Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2013, 5(12):1066-1069

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

A new molecular docking method based on residue groups and PMF scoring function

Zhengfu Li[1,2*], Xicheng Wang[3], Keqiu Li[1], Ling Kang[2] and Quan Guo[2]

[1]*School of Computer Science and Technology, Dalian University of Technology, Dalian, P.R. China*
[2]*Department of Computer Science and Technology, Dalian Neusoft University of Information, Dalian, P.R. China*
[3]*Department of Engineering Mechanics, Dalian University of Technology, Dalian, P.R. China*

ABSTRACT

In this study, a new molecular docking method is presented to improve the docking accuracy. We introduce to docking design a concept of residue groups based on induced-fit and use K Score (a kind of PMF scoring function) to score the docking position. 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. To evaluate the method, we carried out a numerical experiment with 134 protein–ligand complexes of the publicly available GOLD test set. Through the comparison with other popular docking software, the proposed method showed the higher accuracy. The average computing time of this study is 44.1s, that made it has advantages in the virtual screening. Among more than 61% of the complexes, the docked results were within 2.0 Å according to Root-Mean-Square Deviation (RMSD) of the X-ray structure.

Keywords: Molecular docking; Genetic algorithms; Information entropy; Scoring function; Optimization design

INTRODUCTION

Molecular docking is to simulate the interactions of two molecules (such as ligand and receptor)and to predict their binding mode and affinity. It predicts 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].A fundamental problem with molecular docking is that orientation space is very large and grows combinatorially with the number of degrees of freedom of the interacting molecules. Therefore, simpler and efficient methods are continuously being researched into.

Over the past two decades, many automated docking approaches have been developed and can be classified as rigid-docking, flexible ligand-docking and flexible protein-docking methods. The rigid-docking methods, such as the DOCK program [3], treat both ligands and proteins as rigid. In contrast, ligands are considered flexible and proteins rigid for flexible ligand-docking methods, including evolutionary algorithms [4-5], simulated annealing [6], fragment based approach[7], and other algorithms [8–9]. The consideration of protein flexibility is not important than that of ligand flexibility. Protein flexibility has been ignored inmost docking programs since the evaluation of protein-ligand interaction energies at all possible docking configurations isa prohibitively time consuming process. However, it has become increasingly clear that protein flexibility plays a paramount role in protein-ligand complex formation and should be considered during the docking process [10-11].

The knowledge-based scoring methods that emerged in recent years are in essence designed to reproduce the experimental structures (binding poses) of ligands binding to receptors. The potential (more accurately referred to as potential mean force, PMF) of such a method is directly derived, according to the inverse Boltzmann law, from the statistical analysis of different types of atom pairs encoded in available crystal complex structures. These methods capture every interaction term implicitly (including salvation and the entropic effect) with an obvious advantage that it can be constructed without any knowledge of the binding data, and thereby they can be used to score novel ligands that are different from molecules in the training sets. K Score[12], a kind of PMF scoring function, is considered in this study.

Molecular docking is a difficult optimization problem. It contains a large number of design variables. The objective function is a highly nonlinear function, and it is an implicit function of the design variables. To solve it may involve a costly computational effort. 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. In order to evaluate the new optimization model and docking method, we have conducted a numerical experiment with 134 protein–ligand complexes from the publicly available GOLD test set (http://www.ccdc.cam.ac.uk/). Comparisons with six other docking programs show that docking accuracy has been significantly improved by this study.

EXPERIMENTAL SECTION

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

$$\begin{aligned} &\min F(X) \\ &s.t.\ g_i(X) < 0, i = 1, 2, \cdots n \end{aligned} \tag{1}$$

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:

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

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. The constraints $g_i(X)$, i=1, 2… n are shown as follows:

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

In the above model, we only consider the ligand flexibility. However, changes in the receptor structure upon ligand binding are frequently observed [13], and as such, both the structure of the ligand and the receptor change during the binding process. We introduced the concept of the residue groups in the receptor. The residues within the binding site are divided into several residue groups, and the center coordinates of each residue group introduced into the optimization process as design variables. Thus we establish a refined-scale optimization model based on the problem (1), and added the following design variables:

$$\{C_{1x}, C_{1y}, C_{1z}, ..., C_{mx}, C_{my}, C_{mz}\}^T \tag{4}$$

Where m is the number of residue groups, and (C_{ix},C_{iy},C_{iz})(i=1,2,...,m) are the positional coordinates of the center for each residue group. And the constraints added are introduced into $g(X)$as:

$$\begin{cases} \underline{C}_{ix} \le C_{ix} \le \overline{C}_{ix} \\ \underline{C}_{iy} \le C_{iy} \le \overline{C}_{iy} \quad i = 1, ..., m \\ \underline{C}_{iz} \le C_{iz} \le \overline{C}_{iz} \end{cases} \tag{5}$$

The knowledge-based scoring function commonly refers to Potential of Mean Force (PMF). Completely different from force-field scoring, knowledge-based scoring considers docking problem from another point of view. 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 is considered in this study and defined as follows:

$$KScore = \sum_{\substack{pl \\ r<r_{cut-off}}} A_{ij}(r) = \sum_{\substack{pl \\ r<r_{cut-off}}} -K_B T \ln\left[f^{j}_{vol-corr}(r) \frac{\rho^{ij}_{seg}(r)}{\rho^{ij}_{bulk}} \right] \tag{6}$$

Where, K_B is the Boltzmann constant; T is the absolute temperature; $f^{j}_{vol-corr}$ is the ligand volume correction factor; $\rho^{ij}_{seg}(r)$ is the number density of atom pair *ij* that occurs in a spherical shell with a thickness of Δr ranging from r to $r+\Delta r$; ρ^{ij}_{bulk} expresses the number density when no interaction occurs between *i* and *j*.

Eq. (1) is a complex single-objective and multi-constraint optimization problem. Because of the huge searching space, it is very difficult to get the best solution. Genetic algorithms (GA) provide such a capability of adaptation and searching in many optimal design problems. In this study, an improved adaptive GA is adopted[14], in which an entropy-based searching technique with multi-population and the quasi-exactness penalty function is developed to ensure rapid and steady convergence.

For multi-population genetic strategy, the genetic algorithm begins from generating arbitrarily *m* populations with all the same searching space, i.e. design space. For the improved genetic algorithm with narrowing of the search space, we need only to know efficient narrowing coefficients for the searched space. Shannon's theorem has wide-ranging applications in both communications and data storage applications. This theorem is of foundational importance to the modern field of information theory. 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:

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

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[15].The hardware environment is: IntelE5620, quad-core CPU, 2.4GHz, 8GB RAM. Docking accuracy is the primary criterion to evaluate docking methods[16]. 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 docked pose and X-ray crystal structure is less than 2.0 Å. To date, many docking programs are available. Glide[17], GOLD[15], Surflex[16], Flex X[18] and Dock6[19] are the commonly used docking programs. The above programs are based on the assumption of rigid receptor for the assumption is conducive to cut off computing time largely. Docking results of flexible receptor are often better than rigid receptor's. In the study, we selected Dock6 (with flexible default parameters) as a flexible receptor program. Table 1 presents the ratios at different RMSD ranges of these programs.

Table-1RMSD ratios of this study and 6 commonly used docking programs.

RMSD	Percent (%)						
	This study	Glide	GOLD	Surflex	Flex X	DOCK6	Dock6-F
≤0.5	0.20	0.29	0.08	0.16	0.03	0.15	0.09
>0.5, ≤1.0	0.23	0.19	0.27	0.32	0.18	0.15	0.32
>1.0, ≤2.0	0.18	0.23	0.31	0.29	0.28	0.32	0.39
>2.0, ≤3.0	0.16	0.09	0.05	0.06	0.10	0.12	0.10
≥3.0	0.23	0.20	0.28	0.17	0.40	0.27	0.09
Avg. RMSD	2.27	1.98	3.19	2.15	3.69	2.13	1.46

CONCLUSION

With the method proposed in this study, we obtained 27(20%) excellent docking solutions with a RMSD value below 0.5 Å, 55 (41%) good predictions with RMSD between 0.5 and 2.0 Å and only 31 (23%) wrong predictions (RMSD value larger than 3.0 Å). And the average RMSD obtained in this study is 2.27. In the view of RMSD, the method proposed in this study is good among these programs.

Computing time is another important evaluation criterion for a docking method, especially in virtual high throughput screening. GA can find the optimum solution under the probability of 1 if the iteration number becomes large enough. The docking accurate of this study could be better if it has more computing time. However, simply improve the docking accurate is pointless, without considering calculation efficiency. The minimum computing time with the method proposed in this study is 6.7s, maximum computing time is 171.8s. The average computing time of this study is 44.1s, while the average computing time ofDock6 (with flexible default parameters) is 590.6s. Computing time is so good that made it has advantages in the virtual screening.

Acknowledgments
This work was supported by the National Natural Science Funds of China (No. 11202049, NO.11072048, No.61170168 and No.61170169).

REFERENCES

[1] L Kang;Q Guo; XC Wang;*Bioorganic & Medicinal Chemistry Letters*, **2012**, 22, 6568.
[2] BO Villoutreix; R Eudes; MA Miteva;*Comb. Chem. High. Throughput Screen*,**2009**, 12, 1000.
[3] ID Kuntz; JM Blaney; SJ Oatley; R Langridge; TE Ferrin;*J. Mol. Biol.*, **1982**, 161, 269.
[4] G Jones; P Willett; RC Glen; AR Leach; R Taylor;*J. Mol. Biol.*, **1997**, 267, 727.
[5] CM Oshiro; ID Kuntz; J.S. Dixon; J. Comput; *Aided Mol. Des.*, **1995**, 9, 113.
[6]CJ Sherman; RC Ogden; ST Freer;*J. Med. Chem.*, **1995**, 38, 466.
[7] BKramer; MRarey; TLengauer;*Proteins*, **1999**, 37, 228.
[8]PN Palma; L Krippahl; JE Wampler; JJ Moura;*Proteins*, **2000**, 39, 178.
[9] J Wang; PA Kollman; ID Kuntz; *Proteins*, **1999**, 36, 1.
[10] HA Carlson; JA McCammon;*Mol. Pharmacol.*, **2000**, 57, 213.
[11] SJ Teague; N Rev;*Drug. Discov.*, **2003**, 2, 527.
[12] XY Zhao;XF Liu;YY Wang;Z Chen;L Kang;HL Zhang;XM Luo;WL Zhu;KX Chen;HL Li;XC Wang;HL Jiang; *J. Chem. Inf. Model*, **2008**, 48, 1438.
[13]N Brooijmans; ID Kuntz; A Rev; Biophys. Biomol. Struct., 2003, 32, 335.
[14] L Kang;HL Li;HL Jiang;XC Wang; *J. Comput. Aided Mol. Des.*, **2009**, 23, 1.
[15] G Jones;P Willett;RC Glen;AR Leach;R Taylor; *J. Mol. Biol.*, **1997**, 267, 727.
[16] AN Jain.*J. Med. Chem.*, **2003**, 46, 499.
[17] RA Friesner;JL Banks;RB Murphy;TA Halgren;JJ Klicic;DT Mainz;MP Repasky;EH Knoll;M Shelley;JK Perry;DE Shaw;P Francis;PS Shenkin; *J. Med. Chem.*, **2004**, 25, 1739.
[18] B Kramer;M Rarey;T Lengauer; *Porteins: Struct. Funct. Genet.*, **1999**, 37, 228.
[19] PT Lang;SR Brozell;S Mukherjee;EF Pettersen;EC Meng;V Thomas;RC Rizzo;DA Case;TL James;ID Kuntz; *RNA*, **2009**, 15, 1219.