Journal of Molecular Graphics and Modelling 66 (2016) 108–114

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

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

Investigating the importance of Delaunay-based definition of atomic interactions in scoring of protein–protein docking results

Rahim Jafari [a,*], Mehdi Sadeghi [b,c], Mehdi Mirzaie [d]

[a] Department of Nanobiotechnology, Faculty of Biological Sciences, Tarbiat Modares University, Tehran, Iran
[b] National Institute of Genetic Engineering and Biotechnology, Tehran, Iran
[c] School of Biological Sciences, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran
[d] Department of Applied Mathematics, Faculty of Mathematical Sciences, Tarbiat Modares University, Jalal Ale Ahmad Highway, P.O. Box 14115-134, Tehran, Iran

ARTICLE INFO

Article history:
Received 27 October 2015
Received in revised form 8 March 2016
Accepted 1 April 2016
Available online 4 April 2016

Keywords:
Delaunay tessellation
Interaction
Protein–protein docking
Scoring
Potential function

ABSTRACT

The approaches taken to represent and describe structural features of the macromolecules are of major importance when developing computational methods for studying and predicting their structures and interactions. This study attempts to explore the significance of Delaunay tessellation for the definition of atomic interactions by evaluating its impact on the performance of scoring protein–protein docking prediction. Two sets of knowledge-based scoring potentials are extracted from a training dataset of native protein–protein complexes. The potential of the first set is derived using atomic interactions extracted from Delaunay tessellated structures. The potential of the second set is calculated conventionally, that is, using atom pairs whose interactions were determined by their separation distances. The scoring potentials were tested against two different docking decoy sets and their performances were compared. The results show that, if properly optimized, the Delaunay-based scoring potentials can achieve higher success rate than the usual scoring potentials. These results and the results of a previous study on the use of Delaunay-based potentials in protein fold recognition, all point to the fact that Delaunay tessellation of protein structure can provide a more realistic definition of atomic interaction, and therefore, if appropriately utilized, may be able to improve the accuracy of pair potentials.

1. Introduction

Protein-protein interactions are involved in many biological processes. Owing to inherent difficulties in studying them experimentally, computational methods for the structural prediction of protein–protein complexes from their constituent components, referred to as protein–protein docking, are of great value.

These methods typically include two steps: exploring a large number of possible protein–protein configurations in computationally reasonable time to identify a set of structures that contain near-native solutions, followed by refinement step with the aim of improving the rank of the near-native poses. Refining the structures generated in the first step can be performed using combination of several procedures, including: clustering and filtering [1–3], re-ranking with scoring functions [4,5] and structural refinements (energy minimization) [6–8].

There are several classes of scoring functions that have been successfully utilized for docking purpose: Physics-based or force-field-based [9,10], empirical [4,6,11,12] and knowledge-based [4,5,13–17] scoring functions. The functions from the last class are obtained by statistical analysis of structural and physico-chemical features taken from a set of known protein structures. The approach is to convert the observed frequency of features to a set of averaged energy parameters by using the inverse Boltzmann equation.

They were originally developed for single protein structure prediction [18], and over the past years have gained attention in the field of protein docking for their ease of use and computational efficiency.

Regardless of their actual application domain, generally there are several factors that can contribute to the effectiveness of the statistical potentials. One is the method by which the “reference state” is modeled, the state at which the features occur purely by chance. It is used for estimation and removal of random part of the observed frequencies. The next factor is the type of feature(s) cho-

* Corresponding author at: Department of Nanobiotechnology, Faculty of Biological Sciences, Tarbiat Modares University, Jalal Ale Ahmad Highway, Tehran, Iran.
E-mail address: rjafari57@yahoo.com (R. Jafari).

http://dx.doi.org/10.1016/j.jmgm.2016.04.001
1093-3263/

sen for the statistical analysis. Distance between interacting atoms or residues, and relative orientations between them are two well known examples.

Representation of the system is another effective factor. Some representations such as those relying on computation of Voronoi diagram, Delaunay tessellation or alpha shape of protein molecules, are capable of providing more accurate description of protein structure than others.

In fact, these geometric constructs have been extensively used for studying various aspects of both protein structure and protein–protein interaction, including for example: computation of molecular area and volume, study of voids and cavities, identifying protein–protein binding sites and hot-spot residues as well as methods for scoring predicted folds and docking structures [19,20].

Recently, our group has applied an atomic knowledge-based potential derived using Delaunay tessellation (DT) to protein fold recognition, which produced encouraging results [21].

Here, we extend the previous work to the problem of protein–protein docking, centering attention on the role of the DT in performance of the statistical potentials derived specifically for scoring docking predictions.

Based on essentially two different approaches for definition of atomic interaction, we derived two sets of statistical potentials from a sample of native complexes, one based on "Distance threshold" and the other on DT. The ability of these functions for discriminating near-native structures from incorrect predictions were assessed by using two test sets generated through different docking protocols. The results were then compared between the groups of potentials to reveal how Delaunay-based definition of interactions can affect the accuracy of the derived statistical potentials.

2. Materials and methods

2.1. Training dataset

The non-redundant dataset compiled by Huang and Zou [15] was used for obtaining the frequency of pairwise atomic interactions. It contains 851 dimeric protein–protein complexes including 655 homodimers and 196 heterodimers at resolutions of 2.5 Å or better.

2.2. Definition of atom types

From the physico-chemical standpoint, protein atoms differ not only in chemical properties but also the environments they reside in. Considering this view, the total number of heavy atoms of all the 20 standard amino acids is 167. In order to obtain enough observation, this number is commonly reduced by grouping them into fewer atom types according to, for example, their biochemical similarities.

Atom types can be defined in various ways. For this study, four different atom-typing schemes were used for the calculation of atom type pair frequencies, ranging from simple to rather detailed. For the simplest one, the atoms C, O, N and S were divided into 7 groups, according to being in protein backbone or side-chain. The second was from the work of Jiang et al. [22] which classified protein atoms into four types: hydrogen bond donor, hydrogen bond acceptor, both donor and acceptor, and neutral. The next two definitions were based on more elaborate classifications of atoms according to their chemical nature and bond connectivity. They were taken from the works of Melo and Feytmans [23] and Huang and Zou [15] with 40 and 20 atom types respectively. We will refer the above sets of atom types as atomTypes7, atomTypes4, atomTypes20 and atomTypes40 respectively. The details of the atom type definitions are available as Supplemental material.

2.3. Definition of atomic interaction

Classically two atoms/residues are considered to be interacting if they are separated by less than a certain distance threshold. Most well-known docking and scoring methods use this criterion to define atom–atom or residue–residue interactions [4,5,9,11,14,15,24–32].

Though widely used, such a simple definition of interaction neglects the fact that many atomic interactions may actually be interrupted due to the presence of an irrelevant atom in between them. This issue can be largely resolved by the more accurate definition using the Voronoi diagram and its closely related construct, Delaunay tessellation.

For a given set of points in 3D space, Voronoi diagram partitions space into convex polyhedral called Voronoi cells, each corresponding to a point in the set. Each cell determine a region around a point in such a way that every point within this region is closer to this point than to any other point of the set.

DT of a set of points can be obtained by connecting all pairs of points sharing a common Voronoi facet.

To compute DT of protein–protein structure, coordinates of all the heavy atoms were extracted and used as input for the program Qhull [33]. In the resulting tessellation, two atoms that are connected via a Delaunay edge are in direct interaction, that is, they are not insulated from each other by any intervening atom.

Hereafter, for brevity, the interactions that are determined with DT method and those that are defined according to distance threshold will be referred to as DT-based and distance-based interactions respectively. Likewise, we will refer to their corresponding potentials as DT-based and distance-based potentials.

2.4. Calculation of knowledge-based statistical potentials

The intermolecular potentials extracted here are distance dependant. For any pair of atoms, the distance between them was binned into equally spaced intervals. The observed atom pair frequencies were converted into energy terms following the method proposed by Sippl [34]:

$$\Delta E_{ij}(r) = RT \ln\left[1 + M_{ij}\sigma\right] - RT \ln\left[1 + M_{ij}\sigma \frac{f_{ij}(r)}{f_{xx}(r)}\right]$$

Where i and j are interface atoms of types i and j belonging to different molecules, $\Delta E_{ij}(r)$ denotes the potential energy between atoms i and j in distance bin r, M_{ij} is the number of observations for atomic pair ij, $f_{ij}(r)$ is the relative frequency of occurrence for atoms i and j in distance bin r, $f_{xx}(r)$ is the relative frequency of occurrence for all atomic pairs in distance bin r, and σ is the weight given to each observation which was set to 0.02 according to Ref. [34]. The system average temperature, T, was set to 293 K.

For each atom typing, various potential functions were calculated on the basis of different combinations of the bin width and the cutoff distance (beyond which interactions are assumed to be ineffective). These two variables were ranged from 0.1 to 1 Å with 0.1 Å step and from 4 to 14 Å with 0.5 Å step respectively. The complete potentials are available as Supplemental material.

The energy score of each pose is the sum of interaction energies of atoms that are part of interface residues. A residue is defined as interfacial if it has at least one heavy atom separated by at most 4.5 Å from any heavy atom in the other partner molecule.

This definition of the interface region was followed in using both distance-based and DT-based potentials for the calculation of energy scores (Fig. 1).