Applied Soft Computing 26 (2015) 299–302

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Adaptive molecular docking method based on information entropy genetic algorithm

Zhengfu Li [a,b,*], Junfeng Gu [c], Hongyan Zhuang [a], Ling Kang [a], Xiaoyu Zhao [a], Quan Guo [a]

[a] Department of Computer Science and Technology, Dalian Neusoft University of Information, Dalian 116023, PR China
[b] School of Computer Science and Technology, Dalian University of Technology, Dalian 116023, PR China
[c] Department of Engineering Mechanics, Dalian University of Technology, Dalian 116023, PR China

ARTICLE INFO

Article history:
Received 26 February 2013
Received in revised form 3 September 2014
Accepted 8 October 2014
Available online 19 October 2014

Keywords:
Molecular docking
Genetic algorithm
Information entropy
Self-adaptive
Optimization

ABSTRACT

Almost all the molecule docking models, using by widespread docking software, are approximate. Approximation will make the scoring function inaccurate under some circumstances. This study proposed a new molecule docking scoring method: based on force-field scoring function, it use information entropy genetic algorithm to solve the docking problem. Empirical-based and knowledge-based scoring function are also considered in this method. Instead of simple combination with fixed weights, coefficients of each factor are adaptive in the process of searching optimum solution. Genetic algorithm with the multi-population evolution and entropy-based searching technique with narrowing down space is used to solve the optimization model for molecular docking problem. To evaluate this method, we carried out a numerical experiment with 134 protein–ligand complexes of the publicly available GOLD test set. The results show that this study improved the docking accuracy over the individual force-field scoring greatly. Comparing with other popular docking software, it has the best average Root-Mean-Square Deviation (RMSD). The average computing time of this study is also good among them.

1. Introduction

Molecular docking is to predict the conformation of a ligand within the active site of a receptor and search for the low-energy binding modes [1]. Molecular docking is widely used in virtual screen, and some successful cases have been reported [2]. The docking model and scoring functions have received wide concerns in recent years and a lot of scoring functions have been proposed [3]. As the core of molecular docking, scoring function can help a docking program to efficiently explore the binding space of a ligand. It is also responsible for evaluating the binding affinity once the correct binding pose is identified [4]. It is an optimization process of finding the best position of a ligand in the binding site of a receptor.

A lot of comparative studies have been done to evaluate the relative performances of these widely used docking programs and scoring methods [5–18]. However, none of these scoring functions or program is generally applicable for all the situations because the interactions between ligands and receptors are complicated. In addition, it is necessary to simplify docking models to obtain acceptable computing time.

Current scoring functions can be roughly classified into three types: force field-based scoring functions, empirical scoring functions and knowledge-based scoring functions. These models of widespread used docking functions are nearly approximate models. Approximation makes one scoring function inaccurate under some circumstances. Based on force-field scoring function, we also considered hydrophobic and deformation as well in our method. Instead of simple combination of them with fixed weights, coefficients are adaptive in searching procedure. In order to improve accuracy and stability, knowledge-based scoring method was used as another scoring factor with adaptive coefficient. An iteration scheme in conjunction with the multi-population evolution and entropy-based searching technique with narrowing down space was used to solve the optimization model for molecular docking. To evaluate the method, we performed the numerical experiment with 134 protein–ligand complexes from the publicly available GOLD test set (http://www.ccdc.cam.ac.uk/). The results indicated that the scoring function for molecular docking had high accuracy.

2. Optimization model

The process of finding the best pose is an optimization problem. The problem can be described as follows:

$$\begin{aligned} &\text{Min} \quad \{F_1(X) + F_2(X) + F_3(X)\} \\ &\text{s.t.} \quad g_i(X) < 0, \quad i = 1, 2, \ldots, n \end{aligned} \tag{1}$$

where X is a vector of design variables, indicating the orientation and conformation information of a ligand. Due to the computational reasons, it is always assumed that the ligand is flexible and that the receptor is rigid. So X can be defined as follows:

$$X = \left\{T_x, T_y, T_z, R_x, R_y, R_z, T_{b1}, T_{b2}, \ldots, T_{bn}, C_1, C_2, C_3\right\}^T \tag{2}$$

* Corresponding author at: Software Park Road 8, A3-117 Office, Dalian 116023, PR China. Tel.: +86 411 84835202; fax: +86 411 84835202.
E-mail address: lizhengfu@hotmail.com (Z. Li).

http://dx.doi.org/10.1016/j.asoc.2014.10.008
1568-4946/

where T_x, T_y and T_z are the position coordinates of the ligand; R_x, R_y and R_z are the rotational angles of the ligand; T_{b1}, T_{b2}, . . ., T_{bn} are the torsion angles of the rotatable bonds of the ligand; C_1, C_2, C_3 are coefficients of each factor. The constraints $g_i(X)$, $i = 1, 2, \ldots, n$ are shown as follows:

$$\begin{cases} \underline{T_x} \le T_x \le \overline{T_x} \\ \underline{T_y} \le T_y \le \overline{T_y} \\ \underline{T_z} \le T_z \le \overline{T_z} \\ -\pi \le R_{x,y,z}, T_{b1,\cdots,bn} \le \pi \\ 0 < C_{1,2,3} < \overline{1} \end{cases} \tag{3}$$

In Eq. (1), $F_1(X)$ represents the part of Van der Waals; $F_2(X)$ represents empirical-based scoring and $F_3(X)$ represents knowledge-based scoring. $F_i(X)$ is the product of C_j and force-field factor $U_i(X)$.

$$F_i(X) = C_i * U_i(X) \tag{4}$$

The force-field function part of this paper adopts the classical AMBER molecular mechanics energy functions [19,20]. The objective function is the interaction energy between the ligand and protein, consisting of the Van der Waals and Coulomb terms of force field functions:

$$f_1(X) = \sum_{i=1}^{n_{lig}} \sum_{j=1}^{n_{rec}} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + 332.0 \frac{q_i q_j}{D r_{ij}} \right) \tag{5}$$

where each term is a double sum over the ligand atom i and the receptor atom j. n_{lig} and n_{rec} are respectively the number of atoms in the ligand and that in the receptor; A_{ij} and B_{ij} are van der Waals repulsion and attraction parameters; r_{ij} is the distance between atoms i and j; q_i and q_j are the point charges on atoms i and j; D is dielectric function; 332.0 is a conversion factor from the electrostatic energy to kilocalories per mole. The force-field-based scoring function is widely used in popular docking programs, such as DOCK, AutoDock, GoldScore, etc. To simplify the interactions between ligand and receptor, it cannot provide very accurate results in some cases.

Empirical scoring functions is assumed that the van der Waals interaction (E_{vdw}), hydrogen-bonding energy (E_{hb}), hydrophobic (E_{hyd}) and deformation (E_{def}) terms are the primary parts of binding energy. Weights of the above factors are fixed and obtained by training set. X-Score [21] (a kind of empirical scoring) is considered as $f_2(X)$. The knowledge-based scoring function commonly refers to Potential of Mean Force (PMF). According to the inverse Boltzmann law, it can be directly derived from the statistical analysis of different types of atom pairs encoded in available crystal complex structures. The scoring function K-score [22] (a kind of knowledge scoring) is considered as $f_3(X)$.

$U_i(X^k)$ is the normalized objective function. In order to improve the stability, the values of the last two generations are used in Eq. (6). Then, the normalized score $U_i(X^k)$ is represented as follows:

$$U_i(X^k) = \frac{f_i(X^k)}{(f_i(X^{k-1}) + f_i(X^{k-2}))/2} \tag{6}$$

where k is the number of iteration in the optimizing process, and X is the optimal solution of the iteration.

The objective function of Eq. (1) is a complex single-objective and multi-constraint optimization problem. Genetic algorithms provide such a capability of their successful adaptation and implementation in a series of optimal design problems. But genetic search process is a time-consuming work, so that hindered them from applied to molecular docking optimization problem, especially to massively among a virtual library of billions of small molecules for compounds that can bind to known protein binding sites. In this paper, an improved adaptive GA is adopted [23], in which an entropy-based searching technique with multi-population and the quasi-exactness penalty function is developed to ensure rapid and steady convergence.

The crossover and mutation operators (p_c and p_m) are assigned to be the added design variables to overcome the difficulty in confirming the genetic parameters. The lower and upper limits of p_c and p_m can be defined in a reasonable region (here $0.8 \le p_c < 1.0$ and $0.0 \le p_m \le 0.3$). C_1, C_2, C_3 are also design variables of GA.

Shannon's theorem [24] has wide-ranging applications in both communications and data storage applications. This theorem is of foundational importance to the modern field of information theory [25]. There are similarities between the process of optimization and communication of information theory. Information entropy or Shannon entropy H of a discrete set of probabilities $p_1, \ldots, p_n$ is defined by:

$$\begin{aligned} &H = -\sum p_i \ln p_i \\ &\text{s.t.} \quad \sum p_i = 1, p_i \in [0, 1] \end{aligned} \tag{7}$$

Shannon entropy can be used to measure the uncertainty about the realization of a random variable. If p_j is here defined as a probability that the optimal solution of the optimal problem occurs in the population j, then Shannon entropy will be decreased during optimization process of problem.

The $(1 - p_j)$ can be used as the coefficients of narrowing searching space in the modified genetic algorithm. When the optimal solution occurs in the lth population, then $(1 - p_l^*) = 0$, and its searching space is not narrowing. Using multi-population genetic strategy with narrowing down searching space, the M populations with N members are generated in the given space. Design space is defined as initial searching space $D(0)$. M populations with N members are generated in the given space. After a new generation is independently evolved in each population, the searching space of each population is narrowed according to the following equation:

$$\begin{aligned} &D_j(K) = (1 - p_j) D_j(K - 1) \\ &\underline{d_{ji}}(K) = \max\{[d_{ji}^*(K) - 0.5(1 - p_j) D_j(K)], \underline{d_{ji}}(0)\} \\ &\overline{d_{ji}}(K) = \max\{[d_{ji}^*(K) + 0.5(1 - p_j) D_j(K)], \overline{d_{ji}}(0)\} \end{aligned} \tag{8}$$

where $D_j(K)$ is the searching space of the population j at Kth iteration. $\underline{d_{ji}}(K)$ and $\overline{d_{ji}}(K)$ are the modified lower and upper limits of ith design variable in the population j at Kth iteration, respectively. $d_{ji}^*(K)$ is the value of design variable i of the best member in the population j.

Eq. (8) is employed to control the narrowing of searching space for each population. If $(1 - p_l^*) = 0$, the optimal solution occurs in the lth population, and its searching space is not narrowing. Then the convergence criterion of the proposed method can be defined as: when the searching space in the best population has been reduced to a very small area (a given tolerance), the global optimal solution can be obtained approximately. Using narrowed space as the convergence criterion could controls the convergence of the algorithm effectively.

3. Results and discussion

To evaluate the method, we performed the numerical experiment with 134 protein–ligand complexes from the publicly available GOLD test set. This set was originally proposed by Jones et al. [26]. Docking accuracy is the primary criterion to evaluate docking methods [27]. It is based on the RMSD values of the locations of all of the heavy atoms in the crystal structure. In general, the docking accuracy is acceptable if the RMSD value between the