Protein Engineering vol.12 no.8 pp.639–647, 1999

Automated docking of peptides and proteins by using a genetic algorithm combined with a tabu search

Tingjun Hou, Junmei Wang, Lirong Chen[1] and Xiaojie Xu[2]

Department of Chemistry, Peking University Jiuyuan Molecular Design Laboratory and [1]Department of Technical Physics, Peking University, Beijing 100871, China

[2]To whom correspondence should be addressed

A genetic algorithm (GA) combined with a tabu search (TA) has been applied as a minimization method to rake the appropriate associated sites for some biomolecular systems. In our docking procedure, surface complementarity and energetic complementarity of a ligand with its receptor have been considered separately in a two-stage docking method. The first stage was to find a set of potential associated sites mainly based on surface complementarity using a genetic algorithm combined with a tabu search. This step corresponds with the process of finding the potential binding sites where pharmacophores will bind. In the second stage, several hundreds of GA minimization steps were performed for each associated site derived from the first stage mainly based on the energetic complementarity. After calculations for both of the two stages, we can offer several solutions of associated sites for every complex. In this paper, seven biomolecular systems, including five bound complexes and two unbound complexes, were chosen from the Protein Data Bank (PDB) to test our method. The calculated results were very encouraging—the hybrid minimization algorithm successfully reaches the correct solutions near the best binded modes for these protein complexes. The docking results not only predict the bound complexes very well, but also get a relatively accurate complexed conformation for unbound systems. For the five bound complexes, the results show that surface complementarity is enough to find the precise binding modes, the top solution from the tabu list generally corresponds to the correct binding mode. For the two unbound complexes, due to the conformational changes upon binding, it seems more difficult to get their correct binding conformations. The predicted results show that the correct binding mode also corresponds to a relatively large surface complementarity score. In these two test cases, the correct solution can be found in the top several solutions from the tabu list. For unbound complexes, the interaction energy from energetic complementarity is very important, it can be used to filter these solutions from the surface complementarity. After the evaluation of the energetic complementarity, the conformations and orientations close to the crystallographically determined structures are resolved. In most cases, the smallest root mean square distance (r.m.s.d.) from the GA combined with TA solutions is in a relatively small region. Our program of automatic docking is really a universal one among the procedures used for the theoretical study of molecular recognition.

Keyword: automated docking/genetic algorithm/molecular recognition/tabu search

Introduction

Molecular docking can fit molecules together in a favorable configuration to form a complex system. Molecular docking has been shown to be very effective in the study of protein–ligand interactions, and the structural information from the theoretically modeled complex may help us to clarify the mechanism of molecular recognition, and may even suggest how the structure of the receptor or ligand may be changed in order to improve some biological function or for the design of new compounds.

The computational procedure of docking for protein–protein and protein–peptide can be classified into three levels by the degree of approximations (Fraga *et al*., 1995)—RBD (rigid-body docking) (Jiang *et al*., 1991), SFD (semi-flexible docking) (Shoichet *et al*., 1991) and FD (flexible docking) (Hart *et al*., 1992; Luty *et al*., 1995). The RBD computation usually uses the crystal structures of bound complexes to perform docking calculations, but it is also useful just for bound complexes. It is difficult to find the appropriate associated sites for unbound complex systems because this model does not consider the conformational change during the docking process; however, it is very common for minor conformational changes to result when an active molecule associates with its substrate. SD can be viewed as soft-docking which allows for minor conformational changes when a receptor binds with its substrate. One successful method, Fan Jiang's soft-docking procedure (Jiang *et al*., 1991), uses a cube representation of the molecular surface and volume. From this procedure it is possible to design a simple algorithm for a six-dimensional search and to embody implicitly the effects of the conformational changes caused by complex formation. FD is mainly used in the small active molecule–protein systems. In the search, the transnational and rotational degrees of freedom are restricted and some twist angles of the ligand have been treated as variables in the energetic function. In the docking method of Luty *et al*. (1995), molecular dynamics is used to evaluate the ligand–receptor interaction. However, while it is well known that fully considering the conformational changes near the active site is very time-consuming, it is still impossible to carry out full energy minimization at each local position for large complex systems, such as protein–protein systems.

During the binding process between a receptor and its substrate, its energetic potential surface is so complicated that it is impossible to determine the associated site by carrying out minimization using gradient methods such as the steepest descent and Gauss–Newton methods. These methods fall into the local potential wells very easily. Some stochastic methods, including Monte Carlo simulated annealing, have been introduced into the study of molecular association, usually with a more complete potential energetic function. We have introduced the Simplex method into the minimization procedure and we found that it could overcome the local minima more easily than Gradient methods. Combined with a random search, Simplex methods can offer a good set of answers to some systems (Wang *et al.*, 1997). Genetic algorithms, which are regarded as intelligent stochastic methods, were introduced into computational chemistry in the late 1970s (Michalewicz *et al*., 1994). Now genetic algorithms have been used in other fields of computational chemistry. Oshiro *et al*. (1995) started work on docking procedures in 1994. The development of two kinds of GA method was proposed, which were a sphere-based GA method and a explicit orientation-based method. Both of the two GA methods aimed at optimizing orientations and conformations of the ligand. The fitness of each of the 'chromosomes' was the molecular mechanics interaction energy. More recently, the tabu (or taboo, TS) (Glover *et al.*, 1993) search has begun to attract attention as an effective heuristic search procedure for combinatorial optimization problems in the molecular design field. David *et al*. (1997) was the first to apply this search method to a docking procedure and proved it was very effective in finding the proper binding mode.

We have compared several heuristic algorithms in a previous study (Hou *et al*., 1999), which showed that a genetic algorithm and tabu search were both superior to the Monte Carlo simulated annealing algorithm. But from a comparison of the results, we found that these two algorithms did not perform very effectively in all conditions, in some cases both GA and TS showed bad results. It is difficult to solve a docking problem thoroughly when using only a single algorithm. Although genetic algorithms can overcome the potential barriers successfully in some cases, it is still common that GA staggers in local potential wells in most cases. With respect to the escape from local minima, TS seems more superior than GA; however, it converges relatively slower, especially near the best solutions. So according to their merits and shortcomings, a hybrid algorithm (HA) combining GA with TS was proposed. The hybrid algorithm was applied to explore the possible associated sites of protein–peptide and protein–protein complexes. It is expected that the hybrid method may overcome the potential barriers more easily. In our laboratory, we have developed different score functions for the following two stages of conformation searching. In the first stage, surface complementarity is considered, while in the second stage only energetic complementarity is considered. From extensive studies, we found that the steric complementarity was more important than energetic complementarity, especially for protein–protein and protein–peptide systems. Moreover, in order to take account of the conformational flexibility of the ligand and the protein, two strategies were introduced into our docking procedure. The first docking stage is a kind of soft-docking, some degree of surface overlap is tolerated to account for side-chain flexibility of the proteins. The second docking stage is a sort of flexible-docking, to some relatively small ligands, the internal conformational flexibility of the ligands are also taken into account, some torsion angles of the ligand are allowed to rotate freely. Using these two strategies, the conformational changes can be well considered to some extents. The program is written in C language and run under the Unix, Dos or Windows operating systems, and our molecular docking procedure has been embedded into the Peking University Interaction Computational System (PUICS) as a separate module. In this study, all ligands were considered as soft bodies, but their torsion angles cannot be allowed to rotate freely.

Materials and methods

Genetic algorithm

The idea of a genetic algorithm was borrowed from genetics and natural selection. A population of 'chromosomes' encoding solutions to the problem is first generated and then it 'evolves' through a process similar to biological evolution, including genetic crossover, genetic mutation and natural selection. Chromosomes encoding good partial solutions survive, reproduce and combine to generate new chromosomes, which hopefully encode better solutions in the succeeding generations. Chromosomes with small fitness will gradually perish in the succeeding generations.

The strength of the genetic algorithm lies in its ability to handle a large and diverse set of variables. For example, genetic algorithms have been considered as one of the two strongest and hopeful methods in conformational analysis (Fraga *et al.*, 1995) (the other one is molecular dynamics) which involve a large set of variables (twist angles). The genetic algorithm has been widely used in computational chemistry, more detailed introductions can be found in Michalewicz *et al.* (1994).

Tabu search

The tabu search was first suggested by Glover *et al.* (1993) and originally applied in the field of operation research. However, compared with the field of computational chemistry, it may be a somewhat new algorithm. The basic idea of the method, described by Glover *et al*. (1993), is to explore the search space of feasible solutions by a sequence of moves, and, in the mean time, some restrictions will be imposed to enable a search process to rake otherwise difficult regions. The real foundation of the tabu search may be sought in concepts that systematically violate feasibility conditions, as in heuristic procedures based on surrogate constraints, or even in cutting plane algorithms.

A tabu search will only remain one current solution during the course of a search. First, an initial solution is specified or randomly generated at the start of the iterations, then, some moves are generated from the current solution. Each of these moves is evaluated using the evaluation function and they are ranked in order (the best move at the head of the list). Moves are considered as tabu if they are not different enough from those solutions in the tabu list (we will define the criteria to check whether they are tabu). The best move will be accepted if it is better than other solutions in the tabu list. So, only non-tabu solutions will be accepted. If neither of these criteria can be met, the iteration cycle is terminated. If a new current solution can be found, it will be added to the tabu list. If the tabu list is full after several iterations, the current solution will replace one of the solutions in the tabu list. Usually, the tabu list will be managed in a 'first-in, first-out' manner. When a new current solution has been identified and stored, additional moves are generated from it and the search procedure will continue with a new iteration. After a number of iterations, if the best solution cannot be changed, 'convergence' is achieved, the tabu search exits and the best solution will be returned. The restrictions imposed in the tabu search mean that this algorithm can search relatively large areas after many iterations, and it has been proven that tabu searches can efficiently find the global solution to difficult optimization problems.

The hybrid algorithm combined with GA and TS

A comparison of the GA and TS algorithms shows that they both have their merits and limitations. GA converges faster

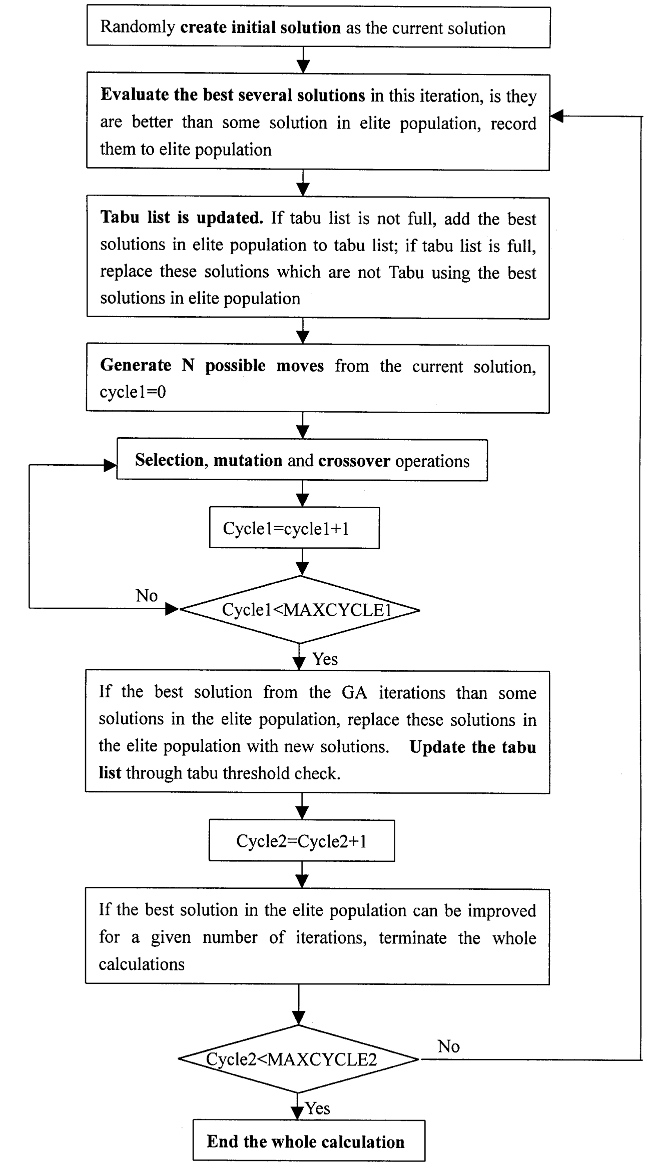
Randomly create initial solution as the current solution
Evaluate the best several solutions in this iteration, is they are better than some solution in elite population, record them to elite population
Tabu list is updated. If tabu list is not full, add the best solutions in elite population to tabu list; if tabu list is full, replace these solutions which are not Tabu using the best solutions in elite population
Generate N possible moves from the current solution, cycle1=0
Selection, mutation and crossover operations
Cycle1=cycle1+1
No
Cycle1<MAXCYCLE1
Yes
If the best solution from the GA iterations than some solutions in the elite population, replace these solutions in the elite population with new solutions. Update the tabu list through tabu threshold check.
Cycle2=Cycle2+1
If the best solution in the elite population can be improved for a given number of iterations, terminate the whole calculations
Cycle2<MAXCYCLE2
No
Yes
End the whole calculation

Fig. 1. The flow chart of the hybrid minimization algorithm. The MAXCYCLE1 represents the number of GA iterations, the MAXCYCLE represents the maximum number of TS iterations.

when near the best solution, and can find it very quickly, but GA can fall into local minima relatively more easily. In contrast with GA, TS can avoid falling into local minima, but it converges relatively more slowly. As a result, a hybrid algorithm was proposed. The basic procedure of the hybrid algorithm is similar to TS, but compared with traditional TS, it is different with respect to two points. The first modification is that after *N* possible moves from the current solution, some extra steps of crossover and mutation operations, which come from GA, are added. The test results showed that this modification offered particular advantages over the traditional tabu search. The second modification is that we use an elite population to remain several best and different solutions in the crossover and mutation operations. After 20–30 steps of crossover operation, *N* new solutions are ranked, and several of the best several solutions in the elite population are compared with the solutions in the tabu list to check if they are tabu; if they are, then the GA iteration will be performed again. In a traditional tabu search only one current solution remains during the search, but in the second modification, several solutions remain, which may help the hybrid algorithm converge faster. The scheme of the hybrid algorithm is illustrated in Figure 1. The new hybrid algorithm combines the advantages of both GA and TS; it not only converges faster, but also does not fall into local minima easily.

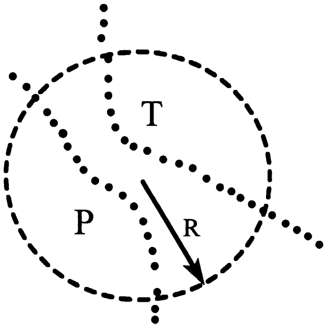

Fig. 2. The definition of seperation angles. The **t** vector is the normal vector of a surface dot on the target molecule and the **p** vector is the normal vector of a surface dot on the probe molecule. The **t** vector is an extension line of the **t** normal vector. The θ angle is the separation angle of **p** normal vector. If the separation angle is smaller than a given criterion, say 30–40°, the surface area of the two dots will be added to the surface score.

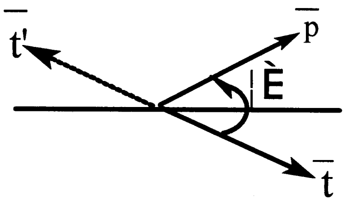

Fig. 3. A picture shows that the local area will be used to calculate the matching complementarity. The center of the ball is a dot of the target molecule and matching complementarity test is performed only on those probe dots which are within the ball. P stands for the probe molecule and T stands for the target molecule. R is usually taken to be a value 1.2–1.5 times that of the sum of the probe atom radius and target atom radius.

Rough searching potential bound sites based on surface complementarity

In the first step, the dot surface is generated using the MS program written by Connolly (1983). The parameters used in this program are discussed later. Then the coordinates of the probe molecule and the target molecule as well as their surface dot are randomly rotated and translated. The surface dot coordinates have also been synchronistically moved with the probe and target molecules. Then an initial solution is randomly generated containing six variables, three translational degrees of freedom and three rotational degrees of freedom. The three rotational variables are described by three Euler angles. The position of the target molecule is fixed and six variables define the orientation of the probe molecule.

The initial solution is evaluated using surface complementarity. The evaluation score is composed of two parts: the matching score and penalty score of atomic overlapping. The matching score is calculated in the following way. For each surface dot of the probe molecule, the matching property with target surface dots within a certain distance is repeatedly tested (Figure 2). The distance usually reaches a value 1.0–2.0 times the sum of radii of the two atoms which hold the two dots. The normal lines of the two interesting dots have a separation angle and if the angle is smaller than a threshold, says 30°, the two dots can be assumed to be 'matched' and the areas shared by the two dots are added. The total value of those areas is used to evaluate the matching complementarity (Figure 3).

The next part of the steric complementarity is the penalty score of atomic overlapping. If the distance between the two atoms is smaller than a given threshold, say 0.8–0.9 times the sum of the van der Walls radii of the two atoms, the two atoms are considered to be overlapped. The penalty score is evaluated by the number of overlapped atomic pairs multiplied by an empirical parameter. The evaluation function at this stage is the total of two parts:

$$\text{Fitness} = \text{Score}_{\text{match}} - \text{const} \times \text{Score}_{\text{overlap}} \qquad (1)$$

where $\text{Score}_{\text{match}}$ is the matching score and $\text{Score}_{\text{overlap}}$ is the penalty score. Const is a coefficient balancing the contributions of the two parts. Const is mainly determined by the dot density, an important parameter of the MS program, which is defined as the average dot numbers per angstrom square area of both probe and target molecules. Const usually takes a value 5–20-fold that of the dot density. Then the hybrid algorithm is performed iteratively, Equation 1 is used to evaluate all solutions. After convergence, a set of solutions is obtained from the tabu list. Using the hybrid algorithm, in most cases we have successfully overcome the tendency of GA to stagger in the local extreme points. Generally, if we performed 50 steps of crossover and mutation operations for one TS iteration, 50–100 iterations were enough to find the appropriate sites. If the best several solutions cannot be improved after 10–20 tabu iterations, 'convergence' is achieved. For a typical system containing more than 1500–3000 atoms, this process takes between 20 and 30 h on a PC with a Pentium II 350 processor. Moreover, cluster analysis is not needed in this stage. Because in the minimization process every solution has been checked whether it is tabu; the threshold measure used in this paper to determine the tabu status has a r.m.s.d. of 3 Å or less between the two solutions being compared. Then, a more detailed searching will be performed for each conformation at the next stage.

Detailed searching the local associated sites based on energetic complementarity

In this stage, a more detailed search is performed for each solution derived from the first stage. The position of the target molecule is also fixed and some changeable variables are defined for the probe molecule. In this stage, only a local search was performed near these binding sites from the surface complementarity. Considering the fast convergence of GA near the best solution, we usually only use GA in the local search. A set of 'chromosomes' are randomly generated and each one represents an orientation. The fitness score of each 'chromosome' is the interaction energy between the probe and target molecules. Only van der Walls energy, electrostatic energy and hydrogen-bond energy are considered. The force field was AMBER (Weiner *et al.*, 1984, 1986). Non-polar hydrogen atoms were omitted for simplification and united atom types were introduced in order to evaluate the interaction energy more precisely. Our purpose of this step is mainly to purge the high-energy conformations after the surface complementarity, and in the mean time to calculate the interaction energy precisely. When the unbonded interaction energy remains stable in a user-defined region after 20–30 iterations, 'convergence' is achieved. At this stage, cluster analysis should be performed and the solution with the largest fitness is selected to calculate the r.m.s. with the crystal structure.

At this stage, different systems are treated in different manners. For protein–protein and some protein–peptide systems, due to the high flexibility of the ligand, it is very difficult to consider their conformational change at this stage; so for relatively large ligands, only three degrees of translation and three degrees of rotation are considered. For protein–small molecule systems, a flexible docking procedure is applied, the internal conformational flexibility of the ligand is taken into account, and some torsional angles are defined as variables in the GA minimizations.

The difference between the two stages lies in the fact that the first one optimizes the orientations in the whole translational space; however, in the second stage, the translational vector is restrained near the associated sites derived from the first stage. The CPU used in this step is only about 5 percent of that in the first stage.

Result and discussion

Seven complexes randomly selected from the Protein Data Bank (PDB) have been used to test the hybrid minimization algorithm and our two-stage soft-docking procedure. All crystallographic water molecules were eliminated from the structures. Some missing hydrogen atoms were added to the complexes using the molecular design software InsightII, with a neutral sp^3 N-terminus and a carboxylic (COOH) C-terminus assigned at neutral pH. Before the calculations, these structures were minimized using the AMBER force field to remove any steric overlap with a restrain of the main chain. Some parameters for these seven complex systems are shown in Table I. Two classes of complex were chosen, including five bound complexes and two unbound complexes. For these five bound complexes, we attempted to regenerate the crystal structure, this part of work is mainly used to test the hybrid minimization algorithm. The other two unbound complexes were more realistic, which can be used to test the capability of the minimization algorithm and docking procedure.

The influence of the parameters

In our docking procedure, so many parameters need to be carefully calibrated. So before every docking calculation, these parameters must be properly defined. Some important parameters and their abbreviations in our docking procedure are listed in Table II. The parameters can be divided into two types: four parameters are concerned with surface and energetic complementarity, the other seven are concerned with the hybrid minimization algorithm. An appropriate selection of these parameters is important, since they affect not only the total computational time but also the quality of the solutions.

For surface complementarity, the two parameters, const and separation angle, seem to be very critical. Const can be the balance of the contributions of two parts of the fitness score in the first stage. It usually takes a value 5–10-fold that of the dot density used in the MS program. This parameter can be varied in a relatively large range without affecting the quality of the last solutions. The separation angle is usually smaller than 60°. For all test systems in this study, const is set to 5-fold that of the dot density and the separation angle is set to 30°.

For the hybrid minimization algorithm, some parameters will greatly affect the calculation results. The parameters of the hybrid minimization algorithm include tabu size, tabu threshold measure, tabu iterations, the number of moves

Table I. Test cases used in our calculations

Molecular names	Probe atoms[a]	Target atoms	Probe dots[b]	Target dots
1CTA	272	271	908	915
1CKA	64	479	342	1275
4DFR	1286	1258	2846	2952
2PTC	454	1629	1184	3347
1PLG	1620	1671	3630	3863
2PTC[c]	454	1629	1184	3347
2KAI[c]	437	1799	1211	2128

[a] The number of probe and target atoms represent only the number of the atoms directly from the protein data bank.
[b] The probe radius to generate the Connolly surface is defined as 1.5 Å
[c] These two test cases are unbound complexes, the receptor and ligand molecules, respectively, come from different protein molecules.

Table II. Definitions of parameters used in our docking procedure

Parameter abbreviation	Parameter meaning	Variable range
1. Speration_Angle	Separation angle defined in Figure 1	30–40°
2. Overlap_const	Coefficient of the total of the overlapping atomic pairs in formula 2. It can balance the contributions of two parts of steric complementarity score	5–20
3. Dot_Density	Number of dots per angstrom square of molecular surface generated by the MS program	0.25–1
4. Probe_Radius	Radius of probe atom in the MS program	1.4–2.0
5. Tabu_List	The number of remaining solutions in the tabu iteration	10–50
6. Tabu_Thresh_Measure	Thresh value used to determine the difference between the present solution with those solutions in the tabu list	2–5 Å
7. Tabu_Iteration	Number of tabu iterations in the calculations	100–500
8. Tabu_Moves	Number of moves produced based on the present solution	50–200
9. Elite_Size	Elite size of chromosomes	2–10
10. Mutation_ Ratio	Ratio of mutation operation	0.05–0.1
11. Crossover_Ratio	Ratio of crossover operation	0.30–0.40
12. Select_Ratio	Ratio of selection operation	0.80–0.90

(population size), elite size, mutation ratio and crossover ratio. The tabu size controls how many different solutions should remain in the tabu list, definition of this parameter by the user is optional. In general, a 10–20 tabu size is recommended in our docking procedure. A larger tabu size did not improve the final solutions by much, but more time was consumed on solution comparison. In this study, the tabu size for all test sets were all defined as 20. From our studies, Tabu threshold measure was a very important parameter, it directly connects with the efficiency of final solutions, this parameter is used to determine if the solutions after the GA interruptions are different enough from those solutions in the tabu list. In this study, the r.m.s.d. between two conformations was used to compare two different docked binding modes, and the tabu threshold measure could vary with specific studied system. For larger complexes, this parameter can be defined larger, but for some smaller complexes, this parameter can be smaller. For example, in systems 1, 2 and 3, the tabu threshold measure was defined as 2 Å, but in other systems, the tabu threshold measure was defined as 4 Å. Another parameter, the number of moves (population size) means that for every tabu iteration, how many moves are generated based on the current solution. This parameter also represents the population size in GA iterations. The larger the number of moves, the greater the chances that global orientations can be found and the more time that will be consumed. The elite size is the number of best solutions directly into the next GA iterations and tabu list in the elitist strategy. In the calculations, the elite size is about 5 percent of the number of moves, the number of moves is 50 and the elite size is 2. The mutation ratio and crossover ratio are defined as 0.05 and 0.35, respectively.

Five bound complexes

The initial five complexes in Table I are bound complexes of protein–protein and protein–peptide. The crystal structure of the complexes are directly from the PDB, the components of each complex were taken apart at an arbitrary relative orientation, then our docking procedure was used to dock together and compare the docked complexes with the crystal structures of the complexes.

In the surface complementarity stage, in order to speed up the calculations, all hydrogens were omitted. Moreover, for energetic complementarity and surface complementarity, two types of reference frames were used. At the surface complementarity stage, the alternate space axes remained the same as the initial axes from PDB. But in the energetic complementarity stage, the alternate space axes were transformed, after transformation, the origin of the coordinate was superposed with the gravity center of the ligand molecule after the surface complementarity. The goal of the transformation was to leave a relatively small rotation vector for the local search stage of energetic complementarity, thus restraining the movement of the ligand within a relatively small region. In the second stage, the translational extents cannot exceed 4 Å. In the energetic complementarity stage, a cut-off of 12 Å was used to calculate the van der Waals interaction, another cut-off at 16 Å was applied in the calculation of the Coulombic interaction.

Table III. The results of molecular docking calculations for six bounded protein complexes

PDB code	Solution number	Rotational Eular angles (radian)			Translational vector (Å)			Surface score	Interaction energy (KJ/mol)	RMS (Å)
1CTA	1	103.44	5.34	260.91	–0.69	0.30	0.09	1981.23	–	0.97
		22.52	334.23	222.13	0.23	–0.12	0.08	–	–345.67	2.37
	2	5.34	21.46	249.95	5.28	11.27	–11.97	1672.71	–	13.36
		14.34	111.26	12.19	0.34	–1.12	0.37	–	–318.76	17.68
	3	260.91	328.04	332.50	–4.32	8.16	12.96	1077.53	–	11.35
		256.33	335.87	12.09	1.76	1.11	–0.87	–	–223.12	10.60
1CKA	1	199.66	12.71	240.85	2.01	–2.48	13.34	895.68		0.56
		103.58	350.69	253.01	–0.31	1.08	–0.13		–1468.00	0.61
	2	121.50	346.05	242.59	–15.31	4.71	–3.89	792.58		9.05
		18.35	338.52	245.74	–1.15	1.02	0.10		–932.16	3.66
	3	142.72	7.85	223.87	18.07	–0.89	3.51	742.21		4.55
		175.44	14.46	189.92	–0.70	0.48	0.08		–1532.83	0.68
4DFR	1	0.69	0.25	0.79	0.92	–0.66	0.68	1674.22		0.59
		0.38	0.11	1.81	0.01	–0.32	0.79		–812.93	0.76
	2	4.72	356.70	1.69	–6.49	–13.92	–10.55	1443.88		16.61
		299.30	34.93	288.30	–1.92	0.68	1.87		–456.91	19.89
	3	99.56	28.04	243.50	11.81	–10.85	18.37	811.44		40.24
		188.39	129.57	12.91	1.91	0.58	–1.17		138.93	41.10
2PTC[a]	1	161.72	359.92	195.45	–3.56	1.18	0.10	1944.83		0.48
		140.49	79.22	249.79	–1.16	0.51	–1.05		–1137.22	1.86
	2	67.04	342.49	294.48	–4.66	–8.62	9.27	1839.80		20.57
		106.24	17.79	12.41	–1.90	1.92	1.92		–747.660	17.02
	3	168.56	210.59	210.59	–16.02	13.92	–2.40	935.60		10.73
		76.35	112.39	12.34	1.16	2.37	–1.12		117.89	10.63
1PLG	1	123.93	123.93	233.83	12.98	1.97	12.33	38187.11		1.18
		34.83	10.20	222.9	1.87	1.87	–0.73		–1145.98	2.50
	2	234.93	33.94	34.93	23.93	17.97	–4.39	2599.83		13.87
		45.83	12.93	139.19	0.27	1.87	–2.00		–786.93	12.91
	3	198.48	45.95	63.93	24.93	10.39	6.38	2345.38		15.07
		45.93	35.93	127.38	0.39	–1.98	0.23		–374.21	19.10

Table III lists the calculated conformations from both the first and second stages. For the surface complementarity, the top three solutions are given in Table III; for the energetic complementarity, only the best solution for each binding mode from the surface complementarity was given. For these five systems, the correct binding model could be found within 50 tabu iterations. The calculated results were very encouraging (see Table III and Figure 4); in most cases, the smallest root mean square distance is smaller than 1.0 Å (besides 1PLG). To these five bound complexes, the correct binding conformation was only determined precisely using surface complementarity. In our laboratory, abundant bound complexes were calculated using our docking procedure, in most cases, the best binding conformation possesses the best surface complementarity. But in a very few cases, another binding conformation may exist with a better surface complementarity than the correct binding conformation; these calculation results will be discussed in the next paper. But, in general, given a reliable surface score function, you can be sure of finding the correct binding conformation.

The calculation results in Table III show that detailed energetic complementarity could not improve the results from the surface complementarity. The values of r.m.s.d. for these five test cases all produce larger changes after energetic minimization. It seems that for bound complexes, surface complementarity is more precise than energetic complementarity. So, the challenge is to develop some simple methods to evaluate the binding free energy between a ligand and a receptor more precisely. But in this study, our main goal is not to calculate the binding free energy, the goal of detailed minimization stage was to deter the high energetic conformation and search relatively low-energy conformations. We believe that our energetic complementarity is precise enough to filter these solutions from the surface complementarity. The docking results in Table III shows that the conformations near the correct binding mode possess relatively low unbonded interaction energy, the interaction energy of the best solution is much lower than the interaction energy of the other two solutions. From the simple interaction energy, the correct and incorrect binding conformation can be clearly partitioned. So, for bound complexes, using only a precise surface complementarity is enough to get the correct binding conformations in most cases.

Two unbound complexes

The ultimate goal of molecular docking is to predict protein–protein and protein–peptide interactions without requiring a complexed crystal structure. Compared with the docking of these complexes with crystal structures, calculations for complexes without crystal structures seem more difficult. During the formation of a complex, some molecules will undergo conformational changes, so the docking procedure must be sufficiently soft to manage conformational changes, yet specific enough to identify the correct solution. In some cases, especially, the binding regions between protein and protein or peptide are unknown, complete search using flexible body is not tractable. Even using rigid-body, it is very difficult to determine the global minimum using conventional minimization algo-

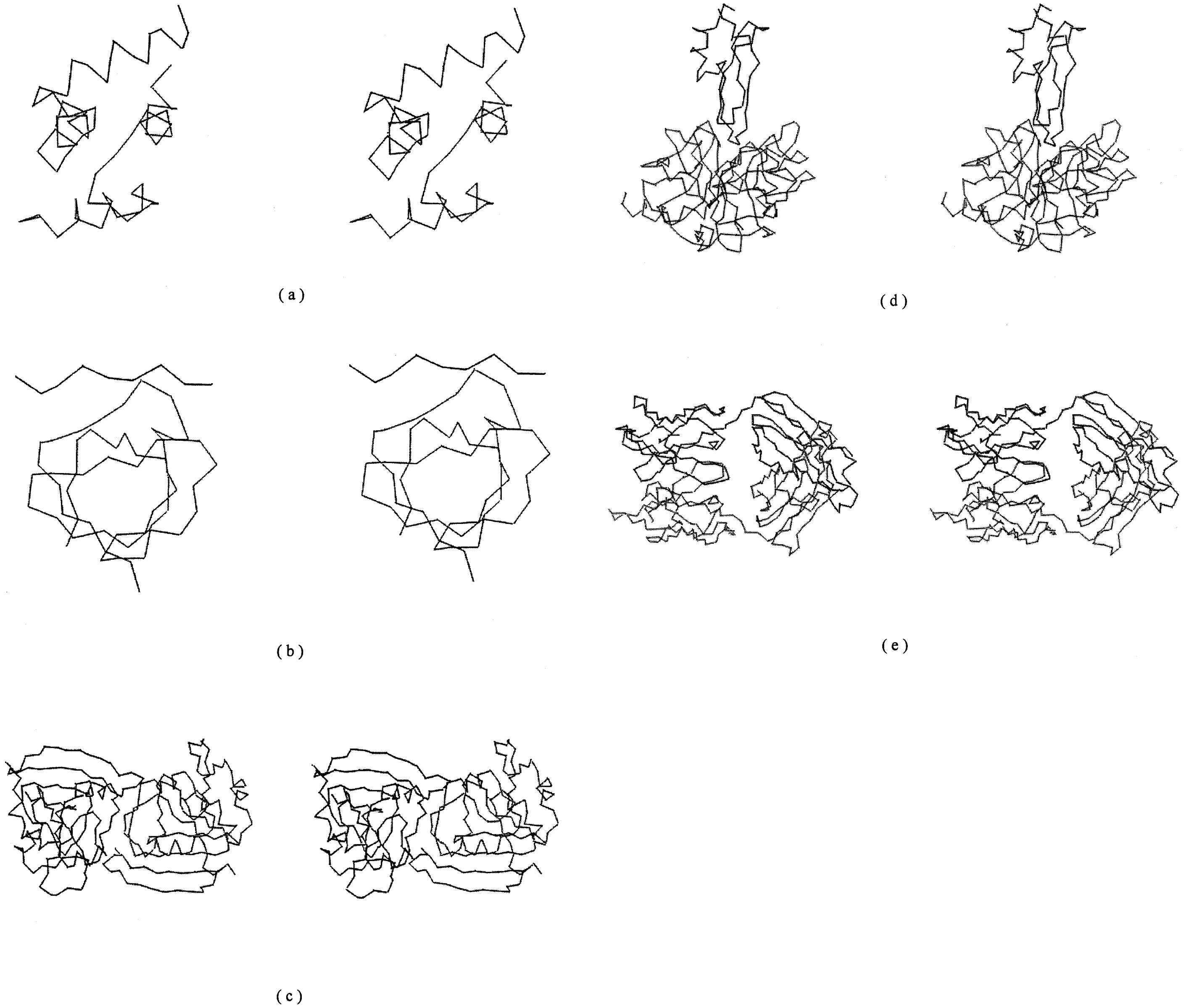

Fig. 4. The fitting structure of five bound complexes. (**a**) Troponin C-site III-site III homdimer (1CTA); (**b**) C-CRK complexes with C_3G peptide (1CKA); (**c**) dihydrofolate reductase complexes with methotrexate (4DFR); (**d**) β-trpsin complexes with pancreatic trypsin inhibitor (2PTC); (**e**) immunoglbulin IGG2A = KAPPA= (1PLG). For every case, the crystal structure is shown on the left, the predicted docking structure is shown on the right.

rithms. In these two cases, the calculation results are listed in Table IV, the highest ranked correct prediction is shown in Figure 5.

In order to test our hybrid minimization algorithm and docking procedure, two uncomplexed systems were chosen. For this, an uncomplexed trypsin inhibitor (4PTI in PDB) and an uncomplexed trypsin (3PTN in PDB) were used in one instance, and an uncomplexed serine proteinase (2PKA) and an uncomplexed boveine pancreatic trypsin inhibitor (2BPI) were used in another instance. The PDB codes for these two cases are 2PTC and 2KAI.

For 2KAI, it can be found that the best solution from the surface complementarity no longer corresponds to the correct docking conformation. After detailed energetic minimization and superimposed with the crystal structure of bound 2KAI, a good solution was found in four out of 10 solutions from the tabu list. So for some unbound complexes, using surface complementarity alone cannot reliably dock unbound complexes, further energetic complementarity is needed to filter these solutions from the surface complementarity. In the mean time, we cannot conclude that the correct solution is sure to have the best energetic complementarity, because in the docking process, we do not really consider the flexibility of the systems. For 2KAI, the second solution has the smallest interaction energy, but its r.m.s.d. was larger than 10 Å.

The crystal structures of uncomplexed trypsin (3PTN) and an uncomplexed trypsin inhibitor (4PTI) have been solved separately in different crystal forms. A comparison of their structures with the corresponding components of a complex has indicated that relatively large conformational changes have occurred, especially in the trypsin inhibitor. After superimposing only the backbone atoms for 3PTN and 4PTI, the r.m.s.d. for 3PTN is only 0.323 Å; but for 4PTI, the conformational change is relatively large, its r.m.s.d. is 1.272 Å. This test case

Table IV. Results of molecular docking calculations for two unbounded protein complexes

PDB code	Solution number	Rotational Eular angles (radian)			Translation vector (Å)			Surface score	Interaction energy (kJ/mol)	R.m.s. (Å)
2PTC	1	30.86	335.75	320.38	14.59	11.36	–12.42	1137.59		2.54
		38.19	9.11	31.66	0.34	–1.15	–1.98		–893.47	2.71
	2	78.03	3.37	325.50	1.16	9.19	16.95	872.78		13.56
		5.16	104.05	312.32	–2.65	1.92	–2.92		–552.01	14.98
	3	128.51	4.13	260.03	–3.62	12.31	16.02	927.55		12.98
		103.25	185.54	105.23	1.50	0.95	0.06		–651.855	11.53
2KAI	1	23.34	9.07	335.17	4.66	–17.26	2.06	973.20		11.98
		34.95	353.03	39.20	–1.01	0.28	1.11		–723.95	10.64
	2	14.97	0.82	316.42	–7.20	7.04	–7.72	726.26		15.98
		134.93	320.42	150.20	–2.00	1.04	–0.85		–965.84	17.83
	3	12.05	1.16	327.56	–1.74	–2.49	–2.48	703.11		14.87
		39.83	334.93	358.23	1.76	0.38	0.33		–232.95	14.99
	4	93.84	1.83	231.47	–10.45	5.13	–1.84	675.05		1.72
		94.83	134.83	338.39	1.06	–0.66	–8.35		–934.87	1.62

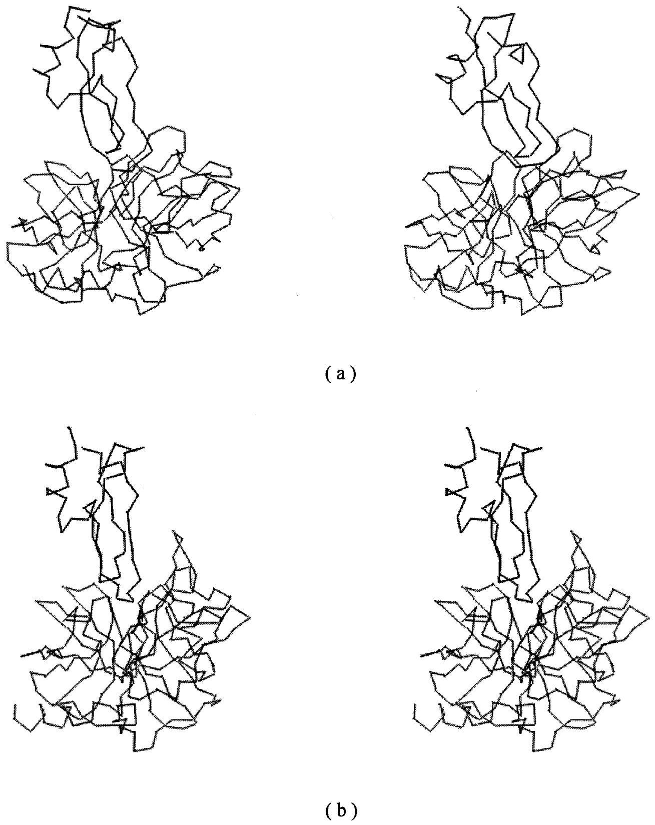

Fig. 5. The fitting structure of two unbound complexes. (**a**) β-Trpsin complexes with pancreatic trypsin inhibitor (2PTC); (**b**) kallikrein A complexes with bovine pancreatic trypsin inhibitor. Since the two structures cannot be distinguished in the superimposed forms, the fitted structure is moved away from the crystal structure. For each case, the left picture is the crystal structure and the right one is the fitted structure.

is very challenging, because it has been extensively studied by several other docking procedures (Kaichalski-Katzir *et al.*, 1992; Gabb *et al.*, 1997). The attempt by Katchalski-Katzir *et al.* (1992) to dock 3PTN and 4PTI was unsuccessful. Our calculation results listed in Table IV show the correct binding conformation was successfully found. When we superimposed the native complexed crystal structure of 2PTC with our docking result, the r.m.s.d. was 2.54 Å (only backbone atoms were considered in the calculation of r.m.s.d.). From Table IV, it is obvious that the best solution from surface complementarity corresponds with the correct solution, but it no longer significantly better than the rest of the solutions. It is relatively difficult for us to determine if the first solution is the correct solution, but after the second stage of energetic complementarity, we found that the first solution was more energetically favorable, in fact, this solution was very close to the correct solution. In order to compare the potential influence of the conformational change and test our minimization algorithm further, we docked receptor and ligand from the complexed protein structure together (see case 4 in Table III). For the bound and unbound structures, the same parameters were used. But we found that the calculation results were so different, the best solution of surface complementarity for the bound system was much better than that for the unbound system. When comparing their r.m.s.d., the result for bound system was significantly superior to unbound system, the r.m.s.d. for bound system is only 0.475, much smaller than that of the unbound system. The reason for these differences between the bound and unbound complexes are mainly derived from the conformational changes during the process of forming the complex. These conformational changes may greatly affect the shape of a molecular surface. Minor changes in the molecular surface, especially near the binding site, will greatly affect the docking results. Our methods only implicitly consider the conformation change for these molecules near the binding site, but the conformations for these molecules do not really change. In this circumstance, the docking results will produce some deviations from the real complex. But for 2PTC, the surface did not change a lot in the process of complex formation, the essential molecular shape does not change greatly. Generally, for most unbound complexes, the best binding conformation may not correspond with the best surface or energetic complementarity, but some binding conformations with a relatively large score of surface and energetic complementarity surely can be found near the correct binding mode. So using surface complementarity and a good minimization algorithm, combined with the filter of energetic complementarity, in most cases, it is possible for us to predict the binding mode for an unbound complex.

Technical information

The complete docking package, named SFDOCK, consists of approximately 5000 lines of C language code, including a soft-docking procedure for protein–protein interactions, a flexible-docking procedure for the small molecules–protein procedure and a database searching procedure for ligand design. All docking experiments were carried out on a PC. The soft-docking procedure has been embedded into the Peking University Interaction System as a separate module and can be easily used through a graphics interface. Some source code from this study can be obtained from the author upon request.

Acknowledgements

This project was supported by grant NSFC 29873003 and was an NSFC key project.

Reference

Connolly,M.L. (1983) *Science*, **221**, 709–713.

David,R.W., David,E.C. and Christopher,W.M. (1997) *J. Comp. Mol. Des.*, **11**, 209–228.

Fraga,S., Parker,J.M. and Pocok,J.M. (1995) *Computer Simulations of Protein Structures and Interaction*. Springer Verlag, Berlin–Heidelberg–New York, p. 2081.

Glover,F. and Laguna,M. (1993) In Reeves,C.R., *Modern Heuristic Techniques for Combinatorial Problem*. Blackells, Oxford, UK, pp. 70–150.

Hou,T.J., Wang,J.M. and Xu,X.J. (1998) *Chinese Chemical Letters*. (In press).

Jiang,F. and Kim,S.H. (1991) *J. Mol. Biol.*, **219**, 79–102.

Katchalski-Katzi,E., Shariv,I., Eisenstein,M., Friesen,A.A., Alfalo,C. and Wodak,S.J. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 2195–2199.

Gabb,H.A., Jackson,R.M. and Sternberg,M.J.E. (1997) *J. Mol. Biol.*, **272**, 106–120.

Luty,B.A., Wasserman,Z.R., Stoutern,P.F.W., Hodge,C.N. and McCammon,J.A. (1995) *J. Comp. Chem.*, **16**, 454–464.

Michalewics,Z. (1994) *Genetic Algorithm + Data Structures = Evolution Programs*. Springer Verlag, Berlin–Heidelberg–New York, pp. 13–30.

Oshiro,C.M., Kuntz,I.D. and Dixon,J.S. (1995) *J. Comp. Mol. Des.*, **9**, 113–130.

Shoichet,K. and Kuntz,I.D. (1991) *J. Mol. Biol.*, **221**, 327–346.

Wang,J.M., Chen,L.R., Jiang,F. and Xu,X.J. (1998) *Proceeding of Chinese Peptide Symposium-96, ESCOM*.

Weiner,S.J., Kollman,P.A., Case,D.A., Singh,U.C., Ghio,C., Alagona,G., Proteta,S.Jr. and Weiner,P. (1984) *J. Am. Chem. Soc.*, **106**, 765–784.

Weiner,S.J., Kollman,P.A., Nguyen,D.T. and Case,D.A. (1986) *J. Comp. Chem.*, **7**, 230–252.

Received September 18, 1997; revised January 22, 1999; accepted May 28, 1999