Genomics 103 (2014) 292–297

Contents lists available at ScienceDirect

Genomics

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

Novel structure-driven features for accurate prediction of protein structural class

Liang Kong [a], Lichao Zhang [b,*]

[a] College of Mathematics and Information Technology, Hebei Normal University of Science and Technology, Qinhuangdao 066004, PR China
[b] College of Marine Life Science, Ocean University of China, Yushan Road, Qingdao 266003, PR China

ARTICLE INFO

Article history:
Received 9 November 2013
Accepted 7 April 2014
Available online 18 April 2014

Keywords:
Protein domains
Secondary protein structure
Protein sequence homology
Support vector machines

ABSTRACT

Prediction of protein structural class plays an important role in inferring tertiary structure and function of a protein. Extracting good representation from protein sequence is fundamental for this prediction task. In this paper, a novel computational method is proposed to predict protein structural class solely from the predicted secondary structure information. A total of 27 features rationally divided into 3 different groups are extracted to characterize general contents and spatial arrangements of the predicted secondary structural elements. Then, a multi-class nonlinear support vector machine classifier is used to implement prediction. Various prediction accuracies evaluated by the jackknife cross-validation test are reported on four widely-used low-homology benchmark datasets. Comparing with the state-of-the-art in protein structural class prediction, the proposed method achieves the highest overall accuracies on all the four datasets. The experimental results confirm that the proposed structure-driven features are very useful for accurate prediction of protein structural class.

1. Introduction

It is commonly believed that the biological function of a protein is essentially associated with its tertiary structure, which is determined by its amino acid sequence via the process of protein folding [1]. Since the structural class of a protein presents an intuitive description of its overall folding process, predicting the structural class of a protein is an important aspect in the identification of tertiary structure. For example, the knowledge of the protein structural class can significantly reduce the search space of possible conformations of the tertiary structure [2]. In addition, the knowledge of the protein structural class also plays an important role in protein function analysis, drug design and many other applications [3].

Based on the type, amount and arrangement of the secondary structural elements, a protein can be classified into several structural classes. Structural Classification of Proteins (SCOP) [4] is a manually annotated database and has been regarded as the most accurate classification of protein structural class. The current version of the SCOP database includes eleven structural classes, and approximately 90% of the protein domains belong to the four major classes (all-α, all-β, α/β, $\alpha + \beta$). With the rapid development of the genomics and proteomics, traditional experimental methods regarding complex and time-consuming apparently cannot cope with the demand for rapid classification of protein structural class. Therefore, it is essential to develop automated and accurate computational methods to help speed up this process.

Numerous computational methods have been developed for identifying protein structural class during the past three decades. These methods typically extract specific features to represent a protein and then perform classification by using different types of machine learning algorithms. One of the most abundant protein information available is its amino acid sequence, and various features have been proposed that turn a varying length of protein amino acid sequence into a fixed length feature vector. This fixed length vector is also known as sequence-driven features [5], such as amino acid composition (AAC) [6], pseudo amino acid composition (PseAA) [7], polypeptides composition [8], functional domain composition [9], and PSIBLAST profile [10]. One of the deficiencies in sequence-driven feature based methods is low accuracy for low-homology datasets, such as the widely-used 25PDB and 1189 datasets with sequence similarities lower than 25% and 40% respectively. Realizing the localization, there have been many attempts to use the secondary structure information to derive features to improve prediction accuracy for low-homology datasets [11,16,15,14,12,13,17]. Accordingly, we denote these features by structure-driven features. The available structure-driven features can be mainly categorized into 3 different types (1) content-related, (2) order-related, and (3) distance-related features. The contents of secondary structural elements and the normalized counts of secondary structural segments are widely used content-related features. Second order composition moment of secondary structural elements can be considered as order-related features. The maximal and average lengths of secondary structural segments are important distance-related features. Novel computational prediction methods with structure-driven features have achieved favorable overall accuracies on several low-homology benchmark datasets.

* Corresponding author.
E-mail address: zhanglichaoouc@126.com (L. Zhang).

http://dx.doi.org/10.1016/j.ygeno.2014.04.002
0888-7543/

However, the predictions for the α/β and $\alpha+\beta$ classes are still of low quality especially for the $\alpha+\beta$ class when compared with the predictions for the all-α and all-β classes. It has been a deficiency in the current protein structural class prediction methods.

In this paper, we focus on the challenging problem of identifying protein structural class solely from the information of the predicted secondary structure. The main contribution is extracting comprehensive structure-driven features especially for distance-related features to reflect general contents and spatial arrangements of the predicted secondary structural elements of a given protein sequence effectively. A 27-dimensional feature vector is selected based on a wrapper feature selection algorithm, and a multi-class nonlinear support vector machine (SVM) classifier is applied to predict protein structural class. The prediction performance is evaluated by a jackknife cross-validation test on four widely-used low-homology datasets (25PDB, 1189, 640, FC699). The experimental results show that the proposed feature vector results in significantly improved ability of the predictor to separate protein structural classes and our method provides better predictions when compared with modern and competing methods.

2. Materials and methods

According to recent research [18], to establish a useful statistical predictor for a protein system, the following procedures should be considered: (1) selection of valid benchmark datasets to train and test the predictor, (2) representation of protein samples to reflect their intrinsic correlation with the target to be predicted, (3) selection of the classification algorithm to operate prediction, and (4) selection of the cross-validation tests to objectively evaluate the anticipated accuracy of the predictor. Below, we will give concrete details about how to deal with these steps.

2.1. Datasets

Sequence homology has a significant impact on the prediction accuracy of protein structural class. Datasets with sequence homology ranging between 20%–40% tend to obtain more reliable and robust results [13]. In this paper, four low-homology datasets (25PDB, 1189, 640, FC699) are used to design and assess the proposed method. All the four datasets have been widely used as benchmark datasets in previous studies [11,16,15,14,12,13,17]. More details of these datasets are shown in Table 1.

2.2. Structure-driven features for protein representation

To be used effectively in our method, every amino acid residue in a protein sequence need to be first transformed into one of the following three secondary structural elements: H(helix), E(strand) and C(coil). The string of secondary structural elements is also known as protein secondary structure sequence (SSS) which can be obtained from protein structure prediction server PSIPRED [19]. In order to sufficiently reflect the general contents and spatial arrangements of the predicted secondary structural elements of a given protein sequence especially for α-helix and β-strand, another two simplified sequences are proposed based on SSS. One sequence is a segment sequence (SS), which is composed of helix segments and strand segments [13,15,17]. First, every H, E and C segment in SSS is respectively replaced by the individual letters H, E and C. Then, all of the letters C are removed and SS is obtained. The other sequence is obtained by removing all of the letters C from SSS, and the new sequence is denoted by E-H [16]. For example, given a secondary structure sequence SSS: EECEEECCEECCCCHHHHCCHHHCCCEEECCHHHCEE, the corresponding SS and E-H are EEEHHEHE and EEEEEEEHHHHHHEEEHHHEE, respectively. Based on the above three sequences, several structure-driven features are rationally constructed. The details of these features are given as follows:

1. The contents of secondary structure elements are the most widely-used structure-driven features, and have been proved significantly helpful in improving prediction accuracy of protein structural class [11]. They are formulated as:

$$p(x)=\frac{N(x)}{N_1},\ x\in\{\mathrm{H},\mathrm{E},\mathrm{C}\} \tag{1}$$

where $N(x)$ is the number of secondary structural element H, E or C in SSS; N_1 denotes the sequence length of SSS. This type of features has been extended to SS [13]. Here we further reuse them in E-H.
2. Biosequence patterns usually reflect some important functional or structural elements in biosequences such as repeated patterns [20]. In SSS, the 2-symbol repeated patterns are considered here, such as HH, EE, HE and EH. Hence, the contents of repeated patterns are proposed as follows:

$$p(xx)=\frac{N(xx)}{N_1},\ xx\in\{\mathrm{HH},\mathrm{EE},\mathrm{HE},\mathrm{EH}\} \tag{2}$$

where $N(xx)$ is the number of 2-symbol repeated patterns HH, EE, HE or EH. Here we extended these features to SS and E-H.
3. The normalized counts of α-helices and β-strands in SSS [16], another important structure-driven features, are given below:

$$\mathrm{NCountSeg}(x)=\frac{\mathrm{CountSeg}(x)}{N_1},\ x\in\{\mathrm{H},\mathrm{E}\} \tag{3}$$

where CountSeg(x) is the number of H or E segments. These features have been reused in E-H [16]. Here we further extended to SS.
The 25 features shown above characterize the contents of the predicted secondary structure from different aspects. They can be categorized into content-related structure-driven features. Below, we will further extract other types of structure-driven features such as order-related and distance-related features.
4. Second order composition moment of H, E and C are specially proposed to reflect the spatial arrangement of secondary structural elements in SSS [14], which are formulated as:

$$\mathrm{CMV}(x)=\frac{\sum_{j=1}^{N(x)} n_{x_j}}{N_1(N_1-1)},\ x\in\{\mathrm{H},\mathrm{E},\mathrm{C}\} \tag{4}$$

where n_{x_j} is the jth order (or position) of the corresponding secondary structural element in SSS. As these features reflect the order-related characteristic of secondary structure, they can be categorized into order-related structure-driven features. This type of features has been reused in E-H [16]. Here we further extended to SS.
5. Classification of protein structures is based on the contents and spatial arrangements of secondary structural elements especially for the α/β and $\alpha+\beta$ classes. While proteins in the α/β and $\alpha+\beta$ classes contain both α-helices and β-strands, they are usually separated in the α/β class but are usually interspersed in the $\alpha+\beta$ class. The distribution information of secondary structure segments will be helpful to inferring spatial arrangement of secondary structural elements. As distance information of secondary structural elements can reflect the distributions of α-helices and β-strands, we propose several distance-related structure-driven features. The length of

Table 1
The number of proteins belonging to different structural classes and homology level of the datasets.

Dataset	All-α	All-β	α/β	α + β	Total	Homology level
25PDB	443	443	346	441	1673	25%
1189	223	294	334	241	1092	40%
640	138	154	177	171	640	25%
FC699	130	269	377	82	858	40%