World Journal of Computer Application and Technology 5(4): 51-55, 2017 http://www.hrpub.org
DOI: 10.13189/wjcat.2017.050401

# Using Semi-supervised Discriminant Analysis to Predict Subcellular Localization of Gram-negative Bacterial Proteins

**Chunming Xu***, **Yong Zhang**

School of Mathematics and Statistical, Yancheng Teachers University, 224051,Yancheng, PR China

Copyright ©2017 by authors, all rights reserved. Authors agree that this article remains permanently open access under the terms of the Creative Commons Attribution License 4.0 International License

**Abstract** In this paper, an effective dimension reduction approach called semi-supervised discriminant analysis (SDA) is employed to deal with the protein subcellular localization problem. Firstly, a novel protein sequence encoding method that consists of pseudo amino acid composition (PseAAC) and dipeptide composition (DC) is introduced to represent a protein. Secondly, the SDA algorithm is applied to extract the essential discriminant features from the combined feature data set consisting of PseAAC and DC. Finally, the K-nearest neighbor (K-NN) classifier is used to identify the subcellular localization of Gram-positive bacterial proteins. The proposed method can effective utilize both manifold information and the class information of the protein samples to guide the produce of protein subcellular localization. To evaluate the prediction performance of the proposed algorithm, a jackknife test based on nearest neighbor algorithm is employed on the gram-negative bacterial proteins data set. The results show that we can get a high total accuracy in a low-dimensional feature space, which indicates that the proposed approach is effective and practical.

**Keywords** semi-supervised discriminant analysis, protein subcellular localization, Gram-positive bacterial proteins, pseudo-amino acid compositions(PseAA), dipeptide composition(DC)

## 1 Introduction

Gram-negative bacteria are a class of bacteria that do not retain crystal violet dye in the Gram staining protocol. Many gram-negative bacteria can cause disease in a host organism. The reliable subcellular localization of a gram-negative bacteria protein based on its sequence information can provide valuable information about its function and is helpful for drug development. Thus it is important to develop methods for accurately predicting protein subcellular localization gram-negative bacteria proteins. Till now, a number of effective computational approaches have been presented for protein subcellular localization prediction[1–3]. However, predict protein subcellular localization in an automatic fashion accurately is remain a challenge.

There are different types of information source such as textual descriptions of proteins, sequence-based features and gene ontology annotation can be used to construct the features for protein subcellular localization prediction, where the sequence-based features are used widely in many biological applications. Representatives of sequence-based features include sorting signals, dipeptide composition, amino acid composition and pseudo amino acid composition. Nakashima and Nishikawa proposed to represent a protein sequence with amino acid composition (AAC)[4–6]. It has been shown AAC is closely related to protein subcellular localizations. Although AAC is a very simple and effective approach for protein sequence encoding which has achieved very promising performance in many applications, it does not consider the sequence order information. To solve this problem, Shen and Chou developed the pseudo-amino acid compositions(PseAAC) which can represent the sample of a protein in a more effective way[7–9]. Since the concept of Chou's PseAAC was introduced, researchers have proposed various PseAAC approaches to enhance the performance of protein classification.

Dipeptide composition(DC) is another sequence-based features that has been successfully used for protein sequence encoding. DC means the occurrence frequencies of two consecutive residues in a protein[10–12]. As a result, we can get feature vectors of dimension 400 for DC of every protein. DC has also been applicated to predict the subcellular localization of proteins.

Despite the sequence-based features has been successful applicated to solve many biological sequence problems, they still contain noise and are not suitable for classifica-

tion. Recently, some dimension reduction methods have been introduced to extract the essential features of protein samples. For example, Ma used principal component analysis to extract PseAAC features from proteins, and then the elman recurrent neural network (ERNN) is employed to identify the protein sequences[13]. Wang used the linear discriminant analysis (LDA) method to to extract the essential low dimensional features of gram-negative bacterial proteins.[14]. In [15], Wang used the kernel based nonlinear dimensionality reduction method to capture the nonlinear characteristics of gram-negative bacterial proteins, and it is shown that the nonlinear dimensionality reduction method is quite promising in dealing with complicated biological problems.

In fact, PCA can provide an efficient way to compress the biological data without losing much information. Different from PCA, LDA aims to extract the most discriminatory features using data label information so that it is more suitable for solving classification tasks. In this paper, a modified version of LDA, i.e., semi-supervised discriminant analysis (SDA)[16] is used to solve the protein subcellular localization problem. The presented method supplies a novel technique for extracting essential discriminant features from combined vectors consisting of PseAAC and DC. To evaluate the prediction performance of the proposed algorithm, a jackknife test based on nearest neighbor algorithm is employed on the gram-negative bacterial protein data set. The results indicate that the proposed approach achieves a high prediction performance and is effective and practical.

# 2 Materials and methods

## 2.1 Dataset

We use the benchmark dataset constructed by Chou[7], which consists of 653 gram-negative bacterial proteins, including 152 cytoplasm, 76 extracell, 12 fimbrium, 6 flagellum, 186 inner membrane, 6 nucleoid, 103 outer membrane and 112 periplasm. In this data set, no two sequences had more than 25% identity.

## 2.2 Sequence encoding methods

### 2.2.1 Pseudo amino acid composition (PseAAC)

In order to make effective use of sequence-order information of proteins, Chou [14] proposed PseAAC approach to represent the protein sequence. According to the concept of PseAAC, the protein $P$ with $L$ amino acid residues $S_1S_2S_3...S_L$, where $S_i$ represents the residue at the sequence position $i$, can be represented as

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

where the $20 + \Lambda$ components are used to reflect the relative frequencies and the sequence order information of the protein which can be expressed as follows:

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

where $f_k$ is the occurrence frequencies of 20 native amino acids and $\tau_j$ is the $j$-tier sequence correlation factor. The weight factor $w$ is used to control the complexity of the sequence order effect and is set at 0.05 as in Ref. [9]. In this paper, the patermater $\Lambda$ is set to be 10 so that we can get a 30-dimensional(D) PseAAC feature vector.

### 2.2.2 Dipeptides composition (DC)

DC is a very effective discrete model which has been successfully applied to predict protein structure information. The main advantage of DC is that it can make full use of the global information about proteins. Suppose the length of the protein $P$ is $L$, which contains $L-1$ dipeptides, i.e. $< S_1, S_2 >, < S_2, S_3 >, ..., < S_{L-1}, S_L >$. Then the feature vector of DC can be calculated as:

$$Comp(i) = \frac{n_i}{L-1} \tag{2}$$

where $i$ represents the 400 dipeptides and $n_i$ denotes the number of each dipeptide.

After obtaining the PseAAC and DC features of the protein, we can fuse them to form a combined feature vector. As a result, a protein sequence is encoded by a 30+400=430D feature vector. In this article, we will continue to use SDA to extract the more discriminant features.

## 2.3 Semi-supervised discriminant analysis (SDA)

SDA can be seen as a extension of LDA, which aims to capture the global and local structure of the given data simultaneously. As we all known, LDA is a supervised method which seeks the optimal transformation that maximizing the between-class scatter while at the same time minimizing the within-class scatter. However, it can only find the global Euclidean structure of the data. To solve the problem, the SDA algorithm was developed by Cai[16]. SDA extends LDA to incorporate a locality preserving regularizer illustrated by training samples.

Given a training set $X = [x_1, x_2, x_2, \ldots, x_n] \in R^{m\times n}$, each column of $X$ is a sample vector. Suppose there are $c$ known pattern classes, and the number of training samples in the $i-th$ class is $n_i$ and the total number of training samples is $n$. The between-class scatter matrix and total scatter matrix can be defined as follows:

$$S_b = \sum_{i=1}^{c} n_i (u_i - u)^T (u_i - u) \tag{3}$$

$$S_t = \sum_{i=1}^{n} (x_i - u)^T (x_i - u) \tag{4}$$

**Table 1.** The total recognition rates(%) of the proposed method with different dimensionality

| Dimensionality | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Recognition rate(%) | 59.41 | 79.02 | 84.83 | 88.20 | 95.25 | 96.93 | 97.70 | 97.70 | 83.76 |

where $u_i$ denotes the mean vector of training samples in $i-th$ class and $u$ denotes the mean vector of all training sample.

The objective function of SDA is as follows:

$$P = \arg\max_{P} \frac{P^T S_b P}{P^T S_t P + \alpha H(P)} \tag{5}$$

where $P$ is the projection matrix, $H(P)$ is the regularizer which incorporates intrinsic geometrical structure inferred from unlabeled data points so that the manifold structure can be optimally preserved and $\alpha$ is the regularization parameter that balances the contribution of the model complexity and the empirical loss.

In order to incorporate the manifold structure information of both labeled and unlabeled samples, the regularizer $H(P)$ can be defined as follows:

$$H(P) = P^T X L X^T P \tag{6}$$

where $L$ is the Laplacian matrix defined as $L = D - S$. $D$ is a diagonal matrix with entries $D_{ii} = \sum_j S_{ij}$ and $S$ is the adjacency matrix which is defined by

$$S_{ij} = \begin{cases} 1 & x_i \in N_b(x_j) \quad or \quad x_j \in N_b(x_i) \\ 0 & else \end{cases}$$

where $N(x_i)$ and $N(x_j)$ are the $k$ nearest neighbors of $x_i$ and $x_j$, respectively.

Then the objective function of SDA can be expressed as:

$$P = \arg\max_{P} \frac{P^T S_b P}{P^T S_t P + \alpha P^T X L X^T P} \tag{7}$$

By means of Lagrangian multiplier method, the projection matrix $P$ can be constructed by the eigenvectors of $(S_t + \alpha X L X^T)^{-1} S_b$ associated with the first $d$ largest eigenvalues $p_1, p_2, \ldots, p_d$, i.e. $P$ can be constructed as $P = [p_1, p_2, \ldots, p_d]$. Therefore the new data representation of $x_i$ can be expressed as:

$$y_i = P^{\mathrm{T}} x_i \tag{8}$$

# 3 Evaluation criteria

In order to investigate the performance of the proposed method, the overall prediction accuracy and the subclass accuracy are used to assess the accuracy rate of the prediction system. The overall prediction accuracy is defined as

$$Q = \frac{T}{n} \tag{9}$$

where $T$ is the number of query sequences whose folds have been correctly recognized and $n$ is the total number of sequences in the test dataset.

The subclass accuracy can reflect the accuracy rate of each subclass. Suppose there are $T_i$ query protein sequences correctly recognized in location $i$, then

$$Q_i = \frac{T_i}{n_i} \tag{10}$$

where $n_i$ is the number of sequences in location $i$.

# 4 Results and discussion

The benchmark dataset described in section 2.1 is used to test the performance of the proposed method, and the MATLAB software is used for data analysis. There are three popular cross-validation method: independent dataset test, subsampling test, and jackknife test. Among the three test methods, the jackknife test is the mostly used due to that it can always yield a unique result for a given benchmark dataset. Therefore, the jackknife test has been widely adopted by researchers to examine the accuracy of various prediction methods. In this paper, we also make use of the jackknife test to test the performance.

## 4.1 Results of the proposed method

The 430D combined feature set is used as the input vectors for SDA. Subsequently, the KNN classifier is applied for classification. It should be pointed out that the choice of the number of neighbors K is a crucial problem for the KNN classifier[17]. In this paper, the value of K was experimentally set to 1 because it can get a comparatively better recognition results.

In general, the recognition rates varies with the dimension of the feature subspace generated by SDA. Tabel 1 shows the recognition rates and the corresponding dimensionality of the proposed methods.

As can be seen, the best result obtained in the optimal subspace is 97.70 % and the corresponding dimensionality is 7, which is very lower than the dimension of original combined feature data set. Moreover, it appears that the performance of the proposed method ascends much quickly when the dimensionality increases form 1 to 7, which indicates that SDA can discover the intrinsic structure of the protein sequences. In addition, table 2 shows the number of true positive(TP), false positive(FP), flase negative(FN) and true negative(TN) examples for each subclass when the dimensionality is 7. From table 2 we could find that the

proposed method can deal with both positive samples and negative samples effectively.

**Table 2.** The FP, FN, TP, TN of each subclass

| Subclass | TP | FP | FN | TN |
|---|---|---|---|---|
| Cytoplasm | 152 | 0 | 4 | 497 |
| Extracell | 72 | 4 | 7 | 570 |
| Fimbrium | 12 | 0 | 0 | 641 |
| Flagellum | 6 | 0 | 0 | 647 |
| Inner membrane | 184 | 2 | 0 | 467 |
| Nucleoid | 6 | 0 | 0 | 647 |
| Outer membrane | 96 | 7 | 4 | 546 |
| Periplasm | 110 | 2 | 0 | 541 |

## 4.2 The usefulness of dimension reduction

In order to illustrate that SDA can extract the more effective features and enhance the prediction accuracy, we compare it with the results performed in the original 430D combined featured space. Table 2 shows the subcellular localization type accuracies and the overall accuracies of the two methods, where the dimensionality of SDA is set at 7.

**Table 3.** Comparison of accuracies(%)for each of the 8 subcellular localization type based on original 430D combined feature vector and 430D combined feature vector plus SDA method

| Subclass | Method | |
|---|---|---|
| | 430D combined vector | SDA |
| Cytoplasm | 69.07 | 100 |
| Extracell | 35.52 | 94.73 |
| Fimbrium | 0 | 100 |
| Flagellum | 16.66 | 100 |
| Inner membrane | 82.79 | 98.92 |
| Nucleoid | 50 | 100 |
| Outer membrane | 42.71 | 93.20 |
| Periplasm | 46.42 | 98.21 |
| Overall | 59.11 | 97.70 |

It can be seen from table 3 that the overall accuracy obtained by the approach without using dimensionality reduction is 59.11%, which is very lower than the result of the propose method. So we can conclude that the prediction quality can be improved using dimensionality reduction method. Furthermore, the recognition rates of Fimbrium and Flagellum are remarkably enhanced using our method, which indicates that the proposed method is well for recognition the bacterial proteins that belong to the locations that have only a few samples.

## 4.3 Comparison with other dimension reduction methods

In this subsection, the SDA-KNN predictor is compared with other dimension reduction based classification methods such as PCA and LDA.

**Table 4.** The recognition rates(%) of PCA, LDA and SDA

| Subclass | PCA | LDA | SDA |
|---|---|---|---|
| Cytoplasm | 73.68 | 97.37 | 100 |
| Extracell | 48.68 | 96.05 | 94.73 |
| Fimbrium | 25 | 83.33 | 100 |
| Flagellum | 50 | 83.33 | 100 |
| Inner membrane | 82.26 | 91.94 | 98.92 |
| Nucleoid | 0 | 100 | 100 |
| Outer membrane | 54.37 | 92.23 | 93.20 |
| Periplasm | 59.82 | 91.96 | 98.21 |
| Overall | 66 | 93.57 | 97.70 |

Table 4 gives the prediction results of three methods. From table 4 we can know that the total accuracies for PCA, LDA and SDA are 66%, 93.57% and 97.70%, respectively. The success rate by the proposed approach is 31.70% and 4.13% higher than the success rates by the PCA and LDA algorithms, respectively. As for dopamine and serotonin receptors, the performance of SDA is also superior to PCA and LDA. The experiments show that SDA outperforms PCA and LDA and is more suitable for solving the classification problems of complex biological patterns.

# 5 Conclusions

In this paper, a novel nonlinear dimension reduction method named SDA is utilized for membrane protein type prediction. The proposed method adopts SDA to extract more discriminating features form a combined vector, which consists of pseudo amino acid composition and dipeptide composition. A jackknife test is performed on the protein data set and the experiments illustrate that the proposed method is effective. Also, the SDA method can be combined with other protein sequence encoding and prediction algorithms to become a very useful tool dealing with complicated biological system problems.

# Acknowledgements

The authors are grateful to the anonymous reviewers for their valuable comments and suggestions that have helped to improve the presentation of the paper. This work is partially supported by the National Natural Science Foundation of China(Grant No.11771376) and the Natural Science Foundation of the Jiangsu Higher Education Institutions of China(Grant No.13KJD520010).

# REFERENCES


[1] K.J. Park, M. Kanehisa. Prediction of protein subcellular location by support vector machines using compositions of amino acids and amino acid pairs, Bioinformatics, 19, 2003, 1656-1663.

[2] S Harsh, R Gaurav, D Abdollah, L Sunil, S Alok.Subcellular localization for Gram positive and Gram negative bacterial proteins using linear interpolation smoothing model, Journal of Theoretical Biology, 386, 2015, 25-33.

[3] CS Yu, CJ Lin, JK Hwang. Prediction subcellular localization of Gram-negative bacterial proteins by support vector machine using wavelet denoising and Chou's pseudo amino acid composition, Chemometrics and Intelligent Laboratory Systems, 167(9), 2017, 102-112.

[4] H Nakashima, K Nishikawa, T Ooi. The folding type of a protein is relevant to the amino acid composition. Journal of Biochemistry, 99(1), 1986, 153-162.

[5] H Nakashima, K Nishikawa, T Ooi. Distinct character in hydrophobicity of amino acid compositions of mitochondrial proteins. Proteins-structure Function and Bioinformatics, 8(2), 1990, 173-178.

[6] H Nakashima, Y Saitou, N Usuki. Differences in Amino Acid Composition between and Structural Classes of Proteins, Journal of Biomedical Science and Engineering, 7(11), 2014, 890-918.

[7] KC Chou, Y Cai, Prediction of protein subcellular locations by GO-FunD-PseAA predictor, Biochem. Biophys. Res. Commun. 320, 2004, 1236-1239.

[8] KC Chou, Does the folding type of a protein depend on its amino acid composition, FEBS Lett, 363, 1995, 127-131.

[9] KC Chou. Prediction of protein cellular attributes using pseudoamino acid composition. Proteins, 43(3), 2001, 246-255.

[10] P Petrilli. Classification of protein sequences by their dipeptide composition. Comput Appl Biosci, 9(2), 1993 , 205-209.

[11] X Niu, N Li, D Chen, Z Wang. Interconnection between the protein solubility and amino acid and dipeptide compositions. Protein and Peptide Letters, 20(1), 2013, 88-95.

[12] M Khan, M Hayat, SA Khan, N Iqbal. Unb-DPC: Identify mycobacterial membrane protein types by incorporating unbiased dipeptide composition into Chou's general PseAAC. Journal of Theoretical Biology, 415, 2017, 13-19.

[13] J Ma, H Gu. A novel method for predicting protein subcellular localization based on pseudo amino acid composition, Bmb Reports, 43(10), 2010, 670-676.

[14] T Wang, J Yang. Using the nonlinear dimensionality reduction method for the prediction of subcellular localization of Gram-negative bacterial proteins, Molecular Diversity, 13(4), 2009, 475-480.

[15] SF Wang, SH Liu.Protein sub-nuclear localization based on effective fusion representations and dimension reduction algorithm LDA, International Journal of Molecular Sciences, 16(12), 2015, 30343-30361.

[16] D Cai, X He, J Han. Semi-supervised discriminant analysis, IEEE 11th International Conference on Computer Vision, 1-7, 2007.

[17] Denoeux, T. A k-Nearest Neighbor Classification Rule Based on Dempster-ShaferTheory. IEEE Trans. Syst. Man Cybern., 1995, 25(5), 804-813.