Journal of Computer-Aided Molecular Design, **13:** 435–451, 1999.
KLUWER/ESCOM

MCDOCK: A Monte Carlo simulation approach to the molecular docking problem

Ming Liu & Shaomeng Wang*
The Drug Discovery Program, Georgetown Institute for Cognitive and Computational Science, Georgetown University Medical Center, 3970 Reservoir Road, Research Building EPO7, Washington, DC 20007, U.S.A.

Received 19 October 1998; Accepted 19 January 1999

Key words: flexible ligand docking, ligand and protein interaction, molecular recognition, Monte Carlo simulation, structure-based drug discovery

Summary

Prediction of the binding mode of a ligand (a drug molecule) to its macromolecular receptor, or molecular docking, is an important problem in rational drug design. We have developed a new docking method in which a non-conventional Monte Carlo (MC) simulation technique is employed. A computer program, MCDOCK, was developed to carry out the molecular docking operation automatically. The current version of the MCDOCK program (version 1.0) allows for the full flexibility of ligands in the docking calculations. The scoring function used in MCDOCK is the sum of the interaction energy between the ligand and its receptor, and the conformational energy of the ligand. To validate the MCDOCK method, 19 small ligands, the binding modes of which had been determined experimentally using X-ray diffraction, were docked into their receptor binding sites. To produce statistically significant results, 20 MCDOCK runs were performed for each protein–ligand complex. It was found that a significant percentage of these MCDOCK runs converge to the experimentally observed binding mode. The root-mean-square (rms) of all non-hydrogen atoms of the ligand between the predicted and experimental binding modes ranges from 0.25 to 1.84 Å for these 19 cases. The computational time for each run on an SGI Indigo2/R10000 varies from less than 1 min to 15 min, depending upon the size and the flexibility of the ligands. Thus MCDOCK may be used to predict the precise binding mode of ligands in lead optimization and to discover novel lead compounds through structure-based database searching.

Introduction

There are now more than 7000 three-dimensional (3D) coordinates of proteins or nucleic acids determined by X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy available from the Protein Data Bank (PDB) and the total number of 3D structures is increasing rapidly. Many of these macromolecules serve as potential therapeutic targets, and the availability of their 3D structures offers us unprecedented opportunities for structure-based drug design and discovery. To effectively carry out structure-based drug design, two fundamental questions need to be addressed. The first question is how a drug (or ligand) fits into the receptor binding site (the docking problem), and the second question is how well it binds to its receptor (the binding affinity prediction problem). In this paper, we present a new method to address the docking problem.

Numerous docking methods dating back to the 1980s were developed, and several good reviews were written recently to summarize and evaluate these methods [1–6]. A number of principles have been used to guide the development of these docking methods. First, it was recognized that when a drug binds to its receptor with high affinity, these two molecules are highly complementary to each other in terms of both their shape and charge characteristics. Some docking programs [7–9] were developed based upon this prin-

*To whom correspondence should be addressed. E-mail: wangs@giccs.georgetown.edu

ciple. Secondly, when a drug binds to its receptor at physiological conditions, the equilibrium state must have the lowest free energy. Based upon this theory, scoring functions have been designed to represent the free energy change of the system. Approximations are necessary since it is difficult to precisely calculate the free energy change of the system. For example, if the entropy and the solvent effect contributions to the free energy change of the system are ignored, the sum of the interaction energy between the ligand and the receptor and the internal conformational energies of the ligand may be used as the scoring function. This type of scoring functions was used in a number of docking methods [10–16]. An empirical function was also developed to represent the free energy change of the system [17–19] and has been used in docking studies such as in the program FLEXX [20].

In addition to an adequate scoring function, another crucial factor that determines the success of a docking method is the employment of an effective and efficient search strategy to find the global minimum. Due to the existence of a huge number of local minima, a searching method should be effective in identifying the global minimum and should not be trapped in local minima. For practical purposes, it also should be efficient enough to complete the search within a reasonable amount of CPU time. A number of searching methods have been employed in docking programs, including systematic methods [7–9,12,13,21], Monte Carlo (MC) simulation [10,11,22–25], and in recent years, genetic algorithms (GA) [16,26].

Regardless of the searching methods employed, a crucial factor that determines the searching efficiency is the number of degrees of freedom included in the searching. In a real biological system, the system at least includes the ligand, the receptor and the solvent molecules. Because of the huge number of degrees of freedom associated with the solvent molecules, they were generally excluded from consideration in docking studies, although in some cases the solvent effect was treated implicitly in the scoring function [17–19]. As for the ligand and the receptor molecules, the degrees of freedom include 3 relative translational, 3 external rotational and numerous additional degrees of freedom associated with internal flexibility, i.e. the number of rotatable bonds in the molecules. In the early 1980s, because of very limited computer power, the internal flexibility of both the drug and the receptor had to be completely ignored in docking studies [8]. However, the importance of the flexibility of both the drug and the receptor for their interactions has long been recognized. It has been argued that in many cases, when a receptor binds to different ligands with the same binding site, the binding site may assume a very similar 3D conformation. In these cases, the binding site conformation of the receptor bound to one drug may be used as the template to dock other ligands. However, it is difficult to predict the active conformation of a flexible ligand, because the bioactive conformation of a drug bound to its receptor rarely corresponds to its isolated single crystal X-ray structure nor to its global minimum in the gas phase as calculated using molecular mechanics methods [27]. For these reasons, it is highly desirable to take the ligand flexibility into account when performing docking studies. With rapid improvement in computing power, flexible ligand docking has become possible. Indeed, over the last few years, flexible ligand docking has been the major focus in the development of new docking methods.

Miller et al. [21] developed the program FLOG in which multiple, pre-generated drug conformations were used and for each conformation, rigid body docking similar to that in the DOCK program was performed. SYSDOC used a similar strategy [12,13]. Olson et al. developed the AutoDock program [10,11], in which they used an MC simulated annealing algorithm combined with grid-based energy evaluation to perform flexible ligand docking. Itai et al. developed a systematic searching method to either interactively [14] or automatically [15] perform flexible ligand docking, again with a grid-based energy calculation. The DOCK program had been applied to tackle the flexible ligand docking problem combined with a genetic algorithm (GA) to sample ligand conformations explicitly [16,26]. Recently, the Kuntz group developed another flexible ligand docking method [28], in which a limited backtrack search technique was used to examine torsion angles associated with the pre-determined rigid fragments. This new method has been integrated in DOCK4.0. Rarey et al. also developed a flexible ligand docking program, FLEXX [20]. The overall strategy in FLEXX is to dock a flexible ligand using an incremental method. The scoring function used in FLEXX was an empirical free energy function, similar to that developed by Böhm [17,18]. A total of 19 ligands were tested and excellent results were reported [20]. Using the GA, Jones et al. [16] developed the program GOLD. The scoring function used in the GOLD program is the summation of the hydrogen bonding energy and steric interaction energy between the ligand and the receptor, and the internal

energy of the ligand, which includes the steric and torsion energies. In total, 100 ligands were docked into their receptors with the program GOLD and a root-mean-square (rms) value of 2 Å or better for all non-hydrogen atoms of the ligand was obtained in 71% of the cases.

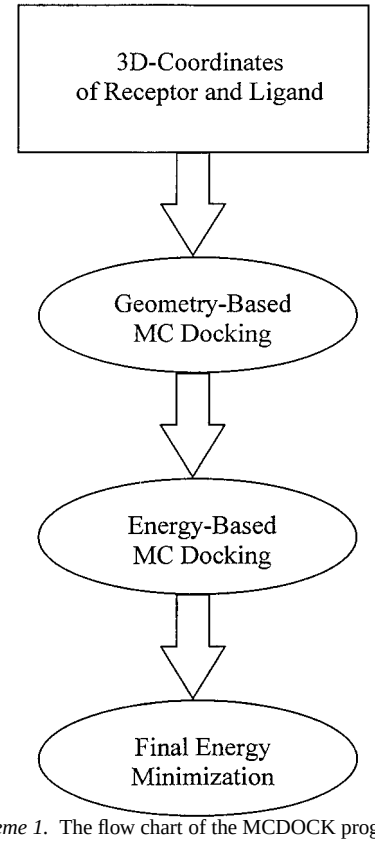
In this paper, we report the development and validation of a new docking program. We named our docking program MCDOCK (*M*onte *C*arlo *DOCK*) since the Monte Carlo simulation method was implemented to search for the global minimum. The final scoring function used in MCDOCK is the detailed atomic interaction energy based upon molecular mechanics, such as that used in the CHARMM program [29] and in AMBER [30,31]. Different MC sampling techniques were implemented in MCDOCK to efficiently search the global minimum. The 19 ligand-receptor complexes used by Rarey et al. [20] were used to validate the MCDOCK program because detailed, excellent docking results were previously reported and, importantly, these examples should provide an unbiased test for the MCDOCK program. Through the validation study, we wish to address two basic questions with regard to the MCDOCK program: (1) whether or not the scoring function is adequate to describe the interactions between the ligand and the receptor when the ligand flexibility is taken into account; and (2) whether or not the sampling strategy employed in MCDOCK can efficiently find the global minimum.

Theory and method

Overview of the MCDOCK method and input data

The flow chart of the MCDOCK program is shown in Scheme 1. The structural input data for the MCDOCK program are the 3D coordinates of the receptor (enzyme) and the ligand (substrate or inhibitor). The MCDOCK program includes three stages of calculations, geometry-based MC docking, energy-based MC docking and final energy minimization. The entire MCDOCK calculations are fully automatic.

The current version of the MCDOCK program was developed for the purpose of determining accurately the binding mode of a ligand to its receptor at a known binding site of the receptor, not for the purpose of identifying possible binding sites on the receptor for a ligand. Hence, the binding site of the receptor should be known. For the sake of computational efficiency,



Scheme 1. The flow chart of the MCDOCK program.

only a portion of the receptor large enough to enclose the binding site is included in the MCDOCK calculation.

An accurate 3D structure of the receptor is essential for the success of docking studies. For this reason, an experimentally determined, high-resolution 3D structure is required for the MCDOCK calculation. Many receptor structures have been experimentally determined and deposited into the Brookhaven Protein Data Bank [32]. In fact, all the 3D coordinates of the receptor-ligand complexes investigated in this paper were obtained from this data bank. The 3D structures of ligands can be obtained from experimental sources such as the Cambridge Structural Database [33], and in a few cases from the structures of ligands in complex with their receptors. The 3D structures of ligands can also be generated using numerous computer programs, such as CONCORD [34] and CORINA [35], or molecular modeling packages such as QUANTA [36]. It should be emphasized that it is not necessary to use the global minimum or the bound conformation of the ligand in complex with its receptor in the MCDOCK calculation. A conformation with good quality of bond

lengths and angles for the ligand should be sufficient for our purpose. In the examples reported in this paper, the initial conformations of ligands used in the docking operations were generated either from the program CORINA or built using the ChemNote module in the QUANTA molecular modeling package with moderate minimization. It is of note that the receptor structure is fixed (rigid) during the entire MCDOCK calculation, and only the relative position and the conformation of the ligand are varied.

The Lennard-Jones and atomic charge parameters for receptors and ligands are needed for the MCDOCK calculation. In the current version of MCDOCK, the parameters in the CHARMM [29] force field (version 24) were used.

With the initial structures of the receptor and the ligand, geometry-based MC docking is first performed. The structure obtained from this step is used as the input structure for the next step, energy-based MC docking. The structure obtained from the energy-based MC docking is finally subjected to an MC energy minimization to further refine the complex structure. It is noted that with the initial structural data and force field parameters, the MCDOCK program runs automatically without any manual intervention.

Geometry-based docking

The first step in the MCDOCK calculation is to simply place the ligand into the binding site of the receptor, which is carried out automatically in the program. One may choose one of the two following simple approaches for this operation. The first approach is to calculate the center-of-mass (COM) of the portion of the receptor which includes its binding site, and the COM of the ligand. The COMs of the receptor and the ligand are then superimposed through a simple coordinate translation of the ligand. For cases in which the COM of the portion of the receptor is not located within the binding site, one should use the following alternative approach. A reference point located within the binding site of the receptor is arbitrarily selected. The receptor and the ligand are then superimposed on each other according to the coordinates of the reference point for the receptor and of the COM of the ligand.

Since the ligand is arbitrarily placed into the binding site of the receptor, some atoms of the ligand and the receptor may overlap (bad contacts). It was found that using this overlapped complex structure to carry out energy-based MC sampling is extremely inefficient. To overcome this difficulty, the geometry-based MCDOCK procedure was developed to eliminate most or all of the bad contacts between the receptor and the ligand, as described below.

First, the binding site of the receptor is divided into grids. If a grid is occupied by an atom of the receptor, an index number of 1 is assigned to this grid and if not, 0 is assigned. MC simulation is carried out to sample randomly the three Euler angles and the three Cartesian coordinates of the COM of the ligand. The ligand is treated as a rigid body, and no attempt is made to dock the ligand accurately at this stage. The scoring function is the total index number of grids that are occupied by ligand atoms. The smaller the total index number is, the less overlap occurs between the ligand and the receptor. By doing so, the ligand moves to space unoccupied by the receptor. Most of the bad contacts can be eliminated through this operation. Boundary conditions were applied to prevent the ligand from escaping from the binding site. Our experience shows that geometry-based MC docking provides a good starting point for the next stage, energy-based docking.

Energy-based docking

Energy-based scoring function

If an adequate scoring function is used, energy-based docking should give a more accurate prediction than simple geometry-based docking. However, energy-based docking involving detailed atomic interaction calculations is expensive. In the development of the MCDOCK program, much effort has been made to improve the computing efficiency and the docking accuracy.

Lennard-Jones and electrostatic non-bonded interaction potentials were used to describe the interactions between a receptor and a ligand, as well as the conformational energy of the ligand. The interaction potential energy between atom i and atom j is:

$$E_{ij} = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \frac{\delta q_i \delta q_j}{r_{ij}} \quad (1)$$

where E_{ij} is the non-bonded interaction energy between atoms i and j, r_{ij} is the distance between atoms i and j, δq_i and δq_j are partial atomic charges for atoms i and j, respectively, and σ_{ij} and ε_{ij} are Lennard-Jones parameters which are computed according to Equations 2 and 3:

$$\sigma_{ij} = \frac{\sigma_i + \sigma_j}{2} \quad (2)$$

and

$$\varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j} \tag{3}$$

where σ_i and ε_i are atomic Lennard-Jones parameters for atom *i*. These parameters were assigned according to the CHARMM force field (version 24), which treats polar and non-polar hydrogen atoms explicitly. The overall interaction energy is:

$$E_{tot} = \sum E_{receptor-ligand} + \sum E_{ligand} \tag{4}$$

It should be pointed out that the torsional terms are not included in the ligand intra-molecular energy and only non-bonding terms are employed to describe the ligand conformational energy.

To prevent a ligand escaping too far from the binding site of its receptor during the simulation, a spherical container was introduced to cover the binding region. During the simulation, if an atom of the ligand moves out of the container, this trial will be rejected. Furthermore, an extra penalty potential energy was introduced as follows,

$$\Omega = \sum \frac{C}{(r - r_0)^n} \tag{5}$$

where Ω is the penalty energy potential, r_0 is the radius of the container, r is the distance of a ligand atom to the center of the container, and C is a constant. The penalty energy potential is the sum over all the ligand atoms. When all the atoms in the ligand are far away from the edge of the container, Ω is negligible, and when any ligand atom is close to the edge, Ω approaches infinity. Therefore, a ligand is forced to move inside the container. In this study, the values of C and n are chosen as 1 and 4, respectively. Thus, the final scoring function with this boundary condition used in MCDOCK is:

$$E_{score} = E_{tot} + \Omega \tag{6}$$

Different cutoff values are used for the calculations of interaction energies at different simulation stages in order to improve the docking efficiency. For instance, 2 Å is used as the cutoff value in the global sampling, 3 Å in the high temperature MC simulation and simulated annealing stages and 4 Å in the final minimization stage. The relatively short cutoff values are chosen based upon our systematic evaluation of the effect of the cutoff value on the accuracy of the docking. Based upon our evaluation, it was found that short-range interactions play a dominant role for the predicted binding mode of a ligand. Very similar final structures are obtained for long cutoff values such as 8 Å and short cutoff values such as 3 Å. It is comprehensible that the shortest cutoff is taken in the global sampling phase since the main aim here is to eliminate the bad contacts (very close to the hard sphere potential) and the relatively longer cutoff is then used in the final minimization. The choice of using different cutoff values at different stages reduces the computational time dramatically without losing much accuracy of docking. A shifted potential is used to smooth the interactions across the cutoff range [29].

Global sampling and Metropolis sampling

If some bad contacts (overlaps) between the atoms of the receptor and the ligand still exist after the geometry-based docking, Metropolis-type MC [37] simulation is inefficient and sometimes even unable to eliminate these bad contacts and to reach low energy states. To overcome this difficulty and to enhance searching efficiency, global sampling of Euler angles and torsion angles is performed. Since global sampling does not depend on previous states at all, it is capable of eliminating bad contacts between the receptor and the ligand. After a sufficient number of steps of global sampling, the Markov chain sampling scheme using the Metropolis algorithm is performed, starting from the lowest energy structure obtained from previous global sampling. The three overall Euler angles, ligand internal torsion angles as well as the coordinates of the COM are sampled. The energy difference (score difference) between trial and previous state is then computed as ΔE. If the Boltzmann weighting factor, $\exp(-\Delta E/kT)$, is greater than a uniformly drawn random number, the trial state is accepted, otherwise rejected. It is of note that k is the Boltzmann constant and T is a pre-selected temperature in the canonical MC simulation.

Two types of neighbor lists are implemented in the MCDOCK program to further speed up the calculation of the atomic potential energy. One is the Verlet neighbor list and the other is the cell-linked list. These two algorithms were well discussed by Allen and Tildesley in their simulation textbook [38]. The Verlet neighbor list is utilized for Metropolis sampling of the Markov chain because the trial state in the simulation is correlated with the previous state, and the cell-linked list is applied for global sampling since there is no correlation between the trial and previous states.

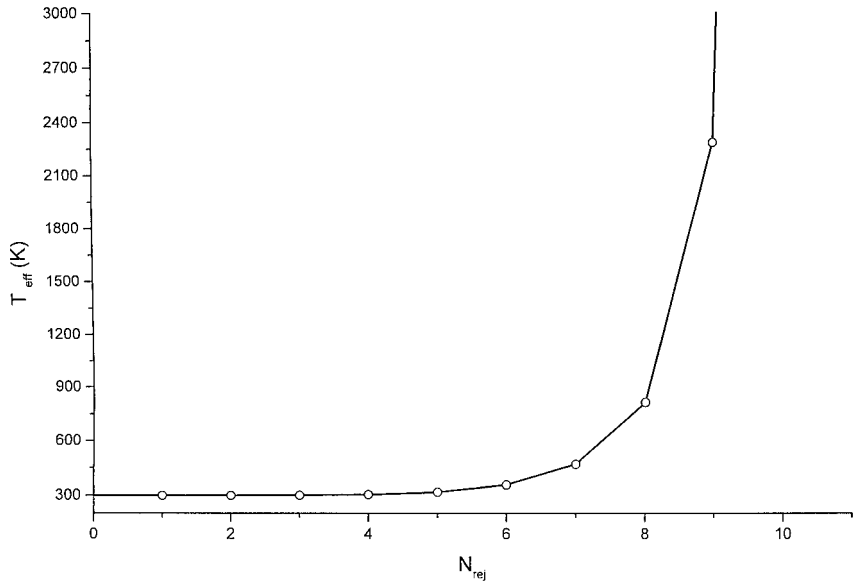


Figure 1. Illustration of the effective temperature used in the MCDOCK program with parameters $T_l = 300$ K and $N_0 = 8$ (see Equation 7).

Table 1. Summary of MCDOCK docking results for a total of 19 ligands

PDB code	N_{rot}	N_{atom}	rms_h (Å)	$\overline{rms_5}$ (Å)	$\overline{rms_{10}}$ (Å)	$\overline{rms_{15}}$ (Å)	$\overline{rms_{20}}$ (Å)	CPU (s)
121p	9	46	0.99	1.59 ± 0.52	1.72 ± 0.56	2.20 ± 1.89	2.55 ± 1.83	190
1dwc	11	71	0.51	0.61 ± 0.07	0.69 ± 0.24	0.74 ± 0.22	0.83 ± 0.36	450
1dwd	8	69	1.81	1.78 ± 0.17	2.14 ± 0.45	2.48 ± 1.31	3.21 ± 1.98	320
1ldm	0	8	0.43	0.48 ± 0.03	0.49 ± 0.02	0.61 ± 0.33	0.82 ± 0.45	12
1rnt	6	37	0.77	0.77 ± 0.05	0.91 ± 0.20	0.91 ± 0.19	1.45 ± 1.39	195
1stp	5	31	0.41	0.49 ± 0.10	1.19 ± 0.94	1.60 ± 0.97	1.76 ± 0.91	115
1tmn	15	66	1.84	1.57 ± 0.18	1.52 ± 0.14	1.55 ± 0.19	2.39 ± 1.88	385
1ulb	0	16	0.32	0.33 ± 0.00	0.33 ± 0.01	0.34 ± 0.01	0.98 ± 1.36	28
2ctc	4	21	0.32	0.38 ± 0.17	0.39 ± 0.16	0.45 ± 0.21	0.59 ± 0.38	80
2phh	0	15	0.73	0.73 ± 0.02	0.84 ± 0.23	0.81 ± 0.20	0.83 ± 0.18	34
3cpa	7	30	0.83	0.78 ± 0.06	0.78 ± 0.09	1.78 ± 1.78	2.82 ± 2.40	111
3ptb	0	18	0.39	0.37 ± 0.06	0.41 ± 0.16	0.40 ± 0.13	0.48 ± 0.21	29
3tpi	11	38	0.25	0.33 ± 0.11	0.98 ± 0.77	1.60 ± 1.37	2.09 ± 1.70	154
4dfr	8	54	0.60	1.36 ± 0.65	1.71 ± 0.67	1.93 ± 0.66	2.73 ± 2.07	420
4phv	15	88	0.55	0.62 ± 0.14	1.59 ± 1.51	2.72 ± 2.08	3.26 ± 2.05	853
4tln	7	24	1.40	1.54 ± 0.20	1.54 ± 0.16	1.61 ± 0.20	1.72 ± 0.33	78
4tsi	5	24	0.76	0.77 ± 0.04	0.78 ± 0.09	0.72 ± 0.05	0.72 ± 0.06	56
5tim	0	5	1.13	1.15 ± 0.04	1.14 ± 0.05	1.15 ± 0.04	1.15 ± 0.05	7
6rsa	3	31	1.11	1.13 ± 0.05	1.14 ± 0.10	1.08 ± 0.15	1.90 ± 1.96	78

Twenty MCDOCK runs were performed for each ligand. N_{rot} and N_{atom} denote the total number of rotatable bonds and the total number of atoms of the ligand, respectively. rms_h is the root-mean-square value of all heavy atoms of the ligand between the binding mode with the lowest interaction energy and the X-ray determined binding mode. $\overline{rms_5}$, $\overline{rms_{10}}$, $\overline{rms_{15}}$ and $\overline{rms_{20}}$ are the average rms values over the top 5, top 10, top 15 and all the 20 runs ranked by score, respectively. The associated standard deviations are presented in the same column. CPU is the average computing time in seconds for each MCDOCK run on an SGI Indigo2 workstation with a single R10000 processor.

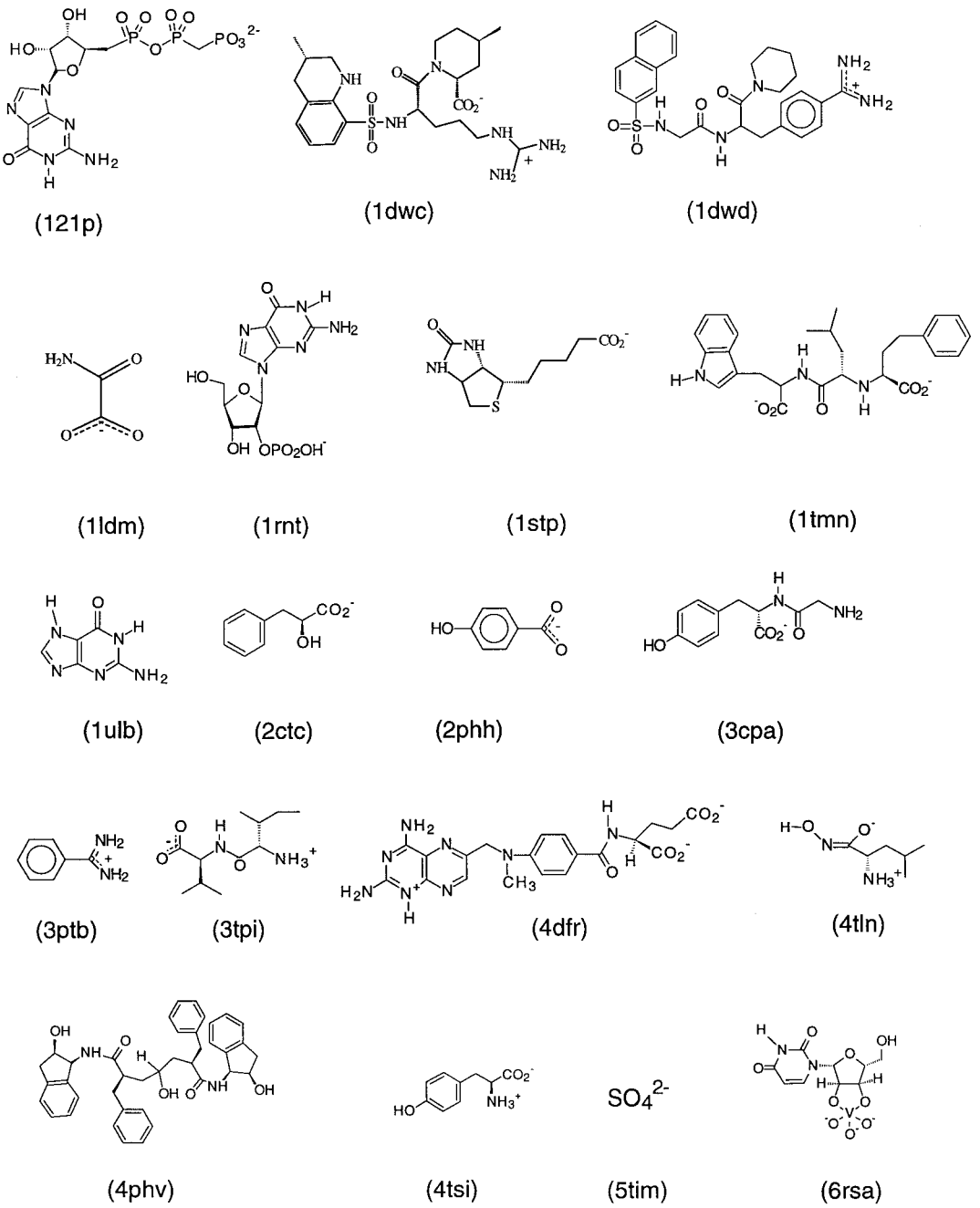


Figure 2. Chemical structures of 19 ligands studied using the MCDOCK program and the associated PDB code for each ligand.

Simulated annealing

In order to widely sample the binding site, a high temperature is initially chosen in the Boltzmann weighting factor during the propagation of the Markov chain. After a sufficient number of steps of simulation with a high temperature, the geometry with lowest energy is saved as the starting state for another Markov chain, in which the temperature is gradually reduced to more precisely locate the local minimum. In each docking simulation, several low energy states are saved from low temperature Markov chains and the one

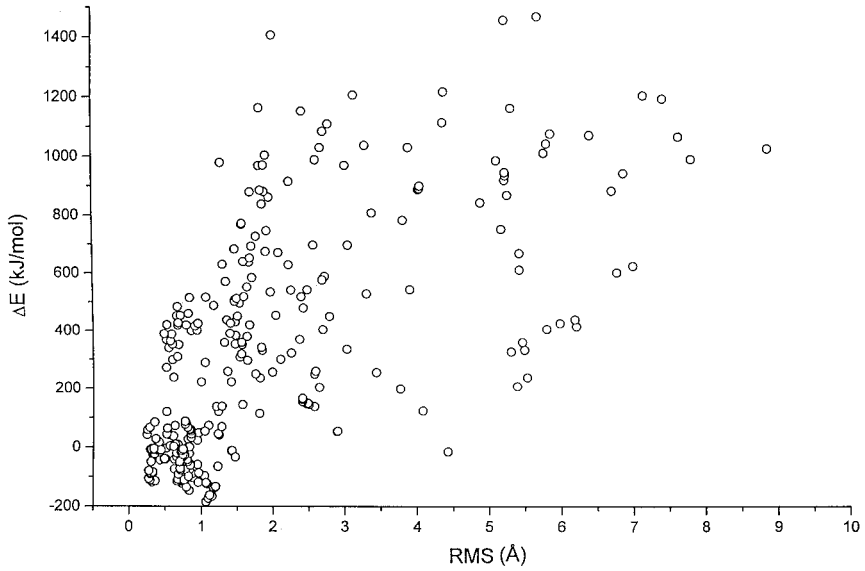


Figure 3. Relative binding energy (ΔE) versus the rms value. ΔE is the energy difference between the MCDOCK predicted binding complex and the X-ray determined binding complex, as calculated using the scoring function implemented in the MCDOCK program. Results of all the 20 MCDOCK runs for each of the 19 ligands are presented. No cutoff was applied in the energy calculation.

with the lowest energy is selected for further energy minimization.

Final energy minimization

In the MC simulation, the system could be trapped in one local minimum or another. If a local minimum is very deep and steep, the chance of escaping from the local minimum is very small. To overcome this problem, a final Markov chain is propagated using an effective temperature, which is defined as:

$$T_{eff} = T_l e^{\left(\frac{N_{rej}}{N_0}\right)^6} \tag{7}$$

where T_{eff} is the effective temperature used in the simulation, T_l is the low temperature used in the simulated annealing, N_0 is a preset number and N_{rej} is the number of continued rejection in the simulation. If N_{rej} is less than N_0, T_{eff} is close to T_l. When N_{rej} is larger than N_0, T_{eff} is increased rapidly with a continued increase of N_{rej}. This is illustrated in Figure 1 using $N_0 = 8$ and $T_l = 300$ K. As can be seen from Figure 1, if the MC simulation stays in a state for more than 8 trial steps, the effective temperature is dramatically increased from 300 K to 2278 K. Therefore, the probability of accepting the next trial is greatly enhanced. If the energy barrier is very large and the next trial is again rejected, the probability of the next trial being accepted will be further increased. Once the trial is accepted, N_{rej} will be reset to zero and the whole procedure will start all over again. Clearly, by introducing this effective temperature, the probability of escaping from local minima and sampling many local minima is greatly enhanced. It is of note that the temperature and the step length employed in the final minimization are often a fraction of the values used in the MC simulations during the global sampling and simulated annealing stages.

Results and discussion

In order to investigate the strength and the weakness of the MCDOCK program, we have studied 19 ligand/receptor complexes. The structures of these 19 ligands in complex with their receptors were previously determined using X-ray diffraction with high resolution, and their coordinates can be retrieved from the Brookhaven PDB [32]. The PDB codes for these 19 complex structures are 121p, 1dwc, 1dwd, 1ldm, 1rnt, 1stp, 1tmn, 1ulb, 2ctc, 2phh, 3cpa, 3ptb, 3tpi, 4dfr, 4phv, 4tln, 4tsi, 5tim and 6rsa, respectively. The chemical structures of the 19 ligands and their associated PDB codes are shown in Figure 2.

The primary reason for selecting these 19 complexes as our test cases is that they were previously chosen for flexible ligand docking studies using the program FLEXX [20] and excellent and detailed re-

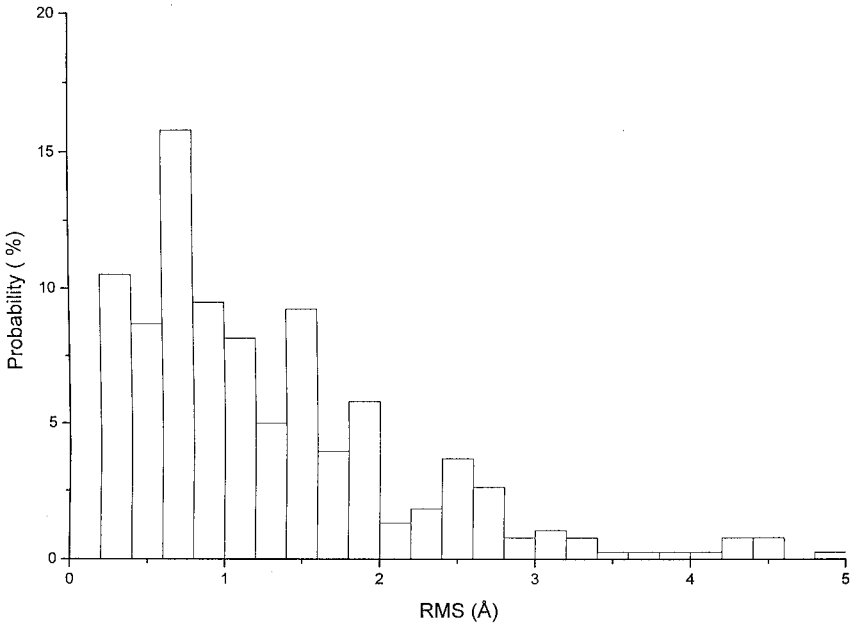


Figure 4. Distribution of MCDOCK docking accuracy. The rms window width is 0.2 Å. The accumulated probabilities of docking a ligand with an rms value within 1 and 2 Å are 45% and 77%, respectively.

sults were reported. An objective and unbiased comparison is thus possible with our docking results.

Statistical analysis

Since the MCDOCK program employs the MC simulation technique, several MCDOCK calculations for the same ligand may not yield identical docking results. To be able to evaluate rigorously the MCDOCK program, 20 MCDOCK runs were performed for each ligand, which resulted in 20 predicted complex structures for each ligand. These 20 runs were ranked by potential energy. From each predicted complex, the predicted ligand position was superimposed onto the experimentally determined ligand position in the receptor, and the rms value was obtained for all the non-hydrogen atoms (heavy atoms) of the ligand.

We first evaluated whether or not the scoring function used in MCDOCK is adequate for the prediction of the experimentally determined binding mode for these 19 ligands. The correlation between the relative interaction energy (scoring function) and the rms is presented in Figure 3. The relative interaction energy is computed as the energy difference between each MCDOCK predicted and the X-ray determined complex structure. Theoretically, if the scoring function is perfect, the rms value should be 0 when the relative score value is 0. As can be seen from Figure 3, a positive correlation was observed between ΔE and the rms, although with a few exceptions. In 48% of MCDOCK runs, the relative potential energy is less than 200 kJ/mol and the rms value is less than 1.5 Å. In 70% of MCDOCK runs, the relative potential energy is less than 800 kJ/mol, and the rms value is less than 2 Å. It is of interest to note that the relative potential energies in some MCDOCK runs are less than zero, indicating that MCDOCK identifies some binding modes with an interaction energy lower than the experimentally determined one. This may reflect the approximations made in the MCDOCK scoring function, which excludes the solvation effect and entropic factors. However, most of these binding modes with negative relative potential energy have a small rms value (less than 1.5 Å), indicating that the predicted ligand binding mode is essentially the same as the X-ray determined one. Taken together, our results suggest that the scoring function employed in the MCDOCK program is reasonably good in the prediction of the ligand binding mode. However, due to the relatively large energy difference between the predicted binding mode and the experimentally determined one, the score may not be used in the binding affinity prediction.

The results of the 20 MCDOCK runs for each ligand are summarized in Table 1, where the rms_h is the rms value of the ligand based upon the predicted

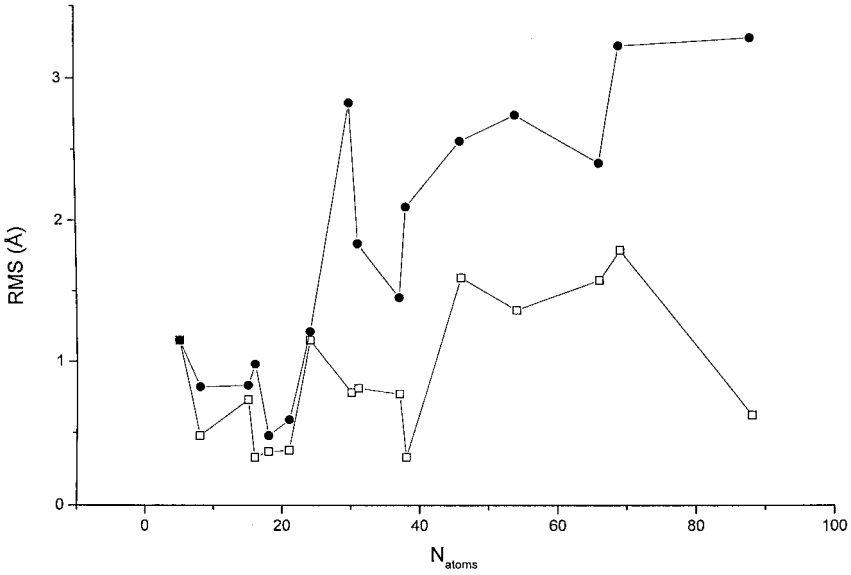


Figure 5. Average rms values versus the number of atoms of the ligands. Curves with open squares and solid circles are the results based upon the top 5 and all the 20 runs, respectively.

binding mode with the lowest potential energy of the 20 MCDOCK runs. As can be seen, for these 19 ligands, the rms_h values are between 0.25 and 1.84 Å. In fact, except for 1dwd and 1tmn, all other ligands have an rms_h value less than 1.50 Å. Using the top 5 predicted binding modes, the average rms value is between 0.33 and 1.78 Å. Using the top 10 predicted binding modes, the average rms value is between 0.33 and 2.14 Å. Only the thrombin inhibitor (1dwd) has an average rms value larger than 2 Å. Our results clearly show that the predicted binding mode for each of these 19 ligands with the lowest potential energy agrees very well with the experimentally determined binding mode. In addition, a significant number of MCDOCK runs converge to the same binding mode, close to the experimentally determined one. The distribution of the rms values from all the 20 MCDOCK runs for each ligand is shown in Figure 4. It is clear that most of the MCDOCK runs have a predicted rms value within 2 Å. The probabilities of obtaining a predicted binding mode for a ligand from a single MCDOCK run with an rms value less than 1 and 2 Å, as compared to the X-ray determined binding mode, are 45% and 77%, respectively. Taken together, our results clearly demonstrate that the MCDOCK program can be used to accurately predict the binding mode of a ligand to its receptor, although multiple MCDOCK runs are clearly necessary.

To investigate the effect of the size of the ligands on the docking accuracy, the average rms values of the top 5 runs and of all the 20 runs are plotted against the total number of atoms of the ligands in Figure 5. As can be seen, when a ligand is small with fewer than 20 atoms, the average rms value of either the top 5 runs or all the 20 runs is less than or close to 1 Å. When a ligand has more than 20 atoms, the average rms value of the top 5 runs increases slightly but is still less than or close to 1.5 Å. However, the average rms value of all the 20 MCDOCK runs has a quite strong dependence on the size of the ligands, suggesting that with increasing size of the ligand, the convergence of different MCDOCK runs to the same binding mode becomes more challenging. Nevertheless, our results showed that more than 25% of these MCDOCK runs converge to a binding mode close to the experimentally determined one, regardless of the size of the ligands.

To examine the impact of the flexibility of the ligands on the docking accuracy, the average rms values of the top 5 MCDOCK runs and of all the 20 runs were plotted against the number of rotatable bonds of the ligands in Figure 6. It is noted that a mean rms value is computed and used in the plot for ligands with the same number of rotatable bonds. As can be seen, the average rms value of the top 5 MCDOCK runs does not strongly depend upon the number of rotatable bonds. However, the average rms value of all the

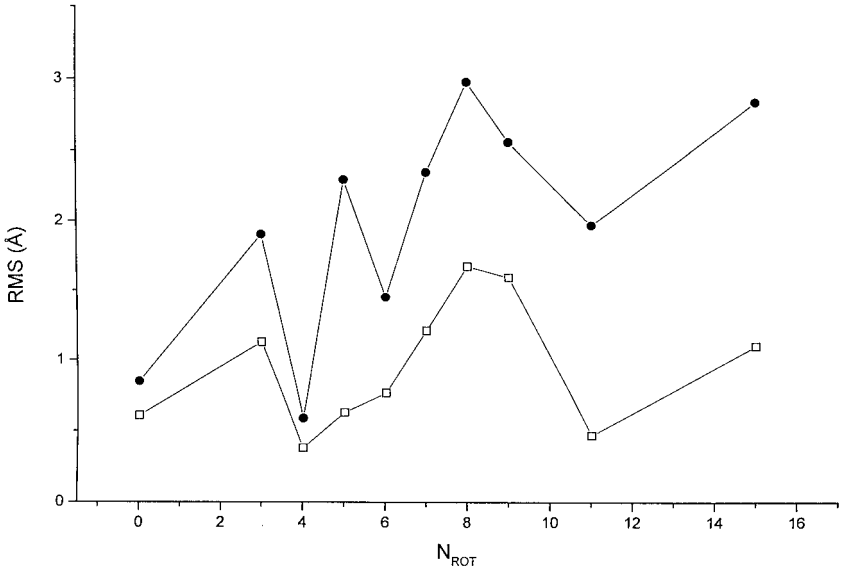


Figure 6. Average rms values versus the number of rotatable bonds of the ligands. Curves with open squares and solid circles are the results based upon the top 5 and all the 20 runs, respectively.

20 MCDOCK runs does have a fairly strong dependence on the number of rotatable bonds in the ligands. These results again demonstrate that MCDOCK can accurately predict the binding mode of ligands, regardless of their flexibility. However, when a ligand is very flexible, the convergence to the experimentally determined binding mode decreases.

It is important to point out that with increasing size and flexibility of ligands, the conformational space that needs to be sampled dramatically increases. Therefore, it is not surprising that the average rms values have a slow increase with increasing size and flexibility of ligands. Our results, however, demonstrated that MCDOCK is able to obtain a fairly accurate binding mode for the majority of the ligands with more than 60 atoms and more than 10 rotatable bonds.

Computational detail analysis

To gain more insight into the MCDOCK calculations, we examined in more detail the docking studies for two ligands, biotin and an HIV protease inhibitor (pdb codes: 1stp and 4phv).

Biotin in complex with streptavidin had been frequently selected in docking studies since biotin is a small molecule, non-covalently bound to its receptor with a very high binding affinity (K_i = ca. 10^{-14} M). Biotin has a moderate flexibility with five rotatable bonds. The conformation generated from the program CORINA was used as the initial conformation. As compared to the active conformation of biotin bound to streptavidin, the CORINA generated conformation has an rms value of 1.75 Å. To be objective, the three Euler angles of the initial conformation of biotin were then randomly rotated, and the resulting rms was 5.20 Å, as compared to the X-ray determined binding mode. It is of note that the rotation of external Euler angles does not alter the conformation of the ligand but rather its relative orientation.

A total of 50 000 steps of MCDOCK simulation were performed to dock biotin into the binding site of streptavidin. In Figures 7 and 8, the rms and interaction energy values are plotted against the number of simulation steps. The biotin structure obtained at the end of each simulation stage was superimposed on the X-ray crystallographically determined biotin binding mode, and the snapshots are illustrated in Figure 9. In the geometry-based docking, a total of 10 000 steps of simulation were performed. As can be seen from Figure 7, the rms value changed wildly from 2 to 7 Å during this stage, which is typical when docking a relatively small ligand in a spacious binding pocket. The structure of biotin finally obtained with the lowest geometry-based score has an rms value of 3.6 Å. This may not be the best structure in terms of rms for 1stp since in a real prediction case, the X-ray determined binding mode is often not known. However, the geometry-based docking does provide a very good initial structure for next stage simulation.

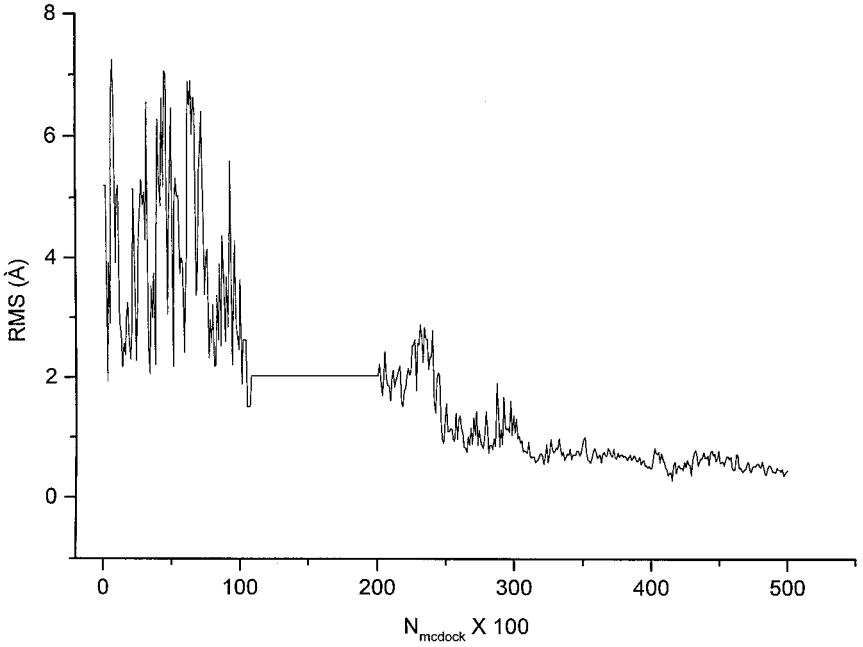


Figure 7. Rms for biotin as a function of the number N_{mcdock} of MCDOCK simulation steps.

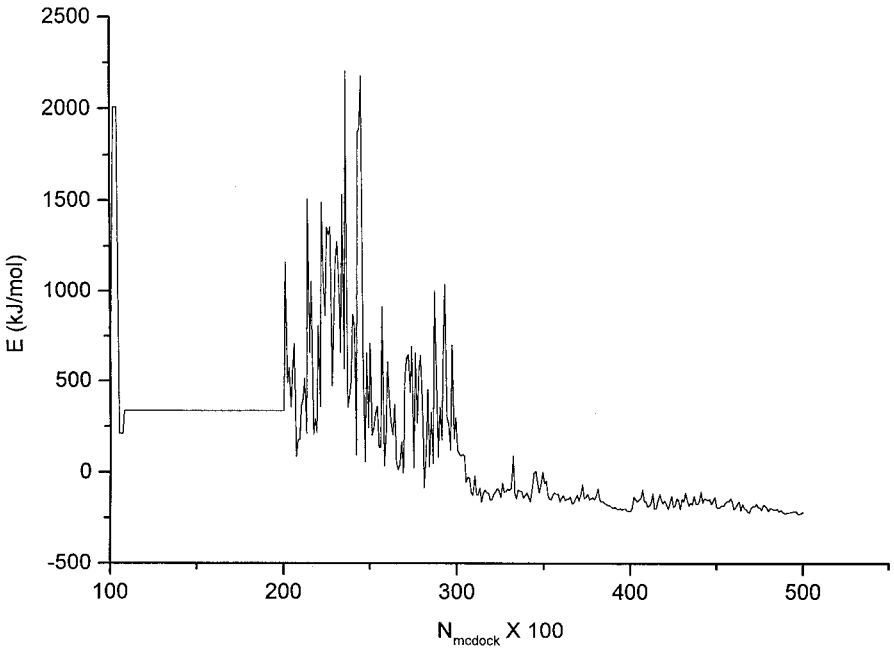


Figure 8. Energy of the biotin/streptavidin complex during an MCDOCK simulation. The geometry-based docking results are not shown because the score is not energy. No cutoff was applied in the energy calculation.

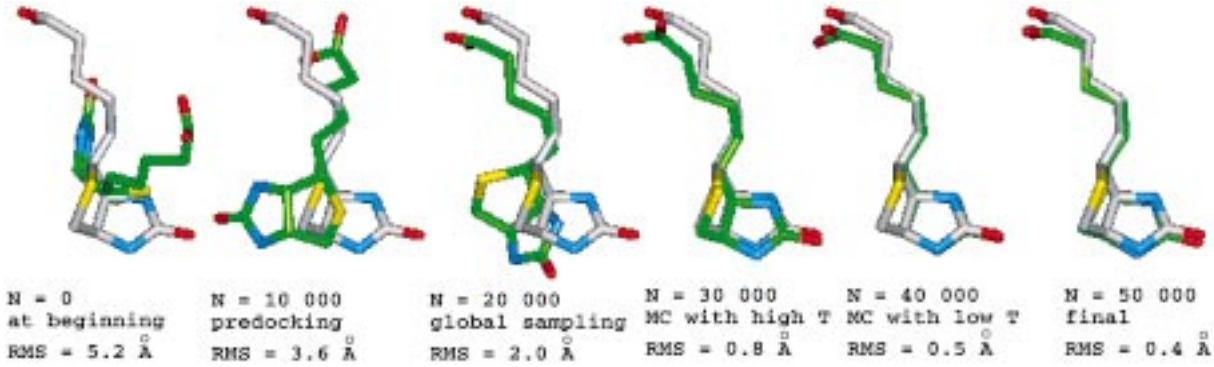


Figure 9. Snapshots of the biotin structure (hydrogen atoms are not shown for clarity; carbon atoms are in green; nitrogen atoms are in blue; oxygen atoms are in red; sulfur atoms are in yellow) at each MCDOCK stage superimposed on the X-ray determined binding mode (carbon atoms are in gray and all other types of atoms are in the same colors as above).

In the energy-based docking, 10 000 steps of global sampling MC simulation were performed and the rms value of biotin decreased from 3.6 to 2.0 Å. The MC Markov chain was then propagated for 10 000 steps with a high temperature, which further improved the rms value to ca. 0.8 Å. Simulated annealing was carried out with a low temperature for 10 000 steps, and an rms value of 0.5 Å was obtained. Finally, energy minimization was performed for another 10 000 steps, which yielded the final predicted binding mode with an rms value of 0.4 Å.

The potential energy changes during the MC simulations were examined (Figure 8). As can be seen, during the global sampling stage (from 10 000 to 20 000 MC steps), the potential energy rapidly decreased from 2 008 to 337 kJ/mol and the rms improved from 3.6 to 2.0 Å (see Figure 7). During the MC Markov chain with high temperature (from 20 000 to 30 000 steps), the energy changed frequently and sometimes even experienced dramatic changes, suggesting the sampling of a fairly large conformational space. The lowest energy obtained was −84 kJ/mol and the rms was further improved to 0.8 Å. In the simulated annealing stage (from 30 000 to 40 000 steps), the energy mainly underwent some small changes but in some steps, some large energy changes did occur. The energy was further decreased to −212 kJ/mol, and the rms was improved to 0.5 Å. In the final stage of energy minimization, the energy only underwent minor changes to −229 kJ/mol, and the rms was 0.4 kJ/mol. As can be seen from Figures 7–9, the global sampling and MC Markov chain with high temperature are most crucial to the overall docking accuracy, and the simulated annealing and energy minimization further refine the binding mode.

The scoring function and the sampling strategy determine the docking accuracy. The 20 MCDOCK runs for 1stp did not always converge to a single binding mode, as evident from the deviation of the rms value in Table 1. In Figure 10, the final energies obtained from these 20 MCDOCK runs for 1stp are plotted against the final rms values. As can be seen, all the binding modes with energies less than −160 kJ/mol have an rms value less than 0.7 Å. This clearly suggests that the scoring function used in MCDOCK is adequate for the prediction of the correct binding mode of biotin to streptavidin. For 6 of the 20 MCDOCK runs, the predicted binding modes have an rms value less than 0.7 Å after only 40 000 steps of simulation, suggesting that the sampling strategy implemented in MCDOCK is fairly efficient. However, 14 runs have a final energy higher than −100 kJ/mol and an rms value more than 1.5 Å, indicating that in these cases 40 000 steps of MCDOCK simulations are not sufficient to find the correct binding mode. However, since these 14 MCDOCK runs have a much higher energy than the other 6 runs, they can easily be distinguished and safely excluded. It is of note that the score for the X-ray determined binding mode is −243.6 kJ/mol, very close to the lowest energy identified through the 20 MCDOCK calculations.

A potent HIV protease inhibitor in complex with the HIV-1 protease (4phv) has been often selected for docking studies because this HIV protease inhibitor has a large size (88 atoms) and is quite flexible (15 rotatable bonds), thus presenting a considerable challenge for flexible ligand docking studies. The initial conformation generated from the program CORINA has an rms value of 2.36 Å as compared to the active conformation of the inhibitor bound to the HIV-1

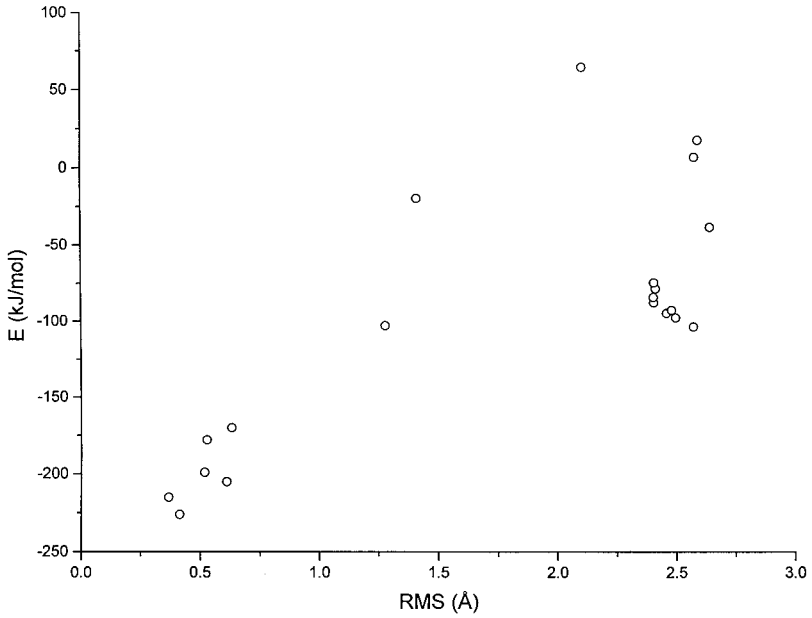


Figure 10. Final energy of the biotin/streptavidin complex obtained from each of the 20 MCDOCK runs versus the rms value of biotin. Note that no cutoff was applied in the energy calculation.

protease. The starting structure used in the MCDOCK study has an rms value of 6.18 Å after random rotation of the three Euler angles of the inhibitor.

Because of the complexity of the inhibitor, a much longer simulation was performed (220 000 steps) for each MCDOCK run. The rms value obtained from one such run was plotted against the number of simulation steps in Figure 11. As can be seen, in the geometry-based docking from 0 to 100 000 steps, the ligand was able to reach a binding mode with an rms value less than 2 Å, as compared to the X-ray determined binding mode. However, a large rms fluctuation from 2 to 11 Å was observed in this stage. The global sampling was performed from steps 100 000 to 150 000 and the rms value was 1.90 Å after this stage. The high temperature Markov chain MC simulation was propagated from steps 150 000 to 175 000 and the simulated annealing was performed from steps 175 000 to 200 000. Final minimization was carried out from steps 200 000 to 220 000. The final binding mode obtained has an rms value of 0.97 Å as compared to the X-ray determined binding mode. In Figure 12, the final potential energies obtained from 20 MCDOCK runs are plotted against the rms values. Eight of the 20 runs converged to a similar binding mode, and each has an rms value less than 1.7 Å as compared to the X-ray determined binding mode. Twelve runs did not converge to a low potential energy state; these runs, however, can be easily distinguished.

Comparison with the FLEXX docking results

The 19 ligands reported in this paper were previously studied by Rarey et al. [20] using the program FLEXX, and good results were obtained. The best rms value and the rms value for the binding mode with the lowest potential energy, obtained from FLEXX and MCDOCK, are presented in Table 2. Generally speaking, both programs are able to dock ligands into their receptors with good accuracy. The best rms value predicted from both programs is less than 1 Å for 15 out of 19 ligands. However, it is probably more meaningful to compare the rms value with the lowest potential energy, because the experimentally determined binding mode is often not available in the real docking application and one will have to rely on the scores of predicted binding modes. Using the rms value with the lowest potential energy, it was found that, except for 1dwd, 1tmn, 2phh and 6rsa, MCDOCK provides consistently better results than FLEXX. It is of interest to note that the difference between the best rms and the rms values for the binding modes with lowest potential energies predicted by MCDOCK is consistently smaller than that obtained with FLEXX. In fact, the binding modes with the best rms and with the lowest potential energy predicted by MCDOCK are very

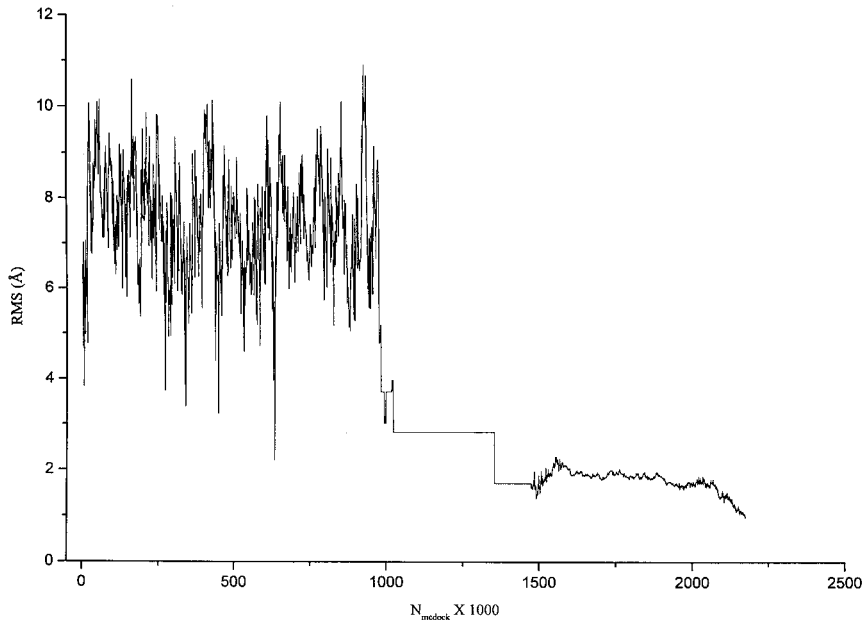


Figure 11. Rms for 4phv (an HIV protease inhibitor) as a function of the number N_{mcdock} of MCDOCK simulation steps.

close to each other in all cases. Therefore, one may use the binding mode with the lowest potential energy as the predicted binding mode. It is noted that FLEXX is faster than MCDOCK for docking operations. However, MCDOCK carries out the docking operation automatically, while FLEXX requires the pre-determination of rigid fragments for each ligand, which may require considerable experience with the program in order to obtain good results.

Utilities of the MCDOCK program

One of the primary utilities of the MCDOCK program is to accurately predict the binding mode of a ligand to its receptor. The MC simulation technique was employed to search for the global minimum of the interaction energy. Based upon our results, we found that the MC searching strategy implemented in the MCDOCK program is efficient, and a significant percentage of MCDOCK runs converged to a binding mode close to the X-ray determined binding mode. However, there is still a certain probability that a single MCDOCK run may not converge to a low energy state, which is the characteristic of MC simulation. Therefore, if the purpose is to predict precisely the binding mode of a ligand to its receptor, it is highly recommended to carry out multiple MCDOCK runs and examine if all these runs converge to a binding mode with similar energy. In cases where some of the MCDOCK runs do not converge to low energy states, those should not be used for the prediction of the binding mode. Complex structures obtained from MCDOCK runs which converged to a low energy should be examined whether or not they are similar to each other in their geometry. We found that for all the 19 cases that have been tested in this paper, those MCDOCK runs which converged in energy for each ligand essentially predicted the same binding mode for the ligand with some minor deviation in details. Our results clearly demonstrate that MCDOCK is able to predict the binding mode of a ligand with an rms value of 2 Å for all cases, or with an rms value of less than 1.0 Å for 79% of the cases if 20 MCDOCK runs are performed (see Table 1).

Another major utility of the MCDOCK program is to perform database searching for lead discovery. The major aim in database searching using the MCDOCK program is not to predict precisely how a potential ligand binds to its target receptor, but rather to access the likelihood of how well the ligand can bind to the receptor. Therefore, only a single MCDOCK run is performed for each ligand in database searching. We found that MCDOCK can screen approximately 10 000 compounds within one week of CPU time on a single processor, an R10000 SGI Indigo2 workstation. This speed makes the MCDOCK program a useful tool to screen large-sized chemical databases such

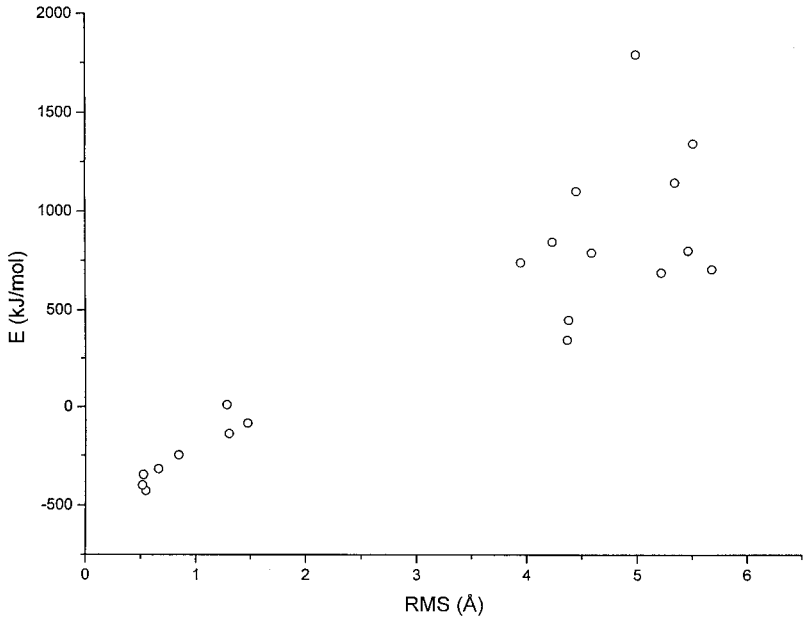


Figure 12. Final energy of the 4phv/HIV protease complex obtained from each of the 20 MCDOCK runs versus the rms value of 4phv. Note that no cutoff was applied in the energy calculation.

as the NCI 3D-database and the Available Chemical Database for lead discovery purposes.

Conclusions

In this paper, we present a new MC simulation approach to the molecular docking problem and a program MCDOCK has been developed. Detailed atomic interactions, similar to those used in the CHARMM program, are used as the scoring function for the prediction of the binding mode. The flexibility of the ligand is fully taken into account by explicitly sampling all the rotatable bonds of the ligand in the docking calculations, and the ligand conformational energy is assessed by the atomic interactions of the ligand. A number of techniques have been integrated to optimize the docking efficiency.

Our results showed that the simple scoring function used in the MCDOCK program is fairly adequate for the accurate prediction of the ligand binding mode. Using the binding mode with lowest potential energy as the predicted binding mode, the rms values for these 19 ligands are between 0.25 and 1.84 Å. Using the top 5 MCDOCK simulations out of a total of 20 runs for each ligand, MCDOCK is able to predict the binding mode with an rms value between 0.33 and 1.78 Å for these 19 ligands.

Table 2. Comparison of MCDOCK results with FLEXX [20] results

PDB code	MCDOCK rms_b (Å)	MCDOCK rms_h (Å)	FLEXX rms_b (Å)	FLEXX rms_h (Å)
121p	0.99	0.99	1.14	2.00
1dwc	0.51	0.51	1.20	2.66
1dwd	1.47	1.81	0.63	0.81
1ldm	0.43	0.43	0.62	0.62
1rnt	0.73	0.77	0.96	1.48
1stp	0.37	0.41	0.81	0.81
1tmn	1.34	1.84	0.87	0.87
1ulb	0.32	0.32	0.65	0.65
2ctc	0.27	0.32	0.65	0.65
2phh	0.70	0.73	0.58	0.58
3cpa	0.70	0.83	1.06	3.08
3ptb	0.29	0.39	0.48	0.48
3tpi	0.24	0.25	0.58	0.58
4dfr	0.60	0.60	0.90	1.34
4phv	0.52	0.55	1.04	1.04
4tln	1.36	1.40	0.93	4.50
4tsi	0.61	0.76	0.71	2.01
5tim	1.06	1.13	0.87	1.99
6rsa	0.88	1.11	0.85	0.85

rms_b is the smallest rms value obtained from multiple runs for each ligand; rms_h is the rms value for the binding mode with the lowest interaction energy among multiple runs.

The CPU time used for each MCDOCK run is from 1 to 15 min for a ligand, depending upon the size and the flexibility of the ligand. The MCDOCK docking operation is fully automatic without any manual operation.

It is noted that the scoring function used in MCDOCK has not been correlated with the binding affinity of the ligand yet. In future developments, we would like to improve the scoring function used in MCDOCK by including the solvation effect and a better description of intra-molecular interactions for the ligands. It is hoped that the improved scoring function will not only allow better prediction of the binding mode of a ligand but also enable the prediction of the binding affinity of a ligand.

Because the MCDOCK program takes account of the full flexibility of the ligands in the docking study and can yield a fairly accurate prediction for the ligand binding mode, it is thus a very useful computational tool in lead optimization. Since the MCDOCK program is also fairly efficient, it can be used in database searching for the discovery of new lead compounds.

Acknowledgements

We would like to thank Dr. Xiong-Wu Wu for many stimulating discussions. The financial support from the Department of Defense (DOD) is gratefully acknowledged.

References

1. Blundell, T.L., Nature, 384 (1996) 23.
2. Ewing, T.J.A. and Kuntz, I.D., J. Comput. Chem., 18 (1997) 1175.
3. Gschwend, D.A., Good, A.C. and Kuntz, I.D., J. Mol. Recogn., 9 (1996) 175.
4. Jones, G. and Willett, P., Curr. Opin. Biotechnol., 6 (1995) 652.
5. Kuntz, I.D., Science, 257 (1992) 1078.
6. Villar, H.O. and Koehler, R.T., The Molecular Modeling e-conference, 1 (1997) 23.
7. Shoichet, B.K. and Kuntz, I.D., Protein Eng., 6 (1993) 223.
8. Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. and Ferrin, T.E., J. Mol. Biol., 161 (1982) 269.
9. Fischer, D., Lin, S.L., Wolfson, H.L. and Nussinov, R., J. Mol. Biol., 248 (1995) 459.
10. Goodsell, D.S. and Olson, A.J., Proteins Struct. Funct. Genet., 8 (1990) 195.
11. Goodsell, D.S., Morris, G.M. and Olson, A.J., J. Mol. Recogn., 9 (1996) 1.
12. Pang, Y.-P. and Kozikowski, A.P., J. Comput.-Aided Mol. Design, 8 (1994) 669.
13. Pang, Y.-P. and Kozikowski, A.P., J. Comput.-Aided Mol. Design, 8 (1994) 683.
14. Tomioka, N. and Itai, A., J. Comput.-Aided Mol. Design, 8 (1994) 347.
15. Mizutani, M.Y., Tomioka, N. and Itai, A., J. Mol. Biol., 243 (1994) 310.
16. Jones, G., Willett, P., Glen, R.C., Leach, A.R. and Taylor, R., J. Mol. Biol., 207 (1997) 727.
17. Böhm, H.-J., J. Comput.-Aided Mol. Design, 6 (1992) 593.
18. Böhm, H.-J., J. Comput.-Aided Mol. Design, 8 (1994), 243.
19. DeWitte, R.S. and Shakhnovich, E.I., J. Am. Chem. Soc., 118 (1996) 11733.
20. Rarey, M., Kramer, B., Lengauer, T. and Klebe, G., J. Mol. Biol., 261 (1996) 470.
21. Miller, M.D., Kearsley, S.K., Underwood, D.J. and Sheridan, R.P., J. Comput.-Aided Mol. Design, 8 (1994) 153.
22. Caflisch, A., Fischer, S. and Karplus, M., J. Comput. Chem., 18 (1997) 723.
23. Caflisch, A., Niederer, P. and Anliker, M., Proteins Struct. Funct. Genet., 13 (1992) 223.
24. Hart, T.N. and Read, R.J., Proteins Struct. Funct. Genet., 13 (1992) 206.
25. Yue, S.-Y., Protein Eng., 4 (1990) 177.
26. Oshiro, C.M., Kuntz, I.D. and Dixon, J.S., J. Comput.-Aided Mol. Design, 9 (1995) 113.
27. Nicklaus, M.C., Wang, S., Driscoll, J.S. and Milne, G.W. Bioorg. Med. Chem., 3 (1995) 411.
28. Makino, S. and Kuntz, I.D., J. Comput. Chem., 18 (1997) 1812.
29. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M., J. Comput. Chem., 4 (1983) 187.
30. Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta Jr., S. and Weiner, P., J. Am. Chem. Soc., 106 (1984) 765.
31. Weiner, S.J., Kollman, P.A., Nguyen, D.T. and Case, D.A., J. Comput. Chem., 7 (1986) 230.
32. PDB. The Protein Data Bank web site is http://www.pdb.bnl.gov.
33. Allen, F.H., Davies, J.E., Galloy, J.J., Johnson, O., Kennard, O., Macrae, C.F., Mitchell, E.M., Mitchel, G.E., Smith, J.M. and Watson, D.G., J. Chem. Inf. Comput. Sci., 31 (1991) 187.
34. Pearlman, R.S., Chem. Des. Autom. News, 2 (1987) 1.
35. Gasteiger, J., Rudolph, C. and Sadowski, J, Tetrahedron Comput. Methodol., 3 (1990) 537.
36. QUANTA is a product of Molecular Simulation, Inc., San Diego, CA.
37. Metropolis, N. and Ulam, S., J. Am. Stat. Assoc., 44 (1949) 335.
38. Allen, M.P. and Tildesley, D.J., Computer Simulation of Liquids, Oxford University Press, Oxford, 1990, pp. 146–152.