Use of Conformational Space Annealing in Molecular Docking

Kyoungrim Lee[1], Cezary Czaplewski[2], Seung-Yeon Kim[1] and Jooyoung Lee[1*]

[1] School of Computational Sciences, Korea Inst itute for Advanced Study, Seoul, Korea

[2] Department of Chemistry, University of Gdańsk, Gdańsk, Poland

*To whom correspondence should be addressed. E -mail: jlee@kias.re.kr

Abstract

Molecular docking falls into the general category of global optimization problems since its main purpose is to find the most stable complex consisting of a receptor and its ligand. Conformational space annealing (CSA), a powerful global optimization method, is incorporated with the Tinker molecular modeling package to perform molecular docking simulations of six receptor-ligand complexes (3PTB, 1ULB, 2CPP, 1STP, 3CPA and 1PPH) from the Protein Data Bank. In parallel , Monte Carlo with minimization (MCM) method is also incorporated into the Tinker package for comparison. The energy function, consisting of electrostatic interactions, van der Waals interactions and torsional energy terms, is calculated using the AMBER94 all-atom empirical force field. Rigid docking simulations for all six complexes and flexible docking simulations for three complexes (1STP, 3CPA and 1PPH) are carried out using the CSA and the MCM methods. The simulation results show that the docking procedures using the CSA method generally find the most stable complexes as well as the native-like complexes more efficiently and accurately than those using the MCM, demonstrating that CSA is a prom ising search method for molecular docking problems.

Introduction

In recent years, a number of computational algorithms have been developed to investigate protein (receptor)-ligand docking. Many of these algorithms share common approaches but contain specific extensions to increase their accuracies and efficiencies in structure-based drug design.[1-3] For a given energy function and molecules under investigation, the docking problem is to find the most stable association of the receptor and ligand molecules. One of the most challenging parts in this problem is to carry out rigorous conformational searches of a receptor-ligand complex system including the flexibility of both molecules. In other words, the molecular docking problem falls into the general category of global optimization pr oblems since its procedure is to optimize the rigid -body intermolecular variables, i.e., the translational vectors for the relative positions and the rotational

This work is supported by the Basic Research Program of the Korea Science & Engineering Foundation

Euler angles for the relative orientations between two molecules (rigid docking) as well as the intramolecular variables including all torsional angles of each molecule (flexible docking)[4-8] to obtain the most stable intermolecular association between them. In ref. 9, varieties of current docking techniques are reviewed with a description of applications for single docking experiments as well as the virtual screening of databases.

Currently most of the widely used conformational search methods are based on either genetic algorithms (GA),[7,8,10] Monte Carlo simulations,[4-6,11] simulated annealing (SA)[12,13] or molecular dynamic simulations.[14,15] Generally, these methods aim for efficient sampling of the receptor-ligand system to find the global minimum energy conformation of the docked complex by overcoming high-energy-conformational barriers.

Here, we present an efficient docking method using the conformational space annealing (CSA) method,[16-19] and its successful application to six receptor-ligand docking systems. The CSA method has been successfully applied for *ab initio* protein structure prediction[20-22] and also used to predict the structures of multi-chain homo-oligomer proteins.[23,24] One of the advantages of the CSA is that it can find many families of low-energy conformations that have distinct structural differences. This makes it possible to search the whole intermolecular space of the receptor-ligand associations for a given energy function. In this method, the sampling diversity is maintained by keeping various conformers of local-energy minima as representatives of structurally similar conformations within hyper spheres centered on them.[16-19] Conformational space annealing is achieved by slowly reducing the radius of these hyper spheres. In this docking study, for rigid docking, the structural similarity between two complexes is determined by considering the relative translational position and rotational orientation of a ligand to its receptor molecule. For flexible docking, the torsional angles of the rotatable bonds of the ligand are also incorporated into the definition of the structural similarity.

The CSA and the Monte Carlo with minimization (MCM) methods[25-27] are implemented into the Tinker package,[28] to perform molecular docking simulations. The energy function used is the AMBER94 all-atom empirical force field[29] without solvation. The major purpose of this study is to investigate the role of efficient conformational search methods in docking simulations. The methods are compared in terms of their docking efficiencies and accuracies for a total of six receptor-ligand complexes. The rest of this paper is organized as follows. First, the computational details are described including implementation, algorithms and docking simulations for both rigid and flexible docking calculations. Then, the results are discussed by comparing the sampling efficiencies of the CSA and the MCM methods. Finally, this work is summarized by highlighting key findings and suggesting modification for further improvement.

Materials and Methods

Adaptation of CSA into docking

Details of the CSA algorithm and its applications can be found in refs. 16-22. Here, we provide only a brief description of the original CSA algorithm and essential changes of the algorithm for its implementation to the docking problem with the

Tinker package program.[28] The CSA unifies the essential ingredients of the three global optimization methods, SA, GA and MCM. First, as in MCM, we consider only the phase space of local minima, that is, all conformations are energy-minimized by a local minimizer. Second, as in GA, we consider many conformations in a *bank* in CSA collectively, which is similar to population in GA, and we perturb a subset of the bank conformations (*seeds*) using information in the remaining other bank conformations to generate new conformational structures. That is, this procedure is similar to mating in GA. However, in contrast to the mating procedure in GA, we replace typically *small* portions of a seed with the corresponding parts of bank conformations since we want to search the conformational neighborhood of the seed. Finally, as in SA, we introduce an annealing parameter D_{cut} (a cutoff distance reflecting the structural difference between the conformations in the phase space of local minima), which plays the role of temperature in SA. In CSA, the diversity of sampling is directly controlled by introducing a distance measure judging the conformational structural difference between two conformations and comparing it with D_{cut}, whereas in SA there are no such systematic controls. The value of D_{cut} is slowly reduced just as in SA, hence the algorithm is named *conformational space annealing*. Maintaining the diversity of the population using a distance measure was also tried in the context of GA, although no annealing was performed. To apply the CSA to an optimization

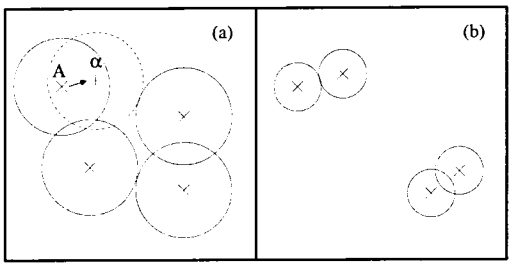

Figure 1. A schematic diagram to describe the search procedure of CSA is shown. The boxes represent the identical phase space. (a) Initially, we cover the phase space by large spheres with a radius of D_{cut} centered on randomly chosen local minima denoted by× symbols, and replace the centers with lower-energy local minima. When an initial conformation A is replaced by a new conformation α, the sphere moves in the direction of the arrow. (b) As the CSA algorithm proceeds and the energies of the representative conformations at the centers of the spheres are lowered, the size of the spheres (D_{cut}) is reduced and the search space is narrowed down to small basins of low-lying local minima.

problem, two things are necessary; a method for perturbing a seed conformation, and a distance measure between two conformations.

The way we picture the phase space of local minima is as follows (see Fig. 1). We assume that most of the phase space of local minima can be covered by a finite number of large spheres with radius D_{cut}, which are centered on randomly chosen minima (*bank*). Each of the bank conformations is supposed to represent all local minima contained in the sphere centered on it. To improve a bank conformation A, we first select A as a seed. We perturb A and subsequently energy-minimize it to generate a trial conformation α. Since α originates from A by small perturbation, it is likely that α is contained in a sphere centered on A. If the energy of α is lower than that of A, α replaces A and the center of the sphere moves from A to α. If it happens that α

belongs to a different sphere centered on B, α can replace B in a similar manner. When α is outside of all existing spheres, a new sphere centered on α is generated. In this case, to keep the total number of spheres fixed, we remove the sphere represented by the highest-energy conformation. Consequently, a larger value of D_{cut} produces more diverse sampling, whereas a smaller value results in quicker search of low-energy conformations at the expense of getting trapped in basins probably far away from the global minimum. Therefore, for efficient sampling of the phase space, it is necessary to maintain the diversity of sampling in the early stages and then gradually shift the emphasis toward obtaining low energy conformations in CSA, by slowly reducing the value of D_{cut}.

When the energy of a seed conformation does not improve after a fixed number of perturbations, we stop perturbing it. When all of the bank conformations are used as seeds (one iteration completed), this implies that the procedure of updating the bank might have reached a deadlock. If this happens we reset all bank conformations to be eligible for seeds again, and we repeat another iteration. After a preset number of iterations, we conclude that our procedure has reached a deadlock. When this happens, we enlarge the search space by adding additional random conformations into the bank and repeat the whole procedure until the stopping criterion is met.

In the application of the CSA method to receptor-ligand complexes, we first randomly generate a certain number of initial conformations (for example, fifty random conformations) whose energies are subsequently minimized using "the optimally conditioned variable metric nonlinear optimization routine without line searches" (OCVM)[30,31] which is an energy-minimizer implemented in Tinker package program.[28] Initial conformations are constructed, first by randomly generating translational vectors (x,y,z) and rotational Euler angles (ϕ,θ,ψ) of a ligand molecule with respect to its receptor protein, and the torsional angles (Ω) of the ligand, then performing local energy minimization of these receptor-ligand complexes. Throughout this work, the term *minimization* is used to refer to the application of the OCVM to a given complex. We call the set of these minimized conformations (complexes) the *first bank*. We make a copy of the first bank and call it the *bank*. The conformations in the bank are updated in later stages, whereas those in the first bank are kept unchanged. Also, the number of conformations in the bank is kept unchanged when the bank is updated. The initial value of D_{cut} is set as $D_{ave}/2$ where D_{ave} is the average distance between the conformations in the first bank. New conformations are generated by choosing a certain number of *seed* conformations (for example, ten or twenty seed conformations) from the bank and by replacing parts of their variables by the corresponding parts of conformations randomly chosen from either the first bank or the bank. The variables of a conformation are defined by three groups: translational vector, Euler angles and torsional angels. New conformations are generated by replacing one of these three groups from a seed conformation by the corresponding group from a conformation in the bank or in the first bank. Then the energies of these conformations are subsequently minimized, and these minimized conformations become trial conformations.

A newly obtained local minimum conformation (trial conformation) α is compared with those in the bank to decide how the bank should be updated. One first finds the conformation A in the bank which is the closest to the trial conformation α with the distance $D(\alpha, A)$ defined by

$$D(\alpha,A)=\sqrt{(x_\alpha-x_A)^2+(y_\alpha-y_A)^2+(z_\alpha-z_A)^2} + \omega_\Theta\Theta(\Delta\phi,\Delta\theta,\Delta\psi)+\omega_\Omega\sum_{\text{torsions}}|\Delta\Omega| \quad (1)$$

where Δx, Δy and Δz are the differences (in angstrom) in the components of the two translational vectors from A and α; $\Delta\phi$, $\Delta\theta$ and $\Delta\psi$ the differences in the components of the two rotational Euler angles from A and α, and the function $\Theta(\Delta\phi,\Delta\theta,\Delta\psi)$ is defined as

$$\Theta(\Delta\phi,\Delta\theta,\Delta\psi)=\cos^{-1}\left\{\frac{\phi_\alpha\phi_A+\theta_\alpha\theta_A+\psi_\alpha\psi_A}{\sqrt{\phi_\alpha^2+\theta_\alpha^2+\psi_\alpha^2}\sqrt{\phi_A^2+\theta_A^2+\psi_A^2}}\right\}.$$

$\Delta\Omega$ is the difference (measured in radian) in two corresponding dihedral angles from A and α, and the summation $\sum_{torstions}$ is taken for all rotatable torsional angles in the ligand. The two weight factors ω_Θ and ω_Ω are determined dynamically during docking simulations as follows. After the first bank is generated, we calculate the average value of each term in eq. (1) considering all pairs of complexes in the first bank. The values of ω_Θ and ω_Ω are chosen so that the three terms from the right-hand side of eq. (1) contribute equally to the distance measure.

The CSA docking follows the search procedure depicted in Fig 1. The algorithm stops when the known global minimum is found, which is examined after the bank is updated by all trial conformations. It should be noted that since one iteration is completed only after all bank conformations have been used as seeds, and we add additional conformations whenever our search has reached a deadlock, there is no loss of generality for using particular values for the number of seeds, the number of bank conformations, etc.

Energy function

The docking in this work is described by an all-atom force field, namely the AMBER94.[29] We do not include solvation energy terms since the primary purpose of this work is to investigate the role of efficient conformational search methods. The energy function used in the calculation of the receptor-ligand interaction, consists of three terms: electrostatic (E_{ele}), van der Waals (E_{vdw}) and torsional (E_{tor}) terms.

$$E_{total}=E_{ele}+E_{vdw}+E_{tor} \quad (2)$$

$$E_{ele}=\sum_{i<j}\frac{q_iq_j}{\varepsilon_{ele}r_{ij}} \quad (3)$$

$$E_{vdw}=\sum_{i<j}4\varepsilon_{ij}^{vdw}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}-\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \quad (4)$$

$$E_{tor}=\sum_{torsions}\frac{V_n}{2}[1+\cos(n\Omega-\Omega_0)] \quad (5)$$

where q_i and q_j represent the atomic partial charges of atoms i and j; ε_{ele} the dielectric constant; r_{ij} the distance between atoms i and j; ε_{ij}^{vdw} and σ_{ij} the van der Waals parameters; V_n the torsional potential force constant; n the periodicity of the torsional potential; Ω_0 a phase for the torsional potential. In the rigid docking study, only the first two energy terms in eq. (2) are used while all three terms are used for the flexible docking experiment.

Preparation of receptor-ligand complexes

We have selected six receptor-ligand complexes (3PTB, 1ULB, 2CPP, 1STP, 3CPA and 1PPH) from the Protein Data Bank (PDB),[32,33] which have been

extensively studied by various docking methods. The structures of these complexes are all determined by X-ray spectroscopy with resolutions better than 2.75Å. The ligands in these complexes are not covalently bonded to the proteins. Schematic structures of six ligands are shown in Figure 2, where rotatable bonds are indicated by curly arrows.

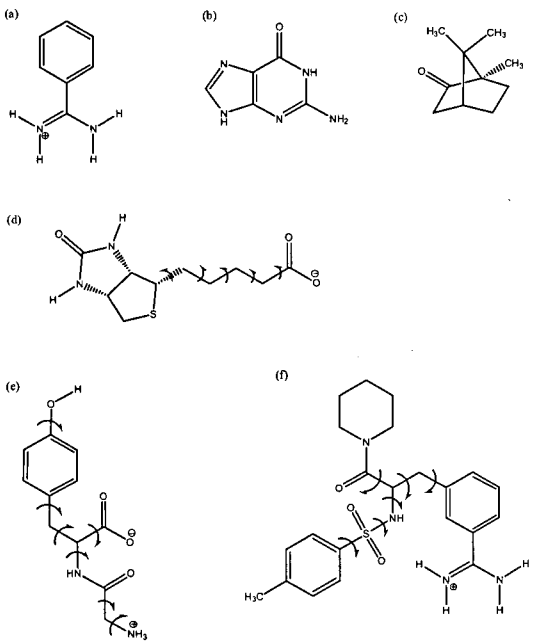

Figure 2. The six ligands chosen for docking experiments are shown: (a) benzamidine; (b) guanine; (c) camphor; (d) biotin; (e) Glycyl-L-Tyrosine; (f) 3-Tapap. Rotatable bonds used in this docking study are marked by curly arrows. The first three molecules are studied by rigid docking calculations, while the last three ligands are studied by both rigid and flexible docking calculations.

The geometries of all six ligands are optimized quantum mechanically starting from their X-ray crystal structures, using the GAMESS program[34] at the level of HF/6-31G(d).[35] Then, the atomic partial charges are determined according to the restrained electrostatic potential fitting procedure (RESP)[36-38] implemented in the AMBER charge fitting program.

Hydrogen atoms and missing atoms of the six X-ray protein structures are generated using the Tinker package. The receptor-ligand complexes prepared in this way with the AMBER94 are locally energy-minimized starting from their X-ray crystal structures. Then, the locally minimized native structures are used as reference structures for the calculation of the root-mean-square-deviation (RMSD) of the receptor-ligand complexes obtained from the docking simulations. In this study, we call these locally energy-minimized X-ray structures as the native-minimum complexes (NMC).

Docking simulations

Two types of docking simulations are carried out in this study, *i.e.*, rigid and flexible docking calculations. Firstly, for the rigid docking simulations, the conformations of a ligand and a receptor are fixed to their crystal structures, and only the rigid-body variables, i.e., the translational vector (x,y,z) and Euler angles (ϕ,θ,ψ) between two molecules are allowed to vary. Secondly, flexible docking simulations are performed only to the ligands with rotatable bonds. The flexibility of the receptor is not taken into account in this study. Three complexes (1STP, 3CPA and 1PPH) are targeted for the flexible docking (see Fig. 2). The CSA and MCM methods are used to find low-energy complexes. We will pay a special attention to compare the efficiencies and accuracies of the two methods.

The search for the ligand positions to form stable complexes is restricted to the inside of a sphere centered on the location of the experimental binding pocket. The radius size of the sphere used is 10Å. A schematic figure illustrating the limited search space of a ligand within a sphere is depicted in Figure 3. The sphere is besieged by a soft wall

represented by a harmonic potential and consequently the movement of a ligand is confined inside the sphere. The energy function of a given receptor-ligand complex is locally energy-minimized.

Figure 3. The search for the optimal association of a receptor-ligand system is carried out around the binding pocket. The yellow sphere represents the search space inside of which the ligands are confined. In this study, the radius size of sphere used is 10Å.

In MCM docking, initial positions of each ligand are randomly placed inside a sphere with the radius of 10Å. The maximum sizes of movement for Metropolis steps are 30° for the rigid-body rotations and 2Å for the rigid-body translations. The MCM docking simulations are carried out until they find a receptor-ligand complex whose energy is lower than or equal to that of the NMC as well as a complex whose RMSD value is less than 1.0Å from the NMC structure (termination condition). For each docking complex, ten separate MCM runs are carried out using separate random numbers.

In CSA docking, the searches for stable docking complexes are initiated by randomly generating the first bank of size "*nbank*" (we have used *nbank* = 20 and 50). Ten conformations from *nbank* = 20 (twenty for *nbank* = 50) are taken as seeds. Seeds are perturbed by replacing a selected group of variables among translational and rotational vectors, and dihedral angles (for flexible docking calculations only) with its corresponding one from conformations in the bank or the first bank. The perturbed conformations are energy-minimized and these are called as trial conformations. Using these trial conformations, the bank is updated. The CSA searches continue until the same termination condition as in the MCM is satisfied. If the termination condition is not satisfied after each CSA round, additional *nbank* randomly-selected and energy-minimized conformations are added to the bank. The maximum bank size is set to 100. Ten independent CSA docking runs using separate random numbers are carried out both for *nbank* = 20 and 50. The results from these ten runs for each receptor-ligand complex are averaged to compare the efficiencies of the MCM and the CSA dockings.

Results

Rigid docking

Six receptor-ligand complexes (3PTB, 1ULB, 2CPP, 1STP, 3CPA and 1PPH) are investigated for rigid docking studies using MCM and CSA methods. The results are summarized in Table 1.

β-Trypsin / Benzamidine (3PTB)

The ligand, benzamidine, has a polar amidine moiety and a hydrophobic benzyl ring.[39] The amidine group is considered to be protonated since its X-ray structure is planar.

The energy of the NMC is -102.82 kcal/mol. Both CSA and MCM methods found the NMC. The NMC for this complex is identical to the global minimum complex (GMC). It takes about 3.2×10^3 energy evaluations on average to find the NMC (GMC) using MCM, while the corresponding numbers from the CSA are 2.5×10^3 (*nbank* = 20)

and 3.4×10^3 (*nbank* = 50).

Purine nucleoside phosphorylase (PNP) Guanine (1ULB)

The guanine is treated only as a rigid body since rotatable bonds are absent in the molecule. The key for the recognition of the guanine binding into the receptor PNP, is based on its favorable hydrogen bonding and hydrophobic interactions.[40]

The energy of the NMC is -81.07 kcal/mol. However, the lowest-energy complexes (GMC's) obtained from the CSA searches are different from the NMC with significantly lower energies than that of the NMC. The RMSD values between the ligand conformations of GMC's and NMC are 11.2 Å with the energy of -111.84 kcal/mol. Examining the GMC's, we find that the ligand binds to the surface polar residues of the PNP with favorable electrostatic and hydrogen-bonding interactions. We believe that this is an artifact that can be fixed by including proper solvation effects[41-43] into the energy function. All runs using the CSA method were successful in finding the GMC as well as the NMC. It took significantly more energy evaluations to find the GMC than the NMC. The average numbers of energy evaluation to find the GMC are 1.4×10^5 (*nbank* = 20) and 2.4×10^5 (*nbank* = 50) while the corresponding numbers for the NMC are 6.5×10^4 (*nbank* = 20) and 2.5×10^4 (*nbank* = 50).

The results from MCM docking calculations were not as successful. Only one out of ten MCM docking runs found the GMC and the other nine runs failed to find it even after 8.3×10^5 energy evaluations. Similarly, the NMC was found in six out of ten runs, and the other four runs failed to find it even after 8.4×10^5 energy evaluations.

Cytochrome $P\text{-}450_{cam}$ / Camphor (2CPP)

A rigid docking procedure is applied to this

Table 1. Summary of rigid docking calculations using MCM and CSA methods.

Complex	Conformer[a]	Energy (RMSD)[b] kcal/mol (Å)	MCM[c]	CSA[c]	
				nbank[d] =20	*nbank*[d] =50
3PTB	GMC·NMC	-102.82	3.0×10^3	2.5×10^3	3.4×10^3
1ULB	GMC	-111.84 (11.2)	5.4×10^5(1)	1.4×10^5	2.4×10^5
	NMC	-81.07	7.7×10^3(6)	6.5×10^4	2.5×10^4
2CPP	GMC·NMC	-39.31	6.6×10^4	1.8×10^4	1.5×10^4
1STP	GMC·NMC	-80.67	5.6×10^5(7)	2.5×10^5	9.7×10^5
3CPA	GMC·NMC	-158.45	5.4×10^4(9)	2.1×10^4	2.9×10^4
1PPH	GMC·NMC	93.16	5.4×10^3	1.6×10^4	1.4×10^4

a. Two types of structures are considered; global minimum complex (GMC) and native minimum complex (NMC). GMC·NMC denotes that the GMC is identical to the NMC.

b. The RMSD of the GMC is shown only when the GMC is not identical to the NMC.

c. The numbers of energy evaluations averaged over ten runs are shown in the corresponding columns. The numbers in parentheses represent successful docking simulations to find the GMC/NMC out of 10 independent runs. The values with parentheses represent the number of energy evaluations averaged over the numbers in parentheses. For the MCM case of 1ULB, GMC is found in only one run, and NMC in six runs. The other nine and four runs failed even after 8.3×10^5 and 8.4×10^5 energy evaluations, respectively. For 1STP, GMC·NMC is found in seven runs and not in the other three runs even after 5.0×10^6 energy evaluations. For 3CPA, GMC·NMC is found in nine runs and not found in the other run even after 7.4×10^6 evaluations.

d. The symbol *nbank* refers to the initial bank size used in the CSA docking

complex system since the camphor is a fused bicyclic molecule consisting of a rigid aliphatic chain and a carbonyl group (see Fig. 2.).

The energy of the minimized X-ray structure (NMC) is -39.31 kcal/mol and the NMC is identical to the GMC. All ten MCM docking simulations found the NMC after the average number of energy evaluation of 6.6×10^4. All CSA runs also succeeded in finding the NMC (GMC). with on average 1.8×10^4 (*nbank* = 20) and 1.5×10^4 (*nbank* = 50) energy evaluations.

Streptavidin / Biotin (1STP)

The biotin is known to bind to the streptavidin[40] with a high value of affinity arising from the multiple hydrogen-bonds and favorable van der Waals interactions between them. The biotin has five rotatable bonds and the results of its flexible docking calculations will be discussed later.

The energy of the NMC is -80.67 kcal/mol. The GMC is identical to the NMC, and the GMC (NMC) was found in all CSA runs. The average numbers of energy evaluation are 2.5×10^5 (*nbank* = 20) and 9.7×10^4 (*nbank* = 50).

In the MCM docking, on the other hand, the GMC was found from seven out of ten runs after 5.6×10^5 energy evaluations. In the other three runs, MCM failed to obtain the GMC (NMC) even after 5.0×10^6 energy evaluations. The docking experiment for this complex clearly demonstrates that the CSA search is more efficient than the MCM in finding both the GMC and the NMC.

Carboxypeptidase / Glycyl-L-Tyrosine (3CPA)

Carboxypeptidase is complexed with glycyl-L-tyrosine that is a dipeptide with six rotatable bonds (see Fig. 2).

The energy of NMC is -158.45 kcal/mol. The NMC is identical to the GMC. All CSA runs found the NMC (GMC) after the average numbers of energy evaluation 2.1×10^4 (*nbank* = 20) and 2.9×10^4 (*nbank* = 50) while nine out of ten MCM runs found the NMC (GMC) with the average number of energy evaluation 5.4×10^4, and the other run could not find it even after 7.4×10^6 energy evaluations.

Trypsin / 3-Tapap (1PPH)

This complex has the same receptor protein as in the 3PTB system. The ligand molecule 3-tapap is a synthetic thrombin inhibitor. It is a modification of the benzamidine from the 3PTB complex with a large substituent which includes a *p*-toluene sulfonate and a piperidine group on the meta-position of the benzyl ring (see Fig. 2). This substitution helps the ligand to bind more tightly to the receptor by occupying the binding site fully.

The GMC is identical to the NMC and its energy is 93.16 kcal/mol. In the 10Å-sphere search, the NMC was found from all MCM runs with the average number of energy evaluations 5.4×10^3. The corresponding numbers from CSA calculations are 1.6×10^4 (*nbank* = 20) and 1.4×10^4 (*nbank* = 50). This, along with the other five rigid docking calculations, clearly demonstrates that the CSA performs more efficiently and more consistently than the MCM.

Flexible docking

Among the six rigid docking complexes studied in the previous section, three receptor-ligand complexes (1STP, 3CPA and 1PPH) contain rotatable bonds in their ligands (see Fig. 2), and have been used for the flexible docking studies. The results of the flexible dockings are summarized in Table 2.

Streptavidin / Biotin (1STP)

In addition to the rigid body variables, all five rotatable bonds of biotin (see Fig. 2) are allowed to vary during the flexible docking searches. The X-ray structure of the complex is locally energy-minimized (NMC), and the energy of the NMC is -112.14 kcal/mol. The structure of the GMC is found to be only slightly different from that of the NMC. Its RMSD value is 1.09Å and its energy is -119.03 kcal/mol. This means that flexible docking calculations have found a lower energy complex according to the energy function. In the GMC, more stable multiple hydrogen-bonds are formed between the receptor and the ligand. In fact, we found that many additional stable complexes, which are similar to the NMC with their RMSD values less than 2.0Å. Therefore, we consider, as the native-like conformer (NLC), the complexes which meet the condition that their energies are less than or equal to that of the NMC and their binding interaction mode is the same with the native binding mode, *i.e.* with the RMSD value less than 2.0 Å. We will take the number of energy evaluations for finding this NLC as the number for the NMC. The numbers of energy evaluation for finding the NLC and the GMC are examined.

The comparison of the results from the CSA and MCM methods are shown in Table 2. First, the CSA docking with a 10Å sphere found the GMC with the average number of energy evaluation

Table 2. Summary of flexible docking calculations using MCM and CSA methods

Complex	Conformer[a]	Energy (RMSD)[b] kcal/mol (Å)	MCM[c]	CSA[c]	
				nbank[d] = 20	*nbank*[d] = 50
1STP	GMC	-119.03 (1.09)	NA	3.4×10^5	2.1×10^5
	NLC	-112.14	NA	2.7×10^5	1.3×10^5
3CPA	GMC	-240.10(1.19)	2.0×10^5	2.4×10^5	2.1×10^5
	NLC	-230.77	1.9×10^5	1.4×10^5	1.4×10^5
1PPH	GMC·NLC	90.05	2.9×10^5(2)	1.4×10^5	2.2×10^5

a., b., c. and d. See the notes under Table 1. The NLC (nativdike complex) represents the complexes with energy less than or equal to the NMCs energy and native-like binding mode. NA stands for"not available" because all ten MCM runs fail to find the corresponding complex conformer. For 1STP docking, GMC or NLC was found in **no** of ten runs even after 4.8×10^6 energy evaluations; for 1PPH, GMCNLC was found in only two runs and the other eight failed even after 1.1×10^6 energy evaluations.

3.4×10^5 (*nbank* = 20) and 2.1×10^5 (*nbank* = 50) while none of the ten MCM runs could find a complex whose energy is less than -112.0 kcal/mol even after 4.8×10^6 energy evaluations. The energy and the RMSD value of the lowest energy complex obtained from the ten MCM runs are -100.79 kcal/mol and 1.1Å, respectively. The CSA finds the NLC (the native-like complex with its RMSD value less than 2.0Å and its energy less than or equal to the NMC) after the average numbers of energy evaluation 2.7×10^5 (*nbank* = 20) and 1.3×10^5 (*nbank* = 50).

Carboxypeptidase / Glycyl-L-Tyrosine (3CPA)

Seven rotatable bonds (see Fig 2) in the ligand including two bonds attached to the hydroxyl and the terminal ammonium groups are varied during the energy minimization. The energy of the NMC is -230.77 kcal/mol. The energy of GMC is -240.10

kcal/mol and its RMSD value is 1.19Å with respect to the NMC. As in the case of 1STP, there are many additional low-energy complexes close to the NMC with their energies less than -230.77 kcal/mol and RMSD values less than 2.0Å. Hence, we applied the same rule as in the flexible 1STP docking experiment to define the native-like complexes (NLC) for this docking simulations. All ten MCM and ten CSA runs found the GMC with the average number of 2.0×10^5 energy evaluations (MCM) and 2.4×10^5 (CSA, *nbank* = 20) and 2.1×10^5 (CSA, *nbank* = 50). For finding the NLC, the average evaluations 1.9×10^5 from ten MCM runs were required, and 1.4×10^5 (for both *nbank* = 20 and 50) for the CSA docking.

Trypsin / 3-Tapap (1PPH)

All six rotatable bonds (see Fig. 2) in 3-tapap are varied during the docking searches. The energy of the NMC is 90.05 kcal/mol and the energy of the GMC from the CSA docking is 90.00 kcal/mol showing a very small amount of energy difference between them. The structure of the GMC is almost identical to the NMC with its RMSD value 0.03Å. Therefore, we have regarded the structures of the GMC and the NMC as the same structure. The CSA found the GMC after the average numbers of energy evaluation 1.4×10^5 (*nbank* = 20) and 2.2×10^5 (*nbank* = 50). In contrast, only two out of ten MCM runs found the GMC after 2.9×10^5 energy evaluations on average and the other eight runs failed even after 1.1×10^6 energy evaluations. Once again, the efficiency and the consistency of the CSA method are clearly demonstrated.

Discussion

Conformational space annealing (CSA) and Monte Carlo with minimization (MCM) methods were implemented into the Tinker package for docking simulations. We have focused our attention especially on the sampling efficiencies and accuracies of the two methods. Six receptor-ligand complexes (3PTB, 1ULB, 2CPP, 1STP, 3CPA and 1PPH) were selected for rigid docking experiments, and three complexes (1STP, 3CPA and 1PPH) among the six were also tested for flexible docking studies. The intermolecular energy function for docking simulations consists of the electrostatic and the van der Waals interactions. We have used the AMBER94 all-atom empirical force field for this purpose. In flexible docking simulations, intra-molecular energy term for conformational changes of the ligand was added into the energy function. For reliable estimates of the sampling efficiencies of the CSA and the MCM, ten independent runs were carried out for each docking complex. For simple systems with low search complexity, the efficiency of the MCM was more or less equivalent to that of the CSA. However, for systems with complicated search spaces, the CSA method was significantly more efficient than the MCM method in finding both the NMC and the GMC.

The results from the rigid docking study have shown that all CSA runs were successful in finding the NMC with less number of energy evaluations on average than MCM runs while a portion of the MCM runs failed to locate the NMC for most receptor-ligand systems (see Table 1). The 3PTB complex was the only one for which all ten MCM runs were successful. In the rigid docking, the GMC corresponds to the NMC in most receptor-ligand complexes. However, the ligands of the complexes, 1ULB, have multiple hydrogen donors and

acceptors, and can interact favorably with the receptor by hydrogen-bonds. Sometimes, this complex has been found with its corresponding substrate bound to the receptor surface, but not in its native binding pocket. This artifact of non-native hydrogen bonds between the ligands and the receptors can be eliminated by adding proper solvation terms to the energy function.

For flexible docking simulations, we obtain many local minimum complexes near the NMC due to the flexibility of ligands. In the flexible docking experiment, the CSA method, in most complexes, was able to find the NMC as well as more stabilized complexes very close to the NMC. On the other hands, only a fraction of the MCM runs were successful in finding the NLC for most complexes.

The comparison between the CSA and the MCM demonstrates that the CSA method is a more promising method for investigating docking problems, especially for flexible docking studies where it finds the NLC and the GMC more efficiently and more accurately. For further improvement of docking accuracies and efficiencies, we should consider a couple of modifications in the future. First, solvation terms should be included in the energy function so that the artifact of non-native ligand binding to the receptor surface can be properly eliminated. Second, the intermolecular energy evaluation between a ligand and a receptor can be expedited by using pre-calculated grid potentials to reduce computation expenses. We leave these for our future studies.

Acknowledgements

This work was supported by the grant No. R01-2003-000-11595-0 from the Basic Research Program of the Korea Science & Engineering Foundation.

References

[1] Kuntz, I. D. *Science* 1992, 257, 1078-1082.

[2] Blaney, J. M.; Dixon, J. S. Perspect Drug Discov Design 1993, 1, 301.

[3] Kuntz, I. D.; Meng., E. C.; Shoichet, B. K. *Acc Chem Res* 1994, 27, 117.

[4] Torsset, J.; Scheraga., H. J. *J Comput Chem* 1999, 20, 244-252.

[5] Torsset, J.; Scheraga., H. J. *J Comput Chem* 1999, 20, 412-427.

[6] Apostolakis, J.; Pluckthun, A.; Caflisch, A. *J Comput Chem* 1998, 19, 21-37.

[7] Jones, G.; Willet, P.; Glen, R.C.; Leach, A. R.; Taylor, R. *J Mol Biol* 1997, 267, 727-748.

[8] Morris, G. M.; Goodsell, D. S. ; Halliday, R. S.; Huey, R.; Hart W. ; Belew, R. K.; Olson, A. J. *J Comput Chem* 1998, 19, 1639.

[9] Taylor, R. D.; Jewsbury, P. J. ; Essex, J. W. *J Comput Aid Mol Des* 2002, 16, 151-166.

[10] Taylor, J. S.; Burnett, R. M. *Proteins* 2000, 41, 173-191.

[11] Abagyan, R.; Totrov, M.; Kuznetsov, D. *J Comput Chem* 1994, 15, 488-506.

[12] Goodsell, D. S.; Olson, A. J. *Proteins* 1990(8), 195-202.

[13] Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. *J Comput Aid Mol Des* 1996, 10, 293-304.

[14] Mongoni, R.; Roccatano, D.; Di Nola, A. *Proteins* 1999, 35, 153-162.

[15] Nakajima, N.; Higo, J.; Kidera, A.; Nakamura, H. *Chem Phys Lett* 1992, 278, 297-301.

[16] Lee, J.; Scheraga, H. A.; Rackovsky, S. *J*

Comput Chem 1997, 18, 1222-1232.

[17] Lee, J.; Scheraga, H. A. *Int J Quant Chem* 1999, 75, 255-265.

[18] Kim, S.-Y.; Lee, S. J.; Lee, J. *J Chem Phys* 2003, 119, 10274-10279.

[19] Lee, J.; Lee, I. H.; Lee, J. *Phys Rev Lett* 2003, 91, 080201.

[20] Lee, J.; Liwo., A.; Ripoll, D. R.; Pillardy, J.; Scheraga, H. A. *Proteins Suppl* 1999, 3, 204-208.

[21] Lee, J.; Liwo., A.; Ripoll, D. R.; Pillardy, J.; Saunders, J. A.; Gibson, K. D.; Scheraga, H. A. *Int J Quant Chem* 2000, 77, 90-117.

[22] Lee, J.; Kim, S.-Y.; Joo, K.; Kim, I.; Lee, J. *Proteins* 2004, 56, 704-714.

[23] Saunders, J. A.; Scheraga, H. A. *Biopolymers* 2003, 68, 300-317.

[24] Saunders, J. A.; Scheraga, H. A. *Biopolymers* 2003, 68, 318-332.

[25] Nayeem, A.; Vila, J.; Scheraga, H. A. *J Comput Chem* 1991, 12, 594-605.

[26] Trosset, J. Y.; Scheraga, H. A. *Proc Natl Acad Sci* 1998, 95, 8011-8015.

[27] Pillardy, J.; Czaplewski, C.; Wedemeyer, W. J.; Scheraga, H. A. *Helv Chim Aata* 2000, 83, 2214-2230.

[28] Tinker 4.0, 2001, http://dasher.wustl.edu/tinker/.

[29] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M. Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J Am Chem Soc* 1995, 117, 5179-5197.

[30] Davidon, W. C. Math Program 1975, 9, 1-30.

[31] Shanno, D. F.; Phua, K-H. *J Optimiz Theory App* 1978, 25, 507-518.

[32] Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F. Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *J Mol Biol* 1977, 112, 535.

[33] Abola, E. E.; Bernstein, F. C.; Byant, S. H.; Koetzle, T. F.; Weng, J. In Data Commission of the International Union of Crystallography; F. H. Allen, G. B., and R. Sievers, Ed.: Bonn/ Cambridge/Chester, 1987, p 107.

[34] Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J Comput Chem* 1993, 14, 1347.

[35] Hariharan, P. C.; Pople, J. A. *Chem Phys Lett* 1972, 66, 217.

[36] Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. *J Phys Chem* 1993, 97, 10269.

[37] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollman, P. A. *J Am Chem Soc* 1993, 115, 9620.

[38] Cieplak, P.; Cornell, W. D.; Bayly, C.; Kollman, P. A. *J Comput Chem* 1995, 16, 1357.

[39] Marquart, M.; Walter, J.; Deisenhofer, J.; Bode, W.; Huber, R. *Acta Crystallogr Sect B* 1983, 39, 480.

[40] Ealick, S. E.; Babu, Y. S.; Bugg, C. E.; Erion, M. D.; Guida, W. C.; Montgomery, J. A., Secrist III., J. A. *Proc Natl Acad Sci* U S A, 1991, 88, 11540.

[41] Gilson, M. K.; Sharp, K. A.; Honig, B. *J Comput Chem* 1988, 9, 327.

[42] Honig, B.; Nicholls, A. *Science* 1995, 268, 1144-1149.

[43] Zou, X.; Sun, Y.; Kuntz, I. D. *J Am Chem Soc* 1999, 121, 8033-8043.