was considered for each test case.

[a]The percentage of the trials that found a docked lowest energy structure within 2.0 Å RMSD with respect to the ligand-containing structure.

[b–f]The results were taken directly from the corresponding original articles.[13,19,45–47]

docking runs, and the docked lowest energy structure was considered for each test case. On average, GEMDOCK took 305 s for a docking run on a Pentium 1.4 GHz personal computer with a single processor. The maximum time was 883 s for 1rne and the shortest time was 102 s for 2pk4.

When the solutions at the first rank were considered, GEMDOCK achieved 79% success in identifying the experimental binding model (Table V). The RMSD values of 59 complexes were < 1.0 Å. This success rate rose to 85% if solutions of any rank were considered. When the structure water molecules in the binding site were retained, the success rates improved to 85% and 89% for solutions at the first rank and any rank, respectively. When GEMDOCK was used as a hybrid docking method, evolving 0.2*N* rigid ligand conformations and 0.8*N* flexible ligand conformations (where *N* is the population size), success rates of 82% and 87% were generated when structure water was retained and removed, respectively, when solutions at the first rank were considered. Figure 3 shows four typical acceptable solutions (the RMSD value < 2.0 Å) in which GEMDOCK predicted correct positions for most of the ligand groups. An RMSD value of < 1.0 Å was calculated for these 4 acceptable conformations: SCH 38057 docked into human rhinovirus 14 (1hri) [Fig. 3(A)][35]; metyrapone

**TABLE IX. Comparing GEMDOCK With GOLD and FlexX on the Data Set of 100 Complexes**

| RMSD (Å) | GEMDOCK | GOLD[a] | FlexX[b] |
|---|---|---|---|
| ≤0.5 | 23% | 8% | 12.5% |
| >0.5, ≤1.0 | 36% | 27% | 38.5% |
| >1.0, ≤1.5 | 9% | 20% | 12.5% |
| >1.5, ≤2.0 | 11% | 11% | 5.5% |
| >2.0, ≤2.5 | 3% | 2% | 7.5% |
| >2.5, ≤3.0 | 5% | 4% | 2.0% |
| >3.0 | 11% | 28% | 21.5% |

The success rate of GEMDOCK was based on solutions having the first rank.

[a]The success rate of GOLD,[13] a steady-state genetic algorithm, was based on solutions having the first rank.

[b]The success rate of FlexX,[17] a fragment-based approach, was based on the solutions having any rank on a data set of 200 complexes extended from the GOLD data set.

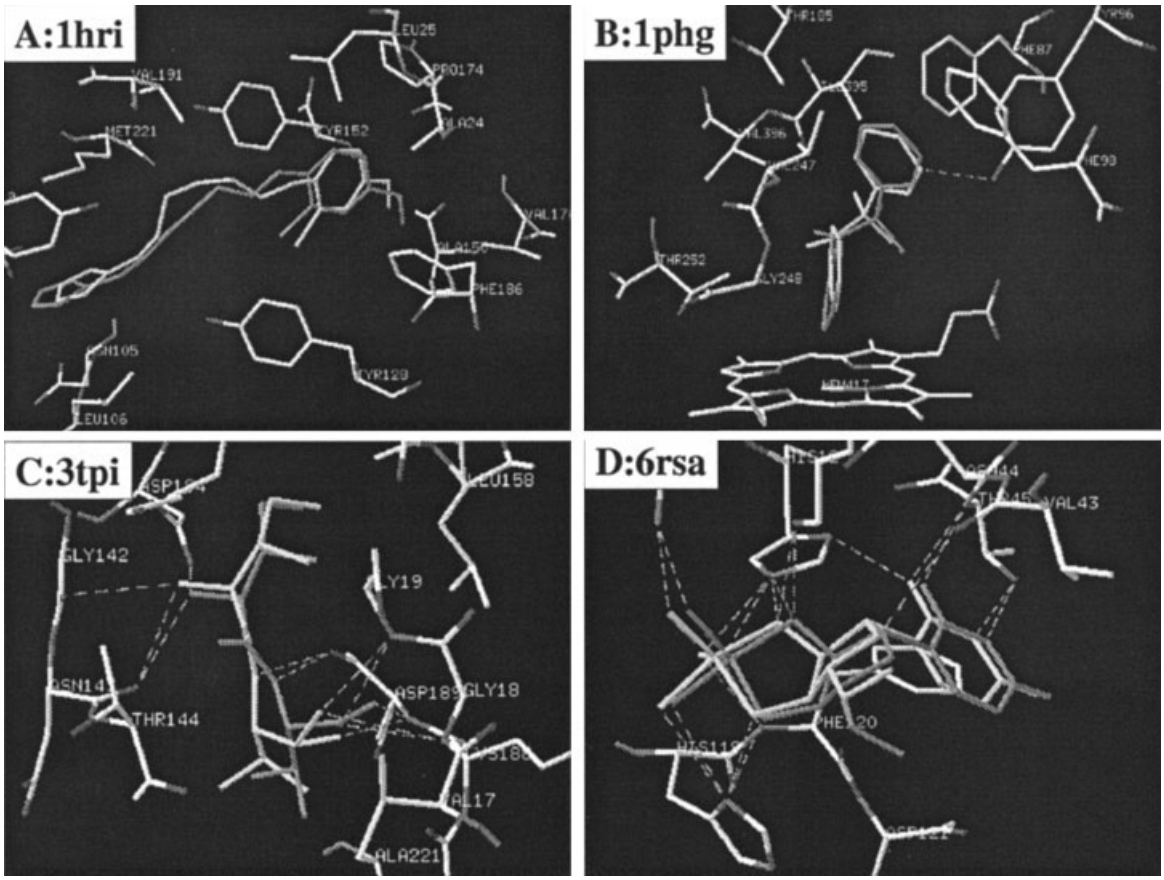


Fig. 3. GEMDOCK results for four typical acceptable complexes (i.e., the RMSD value $< 2.0$ Å). The RMSD values of these four complexes were less than 1.0 Å and most of the docked ligand groups (white) were identical to the crystal ligand structures (gray). The white dotted lines represent hydrogen bonds.

docked into cytochrome P450-cam (1phg) [Fig. 3(B)][36]; a dipeptide (Ile-Val) docked into trypsinogen (3tpi) [Fig. 3(C)][37]; and uridine vanadate docked into ribonuclease A (6rsa) [Fig. 3(D)].[38]

We examined whether GEMDOCK could yield the correct answer in less than 10 runs for the 79 correct conformations of the 100 test complexes. Figure 4 shows that GEMDOCK obtained the correct solutions for 25 complexes after one run (i.e., the percentage of success was 100%), while a total of 54 and 73 complexes were predicted correctly after 2 (i.e., $\geq$ 50%) and 5 ($\geq$ 20%) runs, respectively. When the structure water molecules were retained in the binding area, GEMDOCK yielded the correct conformation for 29 complexes in a single run, and for 62 and 80 complexes after 2 and 5 runs, respectively.

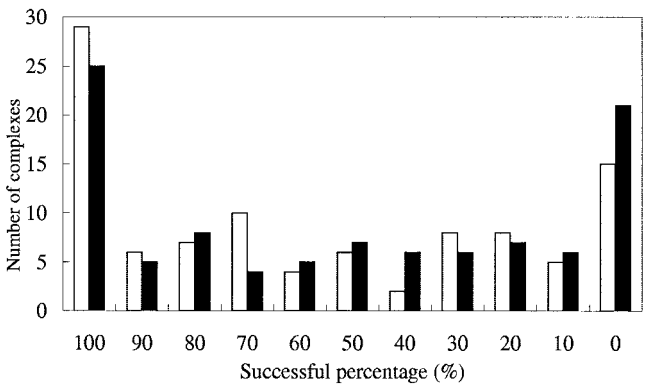


Fig. 4. The successful percentage of GEMDOCK for retaining (white bar) or removing (black bar) structure water molecules in 100 test complexes.

As shown in Figure 5, the factors causing GEMDOCK to generate the 21 unacceptable solutions shown in Tables V and X (i.e., RMSD value $> 2.0$ Å) can be roughly divided into 5 categories. The first category contains solutions when structure water molecules were removed from the binding site [Fig. 5(A), 6rnt).[39] In the second category, the ligands were large (i.e., number of heavy atoms) and highly flexible (i.e., the number of rotatable bonds) [Fig. 5(C), 1rne).[40] The ligand groups in members of the third category had a specific geometry [Fig. 5(C), 1ive).[41] The fourth factor is that some specific protein–ligand interactions were not considered in our energy model, such as the interactions **I..O** in 1eta [Fig. 5(D)][42] and **NH..pi** in 1mcr.[43] The final factor is that our scoring function could not discriminate between native and non-native conformations. Judging by these incorrectly docked solutions and crystal structures, GEMDOCK often inferred more hydrogen bonds than are in the native states to minimize the

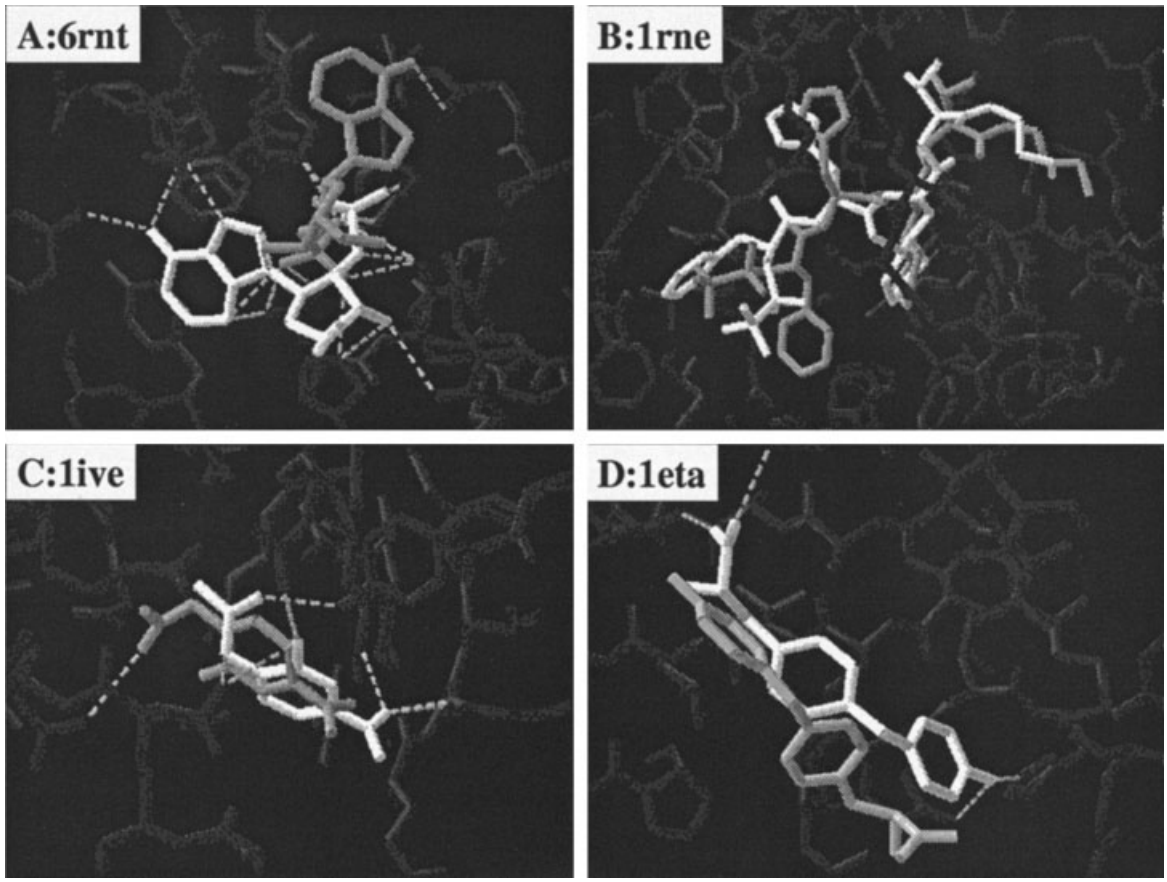


Fig. 5. GEMDOCK results for four factors for unacceptable examples (i.e., RMSD value $>$ 2.0 Å). (**A**) The structure water molecule was removed from the binding site (6rnt). (**B**) The ligand was large and highly flexible (1rne). (**C**) A specific geometry of the ligand functional group in protein structures (1ive). (**D**) A specific protein–ligand interaction was not considered in our scoring function (such as the interaction **I..O** in 1eta). The docked and crystal ligand conformations are white and gray, respectively, and the white dotted lines indicate hydrogen bonds.

**TABLE X. Unacceptable Complexes of GEMDOCK and GOLD on 100 Complexes**

| | GEMDOCK (this study) | |
|---|---|---|
| GOLD[a] (29[b]) | No water[c] (21) | Water[d] (15) |
| 1aaq **1acl** 1acj 1ack 1baf **1did** 1eap 1eed **1eta** 1etr 1hdc 1hri **1icn 1igj** 1lic **1mcr 1mup** 1nis 1rds 1rob 1tdb 2ak3 2mcp **2mth 2plv** 2r07 **3cla** 4fab 6rsa | **1acl** 1apt 1blh **1did 1eta** 1hef **1icn 1igj** 1ive **1mcr 1mup** 1rne 1xie **2mth** 2phh **2plv** 2yhx **3cla** 3gch 4est 6rnt | **1eta** 1hef **1igj** 1ive **1mcr 1mup** 1poc 1rne **2mth** 2phh **2plv** 2yhx 3gch 4est 5p2p |

[a]Results were taken directly from the original article.[13]
[b]Number of unacceptable complexes.
[c,d]Structure water molecules were removed from or retained in the binding site, respectively.

docking energy based on our energy function [Eq. (1)]. Below, we analyze the first 3 factors, and the others will be analyzed in the Discussion section.

Among these 100 complexes, there were 17 complexes with metal ions and 84 complexes with structure water molecules. When the water molecules in the binding site were retained, 6 incorrectly predicted complexes (1blh, 1did, 1icn, 1xie, 3cla, and 6rnt) became acceptable solutions for GEMDOCK (Fig. 6). In general, GEMDOCK consistently improved the docking accuracy when structure water molecules or metal ions are retained. The average RMSD value was 1.18 Å, and the success rate was 85% when the structure water molecules were retained in the binding site, whereas poorer values of 1.78 Å and 79%, respectively, where obtained when structure water were removed. In general, structure water should not be considered when establishing docking benchmarks, because this information is not directly applicable to the modeling of new ligands or high-throughput docking of databases. The treatment of structure water molecules in docking and drug-screening studies is still under investigation.[20,23,44]

Figures 5(A) and 7(A) show the docked conformations of 2′-adenylic acid (2AM) into a ribonuclease (6rnt) following removal and retention of the structure water molecules,

respectively. In the ribonuclease, the adenine of adenosine 2–monophosphate (2′-AMP) was located in a subsite that accommodates the base of the nucleoside downstream of the scissile phosphodiester bond.[39] Although our docked conformation [Fig. 5(A)] had more hydrogen bonds than the native binding structure, the adenine group was located in the opposite orientation when structure water molecules were removed. On the other hand, GEMDOCK generated the correct conformations [Fig. 7(A)] if the water molecules were retained. We observed that structure water molecules often formed hydrogen bonds with ligand atoms and thereby became the search space constraint by which the possible docked orientations was reduced. As shown in Figure 7(A), the ligand 2AM formed hydrogen bonds with the 153rd, 163rd, 182nd, 185th, and 192nd structure water molecules.

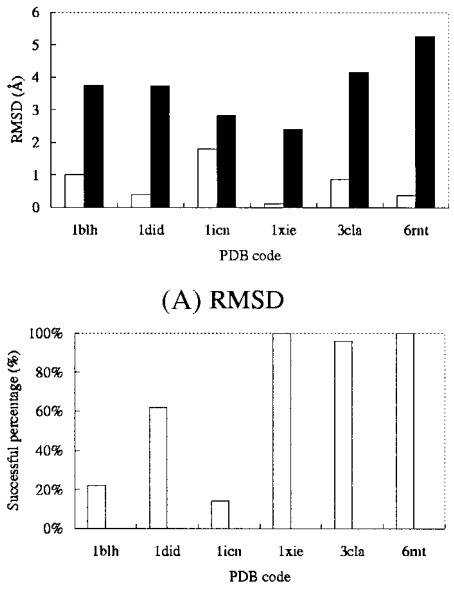


Fig. 6. (**A**) GEMDOCK results for retaining (white bar) or removing (black bar) structure water molecules in 6 complexes for which GEMDOCK yielded a significantly different performance. (**B**) The successful percentages are zero for 6 complexes when structure water molecules were removed.

GEMDOCK performance was somewhat influenced by ligand parameters such as size, flexibility, and polarity (i.e., factors affecting the number of hydrogen bonds and electrostatic interactions with proteins). For large and flexible ligands, GEMDOCK yielded low successful percentage (e.g., 1aqq, 1poc, and 5p2p) or failed to identify correct conformations (1rne, 2phh, and 2plv). All of these complexes have more than 15 single bonds. Fortunately, GEMDOCK generated correct solutions (such as 1rne and 2plv) and improved the successful percentage (such as 1aqq and 5p2p) when the population size was set to 500. Figures 5(B) and 7(B) illustrate the docked conformations for 1rne when the population size was 300 and 500, respectively. The CPU time required was roughly proportional to the population size. For protein structure problems, GEMDOCK was also slightly influenced by the protein resolution.

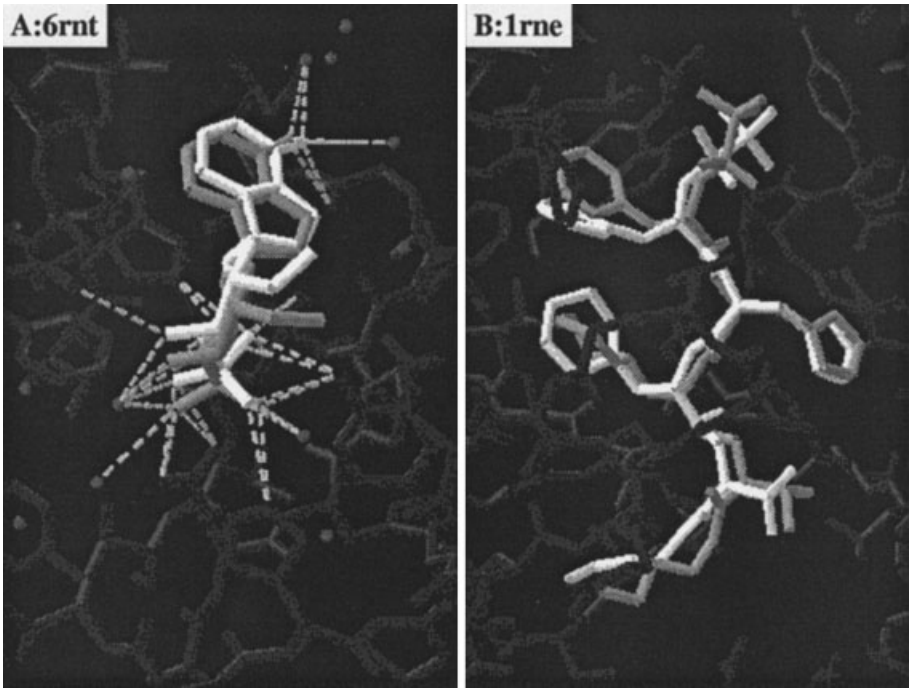


Fig. 7. GEMDOCK results for improving two kinds of unacceptable solutions. (**A**) Structure water molecules were included. (**B**) The population size was enlarged for 1rne. The docked and crystal ligand conformations are white and gray, respectively, and the white dotted lines indicate hydrogen bonds.

| Protein / Ligand | 1hih | 1hvi | 1hvj | 1hvk | 1hvl | 1hvs | 1hvr | 4phv |
|---|---|---|---|---|---|---|---|---|
| 1hih.C20 | 0 | | | | | | | |
| 1hvi.A77 | 0.42 | 0 | | | | | | |
| 1hvj.A78 | 0.44 | 0.2 | 0 | | | | | |
| 1hvk.A79 | 0.47 | 0.25 | 0.22 | 0 | | | | |
| 1hvl.A76 | 0.45 | 0.21 | 0.18 | 0.17 | 0 | | | |
| 1hvs.A77 | 0.44 | 0.25 | 0.27 | 0.52 | 0.23 | 0 | | |
| 1hvr.XK2 | 0.84 | 0.79 | 0.77 | 0.71 | 0.71 | 0.57 | 0 | |
| 4phv.VAC | 0.67 | 0.66 | 0.65 | 0.61 | 0.6 | 0.71 | 0.49 | 0 |

(A)

| Protein / Ligand | 1dbb | 1dbj | 1dbk | 1dbm | 2dbl |
|---|---|---|---|---|---|
| 1dbb.STR | 0 | | | | |
| 1dbj.AE2 | 0.74 | 0 | | | |
| 1dbk.ANO | 0.69 | 0.7 | 0 | | |
| 1dbm.SIH | 0.62 | 0.71 | 0.72 | 0 | |
| 2dbl.S5H | 0.78 | 0.81 | 0.78 | 0.73 | 0 |

(B)

Fig. 8. Cross-RMSD matrices of all paired PDB entries for (**A**) 8 complexes of the HIV-1 protease, and (**B**) 5 complexes of the Fab' fragment.

GEMDOCK most often yielded incorrect solutions in cases of unusual ligand–receptor binding interactions or when complexes (1apt, 1blh, 1hef, 1ive, and 3gch) contained specific geometric functional groups, as described by Jones et al.[13] We will investigate two approaches aimed at reducing this shortcoming in the future. The first involves incorporating the hydrogen-bond strength in the hydrogen-bonding energy calculation,[7] while the second considers some specific functional groups in our scoring function, such as those defined in GOLD[13] that yielded good conformations for these complexes (Table X).

### Comparing GEMDOCK With Other Methods

In general, it is neither straightforward nor completely fair to compare the results of different protein–ligand docking methods given that each employs different accuracy measures, energy functions, and test complexes. Furthermore, with the exception of GOLD[13] and FlexX,[17] these methods have only been tested on small data sets (< 20 complexes). Table VIII compares GEMDOCK with five docking methods[13,19,45–47] using 7 selected test complexes, and Table IX compares GEMDOCK with GOLD and FlexX using the data set of 100 complexes. The unacceptable solutions of GEMDOCK and GOLD are shown in Table X.

We selected a minimal set of 7 complexes (Table VIII) encompassing some of the common systems tested by different methods to compare GEMDOCK with these other methods.[13,19,45–47] GOLD and AutoDock[45] are genetic-based approaches, FlexX is an incremental approach, MSA[19] is a multistep strategy approach, and ConsDock[46] is a consensus docking method combining 3 widely used docking methods (DOCK, FlexX, and GOLD). As shown in Table VIII, GEMDOCK is very comparable to these approaches on this test set. Our results were derived from 10 docking runs, and the solution at the first rank was considered for each test complex. The energy values of docked conformations (Table VIII) obtained with GEMDOCK were often lower than those for native crystal-binding states (Table VI). The successful percentages of all the test complexes exceeded 50% except for the complex 1dwd.

As shown in Table IX, GEMDOCK yielded a 79% success rate based on the top-ranked solutions with RMSD values less than 2 Å. In contrast, GOLD[13] yielded a 71% success rate in identifying the experimental binding model based on their assessment categories, and the rate was 66% if based on the top-ranked solutions with RMSD values less than 2 Å. FlexX[17] achieved 70% and 46.5% success rates for solutions at any rank and the first rank, respectively. A major problem of GOLD is that it was often sensitive to docking hydrophobic ligands,[13] and FlexX was often sensitive to the choice of the based fragment as well as the number of fragments.[17] Experiments show that GEMDOCK was able to reduce the negative effects in these factors. As shown in Table X, GEMDOCK was negatively influenced by protein structures containing poorly determined ligand group geometries, such as 1hef, 1ive, and 3gch. By contrast, GOLD[13] yielded good solutions for these complexes. However, GOLD was unable to make a prediction for complex 1acl, since it has no polar group. In this case, GEMDOCK yielded docked conformations with RMSD values of 2.74 Å and 1.23 Å when the structure water molecules in the binding site were removed and retained, respectively.

### Cross-Docking Results

We evaluated GEMDOCK with respect to unbound complexes in cases in which protein structure undergoes small changes in motion during the process of docking.[48] Two ensembles of protein structures were used, namely, 8 complexes of the HIV-1 protease[23] and 5 complexes of the Fab' fragment protein of monoclonal antibody DB3. The respective complexes within each ensemble differ only by small variations in the side-chains and loops within the active site. Figure 8 shows the cross-RMSD matrices (e.g., protein heavy atoms of the binding site) of all paired PDB

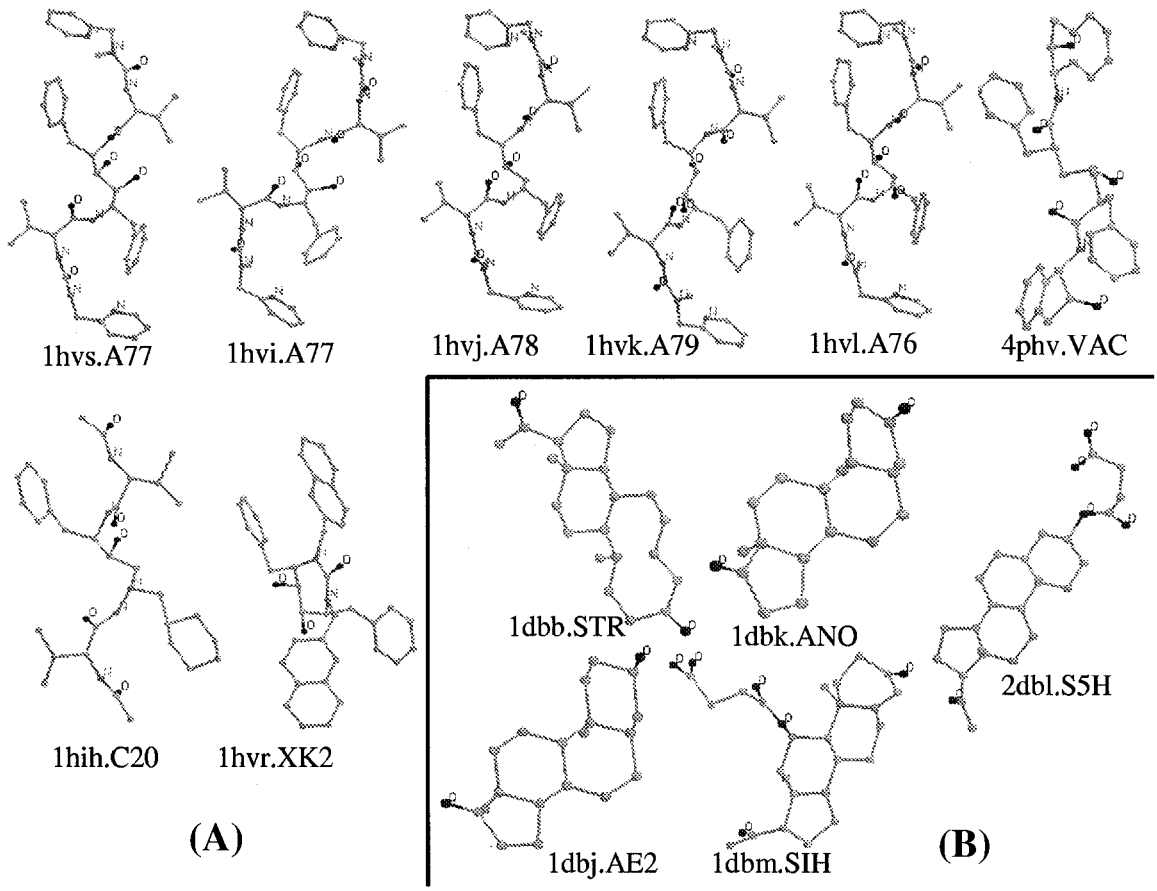


Fig. 9. Cross-docking ligands bound to (**A**) 8 complexes of the immunodeficiency virus type 1 (HIV-1) protease, and (**B**) 5 complexes of the Fab' fragment of the monoclonal antibody DB3. Coordinates for each complex were obtained from the PDB, using the accession codes given here. The 4 characters and 3 characters separated by a period denote the PDB code and the ligand name in the PDB, respectively.

entries that indicate the protein flexibility in the binding site. The largest RMSD is 0.84 Å, and the smallest is 0.17 Å. The number of ligand heavy atoms in the HIV-1 ensemble is between 41 and 58. Figure 9(A and B) shows the ligand structures of the HIV-1 protease and the Fab' fragment, respectively. The 8 inhibitors of the HIV-1 protease can be divided into 2 groups according to the ligand size [i.e., large ligand (1hvi.A77, 1hvj.A78, 1hvk.A79, 1hvl.A76, and 1hvs.A77) and medium ligand (1hih.C20, 1hvr.XK2, and 4phv.VAC)]. We denoted each ligand systematically using 4 characters followed by 3 characters. For example, in the ligand "4phv.VAC," "4phv" denotes the PDB code and "VAC" is the ligand name in the PDB.

HIV-1 protease was identified as a crucial target for designing drugs against acquired human immunodeficiency syndrome.[49] It is an aspartyl protease that acts to cleave the nascent polyproteins into functional proteins during viral replication. The action of the protease is essential for viral maturation and infectivity. The Fab' fragment of the monoclonal antibody DB3 binds a subgroup of progesterone-like steroids that are structurally distinct.[50] The DB3 antibody, a member of the antiprogesterone monoclonal antibodies, can lead to a temporary inhibition of the progesterone-dependent processes during early pregnancy in mice.

Figure 10 shows the results of the cross-docking experiments in which all ligands of each protein ensemble were docked into each complex of the ensemble. For example, we obtained 64 docked results when each of 8 ligands was docked into each of 8 complexes of the HIV-1 protease. The RMSD values of all diagonal docked conformations (docking each ligand back into its respective complexes) are less than 2.0 Å (Fig. 10). GEMDOCK also yielded good results for most of the cross-docking examples (off the diagonal). For the HIV-1 ensemble [Fig. 10(A)], GEMDOCK yielded 45 docked conformations (70.3%) with RMSD values less than 2.0 Å. The 5 large ligands (1hvi.A77, 1hvj.A78, 1hvk.A79, 1hvl.A76, and 1hvs.A77) could not be successfully docked into the complexes (1hih, 1hvr, and 4phv) with medium ligands, whereas all 3 medium ligands could be docked into the 8 complexes. For all 25 docked conformations of the Fab' fragment ensemble, GEMDOCK yielded stable ligand docking results for all 25 docked conformations of the Fab' fragment ensemble. The average RMSD was 0.96 Å, while the largest RMSD value was 1.74 Å [for docking the ligand 1dbm.SIH, the largest ligand of the 5 Fab' ligands, into the complex 2dbl; Fig. 10(B)]. GEMDOCK seemed more stable than FlexX[17] with respect to the Fab' fragment ensemble. These results suggest that GEMDOCK may be useful for addressing the problem

| Protein / Ligand | 1hih | 1hvi | 1hvj | 1hvk | 1hvl | 1hvs | 1hvr | 4phv |
|---|---|---|---|---|---|---|---|---|
| 1hih.C20 | 0.81 | 1.30 | 1.51 | 1.31 | 1.10 | 1.02 | 1.08 | 1.03 |
| 1hvi.A77 | 9.54 | 1.71 | 1.91 | 1.92 | 2.43 | 2.27 | 4.30 | 5.22 |
| 1hvj.A78 | 3.97 | 2.85 | 1.60 | 1.31 | 1.96 | 1.73 | 3.83 | 5.72 |
| 1hvk.A79 | 9.10 | 1.51 | 1.70 | 1.52 | 1.89 | 2.10 | 2.31 | 7.28 |
| 1hvl.A76 | 5.96 | 1.81 | 1.32 | 1.62 | 1.25 | 1.21 | 6.16 | 6.14 |
| 1hvs.A77 | 6.63 | 1.58 | 1.81 | 1.99 | 4.07 | 1.65 | 3.35 | 6.03 |
| 1hvr.XK2 | 1.01 | 1.06 | 1.23 | 1.44 | 1.41 | 1.04 | 0.42 | 0.77 |
| 4phv.VAC | 0.89 | 1.06 | 0.88 | 0.79 | 0.74 | 0.92 | 0.84 | 0.68 |

**(A)**

| Protein / Ligand | 1dbb | 1dbj | 1dbk | 1dbm | 2dbl |
|---|---|---|---|---|---|
| 1dbb.STR | 0.61 | 1.01 | 0.96 | 0.95 | 0.75 |
| 1dbj.AE2 | 1.04 | 0.93 | 0.88 | 0.78 | 0.61 |
| 1dbk.ANO | 1.22 | 0.86 | 1.03 | 0.91 | 0.85 |
| 1dbm.SIH | 1.52 | 1.04 | 1.05 | 1.21 | 1.74 |
| 2dbl.S5H | 0.89 | 0.73 | 0.78 | 1.14 | 0.59 |

**(B)**

Fig. 10. Cross-docking results of all paired experiments for (**A**) 8 complexes of the HIV-1 protease, and (**B**) 5 complexes of the Fab' fragment. The color-coded table shows the gray scale of RMSD values for each ligand (row) docked into each protein (column) of a protein ensemble.

when slight structural changes occur in the protein during the docking process.

## DISCUSSION

One of main objectives of this study was to evaluate whether GEMDOCK, a tool that is almost completely automatic, is robust enough to predict docked structures when the ligand is flexible. GEMDOCK achieved a 79% success rate when tested on a set of 100 complexes selected from the PDB. The effectiveness of GEMDOCK on a problem in which the protein structure changes slightly during the docking process was illustrated by testing this approach on two cross-docking experimental sets. Although GEMDOCK is a very promising tool, it failed on 21 of the 100 test complexes. Upon inspection of the incorrectly predicted ligand structures, we concluded that GEMDOCK was most likely to fail if our scoring function could not discriminate between the native and non-native states. Below, we analyze the characteristics of the scoring function and our evolutionary approach.

Table XI shows the results of GEMDOCK using various scoring functions, including our empirical scoring function [Eq. (1)], a simplified AMBER-based, and RMSD-based scoring functions [Eq. (16)]. For our empirical scoring function, we tested GEMDOCK using various factors (such as the electrostatic energy, $E_{inter}$, $E_{intra}$, and $E_{penal}$) and parameter values (see Materials and Methods section). Since our previous study[29] required additional software to determine AMBER parameter values of ligand atoms, here we used a simplified AMBER-based scoring function derived from AutoDock.[45] Finally, we tested GEMDOCK with respect to the RMSD scoring function [Eq. (16); considered a perfected-fitness function] to evaluate the program's performance and search behavior.

As shown in Table XI, the success rates were 79% and 55% using the empirical and the simplified AMBER-based scoring functions, respectively. GEMDOCK indeed approached perfect prediction when the RMSD scoring function was used. In the empirical scoring function the element $F(r_{ij}^{B_{ij}})$ in the $E_{inter}$ [Eq. (2)] is the main factor that determines GEMDOCK performance. GEMDOCK yields a similar success rate when the term $E_{intra}$ is removed; however, the docked ligand conformations may be unreasonable for some test complexes. The electrostatic energy and $E_{penal}$ are minor factors that influence certain docking cases. It is noteworthy that AutoDock[45] as well as other studies[15,29] that used an AMBER-based scoring function with their tuned parameters achieved good performance with their test systems and in some applications.

GEMDOCK is able to improve the quality of ligand docking by considering the electrostatic energy if electrostatic interactions consist of in the protein–ligand complexes, such as 1tdb, 2mcp, and 1hdc. Figure 11 shows an example (1tdb)[51] that describes the influence of the electrostatic energy. The functional group $PO_3^-$ in the docked conformation (white) was opposite to that in the crystal ligand structure (gray) when the electrostatic energy was excluded. In contrast, GEMDOCK generated the correct conformation when the electrostatic energy was included, because an electrostatic interaction is formed between the $PO_3^-$ group and the receptor.

Figure 12 shows two typical categories of relationships between the energies and the RMSD values based on 100 independent docking runs. In the first category [Fig.12(A)], GEMDOCK robustly obtained correct ligand conformations, and the RMSD value is low when the scoring value is low. When GEMDOCK generates this pattern of results for a given docking system (such as 4dfr), our scoring function discriminates between native and non-native conformations, thereby allowing GEMDOCK to often achieve the correct docked conformations. On the other hand, Figure 12(B) indicates that GEMDOCK attains docked conformations with diverse RMSD values ($> 3.0$ Å) at a similar scoring value. In such docking systems (such as 1ive), our scoring function is often unable to discrimi-

**TABLE XI. GEMDOCK Results Using Different Scoring Functions on 100 Complexes**

| RMSD (Å) | Empirical-based function [Eq. (1)] Completed $E_{tot}$ | $E_{tot}$ without electrostatic energy | $E_{tot}$ without $E_{intra}$ | Simplified AMBER-based function[a] | RMSD [Eq. (16)] as scoring function |
|---|---|---|---|---|---|
| ≤0.5 | 23% | 21% | 22% | 10% | 100% |
| >0.5, ≤1.0 | 36% | 35% | 34% | 25% | 0% |
| >1.0, ≤1.5 | 9% | 10% | 12% | 15% | 0% |
| >1.5, ≤2.0 | 11% | 7% | 9% | 5% | 0% |
| >2.0, ≤2.5 | 3% | 5% | 4% | 4% | 0% |
| >2.5, ≤3.0 | 5% | 7% | 5% | 10% | 0% |
| >3.0 | 13% | 15% | 14% | 31% | 0% |

[a]Using our previous scoring function[29] with simplified AMBER-based parameters.[45]


Fig. 11. The influence of the electrostatic energy in the scoring function [Eq. (2)] for the complex 1tdb. When the electrostatic energy was not considered, GEMDOCK yielded the wrong docked conformation (white) in which the $PO_3^-$ was on the opposite side of the corresponding group in the crystal ligand (gray).

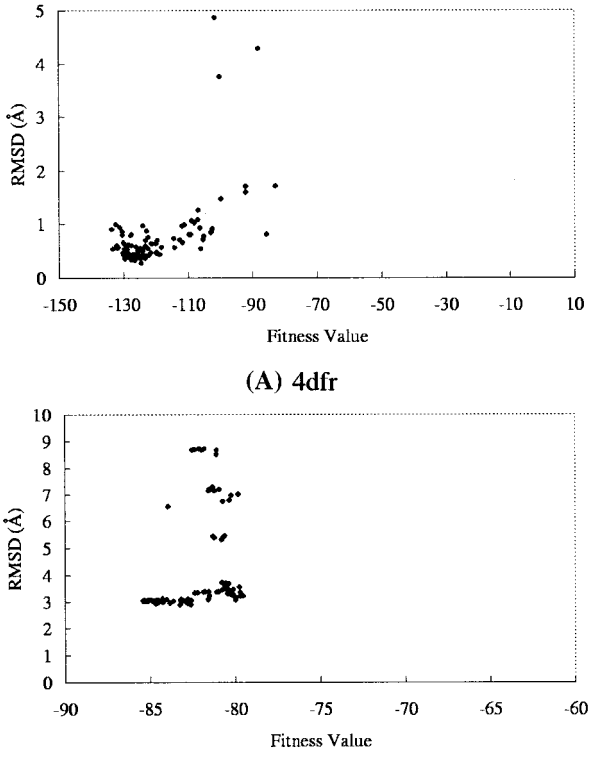


Fig. 12. Typical relationships between the values of the scoring function and the RMSD of (**A**) a correct, and (**B**) an incorrect docking complex with 100 docking runs. For a correct docking complex (4dfr), GEMDOCK yielded 93 solutions with RMSD values of 2.0 Å. The RMSD values were often more than 3.0 Å for incorrect docking complexes (e.g., 1ive).

nate between correct and incorrect binding conformations; that is to say, the lowest energy structure does not promise to produce good docked conformations. Figure 13 shows that the inclusion or removal of structure water molecules affects the scoring function. For the complex 6rnt, GEMDOCK obtained 99 solutions with RMSD values < 2.0 Å among 100 runs when the structure water molecules were retained. On the other hand, the program behaved differently if they were removed.

Given the uncertainty in the scoring function, the robustness of GEMDOCK is difficult to assess. To address this question, we exploited the high adaptability of GEMDOCK by simply replacing the empirical scoring function with a perfect scoring function (i.e., one that would produce zero RMSD in ligand heavy atom positions). As shown in Table XI, when the RMSD scoring function [Eq. (16)] was used, GEMDOCK could indeed approach perfect prediction. Not only was the best RMSD value of docked structures below 0.05 Å, but also the average RMSD value of docked structures was less than 0.23 Å. The successful percentage was 100% for each test complex. It is also noteworthy that GEMDOCK converges much faster with the perfect scoring function (< 20 s for a docking run).

To illustrate the effectiveness of our evolutionary approach for conformational sampling, GEMDOCK was compared with 5 conformational sampling approaches tested using similar empirical scoring functions.[20] These approaches included simulated annealing (SA), evolutionary programming (EP), Tabu search (TS), genetic algorithm (GA), and random search (RS). We followed the specific criteria[20] (i.e., the percentage of success in 500 trials that yielded a solution within 1.5 Å RMSD) to obtain the GEMDOCK results for comparison with those from the other approaches, as reported in the previous study.[20] As shown in Table XII, GEMDOCK performed the best among

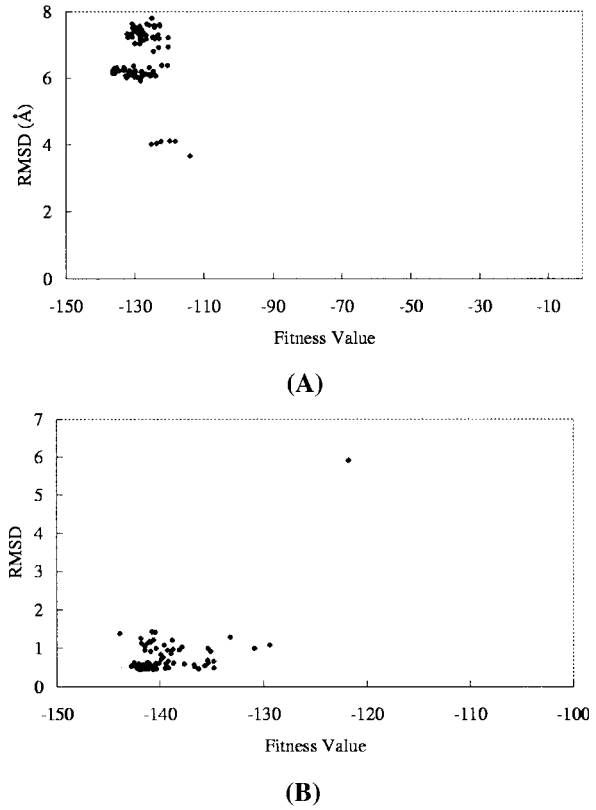


Fig. 13. GEMDOCK results for (**A**) removing, and (**B**) retaining the structure water molecules in the complex 6rnt with 100 independent runs.

**TABLE XII. Comparing GEMDOCK With 5 Conformational Sampling Approaches Based on Successful Percentages**[a]

| PDB code | GEMDOCK | SA[b] | EP[b] | TS[b] | GA[b] | RS[b] |
|---|---|---|---|---|---|---|
| 1etr | 55 | 30 | 21 | 39 | 13 | 3 |
| 1ets | 21 | 3 | 9 | 8 | 11 | 2 |
| 1hvr | 86 | 65 | 54 | 58 | 59 | 2 |
| 1nsd | 98 | 40 | 64 | 88 | 57 | 6 |
| 3dfr | 92 | 90 | 76 | 93 | 76 | 9 |

[a]The percentage of 500 trials that found a solution within 1.5 Å RMSD of the crystal ligand conformation.
[b]These results were taken from Westhead et al.[20]

the approaches with respect to this test set, while the random search was the poorest.

To estimate the orientational and conformational search spaces of GEMDOCK, we assume that unique orientation distances and angles in the search cube differed by 0.3 Å and 0.06 rad, respectively, and that the unique ligand conformations differed by 0.06 rad. The sizes of the orientational search space was $4.70 \times 10^{11}$ for the minimum cube (2mcp) and $2.86 \times 10^{12}$ for the maximum cube (1rne). The size of the conformational search space is $3.02 \times 10^{48}$ for the complex 1rne, which has 24 single bonds. Therefore, the maximum size of the search space (1rne) was $8.65 \times 10^{60}$ among these 100 test complexes. GEMDOCK applied the rotamer-based mutation operator to reduce the search space of ligand structure conformations. It is possible that the GEMDOCK search space may exceed the estimated value, because it continuously evolves the orientations and conformations. Fortunately, GEMDOCK is often able to find correct conformations within a reasonable timeframe when the scoring function can discriminate between native states and non-native docked conformations. For example, our approach yields correct conformations for a large and high flexible ligand [Fig. 7(B)] and for a large protein binding site (Table VII) by enlarging the population size (if our scoring function is satisfactory). At the same time, GEMDOCK yielded perfect docked conformations when the RMSD scoring function was used (Table XI). These results suggest that both our evolutionary approach (see Materials and Methods section) and the scoring function are important factors for determining docking accuracy in GEMDOCK. The versatility of GEMDOCK may allow us to systematically improve the forms and parameters of the energy function for molecular recognition.

Despite its success with the test sets, GEMDOCK still exhibits some limitations. First, the approach is somewhat time-consuming. Second, some protein–ligand interactions and specific ligand geometries were not considered in our fitness function. Third, the binding site is considered to be essentially rigid. Last, the size and location of the active site are manually assigned for unbound docking systems. In the future we will address these shortcomings by (1) developing a rapid energy evaluation using grid-based potentials to speed up convergence; (2) incorporating important functional group interactions between ligands and proteins into our empirical scoring function as in GOLD[13]; (3) incorporating the hydrogen-bond strength for calculating hydrogen-bonding energies[7]; (4) testing GEMDOCK using different scoring functions (e.g., Chemscore,[52] PMF,[53] and the GOLD scoring function[13]) to systematically improve docking accuracy; and (5) considering flexible side-chains and small motions within loops of the protein active site.

In summary, we have developed an automatic tool for flexible ligand docking by applying numerous enhancements and modifications to the original technique. 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. Experiments on 100 test systems and two cross-docking experimental sets verify that the proposed approach is robust and adaptable to flexible ligand docking. The versatility of GEMDOCK may allow us to systematically improve the forms and parameters of the energy function for molecular recognition.

## REFERENCES

1. Kuntz ID. Structure-based strategies for drug design and discovery. Science 1992;257:1078–1082.
2. Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins 2002;47:409–443.
3. Muegge I, Rarey M. Small molecule docking and scoring. In: Boyd DB, editor. Reviews in computational chemistry. Vol. 17. Weinheim: Wiley-VCH; 2001. p 1–60.
4. Gohlke H, Hendlich M, Klebe G. Knowledge-based scoring function to predict protein–ligand interactions. J Mol Biol 2000;295: 337–356.

5. Verdonk ML, Cole JC, Watson P, Gillet V, Willett P. SuperStar: improved knowledge-based interaction fields for protein binding sites. J Mol Biol 2001;307:841–859.
6. Gehlhaar DK, Verkhivker GM, Rejto P, Sherman CJ, Fogel DB, Fogel LJ, Freer ST. Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. Chem Biol 1995;2:317–324.
7. Verkhivker GM, Bouzida D, Gehlhaar DK, Rejto PA, Arthurs S, Colson AB, Freer ST, Larson V, Luty BA, Marrone T, Rose PW. Deciphering common failures in molecular docking of ligand–protein complexes. J Comput Aided Mol Des 2000;14:531–551.
8. Taylor JS, Burnett RM. DARWIN: a program for docking flexible molecules. Proteins 2000;41:173–191.
9. Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, Profeta SJ, Weiner P. A new force field for molecular mechanical simulation of nucleic acids and proteins. J Am Chem Soc 1984;106: 765–784.
10. Shoichet BK, Leach AR, Kuntz ID. Ligand solvation in molecular docking. Proteins 1999;34:4–16.
11. Miller DW, Dill KA. Ligand binding to proteins: the binding landscape model. Protein Sci 1997;6:2166–2179.
12. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. A geometric approach to macromolecular–ligand interactions. J Mol Biol 1982;161:269–288.
13. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. J Mol Biol 1997;267:727–748.
14. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ. Automated docking using a Lamarckian genetic algorithm and empirical binding free energy function. J Comput Chem 1998;19:1639–1662.
15. Oshiro CM, Kuntz ID, Dixon JS. Flexible ligand docking using a genetic algorithm. J Comput Aided Mol Des 1995;9:113–130.
16. Sherman CJ, Ogden RC, Freer ST. De Novo design of enzyme inhibitors by Monte Carlo ligand generation. J Med Chem 1995;38: 466–472.
17. Kramer B, Rarey M, Lengauer T. Evaluation of the FlexX incremental construction algorithm for protein–ligand docking. Proteins 1999;37:228–241.
18. Palma PN, Krippahl L, Wampler JE, Moura JJG. Bigger: a new soft docking algorithm for predicting protein interactions. Proteins 2000;39:178–194.
19. Wang J, Kollman PA, Kuntz ID. Flexible ligand docking: A multistep strategy approach. Proteins 1999;36:1–19.
20. Westhead DR, Clark DE, Murray CW. A comparison of heuristic search algorithms for molecular docking. J Comput Aided Mol Des 1997;11:209–228.
21. Claussen H, Buning C, Rarey M, Lengauer T. FlexE: efficient molecular docking considering protein structure variations. J Mol Biol 2001;308:377–395.
22. Leach AR. Ligand docking to proteins with discrete side-chain flexibility. J Mol Biol 1994;235:345–356.
23. Osterberg F, Morris GM, Sanner MF, Olson AJ, Goodsell DS. Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. Proteins 2002;46:34–40.
24. Schaffer L, Verkhivker GM. Predicting structural effects in HIV-1 protease mutant complexes with flexible ligand docking and protein side-chain optimization. Proteins 1998;33:295–310.
25. Back T. Evolutionary algorithms in theory and practice. New York: Oxford University Press; 1996.
26. Fogel DB. Evolutionary computation: toward a new philosophy of machine intelligence. New York: IEEE Press; 1995.
27. Goldberg DE. Genetic algorithms in search, optimization and machine learning. Reading, MA: Addison-Wesley; 1989.
28. Yang JM, Horng JT, Lin CJ, Kao CY. Optical coating designs using an evolutionary algorithm. Evol Comput 2001;9:421–443.
29. Yang JM, Kao CY. Flexible ligand docking using a robust evolutionary algorithm. J Comput Chem 2000;21:988–998.
30. Yang JM, Kao CY. A robust evolutionary algorithm for training neural networks. Neural Comput Applic 2001;10:214–230.
31. Yang JM, Tsai CH, Hwang MJ, Tsai HK, Hwang JK, Kao CY. GEM: A Gaussian evolutionary method for predicting protein side-chain conformations. Protein Sci 2002;11:1897–1907.
32. Storn R, Price KV. Differential evolution: a simple and efficient heuristic for global optimization over continuous spaces. J Global Optimiz 1997;11:341–369.
33. Knegtel RMA, Antoon J, Rullmann C, Boelens R, Kaptein R. MONTY: a Monte Carlo approach to protein-DNA recognition. J Mol Biol 1994;235:318–324.
34. Wlodawer A, Vondrasek J. Inhibitors of HIV-1 protease: a major success of structure-assisted drug design. Annu Rev Biophys Biomol Struct 1998;27:249–284.
35. Zhang A, Nanni RG, Li T, Arnold GF, Oren DA, Jacobo-Molina A, Williams RL, Kamer G, Rubenstein DA, Li Y, Rozhon E, Cox S, Buontempo P, O'Connell J, Schwartz J, Miller G, Bauer B, Versace R, Pinto P, Ganguly A, Girijavallabhan V, Arnold E. Structure determination of antiviral compound sch 38057 complexed with human rhinovirus 14. J Mol Biol 1993;230:857–867.
36. Poulos TL, Howard AJ. Crystal structures of metyrapone- and phenylimidazole-inhibited complexes of cytochrome P-450cam. Biochemistry 1987;26:8165–8174.
37. Marquart M, Walter J, Deisenhofer J, Bode W, Huber R. The geometry of the reactive site and of the peptide groups in trypsin, trypsinogen and its complexes with inhibitors. Acta Crystallogr B 1983;39:480–490.
38. Borah B, Chen CW, Egan W, Miller M, Wlodawer A. Nuclear magnetic resonance and neutron-diffraction studies of the complex of ribonuclease-a with uridine vanadate, a transition-state analog. Biochemistry 1985;24:2058–2067.
39. Ding J, Koellner G, Grunert HP, Saenger W. Crystal structure of ribonuclease T1 complexed with adenosine 2′-monophosphate at 1.8-Å resolution. J Biol Chem 1991;266:15128–15134.
40. Rahuel J, Priestle JP, Gruetter MG. The crystal structure of recombinant glycosylated human renin alone and in complex with a transition state analog inhibitor. J Struct Biol 1991;107:227–236.
41. Jedrzejas MJ, Singh S, Brouillette WJ, Laver WG, Air GM, Luo M. Structures of aromatic inhibitors of influenza virus neuraminidase. Biochemistry 1995;34:3144–3151.
42. Hamilton JA, Steinrauf LK, Braden BC, Leipnieks J, Benson MD, Holmgren G, Sandgren O, Steen L. The X-ray crystal structure refinements of normal human transthyretin and the amyloidogenic Val-30 → Met variant to 1.7-Å resolution. J Biol Chem 1993;268:2416–2424.
43. Edmundson AB, Harris DL, Fan ZC, Guddat LW, Schley BT, Hanson BL, Tribbick G, Geysen HM. Principles and pitfalls in designing site-directed peptide ligands. Proteins 1993;16:246–267.
44. Pang YP, Perola E, Xu K, Prendergast FG. EUDOC: a computer program for identification of drug interaction sites in macromolecules and drug leads from chemical databases. J Comput Chem 2003;22:1750–1771.
45. Morris GM, Goodsell DS, Huey R, Olson AJ. Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. J Comput Aided Mol Des 1996;10:293–304.
46. Paul N, Rognan D. ConsDock: a new program for the consensus analysis of protein–ligand interactions. Proteins 2002;47:521–533.
47. Rarey M, Kramer B, Lengauer T, Klebe G. A fast flexible docking method using an incremental construction algorithm. J Mol Biol 1996;261:470–489.
48. Knegtel R, Kuntz I, Oshiro C. Molecular docking to ensembles of proteins. J Mol Biol 1997;266:424–440.
49. Vondrasek J, Wlodawer A. HIVdb: a database of the structures of human immunodeficiency virus protease. Proteins 2002;49:429–431.
50. Arevalo JH, Hassig CA, Stura E A, Sims MJ, Taussig MJ, Wilson IA. Structural analysis of antibody specificity: detailed comparison of five Fab'-steroid complexes. J Mol Biol 1994;241:663–690.
51. Perry KM, Carreras CW, Chang LC, Santi DV, Stroud RM. Structures of thymidylate synthase with a C-terminal deletion: role of the C-terminus in alignment of 2′-deoxyuridine 5′-monophosphate and 5,10-methylenetetrahydrofolate. Biochemistry 1993;32:7116–7125.
52. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. J Comput Aided Mol Des 1997;11:425–445.
53. Muegge I, Martin YC. A general and fast scoring function for protein–ligand interactions: a simplified potential approach. J Med Chem 1999;42:791–804.

## APPENDIX: MAIN STEPS OF GEMDOCK FOR MOLECULAR DOCKING

**Main procedure** proceeds in the following steps:

1. Initialize the protein and the ligand as follows:
   (a) Determining the size and location of the ligand binding site and removing the structure water molecules.
   (b) Assigning the atom formal charge (Table II) and the atom type (Table III) of a ligand and a receptor.
2. Fix the location of the receptor and let $g = 1$. Randomly generate initial population, $P(g)$, with $N$ solutions by initializing the orientation and conformation of a ligand related to the receptor.
3. Evaluate the scoring fitness of each solution in the population $P(g)$.
4. Generate a new quasi-population, $P1(g)$, with $N$ solutions by applying FC_Adaptive with $P(g)$ and *decreasing-based Gaussian mutation ($M_{dg}$)*.
5. Generate a new quasi-population, *Pnext*, with $N$ solutions by applying FC_Adaptive with $P1(g)$ and *self-adaptive Cauchy mutation (Mc)*. Let $g = g + 1$ and $P(g) = Pnext$.
6. Repeatedly execute from step 3 to step 5 until the terminal criteria are satisfied.

**FC_Adaptive procedure** proceeds in the following steps with two parameters, working population ($P$) and working mutation ($M_{dg}$ or $Mc$):

1. Let $C$ be an empty set ($C = \phi$). For each solution $a$, called *family father*, in working population ($P$) executes following steps: {*family competition*}
   (a) Generate $L$ docked ligand solutions (the orientation and conformation), denoted as $c1,\ldots,c^L$ by applying the recombination, rotamer mutation, differential evolution, and working mutation.
   (b) Select the one, $c^{best}$, with the lowest scoring value from the union set: $a$ and $c^1,\ldots,c^L$.
   (c) Add the $c^{best}$ into the set $C$.
2. Return the set $C$ with $N$ solutions.