Computers in Biology and Medicine 42 (2012) 1053–1059

Contents lists available at SciVerse ScienceDirect

Computers in Biology and Medicine

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

Prediction of flavin mono-nucleotide binding sites using modified PSSM profile and ensemble support vector machine

Xia Wang [a], Gang Mi [b], Cuicui Wang [a], Yongqing Zhang [c], Juan Li [a], Yanzhi Guo [a,*], Xuemei Pu [a], Menglong Li [a,*]

[a] College of Chemistry, Sichuan University, Chengdu 610064, PR China
[b] College of Life Science, Sichuan University, Chengdu 610064, PR China
[c] College of Computer Science, Sichuan University, Chengdu 610064, PR China

ARTICLE INFO

Article history:
Received 18 March 2012
Accepted 13 August 2012

Keywords:
Flavin mono-nucleotide (FMN)
Binding site prediction
Position specific score matrix (PSSM)
Ensemble classifier
Support vector machine (SVM)

ABSTRACT

Flavin mono-nucleotide (FMN) closely evolves in many biological processes. In this study, a computational method was proposed to identify FMN binding sites based on amino acid sequences of proteins only. A modified Position Specific Score Matrix was used to characterize the local environmental sequence information, and a visible improvement of performance was obtained. Also, the ensemble SVM was applied to solve the imbalanced data problem. Additionally, an independent dataset was built to evaluate the practical performance of the method, and a satisfactory accuracy of 87.87% was achieved. It demonstrates that the method is effective in predicting FMN-binding sites.

1. Introduction

Previous studies show that the most of proteins could not perform their biological functions alone. Cofactors always play pivotal roles to complete protein functions. These protein cofactors are organic compounds or metal atoms which can associate with proteins and strengthen the stability of the compounds [1]. Knowing the interaction sites of proteins with cofactors or ligands would help researchers to understand the protein function and mechanism more specifically.

FMN (flavin mono-nucleotide) is a coenzyme of flavoprotein, which closely takes part in some biological processes such as energy production and cellular respiration (especially in electron transfer process) [2]. FMN is an electron carrier molecule that functions as a hydrogen acceptor. It exists in three oxidation states during catalytic cycle: oxidized (FMN), semiquinone (FMNH+) and hydroquinone (FMNH2). The strong oxidizability of FMN and the capability of transferring one or two electrons make FMN become a critical part in the electron transfer system. For example, flavodoxin is one of the flavoproteins and also an electron transfer protein involved in photosynthetic reactions. All the flavodoxins carry a molecule of flavin mono-nucleotide which confers redox properties to the protein [3]. Flavodoxins have been demonstrated to be critical in some pathogenic organisms and could be used as targets in drug design [4]. Therefore, as an important part of flavodoxins, recognition of FMN binding sites is greatly helpful in investigating the flavin recognition mechanism and the advanced researches about photosynthesis. Moreover, FMN plays a very important role in energy metabolism for several reduction–oxidation enzymes as a coenzyme. And it also exits in processes of metabolism in folate, vitamin B12, and other vitamins [5]. So identifying the FMN site is of great significance for designing relative inhibitors or antagonists [6].

However, identification of FMN binding sites by experimental methods is expensive and time consuming. Now computational methods have been developed. For example, Saito et al. reported an empirical approach to identify the nucleotide-binding sites [7], and Kelly et al. developed a machine learning approach to discriminate flavin adenine dinucleotide binding sites from nicotinamide adenine dinucleotide binding sites [8]. In this paper we specifically aimed at FMN binding sites, and a new method was developed to recognize the FMN binding sites using support vector machine. Instead of examining the structure of proteins, the FMN sites are identified based on amino acid sequences of proteins. The SVM models are constructed based on physicochemical properties and evolutionary information, respectively. For the evolutionary information, not original PSSM profile but the modified PSSM profile scaled by the sliding window and smoothing window was used to characterize the local environmental information of each FMN site. The data imbalanced problem commonly existing in the protein binding sites' researches was effectively resolved by the two most conventional

* Corresponding authors. Tel.: +86 28 85413330; fax: +86 28 85412356.
E-mail addresses: yzguo@scu.edu.cn (Y. Guo), liml@scu.edu.cn (M. Li).

0010-4825/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiomed.2012.08.005

approaches, the ensemble support vector machine classifier [9] and synthetic minority over-sampling technique (SMOTE) [10]. Therefore, these two approaches were introduced to deal with the imbalanced dataset problem in this paper, respectively.

2. Material and method

As to generate a really useful statistical predictor to identity the FMN binding sites, according to Chou's recent review [11], the following procedures was considered: (i) a valid dataset is constructed to train and test the predictor; (ii) an effective mathematical expression should be formulated that can truly reflect protein samples' intrinsic correlation; (iii) a powerful algorithm is developed to operate the prediction; (iv) the proper cross-validation tests is used to evaluate the performance of the predictor; (v) a user-friendly web-server of the predictor which is accessible to the public should be established.

2.1. Data mining

In this paper, 467 FMN binding protein IDs were extracted from SuperSite documentation [12] and the protein chains were downloaded from Protein Data Bank (PDB) [13]. To avoid redundancy and homology bias, proteins that have the mutual identity of over 25% with others in the dataset were removed by blastclust [14]. Then, LPC (Ligand Protein Contact) [15] was used to check the FMN binding sites whether on the chosen proteins or not and to confirm the position. Finally, 111 proteins including 2369 binding sites and 28,136 non-binding sites were retrieved. Thirty proteins containing 628 positive and 7712 negative samples, named as dataset P30, were randomly selected as an independent set to evaluate the practical performance of the method. The other 81 proteins containing 1741 positive and 20,424 negative samples, named as dataset P81, were used to construct the model. The PDB IDs of all the 111 proteins used in this paper have been listed in Supplementary Table S1.

2.2. Feature extraction

2.2.1. Evolutionary information

Biology is a natural science. All biological species have developed beginning from a very limited number of ancestral species, as well as protein sequences [16]. The evolution includes mutations, insertions and deletions of residues [17] and so on. With these changes accumulated for a long period of time, many similarities between initial and resultant amino acid sequences are gradually eliminated, but the corresponding proteins may still share many common attributes, such as having basically the same biological function and similar binding site. In order to incorporate this kind of evolution information into the feature vector, we employed the data derived from the position specific scoring matrix (PSSM) [18]. Currently, evolutionary information by PSSM has been widely used as the feature to characterize the functional sites of proteins [19,20].

In this paper, the PSSM profile was generated by PSI-BLAST. So, a protein of n amino acids was transformed into a $20\times n$ dimensions matrix. However, the original PSSM only expresses the evolutionary information of the binding site. We know that the surrounding residues of the binding site usually affect the binding process. So it is necessary to incorporate the surrounding residues' evolutionary information. In this paper, a modified PSSM by adding the sliding window and smoothing window was used to represent the local environmental information of the binding sites [21–23]. The details of sliding and smoothing window have been elaborated in [24]. After optimizing sliding window size and smoothing window size, the improved PSSM profile is constructed. For a protein of n residues, PSSM could be expressed as follows:

$$P_{PSSM}=\begin{bmatrix} A_{1\to1} & A_{1\to2} & A_{1\to3} & \cdots & A_{1\to20} \\ A_{2\to1} & A_{2\to2} & A_{2\to3} & \cdots & A_{2\to20} \\ A_{3\to1} & A_{3\to2} & A_{3\to3} & \cdots & A_{3\to20} \\ \vdots & \vdots & \vdots & A_{i\to j} & \vdots \\ A_{n\to1} & A_{n\to2} & A_{n\to3} & \cdots & A_{n\to20} \end{bmatrix} \quad (1)$$

where $A_{i\to j}$ means the evolutionary score of ith amino acid mutates to the j type amino acid. Here, the general form of PseAAC [11,25] was introduced to present the input feature more clearly and elegantly. So for a FMN binding site (A_i), the PSSM could be formulated as the following vector:

$$P_{PSSM(A_i)}=\begin{bmatrix} \Psi_1 & \Psi_2 & \cdots & \Psi_j & \cdots & \Psi_{20} \end{bmatrix}^T \quad (2)$$

where $\Psi_j=A_{i\to j}$ $(1\le j\le 20)$.

2.2.2. Physicochemical properties

For binding site prediction, researches using physicochemical properties of amino acids have been reported [26–28]. According to the work of Mishra [28], 544 physicochemical properties were downloaded from AAindex [29–32]. After removing the amino acid indices with the value "N/A", 531 physicochemical properties, named as PP531, were remaining to express the protein sequence information [33]. Each amino acid was substituted by value of 531 physicochemical properties, and then the matrix with size of $531\times n$ was generated, n was the length of a protein sequence. It could be expressed as follows:

$$P_{PP531}=\begin{bmatrix} A_{1,1} & A_{1,2} & A_{1,3} & \cdots & A_{1,531} \\ A_{2,1} & A_{2,2} & A_{2,3} & \cdots & A_{2,531} \\ A_{3,1} & A_{3,2} & A_{3,3} & \cdots & A_{3,531} \\ \vdots & \vdots & \vdots & A_{i,j} & \vdots \\ A_{n,1} & A_{n,2} & A_{n,3} & \cdots & A_{n,531} \end{bmatrix} \quad (3)$$

where $A_{i,j}$ is the value of jth physicochemical property of ith amino acid. Here, the general form of a FMN binding site (A_i) can be described as the following vector:

$$P_{PP531(A_i)}=\begin{bmatrix} \Psi_1 & \Psi_2 & \cdots & \Psi_j & \cdots & \Psi_{531} \end{bmatrix}^T \quad (4)$$

where $\Psi_j=A_{i,j}$ $(1\le j\le 531)$.

2.3. Ensemble support vector machine and synthetic minority over-sampling technique (SMOTE)

For the functional site prediction of proteins, the data imbalanced problem commonly exists. The large ratio between the negatives and positives would result in high prediction accuracy for the majority class but poor prediction accuracy for the minority class [34–36]. In the training set, there were 1741 binding sites and 20,424 non-binding sites. So two methods, ensemble SVM and SMOTE, were used to resolve the imbalanced problem.

SVM is a kind of machine learning approach based on structural risk minimization principle of statistical learning theory proposed by Vapnik [37]. The software, LIBSVM (version 3.0) was freely downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/ [38]. SVM has been widely applied in various biological areas, such as in predicting protein secondary structures [39,40] and classifying functions of proteins [41–44]. In this paper, the radial basis function (RBF) was chosen as the kernel function. The regularization parameter C and the kernel width parameter γ were optimized until an optimal SVM model was obtained.