Chemometrics and Intelligent Laboratory Systems 131 (2014) 16–21

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Prediction of hot spots residues in protein–protein interface using network feature and microenvironment feature

Ling Ye, Qifan Kuang, Lin Jiang, Jiesi Luo, Yanping Jiang, Zhanling Ding, Yizhou Li *, Menglong Li *

College of Chemistry, Sichuan University, Chengdu 610064, PR China

ARTICLE INFO

Article history:
Received 21 July 2013
Received in revised form 17 October 2013
Accepted 2 November 2013
Available online 6 December 2013

Keywords:
Hot spots
Protein interface
Residue–residue interaction network
Microenvironment

ABSTRACT

Hot spots residues in protein–protein interface play crucial roles in protein binding. In the present study, complex network method was applied to uncover influence of neighboring residues on hot spots and then several network and microenvironment features were designed to describe the diversity of environment of hot spots. After feature analysis by permutation importance in Random Forest (RF), an optimal 58-dimensional feature set including ten network and microenvironment features was selected and then applied to construct a Support Vector Machine (SVM) prediction model for hot spots. A satisfactory accuracy (ACC) value of 79.0% and a Mathew's correlation coefficient (MCC) value of 0.470 were obtained for independent test set. The novel network features and microenvironment features were proved to be promising in discovering hot spots in interfaces. A further microenvironment analysis was also performed. Amino acid residues directly contacting with hot spots in residue–residue interaction network exhibit significant importance for the microenvironment of hot spots. Amino acid alanine (A), aspartic acid (D), glycine (G), histidine (H), isoleucine (I), asparagine (N), serine (S) and tyrosine (Y) are more likely to occur in the vicinity of hot spots than in the vicinity of non-hot spots. These amino acid residues probably cluster together to construct a proper microenvironment for hot spots.

1. Introduction

Protein–protein interactions are involved in different kinds of cellular functions such as metabolism and signal transduction [1,2]. Proteins rarely perform one biological function alone, but tend to cooperate with other proteins [3–5]. Studies on protein interfaces have revealed that the free energy contributions of interfacial residues to binding are not uniformly distributed. A small set of interfacial residues is defined as hot spots accounting for the majority of the binding free energy [6]. At present, hot spots could be detected by alanine scanning mutagenesis [7]. Hot spots proved by experiments can be achieved from several databases such as the Alanine Scanning Energetics database (ASEdb) [8] and the binding interface database (BID) [9]. Unfortunately, experimental methods are time-consuming, labor-intensive and high economic costs. Therefore, developing reliable computational methods to identify hot spots is of significant current interest.

Computational methods for the prediction of hot spots fall into the following two categories. The one is energy-based method such as Robetta [10] and FOLDEF function [11], in which the energy contribution to the interface binding is computed for every residue. The other is knowledge-based method such as the KFC2 [12] that infers hot spots by using a model constructed on the features of training data. Various features were proposed to represent hot spots. Cho *et al.* predicted hot spots based on protein structure, sequence and molecular interaction, in which the interactions between hot spots were found to be π-related interactions [13]. Salam *et al.* used interaction engagement index, topographical index, sequence conservation index, 3D regional conservation index features and Bayesian network to predict hot spots [14]. Xia *et al.* combined protrusion-based features with solvent accessibility to predict hot spots [15]. Wang defined physicochemical features of a residue by itself and its interacting residues of the opposite protein chain, where RF was used to train the model and predict hot spots [16]. Other machine learning methods and biological significance analysis were also available for the study of hot spots [17,18].

Moreover, the small world network [19–21] was used to represent interactions of key residues in protein–protein interface in a few reports [22]. Graph theory and its applications in biology were introduced in [23,24], which has been widely used in proteomics [25–28], genomics [29–31] and drug discovery [32–34]. Here, graph theory was employed to represent interface. Several network topological features were then calculated to describe hot spots. These studies show that rational design of proper features is quite important in an inferring tool.

Hot spots are not isolated in protein–protein interaction interfaces [35,36]. "O-ring theory" indicates that hot spots are surrounded by energetically less important residues that occlude bulk solvent from hot spots [35]. "Double water exclusion" theory follows the "O-ring theory" and also reveals that hot spots themselves are water-free [37]. Bagley used the radial distributions of properties to describe the protein sites [38]. In the present study, several microenvironment features were designed for hot spots, which were capable of reflecting different physicochemical properties of amino acid residues in the vicinity of

* Corresponding authors. Tel.: +86 28 89005151; fax: +86 28 85412356.
E-mail addresses: liyizhou_415@163.com (Y. Li), liml@scu.edu.cn (M. Li).

0169-7439/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2013.11.010

hot spots. Then, a prediction model was constructed by combining these features with sequence features and structure features. The SVM was applied to construct prediction model based on an optimal feature set, in which 58 features were contained with 10 novel features. A satisfactory prediction result was obtained (ACC: 79.0%, MCC:0.470) for independent test set.

2. Materials and methods

2.1. Datasets

CSU program [39] was used to define atomic contacts between residues. If a residue has atomic contacts with residues that belong to any other chain in the complex, it is described as interfacial residue. The training data consists of 20 protein complexes taken from the ASEdb database [8] with 77 hot spots and 241 non-hot spots. 18 protein complexes with 38 hot spots and 86 non-hot spots from the BID database constitute the independent test set [9]. The same training and independent data set were used as in the study by Wang [16], where one repeated amino acid residue was removed. In order to avoid redundancy and homology bias, protein sequence in the training set and the independent test set were aligned by CD-HIT with sequence identity <35% [40]. The interfacial residue substitution with change of binding free energy ≥2.0 kcal/mol was defined as hot spots in the training set, while such positions labeled as "strong" in the BID database were identified as hot spots in the independent test set. The training and the independent data set are listed in the Supplementary material 1.

2.2. Classification method

In this study, SVM [41] with the radial basis function as kernel is the classifier, implemented in LIBSVM package which is available at website http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html#nuandone. Grid search was performed to search for the optimal parameters cost (C) and gamma (g). The search interval for cost (C) is $[2^0, 2^{12}]$ and for gamma (g) is $[2^{-15}, 2^5]$. The optimal value of C and g is 4096 and 0.065, respectively.

2.3. Evaluation of classifier

The training set was randomly divided into ten subsets of approximately equal size. Nine subsets are used as the training set for developing SVM model and the remaining as the testing set for evaluating it. This process is repeated ten times until every subset is used for testing once. All features are standardized before modeling.

The sensitivity (SE), specificity (SP), prediction accuracy (ACC) and Mathew's correlation coefficient (MCC) are the evaluation criterions. These measurement criterions are defined as follows

$$SE = \frac{TP}{TP + FN} \tag{1}$$

$$SP = \frac{TN}{TN + FP} \tag{2}$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$

Where TP, FP, TN, FN are the numbers of true-positive, false-positive, true-negative and false-negative, respectively.

3. Calculation

In order to well describe hot spots, a feature set combining conventional features, amino acid residues interaction network features and microenvironment features was constructed.

3.1. Network parameters

First, a residue–residue interaction network was constructed based on the distance between residues. Residues were defined to be connected in residue–residue interaction network if the distance between any two residues was smaller than the sum of their Van der Waals radii plus a threshold value of 0.5 Å [42]. Every protein complex was modeled as an undirected network graph, named bound network, in which vertex represents residue, edge represents residue–residue interaction. Similarly, undirected graph was used to represent the residue–residue interaction in single chain PDB structure, named unbound network. Igraph package (version 0.5.5-4) in R [43] was used to calculate the network parameters such as degree, closeness, betweenness and bonpow in both bound networks and unbound networks. The introduction and calculation methods of degree, closeness, betweenness and bonpow are in ref. [23].

3.2. Microenvironment features

Microenvironment features refer to different environment-based physicochemical features, which suggest the diversity of microenvironment between hot spots and non-hot spots. Based on the residue–residue interaction networks, the amino acid residues in the vicinity of hot spots and non-hot spots named neighboring residues were detected. The neighboring residues were labeled as neighboring residues in the first shell, the second shell and the third shell according to the distance between residue and central residue. An example is illustrated in Fig. 1.

Environment-based physicochemical features include environment-based hydrophobicity (ENPHO), environment-based hydrophilicity (ENPHI), environment-based isoelectric point (ENIP), environment-based mass (ENM), environment-based polarity (ENP), environment-based polarizability (ENPOZA) and environment-based propensity to be buried inside (ENPBBI). They are the sum value of hydrophobicity, hydrophilicity, isoelectric point, mass, polarity, polarizability and propensity to be buried inside of neighboring residues, respectively. They were calculated for the neighboring residues in each shell respectively, e.g. environment-based hydrophobicity (ENPHO) is specified as environment-based hydrophobicity in the first shell (ENPHO I), environment-based hydrophobicity in the second shell (ENPHO II) and environment-based hydrophobicity in the third shell (ENPHO III).

They are defined as

$$ENPHO\,\mathbf{I} = \sum_{1 \le i \le n} hydrophobicity(\mathrm{i}) \tag{5}$$

$$ENPHO\,\mathbf{II} = \sum_{1 \le j \le n} hydrophobicity(\mathrm{j}) \tag{6}$$

$$ENPHO\,\mathbf{III} = \sum_{1 \le m \le n} hydrophobicity(\mathrm{m}) \tag{7}$$

Where hydrophobicity (i) is the hydrophobicity of residue in the first shell, hydrophobicity (j) represents the hydrophobicity of residue in the second shell and hydrophobicity (m) describes the hydrophobicity of residue in the third shell. The definitions of ENPHI, ENIP, ENM, ENP, ENPOZA and ENPBBI are on the analogy of ENPHO. The influences of neighboring residues in different shells on hot spots are different. Taking the influence into consideration, different weights are specified for environment-based physicochemical features of neighboring