Chemometrics and Intelligent Laboratory Systems 45 (1999) 281–286

Chemometrics and intelligent laboratory systems

Automated docking of peptides and proteins by genetic algorithm

Junmei Wang [a], Tingjun Hou [a], Lirong Chen [b], Xiaojie Xu [a,*]

[a] *Department of Chemistry, Peking University Jiuyuan Molecular Design Laboratory, Peking University, Beijing 100871, China*
[b] *Department of Technical Physics, Peking University, Beijing 100871, China*

Abstract

Genetic algorithm (GA) combined with random search has been applied to thoroughly search the appropriate associated sites for both peptide and protein complexes. Steric complementarity and energetic complementarity of ligand with its receptor have been separately considered in our two-stage automated docking. Eight complexes have been randomly selected from the Protein Data Bank to test our procedure. Conformations and orientations close to the crystallographically determined structures are obtained. For most cases, the smallest RMS (root mean square of distance) of the GA solutions is smaller than 1.0. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Automated docking; Genetic algorithm; Random searching

Contents

1. Introduction

Automated docking studies of complex can clarify the mechanism of molecular recognition, and so allow us to design new compounds and estimate their activities. Furthermore, docking associated with database screening offers an efficient and practical way to generate the leading compounds which play an important role in drug design. The energetic potential surface of ligand interacting with its receptor is so complicated that it is impossible to determine the associated site by carrying out minimization using gradient methods such as the steepest descent method,

[*] Corresponding author.

PII: S0169-7439(98)00132-4

the Gauss–Newton method. These methods fall easily into the local potential wells and escaping from them is rather difficult. So some stochastic methods including the Monte Carlo simulations have been introduced into the studies of molecular recognition, usually with more complete potential energetic functions in an automatic fashion.

Genetic algorithm, which is regarded as an intelligent stochastic method, was introduced into computational chemistry in the late 1980s [1,2]. The idea of genetic algorithm is borrowed from genetics and natural selection. A population of ‘chromosomes’ encodes solutions to the problem has been first generated and then it ‘evolves’ through a process similar to biological evolution, including genetic crossover, genetic mutation and natural selection. The strength of genetic algorithm lies in its ability to handle large and diverse set of variables. In 1995, Oshiro et al. [3] introduced this method into their flexible docking procedure. Two kinds of GA methods were proposed, which were sphere-based GA method and explicit-orientation-based method. Both GA methods aimed to optimize orientations and conformations of the ligand. The fitness of each ‘chromosomes’ was the molecular mechanics interaction energy. The flexible docking methods had generally produced structure deviations on the order of 1 Å/atom.

Although genetic algorithm can overcome the potential barriers successfully in most cases, it is still common that GA staggers in local potential wells in some special cases, such as automated docking. In this paper, we combined the genetic algorithm with random search to explore the possible associated sites of protein-small molecule, protein–peptide, protein–protein complexes. We hoped the hybrid method may overcome the potential barriers more easily.

In our two-stage genetic algorithm minimization, steric complementarity and energetic complementarity were separately considered to evaluate the fitness of ‘chromosomes’. The first stage mainly finds the binding sites and the second one adjusts the orientations of ligand around the binding site precisely according to the interaction energy. Our procedure is not only effective for enzyme–ligand systems, but also can optimize the orientations of two domains of proteins which is too difficult to embody conformational changes into the GA optimizations. Although our two-stage docking is not a flexible one, the first docking stage is a sort of soft-docking because it allows atomic overlapping to some degree. The second stage belongs to rigid docking and the fitness is the molecular mechanics interaction energy between probe molecule and the target molecule.

2. Methods

2.1. Rough searching a set of bound sites based on steric complementarity

In the first step, the dot surface is generated using the MS program [4] written by Michael Connolly. The parameters used in this program are discussed in next part. Then, the coordinates of the probe molecule and the target molecule as well as their surface dot are centralized. Next, a set of ‘chromosomes’ is randomly generated and each one contains six variables: three translation degrees of freedom and three rotation degrees of freedom. The three rotational variables are described by three Euler angles. The position of the target molecule is fixed and the six variables define an orientation of the probe molecule. For each ‘chromosome,’ the fitness score is composed of two parts: the matching score and penalty score of atomic overlapping.

The matching score is calculated this way: for each surface dot of probe molecule, the matching property with target surface dots within a certain distance is repeatedly tested. The distance usually gets a value of 1.0–2.0 times the sum of radii of two atoms which hold the two dots. The normal lines of the two overlapping dots have a separation angle and if the angle is smaller than a threshold, say 30°, the two dots can be assumed ‘matched’ and the areas shared by the two dots are added. The total value of those areas is used to evaluate the matching complementarity.

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 threshold, say 0.8–0.9 times the sum of the Van der Wallas radii of the two atoms, the two atoms are considered to be overlapped. The penalty score is evaluated by the number of the overlapped atomic pairs multiplied by an empirical parameter. The fitness score in this stage is the total of the two parts:

$$\text{Fitness} = \text{Score}_{\text{match}} - \text{const} \times \text{Score}_{\text{overlap}} \tag{1}$$

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. The 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. The const usually takes 5–20 fold of the dot density.

Random search is easily introduced into genetic algorithm by replacing the lowest 20–30 'chromosomes' in a population with randomly generated ones. This strategy serves to introduce some new 'chromosomes' into the population which helps maintain the diversity of the population and thus reduces the likelihood of the GA converging on a local-optimal minimum. This strategy is more useful than the normal GA which simply increases the mutation ratio to maintain the diversity of the population. Another strategy in our two-stage docking is elitist strategy which copies a set of the best performing 'chromosomes' from one generation unchanged to the next generation in order to maintain the best individuals of the previous generation.

The three operations of GA—crossover, mutation and selection—are then repeatedly performed. If the crossover probability exceeds a randomly generated number, the selected pair of 'chromosomes' is then bred using the crossover operator. This operator divides both parents at a randomly selected points and joins the pieces together to form a pair of new 'chromosomes' which embody the characteristics of their parents. The mutation operator is performed when the mutation ratio is larger than a randomly generated number. This operator randomly selects a 'gene' in the 'chromosome' chain and replaces it with a randomly generated one. The selection operator chooses the 'chromosomes' in a manner according to their fitness values. The fitter the 'chromosome,' the more chances it is selected. If the average fitness score of a population is not improved after a certain iterations, say 50 iterations, the convergence is achieved. In most cases, 2000 iterations are enough to generate orientations similar to that of crystal structure.

Finally, cluster analysis is performed and the conformation with the largest fitness is selected from each group for further detailed searching. The least RMS difference between two groups is 3.0 Å.

2.2. Detailed searching of the locally associated sites based on energetic complementarity

In this stage, a more detailed searching is performed for each solution derived from the first stage. The position of the target molecule is also fixed and the six variables define an orientation of the probe molecule. A set of 'chromosomes' is 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 Wallas energy, electrostatic energy and hydrogen bond energy are considered. The force field is AMBER. Non-polar hydrogen atoms are omitted for simplification and united atom types are introduced in order to evaluate the interaction energy more precisely. The purpose of this step is not only to purge the high energy conformations, but also to precisely calculate the interaction energy. The procedure of GA minimization is the same as the previous stage and the

Table 1
The result of automated docking calculation. For each complex, only the lowest RMS resolution is listed

Complex no.	PDB brookhaven code	Rotation eular angles (radian)			Translation vector (Å)			Surface score	Interaction energy (kJ/mol)	RMS
1	9HVP [6]	0.59	0.00	5.73	11.26	6.88	18.04	3828.64	−2243.80	0.25
2	2WRP [7]	1.99	0.26	4.39	1.26	5.97	−2.50	2493.94	−268.10	0.02
3	1CGF [8]	2.57	−0.04	3.70	29.44	2.72	20.49	2316.45	−1130.80	0.87
4	1CTA [9]	1.29	−0.01	4.92	8.57	7.29	−1.79	3135.92	−245.77	0.44
5	1CKA	1.54	−0.21	4.86	12.76	−4.10	−3.36	3188.79	−1118.81	0.00
6	1PLG [10]	2.18	0.02	3.88	−17.75	11.00	0.50	1333.09	−1145.92	2.75
7	4DFR [11]	3.20	0.01	3.10	−4.41	12.48	31.55	966.25	−749.89	0.64
8	2PTC [12]	0.58	0.27	−0.76	6.13	23.79	−13.39	1612.85	−1050.38	1.00

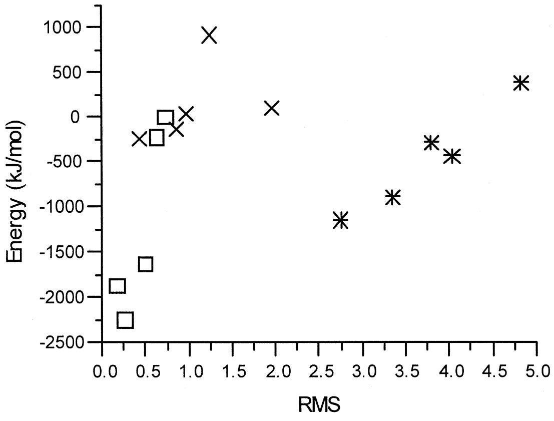


Fig. 1. The scatter plot of RMS vs. interaction energy. The solutions of the second stage are classified into several groups. For each one, the lowest energy solution is picked out as the representative solution of this group. Only the five lowest energy representative solutions are plotted. The general trend is: the lower the interaction energy, the smaller the RMS. ∗ Complex 3, □ Complex 1, × Complex 4.

same convergence criterion is applied. Lastly, cluster analysis is also performed and the solution with the largest fitness is selected to calculate the RMS with the crystal structure.

The difference between the two stages is that the first stage optimizes the orientations in the whole translational space, in the second stage, the translational vectors are restrained around the associated site found from the first stage. The GA optimization usually achieves convergence much faster than that in the first stage.

3. Result and discussion

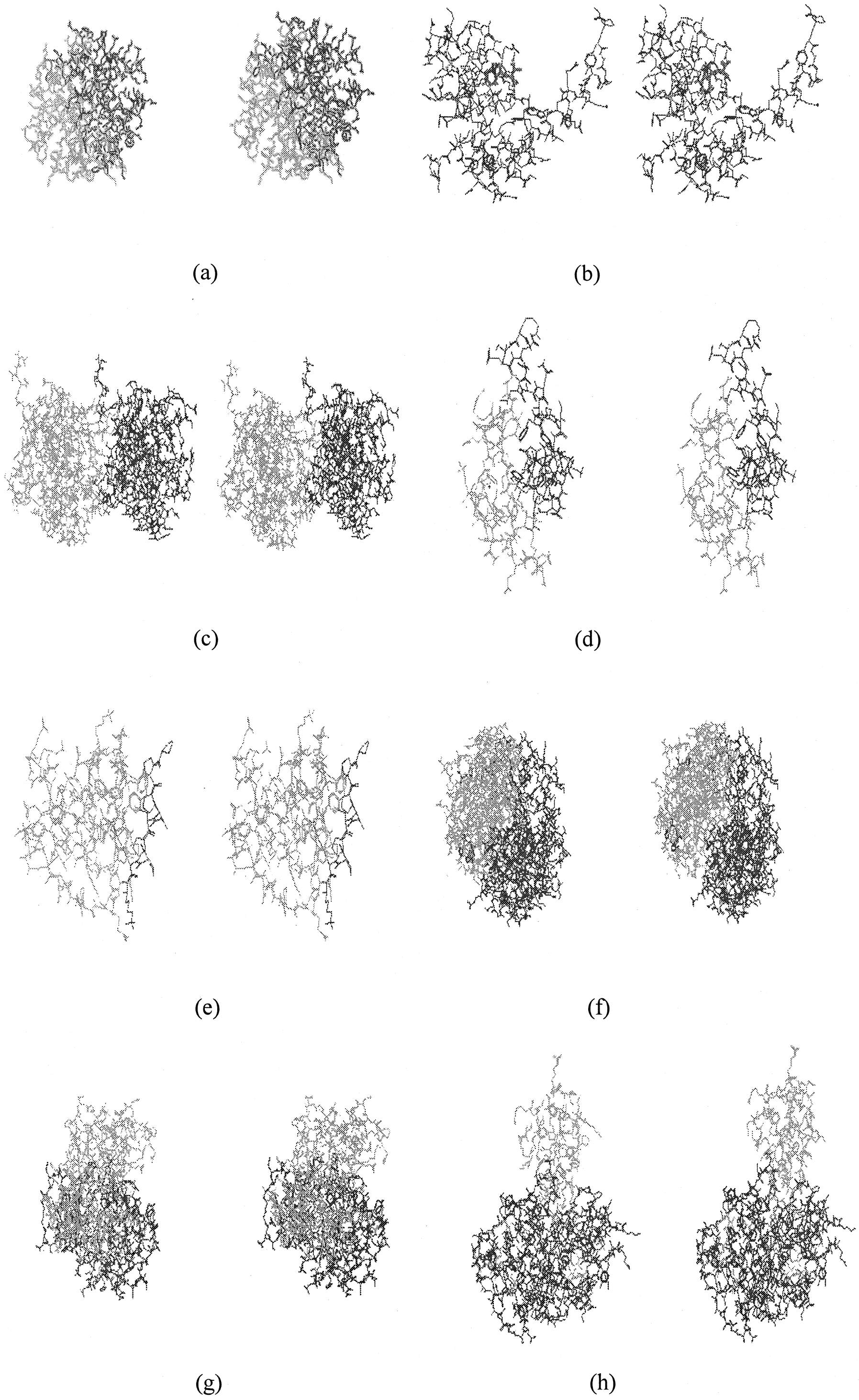
An appropriate selection of the parameters is important, since they affect not only the total computational time but also the quality of the solutions. The parameters of GA optimization include population size, elite size, random size, crossover ratio, and mutation ratio. Population size is defined as the number of ‘chromosomes’ in one generation. The larger the population size, the greater the chances that global orientation can be found and more time should be consumed. The elite size is the number of ‘chromosomes’ survived directly into the next generation in the elitist strategy. The random size is the number of ‘chromosomes’ replaced with new randomly generated ones in the random search strategy. The elite size and the random size are about 5 to 10 percent of the population size. The population size in our program is 100 and the elite size is 5. The mutation ratio and crossover ration are 0.05 and 0.35 in our procedure.

The other two parameters, const and separation angle, are defined in the previous section. const can be the balance of the contributions of two parts of the fitness score in the first docking stage. It usually takes 5–20 fold 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 the test systems, const is set to 10 fold of the dot density and separation angle is set to 30°.

There are two parameters concerns with the MS program which are dot density and the probe radius of water. The probe radius of water is set to 1.4 Å in our procedure. The dot density is usually set to 0.5 dot number per angstrom square area for large systems and 1.0 for small systems. The larger the parameter, the more computational time consumed.

Docking based on the steric complementarity aims to explore the possible binding sites in the whole translational space. Although it is believed that the fitness of this stage has a poorer relationship with the RMS than that of the second stage, the docking based on steric complementarity is necessary because it offers good starting points to perform docking based on energetic complementarity restrained around the binding site. Moreover, in some cases, it is difficult

Fig. 2. The fitted structures of (a) HIV-1 protease complexes with A-74707 (9HVP); (b) Trp receptor (2WRP); (c) fibroblast collagenase (1CGF); (d) Troponin C-site III-site III homdimer (1CTA); (e) C-CRK complexes with C_3G peptide (1CKA); (f) immunoglobulin IGG2A = KAPPA =) (1PLG); (g) dihydrofolate reductase complexes with methotrexate (4DFR); (h) beta–trypsin complexes with pancreatic Trypsin inhibitor (2PTC). 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.


(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)

for docking based on energetic complementarity to find appropriate orientations by thoroughly searching the whole translational space, such as HIV-1 protease, which has relatively small binding site. The possibility for the probe molecule located in the binding site without overlapping is small. The soft-docking based on the steric complementarity is more likely to find the global orientation than the docking based on energetic complementarity, because it allows the atomic overlapping to some degree.

Eight complexes randomly selected from the Brookhaven Protein Data Bank were used to test our two-stage docking procedure. For each complex, Table 1 lists the steric complementarity score and interaction energy of the smallest RMS solution. Fig. 1 shows the scatter plot of interaction energy vs. the RMS for three complexes, which represent the small system, middle system and large system, respectively. The general trend is the lower the interaction energy, the smaller the RMS. Fig. 2 shows the fitted structure of the lowest RMS conformation and the crystal structure for every complex. For most cases, the smallest RMS of the GA solutions is smaller than 1.0 except Complex 6. In the case of Complex 7, the lowest RMS is 0.64, a litter better than the previous reported value [3]. For Complex 8, our result is better than Jiang and Kim's [5] (the lowest RMS they reported is 2.56). So, we can draw the conclusion that our automated docking procedure is a successful one and combined random search with GA can really overcome the potential barriers efficiently. Moreover, our two-stage automated docking procedure is a universal one because it is not only suitable for a small system (e.g., Complex 2), but also for the middle system (e.g., Complex 1) and the large system (e.g., Complex 3).

Acknowledgements

This project was supported by NSFC 29573095.

References

[1] D. Dasgupta, Z. Michalewicz, Evolutionary Algorithms in Engineering Applications. Springer-Verlag, Berlin, 1997.

[2] Z. Michalewicz, Genetic Algorithms + Data Structures = Evolution Programs. Springer-Verlag, Berlin, 1994, pp. 13–30.

[3] C.M. Oshiro, I.D. Kuntz, J.S. Dixon, Journal of Computer-Aided Molecular Design 9 (1995) 113.

[4] M.L. Conally, Science 221 (1987) 709.

[5] F. Jiang, S.H. Kim, J. Mol. Biol. 219 (1991) 79.

[6] J. Erickson, D.J. Neidhart, D.J. Kempf, Science 249 (1990) 527.

[7] R.W. Schevitz, Z. Otwinowski, A. Joachimiak et al., Nature 317 (1987) 782.

[8] B. Lovejoy, A.M. Hassell, M.A. Luther et al., Biochemistry 33 (1994) 8207.

[9] G.S. Shaw, R.S. Hodges, B.D. Sykes et al., Biochemistry 31 (1992) 1992.

[10] S.V. Evans, B.W. Sigurskjold, H.J. Jennings, Biochemistry 34 (1995) 6737.

[11] J.T. Bolin, D.J. Filman, D.A. Matthews et al., J. Biol. Chem. 257 (1982) 13650.

[12] M. Marquart, J. Walter, J. Deisenhofer et al., Acta Crystallogr., Sect. B 39 (1983) 480.