Information Sciences 289 (2014) 206–224

Contents lists available at ScienceDirect

Information Sciences

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

A dynamic niching genetic algorithm strategy for docking highly flexible ligands

Camila Silva de Magalhães [a,b,c,*], Diogo Marinho Almeida [c], Helio José Correa Barbosa [c,d], Laurent Emmanuel Dardenne [c]

[a] Universidade Federal do Rio de Janeiro – UFRJ, Campus Xerém, Duque de Caxias, RJ 25245-390, Brazil
[b] Departamento de Matemática, Universidade Federal Rural do Rio de Janeiro – UFRRJ, Seropédica, RJ 23851-970, Brazil
[c] Laboratório Nacional de Computação Científica – LNCC, Petrópolis, RJ 25651-075, Brazil
[d] Departamento de Ciência da Computação, Universidade Federal de Juiz de Fora – UFJF, Juiz de Fora, MG 36036-330, Brazil

ARTICLE INFO

Article history:
Received 26 March 2013
Received in revised form 25 July 2014
Accepted 1 August 2014
Available online 11 August 2014

Keywords:
Flexible docking
Genetic algorithm
Niching
Real-parameter optimisation

ABSTRACT

Currently, docking methods are very important tools in structure-based drug design (SBDD). However, despite the great advances in the docking area over the last decades, most methods cannot be used to dock highly flexible ligands successfully. It is even harder when the ligand is cross-docked into different ligand-bound receptor structures. In this work, a new multi-solution genetic algorithm method, named Dynamic Modified Restricted Tournament Selection (DMRTS), was developed for the effective docking of highly flexible ligands. The DMRTS method uses an insertion criterion based on similarity and a dynamic tournament size to preserve good, distinct solutions in the genetic algorithm population. The proposed method was implemented in three different versions of a steady-state genetic algorithm and evaluated for the redocking and cross-docking of five HIV-1 protease ligands, with 12–20 rotatable bonds. The DMRTS method was also tested on a more diverse set of 34 protein–ligand complexes covering 18 different protein families. A performance comparison with three of the currently most used docking programs was also done. The proposed method was evaluated for 25 benchmark functions of the CEC2005 test suite. The results indicated that the DMRTS method can adequately sample the conformational search space, producing a diverse set of high quality solutions. Moreover, it might be a powerful tool for docking studies in SBDD practice, increasing the success rate in finding correct ligand conformations and efficiently exploring distinct and valuable ligand binding modes.

1. Introduction

Computational receptor-ligand docking methods are of utmost importance and have been widely used in drug discovery and development projects [31,29]. Receptor-ligand molecular recognition is a very complex process that involves several molecular interactions (e.g., van der Waals interactions, electrostatic interactions, hydrogen bonds, hydrophobic interactions) and a high level of steric and physicochemical complementarity between the receptor and the ligand. The binding process is also driven by important entropic contributions associated with molecular flexibility of the ligand and the receptor as

* Corresponding author. Postal address: Estrada de Xerém, 27, Xerém, Duque de Caxias, RJ 25245-390, Brazil.
E-mail address: camila.mag@gmail.com (C.S. de Magalhães).

http://dx.doi.org/10.1016/j.ins.2014.08.002
0020-0255/

well as solvent effects. The binding affinity can be determined from the experimental binding constant K_i and is related to the Gibbs free energy of binding ΔG:

$$\Delta G = \Delta H - T\Delta S = -RT \ln K_i$$

where T is the absolute temperature, R is the gas constant and ΔH and ΔS are, respectively, the enthalpic and entropic variations between the receptor and the ligand in the bound and unbound states. To simulate the molecular recognition process, the receptor-ligand docking methods focus on the prediction of the position of a small molecule within the three-dimensional structure of a receptor (usually a protein). Docking algorithms must explore the conformational and configurational (i.e., translational and rotational) ligand degrees of freedom and use a scoring function that defines the energy landscape. Ideally, this scoring function must correlate with the receptor-ligand binding free energy to appropriately order different ligand poses and correctly rank different compounds in virtual screening experiments.

Despite the intensive research efforts in the docking area over the last decades, significant challenges still remain, primarily in two crucial aspects [37,14]: (i) development of algorithmic strategies to explore the very complex energy landscape associated with both ligand and receptor molecular flexibility to predict the correct ligand binding mode inside the receptor and (ii) accurate prediction of the ligand-receptor binding affinities i.e., developing a computationally viable free energy evaluation model to correctly discriminate among different binding modes for the same ligand and/or find the best drug candidate in a set of ligands [17]. These two aspects are very active research topics in the structure-based drug design field. In this article, we focus our attention on the first aspect.

A critical issue in molecular docking is properly handling ligand and protein flexibility, considering hundreds of thousands of conformational degrees of freedom. A number of methodologies have been proposed to include partial or full protein flexibility in docking algorithms [28]. This issue remains a very challenging one, and most docking methods include only the ligand flexibility, keeping the protein fixed in a predetermined position. A very promising way to address the receptor flexibility problem is to generate a set of representative conformations of the receptor flexibility and perform a rigid-receptor flexible-ligand docking experiment for each receptor conformation in that set [21]. In the literature, this strategy is known as ensemble docking and has the great advantage of focusing the molecular flexibility problem on the treatment of the ligand degrees of freedom. However, even when the rigid receptor approximation is used, docking large and highly flexible ligands is still a considerable challenge to current docking methods [30]. It is even harder when the ligand is docked in a non-native protein conformation (i.e., as an apoprotein or a structure composed of a distinct ligand complexed with the same protein) [33,37]. This type of docking is usually called cross-docking and corresponds to a more realistic situation. A very interesting study that critically and properly noted this question was performed by Cecchini and co-workers [6], who evaluated an incremental construction method in conjunction with a genetic algorithm in redocking (i.e., docking using the native protein conformation bound to the original ligand) and the cross-docking of five highly flexible HIV-1 protease ligands. Although good performance was observed in the redocking studies, the performance for cross-docking was very poor, with a success rate of 0% in 45% of the experiments and less or equal to 40% in most cases. The highly flexible ligand question and the cross-docking evaluation have been addressed by different search strategies and algorithms, including genetic algorithms [20,24], swarm optimisation [7,22] and Monte Carlo methods [9,23]. Nevertheless, although the fully informed swarm optimisation algorithm (FIPSDock) [22] was successful in the redocking experiments, a significantly lower performance was observed in most of the cross-docking cases for FIPSDock and the other three docking programs evaluated. The Monte Carlo method used by the RosettaLigand [9] and the LGA of Autodock 4 [24] were also unable to successfully dock and cross-dock ligands with more than 10 rotatable bonds in most cases. The effectiveness of genetic algorithms (GAs) in solving complex optimisation problems has been extensively demonstrated in the literature, including for the molecular docking problem. However, due to the high modality of the fitness landscape for the flexible ligand docking problem, a critical issue for GAs is to maintain a useful population diversity to allow for a parallel investigation of several high fitness (i.e., low energy) regions of the search space and to reduce the chances of convergence to lower quality local optima. Niching strategies are suitable techniques that have been proposed to address high modality landscapes [36]. In a previous work [10], we presented a steady-state genetic algorithm (SSGA) with a rank-based selection method for the molecular docking problem. However, the SSGA algorithm showed poor performance when dealing with highly flexible ligands. In the present paper we extend this work, presenting a newly designed SSGA with these principal additional features: (i) an efficient dynamic niching GA strategy; (ii) two other genetic operators and (iii) hybridisation with a local search method. Furthermore, the genetic operator probabilities are dynamically assigned by an adaptive method in this implementation. The proposed niching technique, named the Dynamic Modified Restricted Tournament Selection (DMRTS) method is an improvement of the MRTS method [11] and was inspired by the Restricted Tournament Selection (RTS) method proposed by Harik [18]. The main advantage/difference of the DMRTS method is promoting useful diversity and increasing the search capability using (i) phenotypic information based on the root-mean-square-deviation (RMSD) of generated ligand conformations and (ii) a dynamic tournament size. The DMRTS performance was analysed using three different versions of the implemented genetic algorithm: (1) one without a local search; (2) one in which a local search is applied during the evolutionary process; and (3) one with a local search applied both during the evolutionary process and to the final population. The three algorithm versions were evaluated for the redocking and cross-docking of five highly flexible HIV-1 protease ligands with 12–20 rotatable bonds. The ability of the proposed method to explore distinct and valuable ligand binding modes was investigated. The DMRTS method was also tested on a diverse set of 34 protein–ligand complexes covering 18 different protein families and its performance in finding the experimental ligand binding modes was compared to that of GOLD [20], GLIDE [13] and Autodock Vina [34], three of the most frequently used