Artificial Intelligence in Medicine xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Protein fold recognition based on sparse representation based classification

Ke Yan[a], Yong Xu[a,*], Xiaozhao Fang[a], Chunhou Zheng[b], Bin Liu[a,*]

[a] School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong, 518055, China
[b] College of Electrical Engineering and Automation, Anhui University, Hefei, Anhui, 230039, China

ARTICLE INFO

Article history:
Received 31 December 2016
Received in revised form 6 March 2017
Accepted 7 March 2017

Keywords:
Protein fold recognition
Sparse representation based classification
Protein representation

ABSTRACT

Knowledge of protein fold type is critical for determining the protein structure and function. Because of its importance, several computational methods for fold recognition have been proposed. Most of them are based on well-known machine learning techniques, such as Support Vector Machines (SVMs), Artificial Neural Network (ANN), etc. Although these machine learning methods play a role in stimulating the development of this important area, new techniques are still needed to further improve the predictive performance for fold recognition. Sparse Representation based Classification (SRC) has been widely used in image processing, and shows better performance than other related machine learning methods. In this study, we apply the SRC to solve the protein fold recognition problem. Experimental results on a widely used benchmark dataset show that the proposed method is able to improve the performance of some basic classifiers and three state-of-the-art methods to feature selection, including autocross-covariance (ACC) fold, D-D, and Bi-gram. Finally, we propose a novel computational predictor called MF-SRC for fold recognition by combining these three features into the framework of SRC to achieve further performance improvement. Compared with other computational methods in this field on DD dataset, EDD dataset and TG dataset, the proposed method achieves stable performance by reducing the influence of the noise in the dataset. It is anticipated that the proposed predictor may become a useful high throughput tool for large-scale fold recognition or at least, play a complementary role to the existing predictors in this regard.

1. Introduction

Protein fold recognition is crucial for predicting protein structure and function, which is one of the most important tasks in bioinformatics [1,2]. Fold recognition refers to recognition of structural fold of a protein based on the given sequence information, which is important for protein tertiary structure identification [3]. Most of the computational methods are based on machine learning techniques for fold recognition. There are two important components in these methods, feature extraction and classification. In this regard, several computational predictors have been proposed considering both the two important components.

During the past decades, many powerful feature extraction methods have been proposed. The early methods are based on the primary sequence of amino acids [4]. Some traditional methods utilize the amino acid composition features, such as n-gram composition, dipeptide composition, etc [5,6]. Taguchi and Gromiha [7] propose the syntactical based features such as occurrence and composition to represent the proteins. However, researchers have found that proteins sharing similar structures may only have low sequence similarities. Therefore, these sequence-based methods cannot perform well when the sequence similarity is low. To overcome this disadvantage, several methods incorporate the evolutionary information or structural information into the feature extraction process. Dubchak et al., [8] propose a new feature vector based on the physicochemical properties and structural of amino acids describing the structures, which is associated with the local and global information about amino acid sequence. This method is further improved by some other related studies [9–12]. In order to incorporate the sequence-order information into the predictor, Shen et al., [13] propose a computational method called ensemble classifier, which is based on the pseudo-amino acid composition (PseAAC), physicochemical features, predicted secondary structure of protein. Some popular methods extract the evolution information by using the PSI-BLAST [14] tool and show a better performance on the protein fold recognition. Dong et al., [15]

[*] Corresponding authors.
E-mail addresses: yanke401@163.com (K. Yan), laterfall@hitsz.edu.cn (Y. Xu), xzhfang168@126.com (X. Fang), zhengch99@126.com (C. Zheng), bliu@insun.hit.edu.cn (B. Liu).

http://dx.doi.org/10.1016/j.artmed.2017.03.006
0933-3657/© 2017 Elsevier B.V. All rights reserved.

combine the autocross covariance and PSSM to transform the protein sequences into vectors with fixed-length. Recently, a Hidden Markov Model (HMM) combines the Multiple Sequence Alignment (MSA) to incorporate the evolution information. Remmert et al., [16] propose an effective tool HHblits to perform the remote protein detection. Lyons et al. combine the HHblits and dynamic programming to perform the protein fold recognition [17]. These methods are only based on protein sequence composition information, the physicochemical properties, and evolutionary information. Zakeri et al., [18] propose the functional information, which is effective to improve the performance. But functional domain information is usually extracted by experimental methods or by known structural information [17]. Most of the aforementioned features are complementary. Therefore, several methods combine multiple features into a predictors, and performance improvement can be observed [18–20]. For more information, please refer to a recent review paper on fold recognition [21].

Another important component of computational predictors for fold recognition is the classification algorithm. Some well-known machine learning techniques have been applied to this field, such as Support Vector Machine (SVM) [15,22–32], linear discriminant analysis (LDA) [33], the artificial neural network (ANN) [34–36], k-nearest neighbor (KNN) [37], Bayesian network [38], random forest [39–42], etc. These methods treat fold recognition as a multi-class classification task. Among these methods, the SVM-based method can achieve the state-of-the-art performance. SVM has been successfully applied to the classification and regression tasks, which calculates the maximum margin hyperplane among the training samples to minimize classification error. The kernel function is used to project the data from the original space into a new feature space. The SVM's performance depends on the kernel function, which quantifies the similarity between the protein sequences. The speed of convergence of SVM is faster than some methods, such as ANN [22]. The kernel function is connected with the discriminative features and the prior knowledge of the source data [43]. There are many kernel functions, such as gaussian kernel, polynomial kernel, radial basis function, etc. It is essential to select a suitable kernel in SVM. However, it is difficult to find a suitable kernel function in the applications. Recently, Zakeri et al., [18] combine the geometry means and different kernel matrices to improve the performance of the SVM-based method. Hu et al., [45] combine multi-view feature sets and ensemble classifier to solve the protein crystallization prediction problem.

Sparse Representation based Classification (SRC) [44,46] is a robust machine learning technique, which is stabile for feature selection classification tasks, and outperforms some traditional machine learning methods for some tasks in the field of image recognition and image processing, such as face recognition [47–49], texture classification [50], image denoising, image restoration, etc. Yu-An Huang et al., [51] propose a weighted sparse representation based classification (WSRC) method to solve the problem of protein-protein interactions (PPI). Dong-jun Yu et al. [52] combine the sparse representation technique with SVM, and improve the capability for predicting the binding residues. In these methods, a test sample is expressed by training samples of all classes via a linear representation. The coefficient matrix is sparse, and most nonzero elements in the matrix are essential for fold recognition. The substitution matrices obtained by training dataset and coefficients are used to predict the test sample directly. SRC uses the represent result to perform the final classification [44]. Motivated by its success, in this study, we apply the SRC to protein fold recognition. To improve the performance of the protein fold recognition, we combine some special features through the classifier SRC. Experimental results show that it can improve the predictive performance of some state-of-the-art methods.

2. Materials and methods

2.1. Dataset

Three datasets are used in the study to evaluate the performance of various computational predictors for fold recognition. Three datasets included DD dataset [13], EDD dataset [15] and TG dataset [53]. DD dataset contains 27folds which represent four major structure classes: α, β, α + β, and α/β. The training set has 311 sequences and the testing set contains 383 testing sequences whose sequence similarity is less than 35%. The sequences in the DD dataset were extracted from the Structural Classification of Protein (SCOP) version 1.63 [13].

The EDD dataset contains 3418 protein sequences which belong to the 27 different folds that essentially used in the DD dataset from SCOP (version 1.75), which has more sequences in the each fold [17]. The sequence identify between two proteins is no more than 40%. We use the EDD dataset to further evaluate our proposed method.

The third benchmark which is TG dataset, which contains 1612 protein sequences belonging to 30 different folds from SCOP (version 1.73) constructed by Taguchi and Gromiha [54]. The benchmark has the detailed information of the 30 different fold types is described in [53], and the sequence identify between two proteins is no more than 25%.

2.2. The processes of the competing methods

Three state-of-the-art methods, including ACC fold [15], Bi-gram [55], and D-D [22] are employed to validate whether the proposed SRC framework can improve their performance or not. All these methods are based on SVMs, and they employ different feature extraction methods. Among those methods, ACC fold and Bi-gram are profile-based methods, and the D-D is a sequence-based method. The detailed processes of these methods are shown in the followings.

2.2.1. ACC fold

ACC fold [15] applies the autocross-covariance transformation to extract the features from the PSSM. PSSM is a matrix with dimension of L*20, where L is primary sequence's length. Element $P_{i,j}\,(i \in [1, L], j \in [1, 20])$ of PSSM is interpreted as the probability of the j-th amino acid at the i-th position of protein sequence. The ACC fold transformation method is used to convert the PSSM matrix into a fixed length vector, with dimension of 400*LG (LG represents the distance between the amino acids in the PSSM) [56]. In this study, the value of LG is set as 4.

The process of ACC fold method is as follows. Firstly, the protein sequences' PSSM entries are calculated by the PSI-BLAST tool, which is directly associated with the evolutionary information. Secondly, the corresponding ACC matrix is obtained by the PSSM. The ACC matrix contains the two components: the AC (between the same property) and CC (between two different properties). AC is applied to measure the correlation of two same properties, which have the distances of LG along the sequence, and CC measures the correlation of two different properties between the distances of LG along the sequence [15]. The value of ACC is calculated by Eq. (1) and Eq. (2). Finally, the resulting feature vectors ACC are fed into SVM for classification.

$$AC\,(i, LG) = \Sigma_{j=1}^{L-LG} \left(P_{i,j} - \overline{P_i}\right) \left(P_{i,j+LG}\overline{P_i}\right) / \left(L - LG\right) \tag{1}$$

$$CC\,(i_1, i_2, \mathrm{LG}) = \Sigma_{j=1}^{L-LG} \left(P_{i_1,j} - \overline{P_{i_1}}\right) \left(P_{i_2,j+LG}\overline{P_{i_2}}\right) / \left(L - LG\right) \tag{2}$$

where $\overline{P_i} = \Sigma_{j=1}^{L} P_{i,j}/L,$, $\overline{P_i}$ is the average score of an amino acid i in the total protein sequence [15].