Journal of Theoretical Biology 365 (2015) 32–39

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

A two-layer classification framework for protein fold recognition

Reza Zohouri Aram, Nasrollah Moghadam Charkari *

Faculty of Electrical & Computer Engineering, University of Tarbiat Modares, Tehran, Iran

HIGHLIGHTS

- An individual method and a fusion method are proposed for protein fold recognition.
- The proposed methods are based on two-layer classification.
- The proposed methods improve the prediction accuracy by 2%–10% on a benchmark dataset.

ARTICLE INFO

Article history:
Received 2 June 2014
Received in revised form
9 September 2014
Accepted 19 September 2014
Available online 30 September 2014

Keywords:
Supervised learning
Ensemble classifiers
Fusion system

ABSTRACT

Protein fold recognition is one of the interesting studies in bioinformatic to predicting the tertiary structure of proteins. In this paper, an individual method and a fusion method are proposed for protein fold recognition. A Two Layer Classification Framework (TLCF) is proposed as individual method. This framework comprises of two layers: in the first layer, the structural class of protein is predicted. The classifier in this layer classifies the instances into four structural classes: all alpha, all beta, alpha/beta, and alpha+beta. Then, the classification results will be added as a new feature to further training and testing datasets. Using the results of the first layer, we employ another classifier for predicting 27 folding classes in the second layer. The results indicate that the proposed approach is very effective to improve the prediction accuracy where the measured values of MCC, specificity, and sensitivity are promising. TLCF* is proposed as a fusion method that exploits TLCF as a base model. The experimental results indicate that the proposed methods improve prediction accuracy by 2–10% on a benchmark dataset.

1. Introduction

The fold recognition problem is one of the fundamental problems in molecular biology. It is defined as 'obtain three-dimensional (3D) structure of proteins from their sequences without depending on sequence similarities' (Ding and Dubchak, 2001). Identification of protein tertiary structure is of great importance since the main function of protein is determined by tertiary structure (Shenoy and Jayaram, 2010). Moreover, it plays an essential role in the design of new drugs and therapies. Nowadays, there is an immense gap between the known protein sequence and confirmed protein tertiary structure (Lee et al., 2009). Thus, introducing some efficient computational methods to predicting 3D structures from sequences might be considered as a way to solve the mentioned problem. Computational methods have been used for predicting 3D structures for more than four decades.

There are two popular classes of computational methods for predicting the tertiary structure: (a) Template-Based Methods (TBM) (b) Ab initio methods (Lee et al., 2009). For identifying the tertiary structure of a given sequence, Template-based methods suggest the use of known three-dimensional structures in the Protein Data Bank (PDB) (Kouranov et al., 2006) as a template (Lee et al., 2009). On the other hand, Ab initio not only use any templates but also builds the 3D models from scratch (Lee et al., 2009). Ab initio modeling predicts protein structures using either physical and chemical principles or other techniques (Dong et al., 2007). However, they have a high computational complexity. One of the most common template-based approaches to predict the 3D structure is the machine learning methods. In this paper, we focus on machine learning methods as well.

In this paper, two methods are proposed for protein fold recognition: an individual method and a fusion method. A Two Layer Classification Framework (TLCF) is proposed as individual method. This framework is composed of two layers. In the first layer, we attempt to predict the structural class of protein. The

* Corresponding author. Tel.: +98 2182883301; fax: +98 2182884325.
E-mail addresses: r.zohori@modares.ac.ir (R.Z. Aram), charkari@modares.ac.ir (N.M. Charkari).

http://dx.doi.org/10.1016/j.jtbi.2014.09.032
0022-5193/

classifier in the first layer classifies the instances into four structural classes: all alpha, all beta, alpha/beta, and alpha+beta. Then, we add the classification results of the first layer as a new feature to train and test datasets. Using the results of the first layer, another classifier is employed for predicting 27 folding classes in the second layer. To improve the prediction accuracy, TLCF* is proposed by introducing novel fusion system. Generally, fusion system is the combination of individual classifiers and operates on classifiers outputs. The outputs of all individual classifiers will combine using different techniques such as voting rule.

As discussed in a comprehensive review article (Chou, 2011) and followed up by a series of recent publications (Liu et al., 2014; Qiu et al., 2014; Guo et al., 2014; Ding et al., 2014; Xu et al., 2014), in order to establish a really useful statistical predictor for a biological system, we need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the biological samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform evaluation method to objectively evaluate the anticipated accuracy of the predictor; (v) efforts to establish a user-friendly web-server for the predictor that be accessible to the public. Below, after introducing related work, we describe how to deal with these steps one-by-one.

2. Related work

Support vector machines (SVMs) and neural networks (NNs) are two interesting methods for protein fold recognition. Ding and Dubchak (2001) proposed the Unique One- versus-Others (uOvO) and the all-versus-all methods. They employed SVM and three layers feedforward NNs as base classifiers. Yang et al. (2008) applied the three types of classifiers: k nearest neighbors, class center and nearest neighbor, and probabilistic neural networks. Then the results of the mentioned classifiers were combined using an ensemble voting system.

Ensemble classifiers are frequently used in protein fold recognition. Ensemble method is a supervised learning algorithm that uses multiple classifiers to obtain proper prediction accuracy. Guo and Gao (2008) presented two-layer ensemble classifier. In the first layer, a potential class index for every query protein in the 27-folds is identified. According to this result, a 27-dimension vector is generated in the second layer. Finally, genetic algorithm is adopted to obtain weights for the outputs of the second layer to get the final result. Kavousi et al. (2012) proposed the method upon which an unknown query protein is assigned to a hyperfold rather than a single fold. Each hyper_fold is a set of interlaced folds with a centroid fold and Dempster rule has been used to combine the results. Nanni (2006a, 2006b) proposed ensemble of classifiers, and applied it to protein fold recognition. Another ensemble method applied by Hashemi et al. for protein fold pattern recognition (Hashemi et al., 2009). They could improve the prediction accuracy by using Bayesian Ensemble of RBFN. In Bayesian Ensemble, the normalized confusion matrix of each base classifier (i.e. RBFN) is used to ensemble the outputs. Chmielnicki and Stapor (2012) suggested a hybrid discriminative/generative approach. Accordingly, they utilized RDA (as a generative classifier) and SVM (as a discriminative classifier). Using the results of RDA, SVM classifies the proteins.

Abbasi et al. (2013) made use of an intelligent hyper framework. The existing components in the framework are used to classify proteins under fuzzy conditions. A novel approach named PFP-FunDSeqE is proposed by Shen and Chou (2009). Accordingly, the functional domain information and the sequential evolution information of proteins are combined through a fusion ensemble classifier. PFP-FunDSeqE have improved the prediction rate by fusing five features extracted by Ding and Dubchak and four Pseudo-amino Acid Composition extracted by Chou (2005). Shen and Chou (2006) proposed a different ensemble classifier named PFP-Pred. This ensemble classifier uses Evidence-Theoretic K-Nearest Neighbor (ET-KNN) as base classifier. The evidence-theoretic k-nearest neighbor is a classification method based on the Dempster–Shafer theory (see (Shen and Chou, 2005) for more details). The ET-KNN has been carrying out separately on nine feature sets and nine outputs generated. Finally, the outputs were combined using weighted voting. Jazebi et al. (2009) employed a fusion method for fold pattern recognition. They used the Probabilistic Neural Network (PNN) as base classifier in the fusion method. The fusion method has combined the classification results on six different feature sets by using the weighted voting approach. Leon et al. (2009) presented a taxonometric approach based on different classification techniques such as k-nearest neighbor (K-NN), decision trees, Naive Bayes and neural networks (NNs). In their experiments, they found that the neural network and K-NN have better performance than other techniques. A Multi-Objective Feature Analysis algorithm is proposed in Shi et al. (2004). The objective of this algorithm is to simultaneously selecting the effective features, improving the accuracy and providing bias information of test and train data. To achieve this objective, authors employed an extended wrapper method for feature selection and used SVM for classification task. However, the method suffers from high complexity time.

Huang et al. (2003) proposed a Hierarchical Learning Architecture (HLA) that works in two levels. In the first level, four structural classes (all alpha, all beta, alpha/beta, and alpha+beta) are predicted while in the next level, protein features are classified into 27 folds. The main weakness of HLA is that if the classifier in level 1 makes any mistake, then the classifiers in level 2 will not be able to recover the mistakes. The proposed method in this paper uses two levels of classification like HLA (Huang et al., 2003). However, it is different from various aspects that are discussed in relevant section.

3. The dataset and feature vectors

3.1. Training and test datasets

To compare our method with previous works, the dataset that were introduced in (Ding and Dubchak, 2001) has been used. It contains 313 instances in the training set and 385 instances in the testing set. In training set, two proteins have no more than 35% of the sequence identity for the aligned subsequences longer than 80 residues. The testing dataset of 385 proteins is composed of protein sequences of less than 40% identity with each other. These datasets contain the 27 most populated folds represented by seven or more proteins and corresponding to four major structural classes: α, β, α/β and $\alpha+\beta$. The folds in the dataset and the corresponding number of proteins in two datasets are shown in Table 1.

3.2. Feature vectors

Ding and Dubchak represented the samples based on primary protein sequences. Six features were extracted independently from protein sequences: Amino acids composition (C), predicted secondary structure (S), hydrophobicity (H), normalized van der Waals volume (V), polarity (P), and polarizability (Z). C is the sequence composition of 20 types of amino acids (see (Ding and Dubchak, 2001) for more details). C has the dimensionality of 20