ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 12, No. 1, 2017, *pp*. 026-032

Application of locality sensitive discriminant analysis to predict protein fold pattern

Chunming Xu[1]
1. School of Mathematics and Statistical, Yancheng Teachers University, Yancheng, 224002,PR China, E-mail: ycxcm@126.com.
(Received October 07, 2016, accepted December 17, 2016)

Abstract. Predicting protein-folding patterns is a challenge due to the complex structure of proteins. Many sequence encoding schemes have been proposed to extract the features of pro-tein sequences, and these features are often fused to form a new combined feature set so that it can contain various useful information. However, there usually has redundant information in the combined features. In this paper, a novel approach, LSDA-SVM, is proposed to predict pro-tein fold pattern. Firstly, protein samples are represented by the pseudo amino acid composition (PseAAC), pair wise feature (PF) and the others five types of protein sequence information, and these features are further combined to form a new feature set. Secondly, the locality sensitive discriminant analysis (LSDA) is employed to extract the more discriminant features. Finally, the support vector machine (SVM) is employed to classify the protein sequences. Experimental results demonstrate the effectiveness of the proposed algorithm.

Keywords: protein fold prediction; locality sensitive discriminant analysis (LSDA); support vector machine (SVM); feature extraction.

1. Introduction

Nowadays, with the rapid increasing number of protein sequences, it is urgent to find effective and efficient computational algorithms to find useful information behind these biological sequence data sets. Among these, determination of protein structure from its primary sequence plays a key role because it can help to understand its functions [1, 2]. Moreover, recent research have shown that the knowledge of protein structural class provides useful information towards the development of new drugs [3], cancer research [4], and human immunodeficiency virus therapies [5]. Despite many efforts have been down to protein fold prediction, it is still a hard problem.

Protein fold recognition means the prediction of a protein’s three-dimensional structure based on its amino acid sequence information. The protein sequences usually contain different number of amino acid residues and they are irregular. As a result, the first step of protein fold prediction is to encode the protein sequences such that they can be well classified by a favorable classifier. Till now, various sequence encoding schemes have been applied to represent the features of protein sequences. Representatives of sequence encoding schemes include amino acid composition (AAC) [6], pseudo amino acid composition (PseAAC) [7], polypeptide composition [8], functional domain composition [9] and amino acid sequence reverse encoding[10] et al.

The AAC feature is one of the most fundamental types of information for protein function prediction, and it has been successfully used to encode protein in many applications, such as protein subcellular localization, membrane types and predicting signal peptides. Although AAC is a very effective feature set that have achieved very promising performance in many applications, it neglect the sequence order information. In order to overcome this drawback, pseudo-amino acid compositions (PseAAC) was proposed to represent the sample of a protein in a more effective way. In [16], the authors used the pairwise frequency information about the amino acids to extract the features of the protein sequences. In detail, they considered two types of pairwise frequency information, i.e., the pairwise frequencies of amino acids separated by exactly one residue (PF1) and the pairwise frequencies of adjacent amino acids (PF2). By the way, we can get feature vectors of dimension 400 for both PF1 and PF2. Then we can get a total feature vectors of dimension 800 which is called PF by Yang [11].

In fact, different feature vectors contain different information about the protein sequences, and they are usually fused to form a new combined feature set. As the combined feature set contains more information than single feature set, it is expected to have good discriminating power. However, although the combined

feature set is very effective in solve many biological sequence classification problems, it usually has redundant information. In this study, a novel approach, LSDA-SVM, is introduced to predict protein old pattern. The proposed method is divided into three different stages. Firstly, protein samples are represented by the pseudo amino acid composition (PseAAC), pair wise feature (PF) and the others five types of protein sequence information. Secondly, the locality sensitive discriminant analysis (LSDA)[12] is further employed to extract the more effective discriminant features from the original high-dimensional vectors. Finally, the support vector machine (SVM) is employed to classify the protein sequences. Some advantage of the proposed algorithm are: (1) both manifold information and supervised information of the training samples can be used to guided the produce of feature extraction, so the new feature set is more suitable for protein classification, and the recognition performance can be improved; (2) the redundant information resided in the features can be removed; (3) the dimension of the features are reduced and the classification is performed in a much lower dimensional vector space so that the classification time is accordingly reduced. We demonstrate the usefulness of our approach on the D-B data set and the experiment results show that the proposed algorithm can enhance the recognition accuracies.

2. Materials and methods

2.1. Dataset

We use the D-B dataset constructed by Ding [13], which has 698 proteins. There are 313 protein sequences in the training dataset where two proteins have no more than 35% of the sequence identi-ty for aligned subsequences. On the other hand, the test dataset consists of 385 SCOP sequences hav-ing less than 40% identity with each other. The proteins in both the training and test sets are catego-rized into the following 27-fold types: 1) globin-like, 2) cytochrome c, 3) DNA-binding 3-helical bundle,4) 4-helical up-and-down bundle, 5) 4-helical cytokines, 6) EF-hand, 7) immunoglobulin-like, 8) cupre-doxins, 9) viral coat and capsid proteins, 10) concanavalin A-like lectin/glucanases, 11) SH3-like bar-rel, 12) oligonucleotide/oligosaccharide-binding-fold, 13) β-trefoil, 14) trypsin-like serine proteases, 15)lipocalins, 16) triosephosphate isomerase barrel, 17) flavin adenine dinucleotide (also nicotinamide adenine dinucleotide-binding motif), 18) flavodoxin-like, 19) nicotinamide adenine dinucleotide phosphate-bindingRossmann fold, 20) P-loop, 21) thioredoxin-like, 22) ribonuclease H-like motif, 23) hydrolases, 24) periplasmic binding protein-like, 25) β-grasp, 26) ferredoxin-like, and 27) small inhibitors, toxins, and lectins.These fold types can also be coarse classified into four classes, i.e., types 1-6 belong to the α structuralclass, types 7-15 to the β class, types 16-24 to the α/β class, and types 25-27 to the α + β class.

2.2. Sequence encoding methods

2.2.1. PseAAC

Pseudo-amino acid compositions (PseAAC) was first proposed by Chou [7] to represent the protein sequence. The PseAAC can not only reflect the amino acid composition of the protein but also consider the sequence-order information. To be specially, the protein P with L amino acid residues

$$S_1 S_2 S_3 \cdots S_L \tag{1}$$

where Si represents the residue at the sequence position *i*, can be represented as

$$F_{PseAAC} = [p_1, p_2, ..., p_{20}, p_{20+1}, ... p_{20+\Lambda}] (\Lambda < N) \tag{2}$$

where the 20 + Λ components is represented as follows:

$$p_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\Lambda} \tau_i}, 1 \le k \le 20 \\ \dfrac{w\theta_{k-20}}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\Lambda} \tau_j}, 20+1 \le k \le 20+\Lambda \end{cases}$$

where f_k is the occurrence frequencies of 20 amino acids in sequence and τ_j is the *j*-tier sequence correlation factor which reflect the effect of sequence order. The weight factor w is used to control the complexity of the sequence order effect and is set at 0.05 as in Ref. [7]. In this study, the patermater Λ is set to be 10 so that the PseAAC is corresponding to a 30-D (Dimensionality) vector.