I.J. Wireless and Microwave Technologies **2012, 4, 8-12**
Published Online August 2012 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijwmt.2012. 04.02

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

Identifying Protein Structural Classes Using MVP Algorithm

[a] Tong Wang[*], [a] Xiaoming Hu, [a] Xiaoxia Cao

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

Abstract

A new method for the prediction of protein structural classes is constructed based on MVP (Maximum variance projection) algorithm, which is a manifold learning-based data mining method. DC (Dipeptide Composition) and PseAA (Pseudo Amino Acid) are used as conditional attributes for the construction of decision system. A DR (Dimensionality Reduction) algorithm, the so-called MVP is introduced to reduce the decision system, which can be used to classify new objects. Experimental results thus obtained are quite encouraging, which indicate that the above method is used effectively to deal with this complicated problem of protein structural classes.

Index Terms: protein structural classes; MVP; sequence encoding scheme

1. Introduction

The prediction of protein structural classes is an important topic in molecular biology. This is because there is a gap between sequence and structure. The knowledge of protein structural classes may help to improve the prediction of protein secondary structure prediction, reduce the scope of searching conformational space during energy optimization, and provide useful information for a heuristic approach. In general, protein structure can be classed into four classes: all-↦, all- ↕, ↦/↕ and ↦□ ↕. Many different algorithms and efforts have been made to address this problem so far.

Different clustering and classification algorithms have been used in those prediction approaches, such as the discriminate analysis, the least Euclidian distance, the Mahalanobis distance, covariant discriminant, fuzzy clustering, support vector machines and maximum entropy [1]. A review about prediction of protein structural class by Chou [2] presented this problem systematically, and introduced and compared some existing methods. However, the emphasis in this work is not in identifying the optimal classification scheme but in employing the DR method that can provide the optimal low-dimensional representations so that making the task of structural classes' prediction system becomes simple. In this paper, a DR algorithm, the so-called MVP [3] is introduced to discriminate the protein structural class.

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

Several approaches have been developed to predict protein structural classes based on the AA (amino acid) composition. In this study, a sequence encoding scheme is introduced by fusing PseAA [4] and DC [5]. Our jackknife test results indicate that it is very promising to use the above method to cope with complicated biological problems.

2. METHODS

2.1. Dataset

For clearly presenting the better performance produced by the new method, a well known dataset [6] is used in our work. The dataset was taken from Chou (1999) [6]. It consists of 204 protein chains, of which 52 are all-α proteins, 61 all-β, 45 α/β proteins and 46 $\alpha+\beta$ proteins.

Given a query protein sequence **P**, in order to predict which structural class it is, the first important thing we need to do is to use a proper descriptor to represent it. The descriptor not only contains as much information of the sequence as possible, but also can be handled by powerful prediction algorithms. A powerful descriptor by fusing DC and PseAA introduced in this work is a good one in this regard. Below, a brief introduction of DC and PseAA is given, respectively.

2.2. Dipeptide Composition

DC (Dipeptide Composition) can be considered as a representative form of proteins incorporating sequence neighborhood information. This method extracts and computes the occurrences of two consecutive residues from a sequence string. Therefore, a fixed pattern length obtained is 400-D (dimensional) vector. It can be calculated by (1).

$$\text{Fraction of dipeptide(i)}=\frac{\text{total number of dipeptide(i)}}{\text{total number of all possible dipeptides}} \tag{1}$$

2.3. PseAA composition

The concept of PseAA (Pseudo Amino Acid) composition was proposed by incorporating the sequence order information completely [4]. According to the PseAA composition discrete model, the protein P can be formulated as

$$P_{PseAA} \equiv [p_1, p_2, ..., p_{20}, p_{20+1}, ..., p_{20+\lambda}]^T,\ (\lambda < N) \tag{2}$$

where the 20+λ components are given by

$$p_m \equiv \begin{cases} \dfrac{f_m}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\lambda} \tau_j}, & (1 \le \text{m} \le 20) \\[2ex] \dfrac{w\tau_{m-20}}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\lambda} \tau_j}, & (20+1 \le \text{m} \le 20+\lambda) \end{cases} \tag{3}$$

where w is the weight factor, which was set at 0.05 in Ref. [4]and θ_j is the j th tier correlation factor, which reflects the sequence order correlation between all of the j th most contiguous residues. f_m is the occurrence frequencies of 20 amino acids in sequence. The PseAA is actually corresponding to a 20+20=40-D (Dimensionality) vector, where the first 20-components are the 20 occurrence frequencies of the native amino acids in a protein, the next 20-components are the correlation factors of amino acids.

In present work, a powerful representation method through fusing the DC and PseAA is proposed. Totally, a sequence is encoded by a $400 + 40 = 440$-D vector. The prediction of the protein structural classes can be fairly difficult to make due to the high-dimensional vector space. Below, we will introduce a DR approach to cope with the problem caused by the high dimensional descriptor.

2.4. MVP

A powerful dimension reduction algorithm, MVP algorithm [3], was originally used for the application of pattern recognition. As a new pattern recognition algorithm, MVP has demonstrated its better performance than the conventional dimension reduction algorithms. A brief introduction is given below.

Given a dataset of $X = [\vec{x}_1, ..., \vec{x}_N]$ in a m-dimension real space R^m where $\vec{x}_i \ (i = 1, 2, \cdots, N)$ can be classified into C classes denoted as $\{\omega_1, \omega_2 ..., \omega_C\}$. The objective of MVP is to find a transformation matrix U which maps X into Y in a new d-dimension real space R^d (where $d \ll m$); i.e.,

$$Y = U^{\mathbf{T}} X \tag{4}$$

where $Y = [\vec{y}_1, ..., \vec{y}_N]$ and $\vec{y}_i \ (i = 1, 2, \cdots, N)$ are the corresponding vectors in the new space. According to MVP theory [3], an optimized transformation matrix U can be obtained by minimizing the following objective function:

$$\frac{\sum_{i=1}^{N} \left\| U^T \vec{x}_i - \sum_{j} w_{ij} U^T \vec{x}_j \right\|^2}{\sum_{i=1}^{N} \sum_{j=1}^{N} \left\| U^T \vec{x}_i - U^T \vec{x}_j \right\|^2 s_{ij}} \quad (i, j = 1, 2, \cdots, N) \tag{5}$$

where $W = \left[w_{ij} \right] \ (i, j = 1, 2, \cdots, N)$ is a weight-coefficient matrix and $\mathbf{S} = \left[s_{ij} \right] \ (i, j = 1, 2, \cdots, N)$ is a similarity matrix. For more detailed description about the concept of MVP, see [3].

Thus, according to [3], there is

$$\mathbf{min} \left(\frac{\sum_{i=1}^{N} \left\| U^T \vec{x}_i - \sum_{j} w_{ij} U^T \vec{x}_j \right\|^2}{\sum_{i=1}^{N} \sum_{j=1}^{N} \left\| U^T \vec{x}_i - U^T \vec{x}_j \right\|^2 s_{ij}} \right)$$

$$= \mathbf{arg}\,\mathbf{min}_{U} \left(\frac{\mathrm{tr}\left(U^{\mathbf{T}} \mathbf{X} M \mathbf{X}^{\mathbf{T}} U \right)}{\mathrm{tr}\left(U^{\mathbf{T}} \mathbf{X} L \mathbf{X}^{\mathbf{T}} U \right)} \right) \tag{6}$$

where tr is the trace of the matrix $M = (I - W)^T (I - W)$. I is a unit matrix, $L = D - S$ and D is a diagonal matrix with $d_{ii} = \sum_j s_{ij}$. U in (4) can be obtained by solving the following eigen equation:

$$XMX^{\mathbf{T}} \vec{\alpha} = \ell XLX^T \vec{\alpha} \tag{7}$$

The eigenvectors $\vec{\alpha}_1, \vec{\alpha}_2, ..., \vec{\alpha}_d$ obtained are used to constitute the desired transformation matrix U,

$$U = [\vec{\alpha}_1, \vec{\alpha}_2, ..., \vec{\alpha}_d] \tag{8}$$

Finally, the low-dimension protein dataset can be obtained according to (4).

3. EXPERIMENTAL RESULTS

Eventually, 60-D vector is obtained by using the MVP algorithm to extract the most important features from the 440-D vector. That is to say, the best prediction accuracy is obtained by the MVP algorithm when the dimension of the protein feature vector is 60. Subsequently, the KNN (K-nearest neighbor) algorithm was used to predict the protein structural class based on the 60-D vector. Jackknife test is the most rigorous and objective test method. It is often used for assessing the accuracy of a predictor. In this study, jackknife test has been adopted to evaluate the performance of predictors.

The results thus obtained by the jackknife tests on the benchmark datasets taken from Chou [6] are listed in Table 1. For facilitating comparison, the corresponding results obtained without MVP are also listed. As can be seen from table 1, we can obtain more than 88% success rates in jackknife tests by using KNN algorithm on the reduced 60-D vectors generated by MVP algorithm, which is about 6% higher than the ones obtained without MVP.

Table 1. Success rates in identifying protein structural class by the jackknife tests with two different classifiers

Classifier	Input form	Test method (%)
		Jackknife
KNN (K=1)	Original vector (440-D)	$\frac{180}{204} = 88.2$
MVP & KNN (K=1)	Dimension-reduced vector (60-D) by MVP	$\frac{192}{204} = 94.1$

The overall jackknife success rates by using the fusion sequence encoding approach are higher than the single sequence encoding scheme seen from Table 2. Also, the success rates obtained by MVP are generally higher than the ones obtained without MVP. In summary, base on the observation, it is concluded that the overall jackknife success rate with fusion representation method in reduced 60-D space is the highest relative to the single sequence encoding scheme and original high dimensional vector.

Table 2. The jackknife success rates for protein structural class prediction by using the original high dimensional vector (440-D) and dimension-reduced vector (60-D) with two different sequence encoding schemes

Sequence encoding schemes	Jackknife (%)	
	Original high dimensional vector	Dimension-reduced vector (60-D) by MVP
DPI	$\frac{177}{204}$ □ 86.8	$\frac{188}{204}$ □ 92.2
Fusion of PseAA and DPI	$\frac{180}{204}$ □ 88.2	$\frac{192}{204}$ □ 94.1

4. CONCLUSIONS

The results obtained in discriminating protein structural class by the method proposed in this work are encouraging. The existing predictors mainly focus on finding the optimal classification scheme. In this paper, we apply the MVP algorithm to extract the key information from the high-dimensional space and reduce the original high-dimensional vector to a lower-dimensional one. Here, the application to protein structural class prediction is just as a paradigm to show the advantage of MVP. Besides the application of MVP on the structural class of proteins prediction in this study, MVP method in fact can be used to deal with other complicated biological systems.

Acknowledgment

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]Y.S. Ding, T.L. Zhang, Q. Gu, P.Y Zha, K.C. Chou. Using maximum entropy model to predict protein secondary structure with single sequence. Protein Pept Lett. vol. 16, pp.552-60, 2009.

[2] K.C. Chou. Review: progress in protein structural class prediction and its impact to bioinformatics and proteomics. Curr. Protein Pept. vol. 6, pp.423–436, 2005.

[3] T.H. Zhang, J. Yang, H. H. Wang, C.H. Du. Maximum variance projections for face recognition. Optical Engineering, vol. 46, pp.2007067206-1-067206-8, 2007.

[4] K.C. Chou. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins, vol.43, pp.246-255, 2001.

[5] M.Reczko and H.Bohr. The DEF data base of sequence based protein fold class predictions. Nucleic Acids Res, vol. 22, pp.3616-3619,1994.

[6] Chou, K.C. A key driving force in determination of protein structural classes. Biochem. Biophys. Res. Commun. vol.264, pp.216–224, 1999.