Chinese Chemical Letters 24 (2013) 1001–1004

Contents lists available at SciVerse ScienceDirect

Chinese Chemical Letters

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

Original article

Improving the accuracy of pose prediction in molecular docking *via* structural filtering and conformational clustering

Shi-Ming Peng [a,c], Yu Zhou [b,c], Niu Huang [c,*]

[a] College of Biological Sciences, China Agricultural University, Beijing 100193, China
[b] College of Life Sciences, Beijing Normal University, Beijing 100875, China
[c] National Institute of Biological Sciences, Beijing 102206, China

ARTICLE INFO

Article history:
Received 7 March 2013
Received in revised form 24 May 2013
Accepted 27 May 2013
Available online 16 July 2013

Keywords:
Molecular docking
Pose prediction
Structural descriptor
Conformational clustering

ABSTRACT

Structure-based virtual screening (molecular docking) is now one of the most pragmatic techniques to leverage target structure for ligand discovery. Accurate binding pose prediction is critical to molecular docking. Here, we describe a general strategy to improve the accuracy of docking pose prediction by implementing the structural descriptor-based filtering and KGS-penalty function-based conformational clustering in an unbiased manner. We assessed our method against 150 high-quality protein–ligand complex structures. Surprisingly, such simple components are sufficient to improve the accuracy of docking pose prediction. The success rate of predicting near-native docking pose increased from 53% of the targets to 78%. We expect that our strategy may have general usage in improving currently available molecular docking programs.

1. Introduction

Despite the well-known weaknesses, structure-based virtual screening (molecular docking) is now one of the most practical techniques to leverage target structure for ligand discovery [1–3]. Molecular docking approach is designed to identify small molecules from a large chemical library for shape and physico-chemical complementarity to a macromolecular binding site. Numerous studies have applied such an approach to identify novel ligands for various drug targets [4]. The two major challenges in molecular docking are sampling (*i.e.*, enumerating possible conformations of ligands in the receptor binding pocket) and scoring (*i.e.*, identifying the correct binding orientation and conformation out of an enormous number of alternative modes for each ligand, and ranking different ligands with respect to their estimated binding affinity) [5]. Nevertheless, in order to dock a large compound library, a scoring function has to be simple, fast, and derived from a physically reasonable equation.

Physics-based scoring methods model the protein–ligand interactions based on the law of physics, and have the advantage to be more accurate with systematic improvements [6]. During the past few years, we have developed a hierarchical physics-based virtual screening protocol to integrate different computational methods in an increasing order of complexity and more physically realistic manner, in which a rapid-to-compute docking program (DOCK3.5.54) is used to screen large compound databases, and a more physically rigorous approach (MM-GB/SA) is applied to refine and rescore docking poses [7,8]. However, one major limitation of this protocol is that it relies entirely on the docking algorithm to identify the correct binding pose, and is not effective in rescuing grossly mis-docked ligands. Therefore, we could envision a simple extension by subjecting a small number of dissimilar binding poses for each ligand to the more accurate MM-GB/SA rescoring (arguably more computationally expensive), and use the most favorable binding energy for rank-ordering ligands.

To obtain dissimilar docking poses during docking, an efficient yet effective pose clustering algorithm is essential to process the millions of docking poses on-the-fly. Several clustering algorithms are available, including seeded RMSD (root–mean–square deviation) clustering, greedy RMSD clustering, *K*-means clustering, Jarvis–Patrick clustering and top-first clustering [9,10]. These methods require at least one pre-assigned cutoff value to determine whether two poses belong to the same structural class, which significantly influences the clustering results. In contrast, KGS-penalty function clustering is cutoff-free [11], particularly efficient in clustering diverse poses in global space in an unbiased manner.

In addition, we can also envision a different strategy to integrate docking methods and informatics (*e.g.*, structural analysis) to improve the overall pose prediction performance. The successful identification of near-native poses from decoy poses is challenging for docking scoring function [8]. Therefore, a simple

* Corresponding author.
E-mail address: huangniu@nibs.ac.cn (N. Huang).

1001-8417/$ – see front matter
http://dx.doi.org/10.1016/j.cclet.2013.06.016

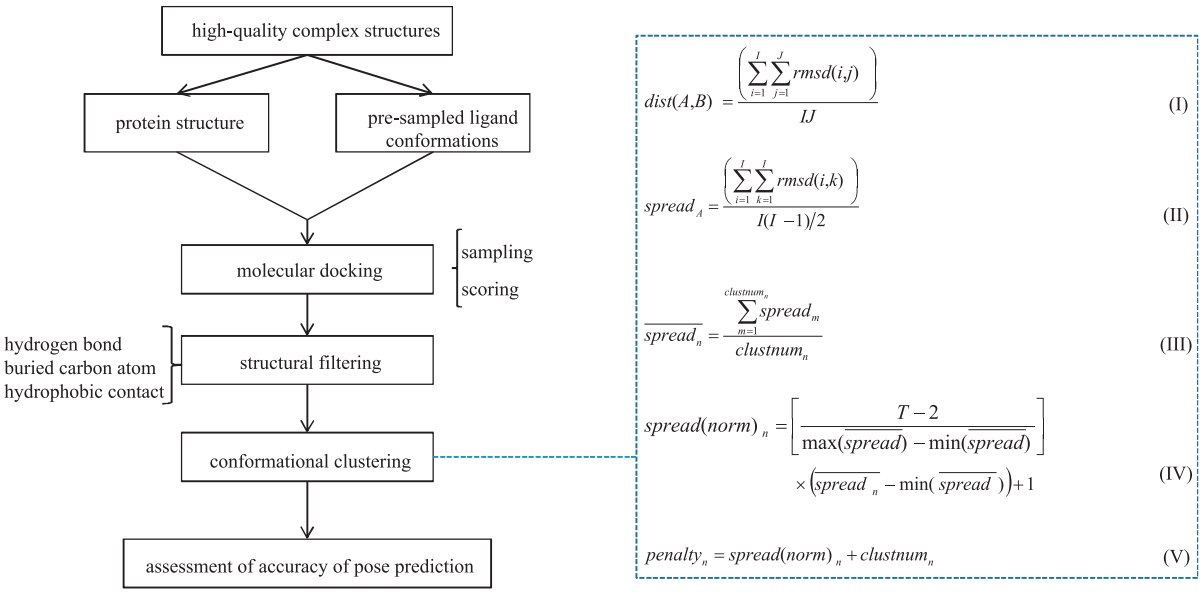

Fig. 1. The scheme of docking process implemented with pose filtering and clustering, and the mathematic equations applied in clustering algorithm.

yet effective way to improve the docking performance is to exclude unreasonable docking poses forbidden by the essential structural criteria [7,12,13], such as the number of hydrogen bonds, the number of buried carbon atoms and hydrophobic contact [14,15].

Here we developed a novel strategy (Fig. 1) to improve molecular docking performance *via* filtering and clustering without significantly sacrificing the calculation speed. We implemented structural descriptor-based filtering and KGS-penalty function clustering in the DOCK3.5.54 program. The docking output is a small subset of dissimilar docking poses satisfying the essential structural criteria, which makes the more rigorous rescoring processes practically feasible. We assessed our approach against 150 high-quality protein–ligand complex structures. The success rate of predicting near-native binding pose was substantially increased from 53% of the targets to 78%.

2. Experimental

Benchmark set preparation: To assess our methodology development, we chose 150 unique protein structures bound with drug-like ligands from CSAR-NRC HiQ data set, where protonation state, tautomeric form and hydrogen atom orientation of the ligands and binding site residues were corrected manually [16].

Molecular docking procedure: The automatic docking procedure was described previously [17]. Briefly, each protein was prepared for docking in the same manner (except for several metalloenzymes), and each ligand was docked back into its corresponding binding pocket. An in-house modified version of program DOCK 3.5.54 was used to dock compounds into the protein binding site. The pre-computed conformational ensemble of each ligand [18] was matched against the docking spheres derived from both the receptor and the crystallographic ligand. The grid-based docking energy components, including van der Waals interactions, electrostatic energy and ligand partial desolvation penalty, were calculated and summed up. The ligand docking poses were scored and ranked based on the total docking energy.

Filtering algorithm: The generated docking poses were filtered based on the three types of structural descriptors calculated for each docking pose, including the number of hydrogen bonds, the number of buried carbon atoms and hydrophobic contact. The buried carbon atom is assigned if the distance between the ligand carbon atom and any heavy atom in the receptor is shorter than 4.0 Å. The hydrophobic contact is defined as follow:

$$f(d) = \begin{cases} 1.0 & d \le d_0 + 0.5\,\text{Å} \\ (1/1.5) \times (d_0 + 2.0 - d) & (d_0 + 0.5\,\text{Å} < d \le d_0 + 2.0\,\text{Å}) \\ 0 & d > d_0 + 2.0\,\text{Å} \end{cases}$$

d is the distance of the two atoms and d_0 is the sum of van der Waals radii of the two atoms.

For each ligand, the unreasonable docking poses are automatically removed if the calculated values of any type of descriptors are below the averaged values throughout the entire conformational ensemble, individually.

Clustering algorithm: The docking poses survived from the filtering step were clustered based on KGS-penalty function [11]. The clustering is an iterated process (Fig. 1). In each step, the distance of every two clusters (A and B) is calculated using equation I and the nearest clusters are merged together. For example, i is the element in A and there are I elements in A. Next, the spread of the cluster and the average spread of clusters in this step are calculated using equation II and III. In n step, there are $clustnum_n$ clusters. Finally, all the data (T) are grouped to a single cluster. The penalty function is defined based on the normalized spread of clusters and the number of clusters ($clustnum_n$) in each step using equations IV and V. The step with the minimum penalty function is regarded as the step with optimal cluster sets.

3. Results and discussion

Control calculation: The docking pose accuracy was assessed based on the RMSD values between the coordinates of the heavy atoms in the ligands in the top scoring poses and those in native crystallographic pose, and a cutoff value of 2 Å was chosen to discriminate the docking success from failure. We firstly computed the RMSD values of the top 1500 docking poses for each ligand, and we plotted the success rates of predicting near-native binding pose (% of the targets with the best RMSD values <2 Å) as the function of the number of top scoring poses (Fig. 2A). Clearly, the success rate reaches 83% at the top 300 ranking poses, and does not increase significantly afterwards by adding more docking poses. Therefore,