Computational Biology and Chemistry 59 (2015) 95–100

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Research article

A highly accurate protein structural class prediction approach using auto cross covariance transformation and recursive feature elimination

Xiaowei Li[a], Taigang Liu[b,*], Peiying Tao[a], Chunhua Wang[b], Lanming Chen[a,**]

[a] College of Food Science & Technology, Shanghai Ocean University, Shanghai 201306, China
[b] College of Information Technology, Shanghai Ocean University, Shanghai 201306, China

ARTICLE INFO

Article history:
Received 14 November 2014
Received in revised form 30 August 2015
Accepted 30 August 2015
Available online 2 September 2015

Keywords:
Low-similarity
Position-specific score matrix
Auto cross covariance
Support vector machine
Recursive feature elimination

ABSTRACT

Structural class characterizes the overall folding type of a protein or its domain. Many methods have been proposed to improve the prediction accuracy of protein structural class in recent years, but it is still a challenge for the low-similarity sequences. In this study, we introduce a feature extraction technique based on auto cross covariance (ACC) transformation of position-specific score matrix (PSSM) to represent a protein sequence. Then support vector machine-recursive feature elimination (SVM-RFE) is adopted to select top *K* features according to their importance and these features are input to a support vector machine (SVM) to conduct the prediction. Performance evaluation of the proposed method is performed using the jackknife test on three low-similarity datasets, i.e., D640, 1189 and 25PDB. By means of this method, the overall accuracies of 97.2%, 96.2%, and 93.3% are achieved on these three datasets, which are higher than those of most existing methods. This suggests that the proposed method could serve as a very cost-effective tool for predicting protein structural class especially for low-similarity datasets.

1. Introduction

Knowledge of protein structural class plays an important role in the prediction of secondary structure and function analysis from the amino acid sequence information (Anand et al., 2008). Nowadays, the most frequently used classifications of protein structural classes can be found in the Structural Classifications of Protein (SCOP) database (Murzin et al., 1995). There are 110,800 protein domains with known structural class in SCOP, and about 90% of them belong to the four major classes: all-α, all-β, α/β and α+β. With the rapid development of genomics and proteomics, the newly discovered protein sequences are growing exponentially, which has made a large gap between the number of sequence-known and structure-known proteins. The current experimental determination of protein structure is costly and time-consuming and thus cannot cope with the demand for rapid classification. Hence there exists a great challenge to develop reliable and accurate computational methods to determine protein structural class.

As a typical pattern recognition problem, computational methods for predicting protein structural class generally consist of two main steps: protein feature representation and algorithm design for classification. For the first step, previous studies have shown that sequence features can be represented in many different ways, including amino acids composition (Chou, 1999 Zhou, 1998), pseudo amino acid (PseAA) composition (Chou, 2001; Li et al., 2009), polypeptide composition (Luo et al., 2002; Sun and Huang, 2006), functional domain composition (Chou and Cai, 2004), amino acid sequence reverse encoding (Yang et al., 2009), position-specific score matrix (PSSM) (Chen et al., 2008; Ding et al., 2014; Liu et al., 2010; Liu et al., 2012), and predicted secondary structure information (Dai et al., 2013; Dehzangi et al., 2014; Kong et al., 2014; Kurgan et al., 2008; Mizianty and Kurgan, 2009; Yang et al., 2010). It is worth mentioning that through quantitative analysis, Dai and his coauthors verify that exploring the position information of predicted secondary structural elements is a promising way to improve the abilities of protein structural class prediction (Dai et al., 2013). For the later step, a wide range of classification algorithms have been used to perform the prediction, such as neural network (Cai and Zhou, 2000), support vector machine (SVM) (Cai et al., 2001; Kong and Zhang, 2014; Li et al., 2008; Nanni et al., 2014), fuzzy clustering (Shen et al., 2005), fuzzy *k*-nearest neighbor (Zhang et al., 2008; Zheng et al., 2010), Bayesian classification (Wang and Yuan, 2000), Logistic regression

* Corresponding author.
** Corresponding author.
E-mail address: tgliu@shou.edu.cn (T. Liu).

http://dx.doi.org/10.1016/j.compbiolchem.2015.08.012
1476-9271/

(Jahandideh et al., 2007; Kurgan and Chen, 2007; Kurgan and Homaeian, 2006), rough sets (Cao et al., 2006), and classifier fusion techniques (Cai et al., 2006; Chen et al., 2006; Chen et al., 2009; Dehzangi et al., 2013). Early methods can achieve prediction accuracies more than 90% when tested on datasets with high sequence identities. However, they perform poorly on low-similarity datasets, with accuracies between 50% and 70% (Kurgan et al., 2008). To solve this problem, by incorporating various features such as PSSM, predicted secondary structure and physical-chemical properties, several methods have been proposed to improve prediction accuracies on low-similarity datasets (Dehzangi et al., 2013; Dehzangi et al., 2014; Kurgan et al., 2008; Liu et al., 2010; Wang et al., 2015; Yang et al., 2010). Nevertheless, most studies which rely only on predicted secondary structure to enhance the accuracy could not reach too far better results than 80% (Kurgan et al., 2008; Yang et al., 2010). This may be due to limited prediction accuracy (about 80%) of protein secondary structure by PSIPRED (Jones, 1999). On the other hand, since the performance of PSIPRED algorithm relies mainly on PSSM, PSSM profile provides more important and original discriminatory information for protein structural class prediction. In our previous study, we extracted auto-covariance variables from the PSSM profile and also obtained favorable prediction accuracy when the predicted secondary structure was not utilized (Liu et al., 2012).

In this study, in order to further improve the prediction accuracy of protein structural class, we extract both auto-covariance variables and cross-covariance variables from the PSSM profile by auto cross covariance (ACC) transformation. The flowchart of the proposed method is depicted in Fig. 1, which presents the pipeline that goes from the query sequence to the final output as well as intermediate steps. Firstly, the PSSM profile generated by PSI-BLAST program (Altschul et al., 1997) is transformed into a fixed-length feature vector by ACC transformation. Secondly, support vector machine-recursive feature elimination (SVM-RFE) is applied for feature selection and reduced vectors are input to an SVM classifier to perform the prediction. Finally, results by the jackknife test on three widely used benchmark datasets suggest that the proposed method yields substantial improvements in prediction accuracies compared with most published results.

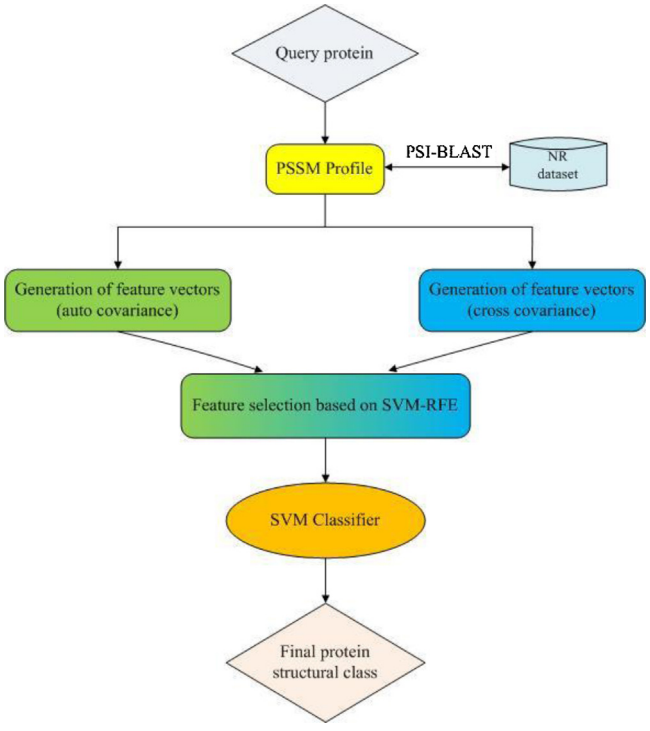

Fig. 1. The flowchart of the proposed method.

2. Materials and methods

2.1. Datasets

In order to evaluate the prediction accuracy of the proposed method and compare it with those of existing methods, three widely used datasets are adopted in our work: D640 (Chen et al., 2008), 1189 (Wang and Yuan, 2000) and 25PDB (Kurgan and Homaeian, 2006), with sequence similarity lower than 25%, 40% and 25% respectively. The D640 dataset contains 640 protein sequences, which consists of 138 all-α proteins, 154 all-β proteins, 177 α/β proteins and 171 α+β proteins. The 1189 dataset includes 1092 protein domains of which 223 belong to the all-α class, 294 belong to the all-β class, 334 belong to the α/β class, and 241 belong to the α+β class. The 25PDB dataset includes 1673 protein domains of which 443 are all-α proteins, 443 are all-β proteins, 346 are α/β proteins and 441 are α + β proteins.

2.2. Protein sequence representation

To develop a powerful predictor for a protein system, one of the key tasks is to formulate the protein samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted (Chou, 2011). In this section, we combine PSSM and ACC transformation to represent a protein sequence by a fixed-length feature vector.

PSSM which can represent the evolutionary information of each sequence is generated by running PSI-BLAST (Altschul et al., 1997) program against the NCBI's non-redundant (NR) dataset (ftp://ncbi.nih.gov/blast/db/nr). The parameters are defaulted except that parameter h and j are set to 0.001 and 3 respectively. The PSSM profile for each protein sequence is an $L \times 20$ matrix, where L is the length of the corresponding sequence. The (i,j) th element in the matrix reflects the score of amino acid in the ith position of the query sequence being mutated to amino acid type j during the evolution process.

For convenience, the PSSM of the query sequence S can be described as follows:

$$P = (P_1, P_2, ..., P_{20})$$

where

$$P_j = \left(p_{1,j}, p_{2,j}, ..., p_{L,j}\right)^T (j = 1, 2, ..., 20),$$

L is the length of the query sequence S, and T is the transpose operator.

Next, we adopt ACC transformation to convert the PSSMs of different lengths into uniform equal-length vectors. ACC is a powerful protein sequence analysis method developed by Wold and his colleagues (Wold et al., 1993), which has been widely applied to the field of bioinformatics such as protein family classification and protein interaction prediction (Dong et al., 2009; Guo et al., 2006; Guo et al., 2008; Liu et al., 2011). Since each residue has many physical-chemical properties such as hydrophobicity, hydrophilicity, etc., ACC can measure the correlation of two properties (or the same property) along the protein sequence. Two kinds of variables, i.e., auto-covariance $A(j,g)$ and cross-covariance $C(j,k,g)$, are calculated according to the following two equations: