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

Published by World Academic Press, World Academic Union

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 fold 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.

2.2.2. PF

In [16], the pairwise frequency information about the amino acids was used as the fold discriminative features. The authors proposed two types of features, i.e., one residue (PF1) and the pairwise frequencies of adjacent amino acids (PF2). The PF1 feature set is calculated using the occurrence of all possible pairs of amino acids separated by one residue in the protein, while the PF2 feature set is formed by considering only adjacent pairs. The formed feature set is also called PF [11]. In this way, we get feature vectors of dimension 800 for PF.

2.2.3. Some other feature sets

We also considered some other popularly feature sets such as predicted secondary structure (S), hydrophobicity (H), normalized van der Waals volume (V), polarity (P) and polarizability (Z) [13]. All the features have dimensionality 21.

In present study, we have introduced several sequence encoding schemes, which can reflect various information of the protein sequences. When we combine them, we can get a high dimensional feature set. However, there usually has redundant information in such a high dimensional vector space. In order to solve this problem, we will introduce a nonlinear and supervised feature extraction method, i.e. locality sensitive discriminant analysis (LSDA) to further extract more effective discriminant fold features.

2.3. Locality sensitive discriminant analysis (LSDA)

Locality sensitive discriminant analysis (LSDA)[12] aims to find a good data representation so that nearby objects with the same labels in the input space also are close to each other in the new representation; while nearby objects with different labels in the input space should be far apart. LSDA can effective use of both the label information and the local manifold structure information of labeled samples to aid the dimensionality reduction process.

Given l data points $x_1, x_2, ..., x_l \in R_p$ that are distribduted on a underlying submanifold. Let *l*(*xi*) be the class label of x_i, let the k nearest neighbors of x_i be $N(x_i) = \{x_i^1, x_i^2, ..., x_i^k\}$. By the label information, the set $N(x_i)$ can be further split into two subsets, $N_b(x_i)$ and $N_w(x_i)$.$N_w(x_i)$ contains the neighbors having the same label with x_i, while $N_b(x_i)$ contains the neighbors sharing different labels. Specifically,

$$N_w(x_i) = \{x_i^j \,|\, l(x_i^j) = l(\chi_i), 1 \le j \le k\} \tag{3}$$

$$N_b(x_i) = \{x_i^j \,|\, l(u_i^j) \ne l(\chi_i), 1 \le j \le k\} \tag{4}$$

Define the weight matrices W_b and W_w respectively as follows:

$$W_{b,i,j} = \begin{cases} 1, x_i \in N_b(x_j) \quad or \quad x_j \in N_x(u_i) \\ 0, else. \end{cases}$$

$$W_{w,ij} = \begin{cases} 1, x_i \in N_w(x_j) \quad or \quad x_j \in N_w(x_i) \\ 0, else. \end{cases}$$

The main idea of LSDA is to maximize $\sum (y_i - y_j)^2 W_{b,ij}$ while at the same time minimize $\sum (y_i - y_j)^2 W_{w,ij}$, where $y_i = A^T x_i$. Because $\sum (y_i - y_j)^2 W_{b,ij}$ = $\sum (A^T x_i - A^T x_j)^2 W_{b,ij}$ = $A^T X(D_b - W_b)X^T A$, where X = [*x*1, *x*2; ..., *x*l] and D_b is a diagonal matrix with entries $D_{b,ii} = \sum_j W_{b,ij}$. On the other hand, $\sum (y_i - y_j)^2 W_{w,ij}$ = $\sum (A^T x_i - A^T x_j)^2 W_{w,ij}$ = $A^T X(D_w - W_w)X^T A$. Also, D_w is a diagonal matrix with entries $D_{w,ii} = \sum {}_j W_{w,ij.}$ Note that LSDA exploits not only the discriminant structure information but also the manifold information of the samples. As a result, LSDA will apart the data samples from different classes at each local area well.

Formally, the objective function of LSDA can be written as follows:

$$J(A) = \arg\max_A \alpha A^T(X\ {}_b D - ({}_b W\ {}^T X) - A - \alpha\ (1^T\ A)\ U_W - D({}_w\ W^T \tag{5}$$

where α is a positive parameter and 0 < α < 1.

Define $L_b = D_b - W_b$ and $L_w = D_w - W_w$, then *J*(*A*) can be rewritten as:

$$J(A) = \arg\max_{A} A^T X(\alpha L_b - (1-\alpha)L_w)X^T A \tag{6}$$

By means of Lagrangian multiplier method, the coefficient matrix A can be constructed by the eigenvectors of $X(\alpha L_b - (1-\alpha)L_w)X^T$ associated with the first d largest eigenvalues $a_1, a_2, ..., a_d, i, e.$ A can be constructed as $A = (a_1, a_2, ..., a_d)$. Therefore the new data representation of xi can be expressed as:

$$y_i = A^T x_i \tag{7}$$

3. Evaluation criteria

Three evaluation criterias are used in this paper to assessment of the prediction system. The first is the overall prediction accuracy, which can be expressed as

$$Q = \frac{c}{n} \tag{8}$$

where c is the number of query sequences whose folds have been correctly recognized and n is the total number of sequences in the test data set. In addition, the prediction accuracies in different folds are also employed to evaluate the proposed method. Suppose there are *ni* query protein sequences correctly recognized as belonging to fold *i*, and *ci* is the number of query sequences whose folds have been correctly recognized belonging to fold i, then $Q_i = \frac{c_i}{n_i}$.

The last one is the Matthew's correlation coefficient (MCC). The formula for MCC measurements are given below:

$$MCC = \frac{(TP)(TN) - (FP)(FN)}{TN + FP} \tag{9}$$

where *TP*, *TN*, *FP* and *FN* are the number of true positives, true negatives, false positives and false negatives.

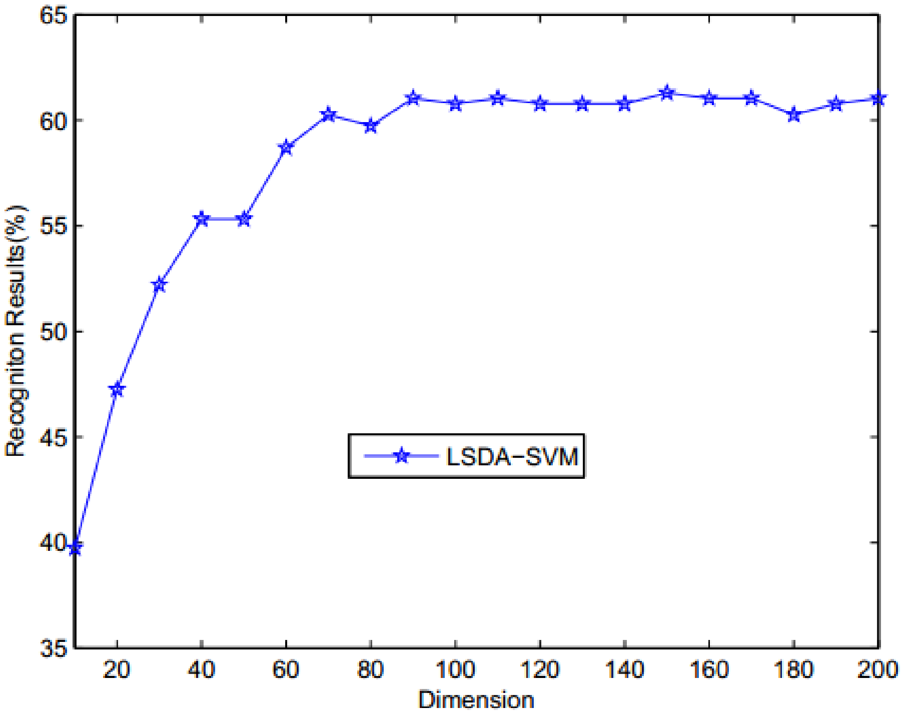

Figure 1. The performance of LSDA-SVM

4. Results and discussion

The benchmark D-B dataset constructed by Ding [13] was used to test the performance of the proposed method, which contains 698 protein sequences. Since the D-B dataset has both the training and test datasets, the tested recognition methods are modeled on the training set only and their recognition accuracies are

DOI https://doi.org/2024-JICS-22495 | Generated on 2025-04-09 07:43:40

calculated on the test dataset. By the sequence encoding methods introduced in this paper, we can get a combined feature set of 935D that used as are the input vectors for LSDA. Then LSDA is used to extract the discriminant classification features. Lastly, the SVM classifier is employed for classification. In this study, we used the MATLAB OSU-SVM Toolbox for learning SVM model, which can be obtained from http://sourceforge.net/projects/svm/. In our experiment we select polynomial kernel which is defined as $k(x,y)=(x^T y+1)^d$ to perform the classification. There are two important parameters should be defined, i.e. the kernel parameter d and the regularity parameter C. In this study, they are optimized on the training set using a grid search strategy and are set at *d*=3.2, *C*=10.

Table 1. The total recognition rates of SVM and LSDA-SVM

Method	SVM	LSDA-SVM
Accuracy(%)	55.06	61.30

Table 2. Comparison of fold accuracies and MCC between SVM and LSDA-SVM

Fold types	Accuracy(%)		MCC	
	SVM	LSDA-SVM	SVM	LSDA-SVM
1	83.33	83.33	0.64	0.64
2	77.78	88.89	0.73	0.83
3	55	45	0.64	0.59
4	62.5	62.5	0.58	0.72
5	77.78	100	0.66	0.86
6	11.11	44.44	0.98	0.60
7	63.64	84.09	0.52	0.68
8	33.33	50	0.42	0.60
9	53.85	69.23	0.47	0.60
10	33.33	50	0.42	0.57
11	50	62.5	0.49	0.72
12	21.05	34.78	0.2	0.35
13	75	50	0.19	0.57
14	25	50	0.19	0.49
15	28.57	57.14	0.53	0.61
16	85.41	68.75	0.68	0.53
17	66.67	75	0.73	0.52
18	53.85	53.85	0.48	0.55
19	29.63	48.15	0.28	0.53
20	41.67	41.67	0.47	0.53
21	50	50	0.43	0.41
22	64.29	64.29	0.63	0.66
23	71.43	42.86	0.49	0.42
24	25	25	0.35	0.35
25	25	25	0.34	0.28
26	33.33	33.33	0.39	0.37
27	81.48	96.30	0.90	0.98

4.1. Results of LSDA-SVM

In general, the recognition rates varies with the dimension of the feature subspace. Figure 1 gives the plots of recognition rates versus the corresponding dimension of LSDA-SVM.

As can be seen, the best result obtained in the optimal subspace are 61.30% and the corresponding dimensionality is 150, which is very lower than the dimension of original combined feature data set. In addition, the classification time in such a lower dimension space will be also reduced. Moreover, it appears that the recognition rate of LSDA-SVM increases much quickly when the dimension increases from 10 to 90, which indicates that LSDA can discover the intrinsic structure of the protein sequences and the extracted features have powerfully representative and generalization ability.

4.2. Comparison with SVM

To verify that the use of LSDA can extract the more effective features and improve the predict performance, we compare the presented LSDA-SVM method with SVM in this section. Table 1 show the total recognition rates of SVM and LSDA-SVM at the dimension 150. From table 1 we could find that the performance of LSDA-SVM is better than SVM.

In addition, table 2 gives the prediction accuracies, specificity and MCC in different folds of SVM and LSDA-SVM. From table 2 we could find that for many detail folds, the performance of LSDA-SVM is superior to SVM. This further proves that LSDA-SVM is a very effective protein fold prediction algorithm.

4.3. Comparison with other methods

In this subsection, the LSDA-SVM predictor is compared with other classification methods that based on single classifier such as ALH, MLP, SVM, HKNN and RBFN.

From Table 3, we can find that the prediction capacity of LSDA-SVM is stronger than that of other existing algorithms. This also shows that when using single classifiers, our algorithm has a good predictive quality.

Table 3. Comparison with other methods

Method	Accuracy(%)	Source
ALH	60.8	Kecman and Yang[17]
MLP	57.1	Ghanty and Pal [16]
SVM	56.0	Ding and Dubchak[13]
HKNN	57.1	Okun [15]
RBFN	56.4	Huang et al.[14]
LSDA-SVM	61.3	This paper

5. Conclusions

In this paper, a novel method called LSDA-SVM is presented to predict protein fold pattern. The proposed method considers not only manifold structure information but also label information of the protein samples, so that it can have more discriminating power. Experimental results on the D-B dataset show that LSDA-SVM is an effective protein fold prediction method. In the next study, we will study how to utilize other information of the protein samples to further improve the performance. It is also worth investigating the use of LSDA to predict other protein attributes such as protein subcellular and membrane types.

6. Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (No. 61273106 and No.11301457) and the Natural Science Foundation of the Jiangsu Higher Education Institutions of China(No.12KJD110007 and No.13KJD520010).

7. References

[1] R.Z. Aram, N.M. Charkari, A two-layerclassification framework for protein fold recognition, Journal of Theoretical Biology. 365(9):32-39(2015).

[2] G. Raicar, H. Saini, A. Dehzangi, S. Lal, A. Sharma, Improving protein fold recognition and structu-ral class

prediction accuracies using physicochemical properties of amino acids, Journal of Theoret-ical Biology. 402(4):117-128(2016).

[3] M. Vendruscolo, C. M. Dobson, A glimpse at the organization of the protein universe, PNAS. 102:5641-5642(2005).

[4] M. Honda, H. Kawai, Y. Shirota, T. Yamashita, cDNA microarray analysis of autoimmune hepatitis, primary biliary cirrhosis and consecutive disease manifestatio, J. Autoimmun. 25(2):133-140(2005).

[5] S. Boisvert, M. Marchand, F. Laviolette, J. Corbeil, HIV-1 coreceptor usage prediction without mult-iple alignments: an application of string kernels, Retrovirology. 5:110-115(2008).

[6] K. C. Chou, A novel approach to predicting protein structural classes in a (20-1)-D amino acid com-position space, Proteins. 21(3):319-344(1995).

[7] K. C. Chou, Prediction of protein cellular attributes using pseudoamino acid composition, Proteins. 43(1):246-255(2001).

[8] R. Y. Luo, Z. P. Feng, J. K. Liu, Prediction of protein structural class by amino acid and polypeptide composition, Eur. J. Biochem. 269(6):4219-4225(2002).

[9] K. C. Chou, Y. D. Cai, Predicting protein structural class by functional domain composition,Biochem Biophys Res Commun. 329(4): 1007-1009(2002).

[10] P. Deschavanne, P. Tuffery, Exploring an alignment free approach for protein classification and str-uctural class prediction, Biochimie. 90(4):615-625(2008).

[11] T. Yang, V. Kecman, L, B, Cao, C. Q. Zhang, Margin based ensemble classifier for protein fold reco-gnition, Expert Syst. Appl. 38(10):12348-12355(2011).

[12] D. Cai, X. F. He, K. Zhou, J. W. Han, Locality sensitive discriminant analysis. Proc. 2007 Int.Joint Conf. on Artificial Intelligence:1713-1726(2007).

[13] C. Ding, I. Dubchak, Multi-class protein fold recognition using support vector machines and neural networks, Bioinformatics. 17(4):349-358(2001).

[14] C. D. Huang, Hierarchical learning architecture with automatic feature selection for multiclass prot-ein fold classification, IEEE Transactions on NanoBioscience. 2(4):221-232(2003).

[15] Q. Okun, K-local hyperplane distance nearest-neighbor algorithm and protein fold recognition. Patt-ern Recognition and Image Analysis, 16(1):19-22(2006).

[16] P. Ghanty, N. R. Pal, Prediction of protein folds: Extraction of new features, dimensionality reducti-on, and fusion of heterogeneous classifiers, IEEE Transactions on NanoBioscience. 8(1):100- 110 (2009).

[17] V. Kecman, T. Yang, Protein fold recognition with adaptive local hyperplane algorithm. Proceedin-gs of IEEE symposium on computational intelligence in bioinformatics and computational biology. TN, USA. 75-78(2009).