Computers in Biology and Medicine 61 (2015) 127–137

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Prediction of hot regions in protein–protein interaction by combining density-based incremental clustering with feature-based classification

Jing Hu [a,b], Xiaolong Zhang [a,b,*], Xiaoming Liu [a,b], Jinshan Tang [c,*]

[a] School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, Hubei, China
[b] Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan 430065, Hubei, China
[c] School of Technology, Michigan Technological University, Houghton, MI 49931, USA

ARTICLE INFO

Article history:
Received 3 October 2014
Accepted 20 March 2015

Keywords:
Hot region
Clustering
Feature-based classification
Protein–protein interaction
Density-based

ABSTRACT

Discovering hot regions in protein–protein interaction is important for drug and protein design, while experimental identification of hot regions is a time-consuming and labor-intensive effort; thus, the development of predictive models can be very helpful. In hot region prediction research, some models are based on structure information, and others are based on a protein interaction network. However, the prediction accuracy of these methods can still be improved. In this paper, a new method is proposed for hot region prediction, which combines density-based incremental clustering with feature-based classification. The method uses density-based incremental clustering to obtain rough hot regions, and uses feature-based classification to remove the non-hot spot residues from the rough hot regions. Experimental results show that the proposed method significantly improves the prediction performance of hot regions.

1. Introduction

Protein functions can be expressed by protein–protein interactions which are very useful to understand the origination of diseases, but the principles that govern the interaction of two proteins and the general properties of their interaction interfaces remain unknown, resulting in difficulties when predicting interface regions. Hot spots [1–6] of protein–protein interactions play important roles in the functions and stability of protein complexes. Instead of being distributed along the protein interfaces homogeneously, hot spot residues are clustered within tightly packed regions [3,5,7,8], which are called hot regions. These are more important than hot spots in maintaining the stability of protein complexes and exerting the molecular mechanism of biological functions.

In the past, many attempts have been made to predict hot regions. The research group [3,6,8–14] in Koc University, Turkey, made contributions to the prediction of hot regions. Keskin developed an algorithm [3] to cluster hot spots into hot regions after studying the organization and contribution of structurally conserved hot spot residues. Tuncbag proposed a method [13] which combined the conservation of residues, accessible surface area and pair potential for prediction of hot regions. In [12,14] they predicted hot regions by the rule in [3] and the method of predicting hot spots in [8], then built a database called Hot Region [11]. But this method requires the structure of the protein, and is therefore limited by the available protein structures. In 2007, Hsu [15] presented a pattern-mining approach for the identification of hot regions in protein–protein interactions. The proposed method aimed to demonstrate that the important residues associated with the interface of protein–protein interactions may be discovered by sequential pattern-mining automatically. In [16], Pons studied a network-based method and used small-world residue networks to predict protein-binding areas. Although the proposed method has potential applications for protein docking as a complement to energy-based approaches, it shows limitations in many cases with certain topological features, like spherical or very large proteins. In [17], Nan proposed a method to predict hot regions based on complex network and community detection. By revising false positive and false negative during the detection process, the proposed method can improve the reliability in the recognition of hot regions. However, the prediction accuracy needs further improvement.

In this paper, we propose a method called Density-based Incremental Clustering with Feature-based Classification (DICFC), which can predict hot regions in protein–protein interactions by combining density-based incremental clustering with feature-based classification.

[*] Corresponding authors. Tel.: +86 27 68893295.
E-mail addresses: xiaolong.zhang@wust.edu.cn (X. Zhang), jinshant@mtu.edu (J. Tang).

http://dx.doi.org/10.1016/j.compbiomed.2015.03.022
0010-4825/© 2015 Elsevier Ltd. All rights reserved.

DICFC first forms the primary clusters by applying the density-based incremental clustering method to remove outliers, and then forms final hot regions, where a feature-based classification method is presented to remove the non-hot spot residues in the clustering results. In order to get the best features for classification, a feature selection method is studied. Experimental results show that the proposed method significantly improves the prediction performance for hot regions.

2. Method

In the proposed method, firstly, standard hot regions can be constructed for comparison using hot spots with the experimental data from the alanine mutation energy database [27]; then we will make some hot region predictions of both the hot spots and non-hot spots using the proposed method which combines density-based incremental clustering and feature-based classification; finally the prediction accuracy will be compared to the standard hot regions constructed above, from which the superiority of the proposed method can be drawn.

2.1. Definition of standard hot regions

In this paper, we adopted the standard definition of hot regions from Ozlem Keskin [3]. A hot region is defined as follows: every hot region contains at least three hot spots, and each hot spot is assumed to be within a hot region if it has at least two hot spot neighbors, and each hot spot residue is assumed to be a perfect sphere with a specific volume. The C^{α}-atoms of the hot spot residues are the centers of these spheres. The radii of the spheres are extracted from their sphere volumes. If the distance between the centers of two spheres (two C^{α}-atoms of two hot spots) is less than the sum of the radii of the two spheres plus a tolerance distance (2 Å), the two hot spot residues are flagged to be clustered and to form a network in the hot region.

Table 1
Standard hot region.

Complex	Hot region	Residues in the hot region
1A22	1	(A 172[a]) (A 175) (B 304) (A 178) (B 369) (B 243) (B 365)
1BRS	2	(A 73) (A 87) (A 102) (D 29) (D 35) (A 59) (D 39)
1BXI	3	(A 41) (A 50) (A 51) (A 55)
1DVF	4	(A 32) (B 101) (B 98) (B 100) (B 52) (B 54)
1F47	5	(A 8) (A 11) (A 12)
1FCC	6	(C 27) (C 31) (C 35) (C 43)
1JRH	7	(L 92) (I 49) (I 52) (I 53) (I 47) (I 82) (H 52) (H 53)
3HFM	8	(H 32) (H 33) (H 53) (H 50)
	9	(Y 20) (Y 96) (Y 97)
	10	(L 31) (L 32) (L 50)

[a] (A 172), ‘A’ is chain ID and ‘172’ is residue ID.

The coordinates of a C^{α} atom are obtained from the Protein Data Bank (PDB) [18], and the volume of a hot spot is as described in Appendix 1.

Based on the above definitions, the 65 hot spots in the data set (see Table 6 of Section 3.1) are organized into 10 hot regions, which contain the 49 hot spots shown in Table 1. Eight complexes out of 16 (see Table 7) have formed hot regions while the other eight complexes are excluded. The hot spots outside the hot regions are unable to form standard hot regions since they are not physically close enough to other hot spots. Table 2 lists all the hot spots of the eight complexes in standard hot regions and the hot spots outside standard hot regions are signified in bold.

2.2. Density-based incremental clustering

Similar to density-based clusters, hot spot residues are packed tightly within local regions rather than distributed along the protein interfaces homogeneously. Thus the hot spot residues can be clustered using some clustering methods. Clustering is a process to group data into multiple sub-groups or clusters so that objects within a cluster may have strong similarities [19]. The density of a residue O in the space can be measured by the number of residues close to it. Thus clustering is used to find the core residues, which are defined as the residues that have dense neighborhoods [19]. The proposed algorithm connects core residues and their neighborhoods to form dense regions as clusters. In order to use the clustering method to cluster the hot spot residues, we need to adopt several concepts from [19] (all distances in this paper are Euclidean distance):

- *Neighborhood*: a user-specified parameter $\varepsilon > 0$ is used to specify the radius of a neighborhood for every residue. The ε-neighborhood of a residue O is the space within radius ε centered at O.
- *Density of neighborhood*: due to the neighborhood size determined by ε-neighborhood, the density of any neighborhood can be measured simply by the number of residues in the corresponding neighborhood.
- *Dense region*: to determine whether a neighborhood is dense or not, another user-specified parameter “Min” is used to specify the density threshold of dense regions. “Min” is a variable that can be specified by the user.
- *Core residue*: a residue is a core residue if the ε-neighborhood of that residue contains at least “Min” residues.

For a dataset D composed of residues, we will identify all core residues in it with respect to the given parameters “ε” and “Min” by checking the number of residues in the neighborhood of a residue. Thus, the clustering task is reduced to using core residues and their neighborhoods to form dense regions, which are the clusters we need.

The process of density-based incremental clustering is described as follows: Initially, all residues in D are marked as “unvisited”. Then an

Table 2
Hot spots of the 8 complexes in standard hot regions.

Complex	Hot Spot residues
1A22	(A 172[a]) (A 175) (B 304) (A 178) (B 369) (B 243) (B 365)
1BRS	**(A 27)** (A 73) (A 87) (A 102) (D 29) (D 35) (A 59) (D 39)
1BXI	**(A 33) (A 34)** (A 41) (A 50) (A 51) (A 55)
1DVF	(A 32) (B 101) (B 98) (B 100) (B 52) (B 54)
1F47	(A 8) (A 11) (A 12)
1FCC	(C 27) (C 31) (C 35) (C 43)
1JRH	(L 92) (I 49) (I 52) (I 53) (I 47) (I 82) (H 52) (H 53)
3HFM	(H 32) (H 33) (H 53) (H 50) (Y 20) (Y 96) (Y 97) (L 31) (L 32) (L 50) **(L 96)**

The residues in bold are the hot spots outside standard hot regions.

[a] (A 172), ‘A’ is chain ID and ‘172’ is residue ID.