I.J. Wireless and Microwave Technologies, **2012, 5, 28-33**
Published Online October 2012 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijwmt.2012.05.04

Available online at http://www.mecs-press.net/ijwmt

Feature Dimension Reduction Algorithm Based Prediction Method for Protein Quaternary Structure

Tong Wang[a,*], Tian Xia[a], Xiaoxia Cao[a]

[a] *Institute of Computer and Information, Shanghai Second Polytechnic University, Shanghai, 201209, China*

Abstract

Knowing the quaternary structure of an uncharacterized protein often provides useful clues for finding its biological function and interaction process with other molecules in a biological system. Here, dimensionality reduction algorithm is introduced to predict the quaternary structure of proteins. Our jackknife test results indicate that it is very promising to use the dimensionality reduction approaches to cope with complicated problems in biological systems, such as predicting the quaternary structure of proteins.

Index Terms: PSSM; Protein Quaternary Structure; Feature Dimension Reduction

1. Introduction

In a number of biological processes, the quaternary structure is an interesting field in bioinformatics. It is generally accepted that the amino acid sequence of most proteins contains all the information needed to fold the protein into its correct three-dimensional structure. The quaternary structure of proteins, which is the association of tertiary structure subunits, depends on the existence of complementary "patches" on their surfaces [1]. Therefore, the patches that are buried in the interfaces formed by the subunits play a vital role in both tertiary and quaternary structures. This suggests the possibility to predict the quaternary structure from primary sequences [2].

Given a polypeptide chain, will it form a dimmer, trimer, or any other oligomer, or exist only as a monomer? Some efforts have been made in developing computational tools to predict protein quaternary structure from its sequence. In a pioneer study, Chou and Elrod [3] introduced the covariant discriminant algorithm to predict the quaternary structure of proteins based on the pseudo amino acid (PseAA) composition.

The present study was initiated in an attempt to propose a completely different approach, the comprehensive comparative study of different DR methods in terms of their ability to predict the quaternary structure. Moreover, protein sequences are represented by PSSM (Position-Specific Score Matrix) [4] which incorporate the evolution information. The result thus obtained is quite encouraging, indicating that the above approach can also be effectively used to deal with other complicated biological systems.

* Corresponding author:
E-mail address: tongwang0818@yahoo.cn

2. Methods

2.1. Dataset

The data set taken from Chou and Cai [3] are used to test the current method. The data set consists of 3,174 protein sequences, of which 382 are with annotation of monomer, 817 of dimer, 593 of trimer, 884 of tetramer, 54 of pentamer, 287 of hexamer, and 157 of octamer.

2.2. Position-Specific Scoring Matrix

In this study, a powerful sequence encoding scheme PSSM is introduced. It is useful to summarize the main definitions associated with this method here.

A protein sequence containing N amino acids can be represented by a 420-D (Dimensional) vector, i.e.,

$$\mathbf{P}_{\text{PSSM-420}} = \begin{bmatrix} \overline{\mathrm{A}}_1 & \overline{\mathrm{A}}_2 & \cdots & \overline{\mathrm{A}}_{20} & \mathrm{S}_1 & \mathrm{S}_2 & \cdots & \mathrm{S}_{400} \end{bmatrix}^{\mathrm{T}} \tag{1}$$

where the first 20 components are the average scores of every column in $\mathbf{P}_{PSSM}$ matrix. $\mathbf{P}_{PSSM}$ is shown as below:

$$\mathbf{P}_{PSSM} = \begin{pmatrix} \mathrm{A}_{1\to1} & \mathrm{A}_{1\to2} & \cdots & \mathrm{A}_{1\to20} \\ \mathrm{A}_{2\to1} & \mathrm{A}_{2\to2} & \cdots & \mathrm{A}_{2\to20} \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{A}_{N\to1} & \mathrm{A}_{N\to2} & \cdots & \mathrm{A}_{N\to20} \end{pmatrix} \tag{2}$$

where $\mathrm{A}_{i\to j}$ represents the score of amino acid residue at the i -th position of the protein sequence being substituted to the amino acid type $j(1 \le j \le 20)$ during evolution process. Here, the numerical codes 1, 2,…, 20 represent the 20 native amino acid types according to the alphabetical order of their single-residue codes. N denotes the length of the protein. In this study, $\mathbf{P}_{PSSM}$ is generated by carrying out PSI-BLAST. This process will search the Swiss-Prot database through three iterations for multiple sequence alignment against the protein $\mathbf{P}$. Every element in $\mathbf{P}_{PSSM}$ was scaled by a standardization procedure. The components S_1 , S_2 ,…, S_{400} in (1) are obtained by summing up all rows in the $\mathbf{P}_{PSSM}$, each of which corresponds to the same amino acid in the primary sequence $\mathbf{P}$. It means for each column in $\mathbf{P}_{PSSM}$, there are 20 values instead of N . Hence, we will have a vector of dimension 20×20 for a $\mathbf{P}_{PSSM}$.

2.3. PCA

Principal Components Analysis (PCA) constructs a low-dimensional representation of the data that describes as much of the variance in the data as possible. This is done by finding a linear basis of reduced dimensionality for the data, in which the amount of variance in the data is maximal[5].

2.4. LDA

Linear Discriminant Analysis (LDA) attempts to maximize the linear separability between datapoints belonging to different classes. In contrast to most other dimensionality reduction techniques, LDA is a supervised technique[5]. LDA finds a linear mapping M that maximizes the linear class separability in the low-dimensional representation of the data. The criteria that are used to formulate linear class separability in LDA are the within-class scatter S_W and the between-class scatter S_B.

2.5. Kernel PCA

Kernel PCA (KPCA) is the reformulation of traditional linear PCA in a high-dimensional space that is constructed using a kernel function [5]. Kernel PCA computes the principal eigenvectors of the kernel matrix, rather than those of the covariance matrix.

The reformulation of traditional PCA in kernel space is straightforward, since a kernel matrix is similar to the inproduct of the datapoints in the high-dimensional space that is constructed using the kernel function. The application of PCA in kernel space provides Kernel PCA the property of constructing nonlinear mappings.

Kernel PCA computes the kernel matrix K of the datapoints x_i. The entries in the kernel matrix are defined by

$$k_{ij} = k(x_i, x_j) \tag{3}$$

where k is a kernel function. Subsequently, the kernel matrix K is centered using the following modification of the entries

$$k_{ij} = k_{ij} - \frac{1}{n}\sum_l k_{il} - \frac{1}{n}\sum_l k_{jl} + \frac{1}{n^2}\sum_{lm} k_{lm} \tag{4}$$

The centering operation corresponds to subtracting the mean of the features in traditional PCA. It makes sure that the features in the high-dimensional space defined by the kernel function are zeromean. Subsequently, the principal d eigenvectors v_i of the centered kernel matrix are computed. It can be shown that the eigenvectors of the covariance matrix α_i are scaled versions of the eigenvectors of the kernel matrix v_i

$$\alpha_i = \frac{1}{\sqrt{\lambda_i}} v_i \tag{5}$$

In order to obtain the low-dimensional data representation, the data is projected onto the eigenvectors of the covariance matrix α_i. The result of the projection is given by

$$Y\left\{\sum_j \alpha_1 k(x_j, x), \sum_j \alpha_2 k(x_j, x), ..., \sum_j \alpha_d k(x_j, x)\right\} \tag{6}$$

2.6. Kernel LDA

By introducing a kernel function which corresponds to the non-linear mapping, all the computation can conveniently be carried out in the input space. The problem can be finally solved as an eigen-decomposition problem like PCA, LDA and KPCA. From the theory of reproducing kernel we know that any solution $w \in F$ must lie in the span of all training samples in F. Let ϕ be a nonlinear mapping to some feature space F. To find the linear discriminant in F we need to maximize [6]

$$J(w) = \frac{w^T S_B^\phi w}{w^T S_W^\phi w} \tag{7}$$

where S_B is between-class scatter matrix and S_W is within-class scatter matrix. Therefore we can find an expansion for w of the form

$$w = \sum_{i=1}^{l} \alpha_i \phi(x_i) \tag{8}$$

Using the expansion (8) and the definition of m_i^ϕ we write [6]

$$\begin{aligned} w^T m_i^\phi &= \frac{1}{l_i} \sum_{j=1}^{l} \sum_{k=1}^{l_i} \alpha_j k(x_j, x_k^i) \\ &= \alpha^T \mathrm{M}_i \end{aligned} \tag{9}$$

Where we defined $(M_i)_j := \frac{1}{l_i} \sum_{k=1}^{l_i} k(x_j, x_k^i)$ and replaced the dot products by the kernel function. Now consider the numerator of (7). Be using the definition of S_B^ϕ and (9) it can be rewritten as

$$w^T S_B^\phi w = \alpha^T M \alpha \tag{10}$$

where $M := (M_1 - M_2)(M_1 - M_2)^T$. Considering the denominator, using (8), the definition of m_i^{ϕ} and a similar transformation as in (10) we find:

$$w^T S_W^{\phi} w = \alpha^T N \alpha \tag{11}$$

Where we set $N := \sum_{j=1,2} K_j (I - 1_{l_j}) K_j^T, K_j$ is a $l \times l_j$ matrix with $(K_j)_{nm} := k(x_n, x_m^j)$ (this is the kernel matrix for class j), I is the identity and 1_{l_j} the matrix with all entries $1/l_j$.

Combining (10) and Eq.11 we can find linear discriminant in F by maximizing

$$J(\alpha) = \frac{\alpha^T M \alpha}{\alpha^T N \alpha} \tag{12}$$

This problem can be solved (analogously to the algorithm in the input space) by finding the leading eigenvector of $N^{-1}M$. We will call this approach (nonlinear) Kernel LDA. The projection of a new pattern x onto w is given by

$$(w \cdot \phi(x)) = \sum_{i=1}^{l} \alpha_i k(x_i, x) \tag{13}$$

Thus, using (13) we can map a protein sample into some high-dimensional feature space as desired.

3. Experimental Results

The performance of four different DR methods from the perspective of identifying quaternary structure of proteins was compared. The accuracy of the low dimensional representations of the high dimensional data obtained by the different DR methods was evaluated via KNN algorithm. Accordingly, the jackknife test has been increasingly and widely adopted by investigators to test the power of various predictors. Therefore, in this study, jackknife test was performed with the current approach in predicting the quaternary structure of proteins.

Table 1. Success rates in identifying quaternary structure of protein by the jackknife test

Method	Sequence encoding schemes	Test method (%)
		Jackknife
K-NN(K=1)	PSSM	82
PCA& K-NN(K=1)	PSSM	84.65
KPCA& K-NN(K=1)	PSSM	85.73
LDA& K-NN(K=1)	PSSM	89.28
KLDA& K-NN(K=1)	PSSM	93.19

As shown in **Table 1,** the overall jackknife success rates obtained by DR methods in identifying the quaternary structure of proteins are higher than the ones obtained without using linear DR methods. Meantime, it indicates that supervised DR methods (LDA and KLDA) outperform unsupervised DR methods (PCA and KPCA) and the nonlinear DR methods (KPCA and KLDA) outperform linear DR methods (PCA and LDA). In summary, base on the observation, it is concluded that the overall jackknife success rate with KLDA is the highest relative to the other DR methods.

4. Conclusions

In this paper, we compared the performance of four different DR methods from the perspective of discriminating quaternary structure of proteins. The results obtained are encouraging, which are higher than the ones obtained without DR methods. The application of DR approach to the prediction of protein quaternary structure is just an example to demonstrate its advantages. It has not escaped our notice that the DR approach can also be used to deal with many other complicated biological systems.

Acknowledgements

This work was supported by shanghai university scientific selection and cultivation for outstanding young teachers in special fund (EGD10003) and This work was supported by Chenguang Program of Shanghai Municipal Education Commission(10CG61).

References

[1] R. Garian. Prediction of quaternary structure from primary structure. Bioinformatics, vol. 17, p. 551-556, 2001.

[2] X. Xiao, P. Wang, K.C. Chou. Quat-2L: a web-server for predicting protein quaternary structural attributes. Molecular Diversity, 2010, in press.

[3] K.C. Chou and Y.D. Cai, Predicting protein quaternary structure by pseudo amino acid composition. Proteins, vol. 53, p. 282-289, 2003

[4] H. Kaur and G.P. Raghava. A neural-network based method for prediction of gamma-turns in proteins from multiple sequence alignment. Protein Sci, vol. 12, p. 923-929, 2003.

[5] L.J.P.v.d. Maaten, E.O. Postma, and H.J.v.d. Herik, Dimensionality Reduction: A Comparative Review. 2007.

[6] S. Mika, G. Ratsch, J. Weston, B. Scholkopf. "Fisher discriminant analysis with kernels," in: Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop Madison, WI, US, pp.41–48, 1999.