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

freepaper.me FREE paper

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.

© 2015 Elsevier Ltd. All rights reserved.

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.

freepaper.me FREE paper

unvisited residue p is selected randomly and marked as "visited". After that, the ε-neighborhood of p is checked to see whether it contains at least "Min" residues. If not, p is marked as a noise point.

Otherwise, a new cluster C is created for p, and all the residues in the ε-neighborhood of p are added to a candidate set, N. The algorithm iteratively adds to C those residues in N that do not belong to any cluster. In this process, for a residue p in set N that is labeled "unvisited", we mark it as "visited" and check its ε-neighborhood. If the ε-neighborhood of p' has at least "Min" residues, those residues in the ε-neighborhood of p' are added to N. Clustering continues adding residues to C until C can no longer be expanded, that is, until N is empty. At this time, cluster C is completed, and then we randomly select an unvisited residue from the remaining ones. The clustering process continues until all residues in the set D are visited.

2.3. Feature-based classification

After clustering, several clusters are generated, which contain both hot spots and non-hot spots. However, the hot region is just composed of hot spots [3], thus the non-hot spots should be removed from the obtained clusters. In this paper, a feature-based classification method [20] is introduced to identify the non-hot spots, which can then be removed from the density-based incremental clusters

SVM has been widely applied in the field of bioinformatics. Based on the research of Xia et al. [20], SVM, Bayes NET, Naive Bayes, RBF Network, decision tree and decision table are all involved in hot-spots classification but SVM outperforms the others, as can also be validated by the results from many other studies [20,21]. In this paper, an extended version of SVM called LIBSVM [22] is adopted. So the clustered residues are applied for classification using LIBSVM, in which Leave-one-out cross-validation is followed. First we divide all the clustered residues into 10 subsets, and then we randomly select a subset as the test set, and the other 9 subsets as the training set. After repeating the process ten times, the parameters c and g can be optimized.

For classification, feature selection is essential. There are many physical and chemical features for protein complexes and we cannot use all of them because some redundant and irrelevant features need to be removed to improve the classification performance. In our work, in order to explore the optimal features, first all the features by SVM-RFE with NMIFS filter (SRN) [23] are sorted and then the F-score is introduced to find the optimal combination of those features.

When sorting the features, the SRN [23] feature selection is a combination of SVM-RFE [24] and NMIFS [25]. Features are selected by embedding a SVM classifier, where normalized mutual information is used to balance the relevance between a feature and class label and the redundancy among different features.

In a linear SVM, the final classification function for a pattern $\boldsymbol{x} \in R^D$ is $f(x) = \sum_{i=1}^{D} w_i x_i + b$, and the component w_i in the weight vector $w \in R^D$ is used to measure the effectiveness of the i-th feature on the final classifier.

Table 3
Structure feature abbreviation of protein.

No	Symbol	Feature description
1	RctASA	Relative change in total ASA upon complexation
2	RcsASA	Relative change in side-chain ASA upon complexation
3	BsASA	Bound side-chain ASA
4	UsASA	Unbound side-chain ASA
5	RctmPI	Relative change in total mean PI upon complexation
6	UtmPI	Unbound total mean PI
7	BtASA	Bound total ASA
8	UtASA	Unbound total ASA
9	BtmPI	Bound total mean PI
10	RcsmPI	Relative change in side-chain mean PI
11	UsmPI	Unbound side-chain mean PI
12	BsmPI	Bound side-chain mean PI
13	RcpASA	Relative change in polar ASA upon complexation
14	UpASA	Unbound polar ASA
15	BpASA	Bound polar ASA
16	SRASA	Side-chain relative ASA
17	TRASA	Total relative ASA
18	Na	Number of atoms
19	B-factor	Temperature potential

ASA: Accessible Surface Area.
RASA: Relative Accessible Surface Area.
PI: Protrusion Index.

$$\text{RctASA} = ([\text{unbound total ASA}] - [\text{bound total ASA}])/([\text{unbound total ASA}]) \quad (3)$$

$$\text{RcsASA} = ([\text{unbound side} - \text{chain ASA}] - [\text{bound side} - \text{chain ASA}])/[\text{unbound side} - \text{chain ASA}] \quad (4)$$

$$\text{RctmPI} = ([\text{unbound total mean PI}] - [\text{bound total mean PI}])/[\text{unbound total mean PI}] \quad (5)$$

$$\text{RcsmPI} = ([\text{unbound side} - \text{chain mean PI}] - [\text{bound side} - \text{chain mean PI}])/[\text{unbound side} - \text{chain mean PI}] \quad (6)$$

$$\text{RcpASA} = ([\text{unbound polar ASA}] - [\text{bound polar ASA}])/([\text{unbound polar ASA}] \quad (7)$$

Table 4
Sorting result of 19 features.

Sorting	Feature
1	RctASA
2	RcsASA
3	BsASA
4	UsASA
5	RctmPI
6	UtmPI
7	BtASA
8	UtASA
9	BtmPI
10	RcsmPI
11	UsmPI
12	BsmPI
13	RcpASA
14	UpASA
15	BpASA
16	SRASA
17	TRASA
18	Na
19	B-factor

Table 5
F-score for different combinations.

Combination of features	F-score
1	0.756
1–2	0.769
1–3	0.781
1–4	0.811
1–5	**0.814**
1–6	0.813
1–7	0.811
1–8	0.807
1–9	0.807
1–10	0.806
1–11	0.803
1–12	0.801
1–13	0.801
1–14	0.801
1–15	0.781
1–16	0.769
1–17	0.769
1–18	0.756
1–19	0.75

The highest F-score value is shown in bold.

FREE paper freepaper.me دانلود کننده مقالات علمی

In the SRN method, the rank of the i-th feature is calculated as follows [23]:

$$r_i = \beta|w_i| + (1-\beta)\left[R_{s,f_i} - Q_{s,f_i}\right]$$
$$= \beta|w_i| + (1-\beta)[I(c;f_i) - \frac{1}{|S|}\sum_{f \in s} NI(f_i;f_s)] \quad (1)$$

where the parameter $\beta \in |0,1|$ is a user-tuned parameter, $I(c;f_i)$ is the relevance of feature f_i and the class label $c \in C = \{+1, -1\}$, $NI(f_i,f_s)$ is the normalized mutual information between features f_i and f_j, where $|S|$ is the number of features in the currently selected feature set S. Eq. (1) is used to rank features by combining the weight in the SVM classifier, relevance and redundancy.

The SRN works in an iterative way. The method uses a backward feature elimination strategy to obtain the important features and remove insignificant features (i.e., the feature having the smallest impact on the object function is excluded) at each step. It starts with all the features selected, then at each step the features are ranked with the above rank function, and those with the smallest rank value will be deleted. Finally, the rest will be features that are expected.

We apply the SRN [23] to select 19 physical and chemistry features [20,21,26], which are displayed in Table 3.

Then, features can be selected using the F-score [20], which assesses the discriminatory power of each individual feature. The larger the F-score is, the more discriminative ability it has for the feature. The F-score was calculated as

$$F - score = \frac{xni - xhi}{\sigma ni + \sigma hi} \quad (2)$$

where x_{ni} and x_{hi} denote the averages of both non-hot spots and hot spots, and σ_{ni} and σ_{hi} are the corresponding standard deviations. The F-score can measure the separation of the means for the two populations (hot spots and non-hot spots) in terms of their variances, which is closely related to the F-statistics and commonly used to evaluate the separation of the means for two random variables.

To discover the highest F-score combination, we list the 19 features in descending order in Table 4. First we combine the feature ranked the top with its next one to obtain the F-score; the same process is continued by grouping the last combined features with its next one until all the features are combined. Finally, the corresponding F-scores can be calculated, as shown in Table 5. It is obvious that the F-score of "combination 1–5" is the highest, and these five features are RctASA, RcsASA, BsASA, UsASA and RctmPI.

Table 6
Number of interface residues, hot spots, non-hot spots, and other unlabeled residues.

Complex	Interface residues	Labeled residues		Unlabeled residues
		Hot spots	Non-hot spots	
1A4Y	23	3	11	9
1A22	51	7	28	16
1AHW	7	1	3	3
1BRS	11	8	1	2
1BXI	17	6	3	8
1CBW	6	1	4	1
1DAN	13	2	8	3
1DVF	18	6	1	11
1F47	8	3	1	4
1FC2	3	1	0	2
1FCC	7	4	2	1
1GC1	17	0	11	6
1JRH	27	8	5	14
1VFB	23	3	6	14
2PTC	1	1	0	0
3HFM	23	11	6	6
TOTAL	**255**	**65**	**90**	**100**

Interface residues include labeled residues and unlabeled residues.
Labeled residues include hot spots and non-hot spots.
Hot spots are the interface residues with binding free energy higher than or equal to 2.0 kcal/mol.
Non-hot spots are the interface residues with binding free energy less than or equal to 0.4 kcal/mol.
Unlabeled residues are the interface residues with binding free energy between 0.4 kcal/mol and 2.0 kcal/mol.

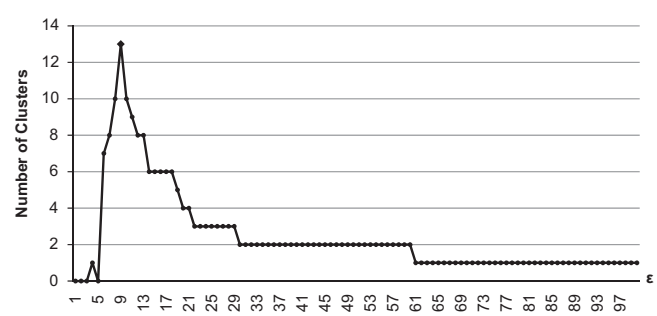

Fig. 1. Relationship between neighborhood radius ε and the number of clusters. The horizontal axis represents neighborhood radius and the vertical axis represents number of clusters. The dots show the value of the number of clusters, and the diamond shows the maximum value of the number of clusters.

Table 7
The 16 protein complexes in the dataset.

PDB ID of complex	First molecule	Second molecule
1A4Y	Angiogenin	Ribonuclease inhibitor
1A22	Human growth hormone	Human growth hormone binding protein
1AHW	Immunoglobulin Fab 5G9	Tissue factor
1BRS	Barnase	Barstar
1BXI	Colicin E9 Immunity Im9	Colicin E9 DNase
1CBW	BPTI Trypsin inhibitor	Chymotrypsin
1DAN	Blood coagulation factor VIIA	Tissue factor
1DVF	Idiotopic antibody FV D1.3	Anti-idiotopic antibody FV E5.2
1F47	Cell division protein ZIPA	Cell division protein FTSZ
1FC2	Fc fragment	Fragment B of protein A
1FCC	Fc (IGG1)	Protein G
1GC1	Envelope protein GP120	CD4
1JRH	Antibody A6	Interferon-gamma receptor
1VFB	Mouse monoclonal antibody D1.3	Hen egg lysozyme
2PTC	BPTI	Trypsin
3HFM	Hen egg lysozyme	Ig FAB fragment HyHEL-10

FREE paper freepaper.me دانلود کننده مقالات علمی

2.4. The DICFC algorithm

The proposed algorithm is a hybrid algorithm which combines density-based incremental clustering with feature-based classification (DICFC) for hot region prediction. First, the algorithm visits all the residues in the data set D to find the core residues set C, and then LIBSVM is used to remove the non-hot residues from set C to obtain predicted hot regions. The algorithm can be described as follows:

Input:

- D: a data set containing 155 residues and their three-dimensional coordinates,
- ε: the radius parameter,
- Min: the neighborhood density threshold.

Output: predicted hot regions from DICFC.
Method:

```
(1)  mark all residues as unvisited;
(2)  do
(3)    randomly select an unvisited object p;
(4)    mark p as visited;
(5)    if the ε-neighborhood of p has at least "Min" residues
(6)      create a new cluster C, and add p to C;
(7)      let N be the set of residues in the ε-neighborhood of
  p;
(8)      for each point p' in N
(9)          if p' is unvisited
(10)             mark p' as visited;
(11)         if the ε-neighborhood of p' has at least "Min"
  residues, add those residues to N;
(12)         if p' is not yet a member of any cluster, add p' to
  C;
(13)     end for
(14)   output clusters C;
(15) else mark p as outlier;
(16) until no residue is unvisited;
(17) run PSAIA with all residues in clusters C;
(18) calculate RctASA, RcsASA, BsASA, UsASA and RctmPI of
  residues in clusters C;
(19) run LIBSVM with the above five structure feature values;
(20) mark h as hot spot residue, n as non-hot residue;
(21) remove residues of marked n in clusters C,
(22) mark formed new cluster as hot region R;
(23) output hot region R.
```

Phase I: Clustering

(1)–(6): Create the core residues set C.

(7)–(13): Visit residues in the ε-neighborhood of set C, and add all residues that match the condition into set C.

(14)–(16): The clustering process continues until all residues in set D are visited, then output set C and mark all residues outside set C as outliers.

Phase II: Classification

(17)–(18): Calculate the five structure feature values of the residues in C using PSAIA.

(19)–(21): Classify all residues in set C using LIBSVM to remove non-hot spot residues.

(22)–(23): Mark newly formed clusters from (21) and output as hot regions.

3. Experiment results and evaluation

3.1. Datasets

In the experiments, we used the 16 protein complexes listed in Table 7. These complexes are from the database called ASEdb [27]. In the experiments, each complex is composed of a bunch of interface residues. These interface residues [3] are the residues that the decrement of the accessible surface area are more than 1 Å during the process of forming the complex. The energy values binding for each residue were obtained by alanine mutation experiments.

We removed the protein chains which are repetitive using the CATH [28] query system with the sequence identity less than 35% and the SSAP [29] score less than or equal to 80. Moreover, we also removed the protein chains that do not interact with each other from the original data set. After that, we had 255 interface residues. An interface residue [3,20] will be defined as a hot spot if its corresponding binding free energy is higher than or equal to 2.0 kcal/mol, while an interface residue with binding free energy less than or equal to 0.4 kcal/mol is considered as a non-hot spot, according to which there are 65 hot spots and 90 non-hot spots in the dataset. The other 100 unlabeled residues with binding free energy between 0.4 kcal/mol and 2.0 kcal/mol are excluded from the training set because their hot spot or non-hot spot features are not very clear [20]. Table 6 summarizes the interface residue labeling for each complex, and illustrates the number of interface residues, hot spots, non-hot spots, and other unlabeled residues separately. In the experiments, we got 155 interface residues as the training data set, whose structure features are obtained by PSAIA [30].

Table 8
Feature-based incremental cluster result in each protein complex when $\varepsilon=9$, Min=3.

Complex	Cluster	Residues in cluster
1A22	1	(A 18[a]) (A 21) (A 22) (A 174) (B 417) (B 418) (A 25) (A 167) (A 168) (B 419) (A 172) (A 175) (A 178) (B 371) (A 164) (B 304) (B 369) (B 367) (B 365) (A 51) (A 62) (A 63) (B 302) (B 324) (B 243) (B 366) (A 65) (B 298)
	2	(A 42) (A 46) (B 320) (B 321) (B 275) (B 273) (B 274)
1BRS	3	(A 102) (A 87) (D 29) (D 35) (D 39) (A 59) (A 60)
1BXI	4	(A 48) (A 41) (A 50) (A 51) (A 55)
1CBW	5	(I 11) (I 15) (I 34) (I 39)
1DAN	6	(T 17) (T 18) (T 20) (T 21) (T 58)
1DVF	7	(B 98) (B 52) (B 100) (B 101) (A 32) (A 92)
1F47	8	(A 11) (A 8) (A 12) (A 15)
1GC1	9	(C 23) (C 25) (C 42) (C 63) (C 27) (C 40) (C 35) (C 45) (C 32) (C 33)
1JRH	10	(L 92) (L 94) (I 49) (I 53) (I 54) (I 55) (I 52) (I 47) (I 82) (H 52) (I 84) (I 98)
3HFM	11	(H 31) (H 32) (H 33) (H 53) (Y 73) (H 50) (Y 63)
	12	(Y 96) (Y 97) (Y 100)
	13	(Y 15) (Y 89) (L 31) (L 32) (L 50) (Y 20)

[a] (A 18), 'A' is chain ID and '18' is residue ID.

freepaper.me FREE paper

3.2. Experimental results

We first conducted the clustering experiments, where 3-D coordinates of the residues were used for clustering. In the experiments, we used the coordinates of the Cα atom of a protein residue to represent the residue's coordinates. All the values of coordinates of the residues were from the protein data bank (PDB) [18]. In the clustering experiments, the density threshold "Min" was set to three when performing the density-based incremental clustering because every hot region contains at least three hot spots [3]. In order to determine the value of the neighborhood radius ε, we studied the relationship between neighborhood radius ε and the number of clusters. The relationship is shown in Fig. 1, from which we can find that different ε values will result in different cluster numbers. When $\varepsilon=9$, the cluster number reaches the maximum value, which is 13. When $\varepsilon \geq 61$, the number of clusters is 1. And the experiment result indicates that when $\varepsilon < 9$ or $\varepsilon > 9$, the hot spot residues are not enough to form any hot region. Thus we set $\varepsilon=9$.

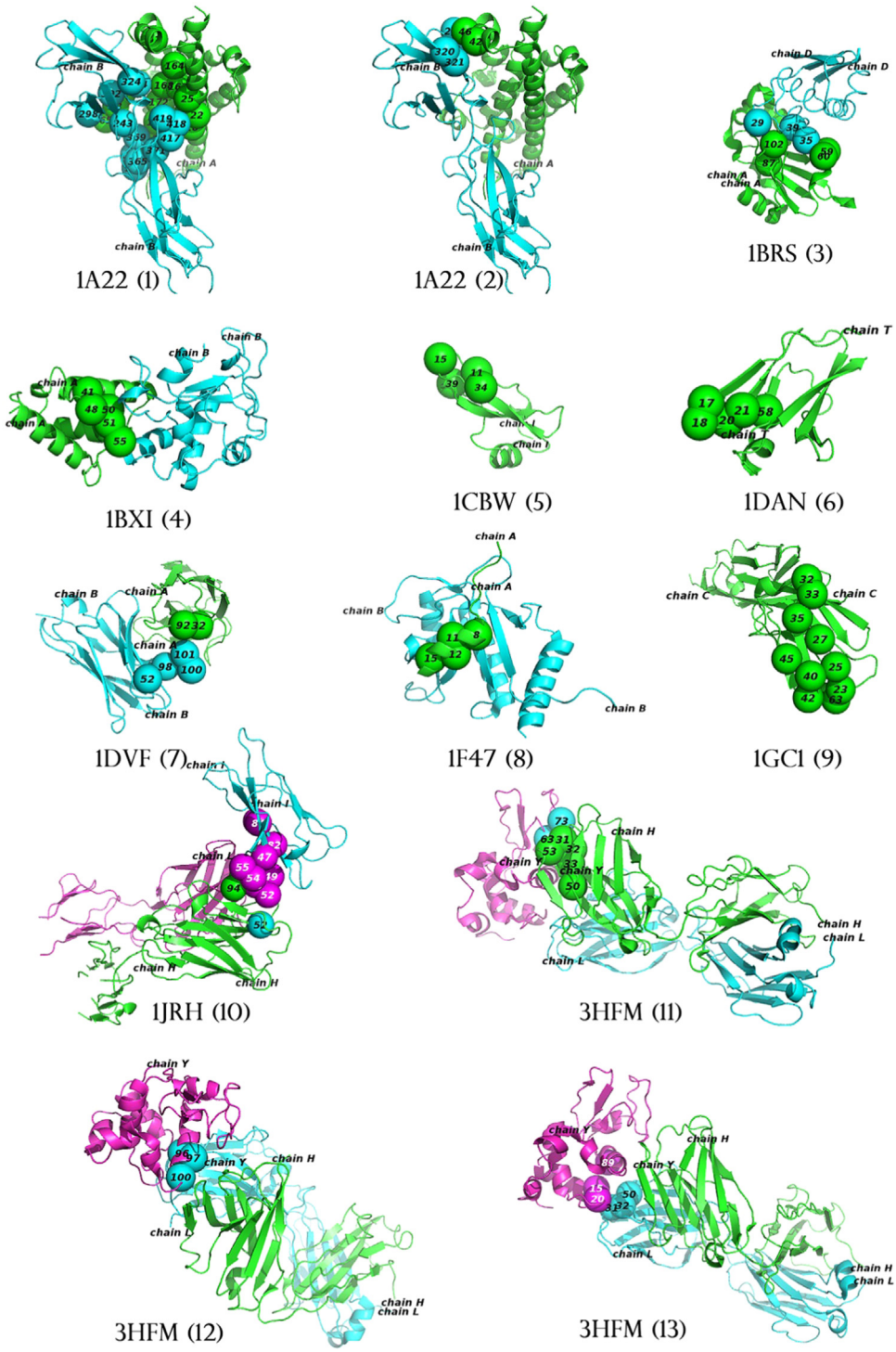

Fig. 2. Visualization of clustering results in each complex when $\varepsilon=9$ and Min = 3. The spheres labeled with their residue IDs represent the residues with IDs listed in Table 8. Spheres and their located chain are the same color, while different colors represent different chains. The numbers in brackets correspond to the clusters in Table 8. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

دانلود کننده مقالات علمی freepaper.me FREE paper

Table 9
Prediction result of DICFC in this paper.

Complex	Hot region	Corresponding cluster[a]	Residues in cluster
1A22	1	1	(A 21[b]) (B 418) (A 168) (A 175) (A 164) (B 304) (B 369) (B 243)
1BRS	2	3	(A 102) (A 87) (D 29) (D 35) (D 39) (A 59)
1DAN	3	6	(T 18) (T 20) (T 21) (T 58)
1DVF	4	7	(B 98) (B 52) (B 101) (A 32)
1GC1	5	9	(C 42) (C 40) (C 35)
1JRH	6	10	(L 92) (L 94) (I 49) (I 53) (I 52) (I 47) (H 52)
3HFM	7	11	(H 31) (H 33) (H 53) (H 50)
	8	12	(Y 96) (Y 97) (Y 100)
	9	13	(L 31) (L 32) (L 50) (Y 20)

After using feature-based classification to remove the non-hot spot residues from the clusters obtained in Table 8, the numbers of hot spots in cluster 2, cluster 4, cluster 5 and cluster 8 are less than 3 to form a hot region, so these clusters are removed.

[a] Corresponding clusters are the clusters in Table 8.
[b] (A 21), 'A' is chain ID and 21 is residue ID.

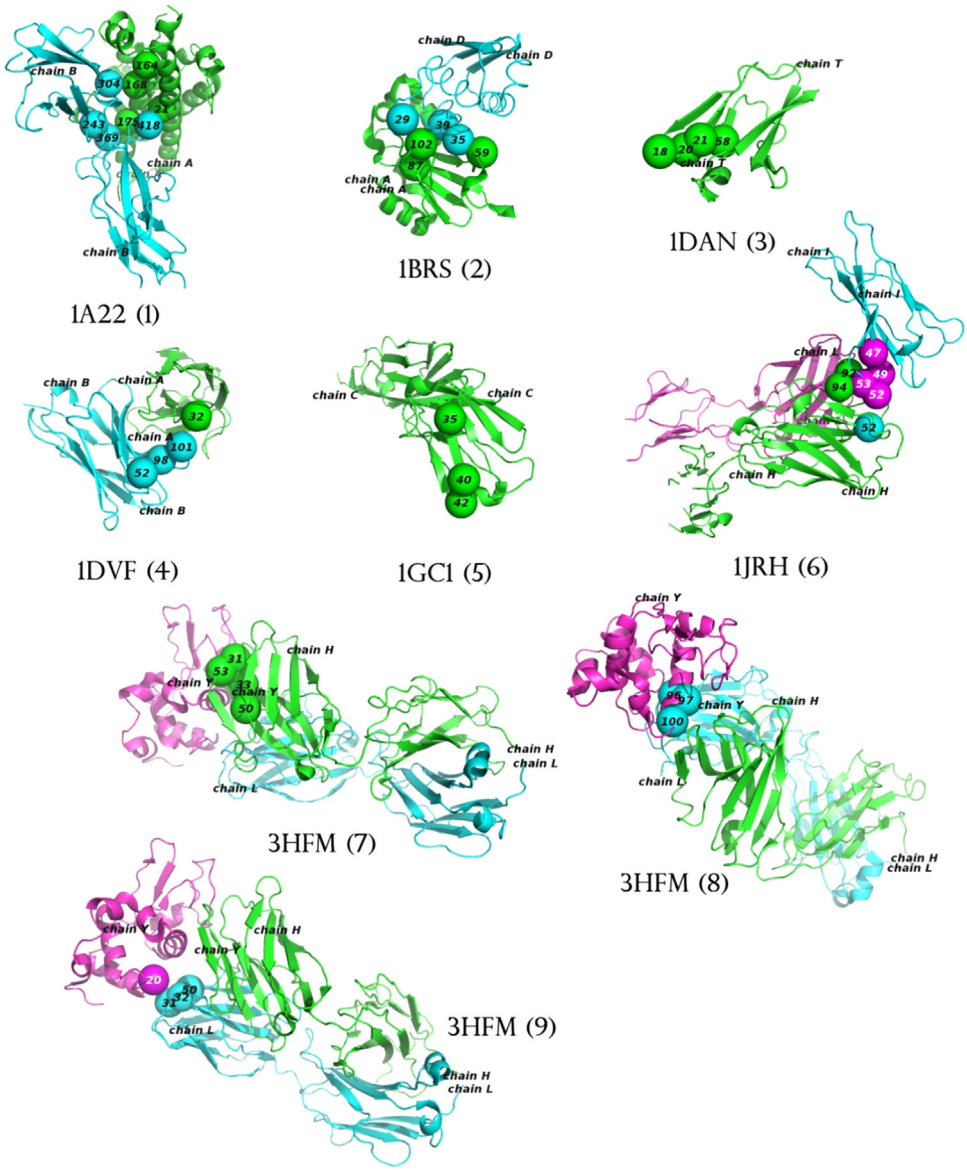

Fig. 3. Visualization of prediction results of DICFC. The spheres labeled with their residue IDs represent the residues with IDs listed in Table 9. Spheres and their located chain are the same color, while different colors represent different chains. The numbers in brackets correspond to the clusters in Table 9. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

FREE paper freepaper.me دانلود کننده مقالات علمی

Table 8 provides the clustering results in each complex when $\varepsilon=9$ and Min=3. Fig. 2 shows the residues highlighted on the structure and their IDs are stated in Table 8. After clustering, there are 104 residues left when 51 outliers from the dense region are eliminated. There are 13 clusters in 10 proteins of the 16 complexes, of which there are 3 clusters in 3HFM, 2 clusters in 1A22, and 1 cluster each in 1BRS, 1BXI, 1CBW, 1DAN, 1DVF, 1F47, 1GC1 and 1JRH, but no clusters in 1A4Y, 1AHW, 1FC2, 1FCC, 1VFB and 2PTC. Table 9 shows the final results after using feature-based classification to remove non-hot spot residues from the clusters obtained in Table 8. Fig. 3 shows the residues highlighted on the structure and their IDs are stated in Table 9. There are 9 hot regions compounded from 43 residues. The numbers of the hot regions in each protein complex are given in Table 10.

3.3. Evaluation

In order to evaluate the performance of the proposed method, three criteria are used for predicting both hot spots and hot regions [17,20]:

$$Recall = \frac{TP}{TP+FN} \tag{8}$$

$$Precision = \frac{TP}{TP+FP} \tag{9}$$

$$F-measure = \frac{2*Recall*Precision}{Recall+Precision} \tag{10}$$

When predicting hot spots, the following notations are used:

True Positive (TP): the number of hot spots in predicted hot regions and also in standard hot regions.

False Negative (FN): the number of hot spots that are not in predicted hot regions but in standard hot regions.

False Positive (FP): the number of hot spots in predicted hot regions but not in standard hot regions.

Precision represents the accuracy of the hot spot prediction, and Recall represents the coverage of predicted hot spots in standard hot regions. With a good balance between Precision and Recall, the *F*-measure offers a better overall accuracy of hot spot prediction.

However, for prediction of hot regions, the above notations assume different meanings, as follows:

True Positive (TP): the number of hot regions in predicted hot regions and also in standard hot regions.

False Negative (FN): The number of hot regions that are not in predicted hot regions but in standard hot regions.

False Positive (FP): The number of hot regions in predicted hot regions but not in standard hot regions.

Similarly, Precision represents the accuracy of the hot region prediction, and Recall represents the coverage of predicted hot regions in standard hot regions. With a good balance between Precision and Recall, the *F*-measure offers a better overall accuracy in predicting hot regions than solely using either Precision or Recall.

3.4. Performance comparison with the previous methods

3.4.1. Comparison of prediction performance using evaluation criteria

In this study, we compared the proposed method with the previous methods including those of Tuncbag [14] and Nan [17].

Table 11 summarizes the performance of the different methods on the same data set. Among these approaches, Tuncbag's [13,14] method is physicochemical and structure feature-based, while Nan's method [17] is based on complex network and community detection. Regarding the number of hot regions, DICFC is able to correctly predict 7 (recall=0.700) hot regions from 10 standard hot regions while Nan and Tuncbag only predict 4 (recall=0.400) and 2 (recall=0.200). On the hot spot coverage of predicted hot regions in standard ones, the hot spot recall for DICFC, Nan's method, and Tuncbag's method are 0.571, 0.285 and 0.122, respectively. Therefore, the proposed method predicts the largest proportion of hot regions, which reaches 57.1%, while the other two methods achieve no more than 30%. DICFC is able to predict hot regions correctly from the data set with 0.651 and 0.571. This means that DICFC is able to correctly predict 57.1% of the true hot regions in this data set, and 65.1% of the predicted hot regions are

Table 10
Number of hot regions in DICFC of each protein complex.

Complex	Hot region
1A4Y	0
1A22	1
1AHW	0
1BRS	1
1BXI	0
1CBW	0
1DAN	1
1DVF	1
1F47	0
1FC2	0
1FCC	0
1GC1	1
1JRH	1
1VFB	0
2PTC	0
3HFM	3

Table 11
Comparison results with different methods to predict hot regions.

Method	Hot spot			Hot region		
	Recall	Precision	*F*-measure	Recall	Precision	*F*-measure
Tuncbag	0.122	**1**	0.217	0.200	**1**	0.333
Nan	0.285	0.424	0.341	0.400	0.667	0.500
DICFC (this work)	**0.571**	0.651	**0.608**	**0.700**	0.778	**0.737**

The highest value in each column is shown in bold.

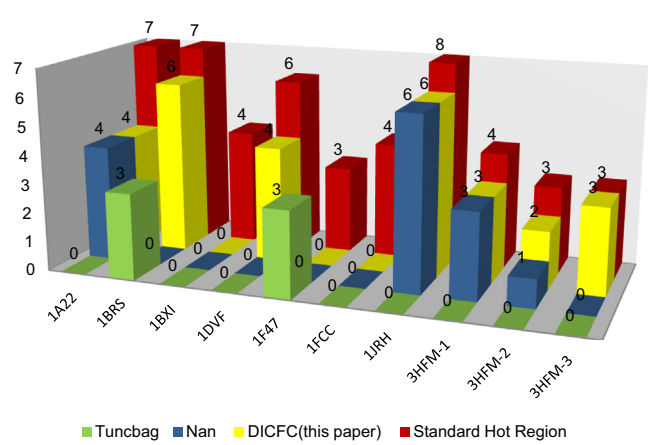

Fig. 4. Number of standard hot regions correctly predicted by different methods. Red bars represent standard hot regions, and yellow/blue/green bars represent the predicted results of DICFC, Nan and Tuncbag respectively. The number above each bar is the number of residues correctly predicted by the different methods. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

freepaper.me FREE paper

identified as true hot regions. By contrast, Tuncbag's method has high accuracy, but the coverage of predicted hot regions in standard hot regions is very low, which means that Tuncbag's method is able to predict only 12.2% of hot regions; although the predicted hot regions are correct, the accuracy is too low. The prediction performance of Nan's method is located between that of our method and that of Tuncbag's. In addition, the comprehensive score *F*-Measure of DICFC is 0.608, which is higher than Nan's and Tuncbag's. From the above comparison, we can conclude that DICFC provides a remarkably better prediction performance than the previous prediction methods.

Fig. 4 describes the coverage status of predicted hot regions in standard ones. Our method is able to predict 7 hot regions from 10 standard hot regions, while Nan is able to predict 4 (Appendix 2) and Tuncbag is able to predict 2 (Appendix 3).

For a more comprehensive study, the results of applying hot spot prediction before the density clustering were also conducted as a comparison with the current one. The prediction results using the same data set are shown in Appendix 4, where the recall, precision and *F*-measure of hot spots are 0.408, 0.645 and 0.5, while the recall, precision and *F*-measure of hot regions are 0.6, 0.857 and 0.706. According to these results, applying density clustering before hot spot prediction (DICFC in this paper) gives better prediction performance than applying hot spot prediction before density clustering. Considering that these two methods are of the same type, we select the one with the better performance as the proposed method.

FREE paper freepaper.me دانلود کننده مقالات علمی

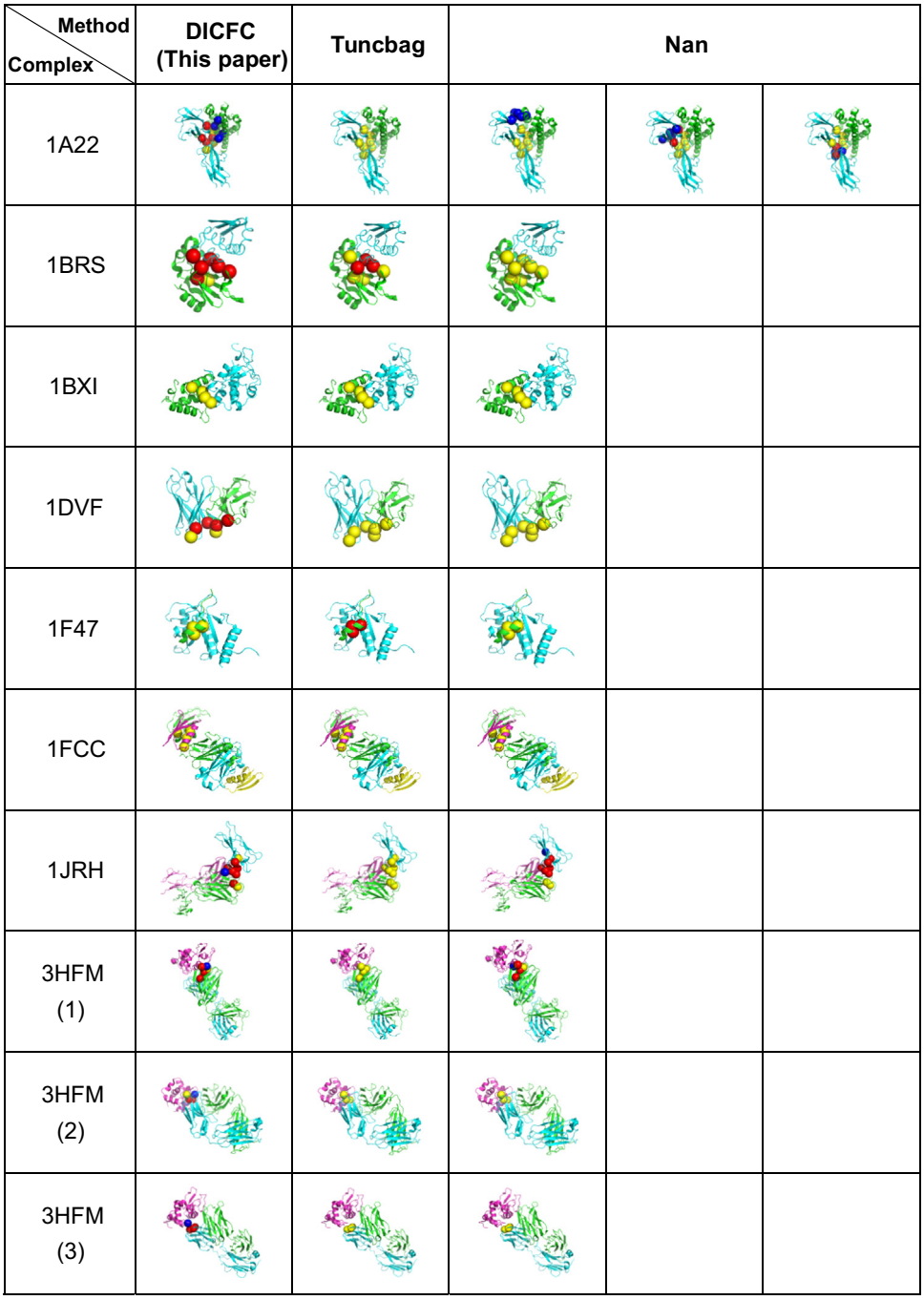

Fig. 5. Visualization of prediction results for 8 protein complexes in standard hot regions using different methods. The related residues are shown as small spheres: red ones are the real hot regions and are correctly predicted; yellow ones are the real hot regions but missed out by the prediction; and blue ones are the fake hot regions predicted as hot regions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.4.2. Comparison of prediction results visualization

In order to show the effectiveness of the proposed method, we provided a visualization of the prediction results obtained by the different methods. The visualization tool we adopted is the molecular visualization software Pymol [31]. Fig. 5 shows the visualization results for 8 complexes.

To further illustrate the effectiveness of the proposed approach (DICFC) for predicting hot regions, we randomly selected two examples for detailed analysis. The two complexes are 1BRS and 1JRH.

The first example is the interaction between chain A and chain D of complex 1BRS, which is shown in Fig. 6. In this complex, there are seven hot spot residues in the standard hot regions and they are (A GLU 73), (A ARG 87), (A HIS 102), (D TYR 29), (D ASP 35), (A ARG 59) and (D ASP 39). DICFC is able to correctly predict six of these seven hot spots, which are (A HIS 102), (A ARG 87), (D TYR 29), (D ASP 35), (D ASP 39) and (A ARG 59). Tuncbag's method is only able to predict three of the seven hot spots correctly, i.e. (A HIS 102), (D ASP 35) and (D ASP 39). Nan's method failed to predict any residues in this hot region. However, hot spots (A GLU 73) and (D TYR 29) cannot be predicted correctly by any of these three methods, which suggests that the mutations of these two residues might contribute to protein destabilization.

The second example is the interaction between chain H, chain I and chain D of complex 1JRH, which is shown in Fig. 6. There are eight hot spot residues in the standard hot regions, which are (L TRP 92), (I TYR 49), (I LYS 52), (I ASN 53), (I LYS 47), (I TRP 82), (H TRP 52) and (H TRP 53). DICFC is able to correctly predict six of these eight hot spots, which are (L TRP 92), (I TYR 49), (I ASN 53), (I LYS 52), (I LYS 47) and (H TRP 52). The unpredicted residues are (I TRP 82) and (H TRP 53), and the incorrectly predicted hot spot is (I LYS 98). The five residues (L TRP 92), (I LYS 47), (I ASN 53), (I TYR 49) and (I LYS 52) are correctly predicted by both DICFC and Nan' methods, which means that the region formed by these five residues plays an important role in the function and the stability of protein complexes. Tuncbag was unable to predict any of the residues in this hot region. Residue (H TRP 53) is not predicted by any of the three methods, suggesting that the mutations of this residue might contribute to protein destabilization.

These prediction results demonstrated that DICFC is able to predict more hot regions than the others in all the hot regions formed by the 16 protein complexes. Moreover, DICFC predicts more hot spots in a single hot region of a single protein complex.

4. Discussion

A deeper understanding of the functions of proteins and the interactions between proteins is the key to designing drugs and proteins. And discovering hot regions in protein–protein interaction is important for understanding the interactions between proteins. Because of the complexity of experimental methods, the computational prediction method is greatly improving the efficiency and accuracy of experimental identification of hot regions. The major contribution of this study is to propose a new model combining clustering with classification that has several advantages as it significantly improves the hot region prediction performance; firstly it imports density-based incremental clustering into hot region prediction; it simplifies the hot region prediction requirements from specific shape and detailed structure information to the residue coordinates

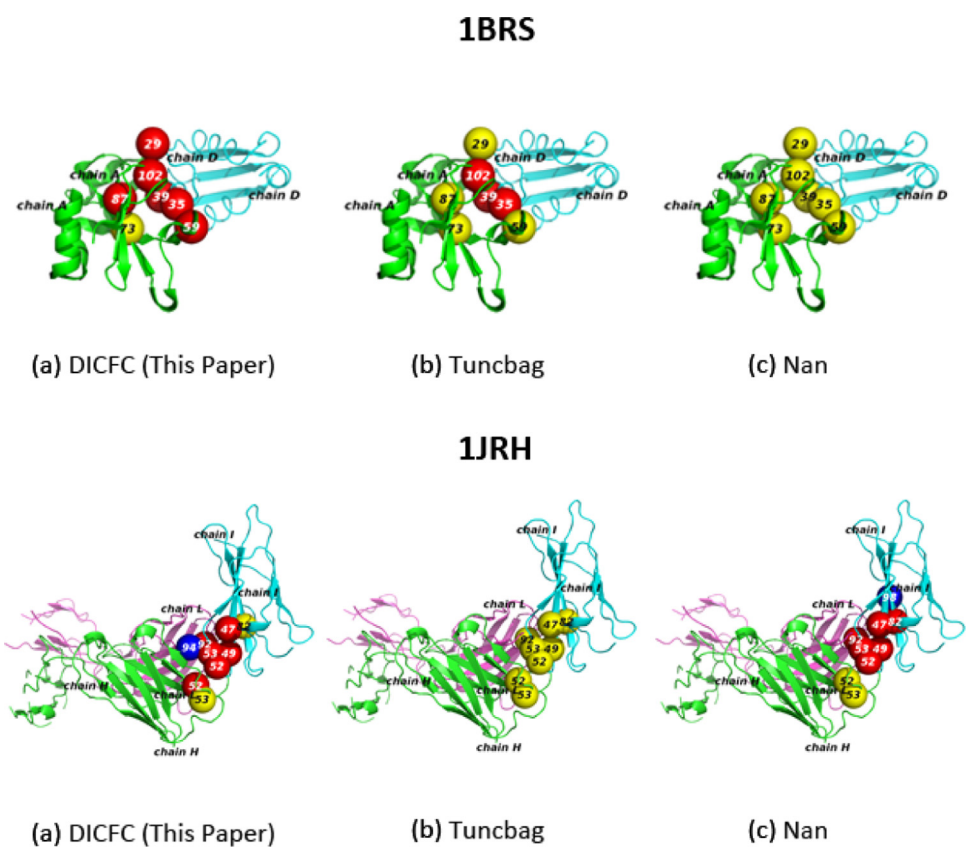

Fig. 6. Visual prediction results of 1BRS and 1JRH from different methods. In complex 1BRS, the green belt represents chain A, and the blue belt represents chain D. In complex 1JRH, the green belt represents chain H, the blue belt represents chain I, and the magenta belt represents chain L. The related residues are shown as small spheres: red ones are the real hot regions and are correctly predicted; yellow ones are the real hot regions but missed out by the prediction; and blue ones are the fake hot regions but predicted as hot regions. The number on each sphere is its residue ID. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

FREE paper freepaper.me دانلود کننده مقالات علمی

only; the SRN feature selection combined with SVM-RFE and NMIFS is applied to efficiently find the best features for classification.

Meanwhile, the limitation of DICFC is that the clustering results probably depend on the choice of the values of the neighborhood density threshold and neighborhood radius. In the future, we will conduct research on better principles or criteria for choosing these two parameters, and we will also continue to collect more experimental and published data for wider tests in order to optimize our prediction model in the future.

Although the proposed model in this paper is density-based incremental clustering combined with feature-based SVM, a better cluster method and novel feature selection methods can easily be incorporated to improve the hot regions prediction performance.

5. Conclusions

In this paper, we propose a method called DICFC, which predicts hot regions by combining density-based incremental clustering with feature-based classification. The first step uses the density-based incremental clustering method to obtain rough hot regions and the second step uses feature-based classification to remove the non-hot spot residues from the clusters obtained by the first step. By removing outliers when clustering, the experimental results show that the proposed method significantly improves the performance of hot regions prediction.

Feature-based classification, as an important and efficient way to predict hot spots, may not be the most efficient method to use for hot region prediction. Our research shows that we have already completed most of the predictions by discovering a large part of the hot region clusters using the density-based incremental cluster before feature classification. Although we go on to use feature classification to further improve the prediction accuracy, compared with feature classification, the density-based incremental clustering plays a more important role in hot region prediction. Importing density-based incremental clustering will greatly improve the prediction accuracy of hot regions.

Conflict of interest statement

We declare that we have no financial and personal relationships with other people or organizations that could inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled “Prediction of hot regions in protein–protein interaction by combining density-based incremental clustering with feature-based classification”.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (Nos. 61273225 and 61201423), the Young Science and Technology Outstanding Training Plan of Wuhan University of Science and Technology (No. 2013xz014). Thanks to Dongfang Nan, Juhong Qi and Jing Ye in our lab, and two college students Maosen Chen and Qi Mo for their meaningful discussion.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.compbiomed.2015.03.022.

References

[1] A.A. Bogan, K.S. Thorn, Anatomy of hot spots in protein interfaces, J. Mol. Biol. 280 (1998) 1–9.
[2] T. Clackson, J.A. Wells, A hot spot of binding energy in a hormone-receptor interface, Science 267 (1995) 383–386.
[3] O. Keskin, B. Ma, R. Nussinov, Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues, J. Mol. Biol. 345 (2005) 1281–1294.
[4] B. Ma, R. Nussinov, Druggable orthosteric and allosteric hot spots to target protein–protein interactions, Curr. Pharm. Des. 20 (2014) 1293–1301.
[5] B. Ma, T. Elkayam, H. Wolfson, R. Nussinov, Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces, Proc. Natl. Acad. Sci. USA 100 (2003) 5772–5777. http://dx.doi.org/10.1073/pnas.1030237100, published online Epub May 13.
[6] X. Li, O. Keskin, B. Ma, R. Nussinov, J. Liang, Protein–protein interactions: hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states: implications for docking, J. Mol. Biol. 344 (2004) 781–795. http://dx.doi.org/10.1016/j.jmb.2004.09.051, published online Epub November 26.
[7] I. Halperin, H. Wolfson, R. Nussinov, Protein–protein interactions: coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking, Structure 12 (2004) 1027–1038.
[8] O. Keskin, B. Ma, K. Rogale, K. Gunasekaran, R. Nussinov, Protein–protein interactions: organization, cooperativity and mapping in a bottom-up systems biology approach, Phys. Biol. 2 (2005) S24–35.6.
[9] O. Keskin, A. Gursoy, B. Ma, R. Nussinov, Principles of protein–protein interactions: what are the preferred ways for proteins to interact? Chem. Rev. 108 (2008) 1225–1244. http://dx.doi.org/10.1021/cr040409x, published online Epub April.
[10] E. Cukuroglu, A. Gursoy, O. Keskin, Analysis of hot region organization in hub proteins, Ann. Biomed. Eng. 38 (2010) 2068–2078.
[11] E. Cukuroglu, A. Gursoy, O. Keskin, Hotregion: a database of predicted hot spot clusters, Nucl. Acids Res. 40 (2012) D829–D833.
[12] E. Guney, N. Tuncbag, O. Keskin, A. Gursoy, Hotsprint: database of computational hot spots in protein interfaces, Nucl. Acids Res. 36 (2008) D662–D666.
[13] N. Tuncbag, A. Gursoy, O. Keskin, Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy, Bioinformatics 25 (2009) 1513–1520.
[14] N. Tuncbag, O. Keskin, A. Gursoy, Hotpoint: hot spot prediction server for protein interfaces, Nucl. Acids Res. 38 (2010) W402–W406.
[15] C.M. Hsu, C.Y. Chen, B.J. Liu, C.C. Huang, M.H. Laio, C.C. Lin, T.L. Wu, Identification of hot regions in protein–protein interactions by sequential pattern mining, BMC Bioinform. 8 (Suppl 5) (2007) S8.
[16] C. Pons, F. Glaser, J. Fernandez-Recio, Prediction of protein-binding areas by small-world residue networks and application to docking, BMC Bioinform. 12 (2011) 378.
[17] D.F. Nan, X.L. Zhang, Prediction of hot regions in protein–protein interactions based on complex network and community, in: Fourth Workshop on Integrative Data Analysis in System Biology (IDASB), vol. 12, 2013.
[18] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, Nucl. Acids Res. 28 (2000) 235–242.
[19] J.W. Han, M. Kamber, J. Pei, Data Mining Concepts and Techniques, third ed., China Machine Press, China, ISBN 978-7-111-39140-1, 2012.
[20] J.F. Xia, X.M. Zhao, J. Song, D.S. Huang, APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility, BMC Bioinform. 11 (2010) 174.14.
[21] J. Song, H. Tan, K. Mahmood, R.H. Law, A.M. Buckle, G.I. Webb, T. Akutsu, J.C. Whisstock, Prodepth: predict residue depth by support vector regression approach from protein sequences only, PLoS One 4 (2009) e7072.
[22] 〈http://www.csie.ntu.edu.tw/∼cjlin/libsvm〉.
[23] X.M. Liu, J.S. Tang, Mass classification in mammograms using selected geometry and texture features, and a new SVM-based feature selecting method, IEEE Syst. J. 99 (2013) 1932–8184.
[24] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (1) (2002) 389–422.
[25] P.A. Estévez, M. Tesmer, C.A. Perez, J.M. Zurada, Normalized mutual information feature selection, IEEE Trans. Neural Netw. 20 (2) (2009) 189–201.
[26] J. Song, Z. Yuan, H. Tan, T. Huber, K. Burrage, Predicting disulfide connectivity from protein sequence using multiple sequence feature vectors and secondary structure, Bioinformatics 23 (2007) 3147–3154.
[27] K.S. Thorn, A.A. Bogan, ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions, Bioinformatics 17 (2001) 284–285.
[28] F. Pearl, A. Todd, I. Sillitoe, M. Dibley, O. Redfern, T. Lewis, C. Bennett, R. Marsden, A. Grant, D. Lee, A. Akpor, M. Maibaum, A. Harrison, T. Dallman, G. Reeves, I. Diboun, S. Addou, S. Lise, C. Johnston, A. Sillero, J. Thornton, C. Orengo, The CATH domain structure database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis, Nucl. Acids Res. 33 (2005) D247–D251.
[29] G. Toth, C.R. Watts, R.F. Murphy, S. Lovas, Significance of aromatic-backbone amide interactions in protein structure, Proteins 43 (2001) 373–381.
[30] J. Mihel, M. Sikic, S. Tomic, B. Jeren, K. Vlahovicek, PSAIA – protein structure and interaction analyzer, BMC Struct. Biol. 8 (2008) 21.
[31] 〈http://www.pymol.org/〉.

دانلود کننده مقالات علمی freepaper.me FREE paper