A Data Mining Method For Improving the Prediction Of Bioinformatics Data

Tong Wang*, Wenan Tan, Jianxin Xue
Institute of Computer and Information Engineering, Shanghai Polytechnic University, Shanghai, 201209, China
*Corresponding author e-mail: wangtong@sspu.edu.cn

Abstract. The composition of proteins nearly correlated with its function. Therefore, it is very urgently important to discuss a method that can automatically forecast protein structure. The fusion encoding method of PseAA and DC was adopted to describe the protein features. Using this encoding method to express protein sequences will produce higher dimensional feature vectors. This paper uses the algorithm of predigesting the characteristic dimension of proteins. By extracting significant feature vectors from the primitive feature vectors, eigenvectors with high dimensions are changed to eigenvectors with low dimensions. The experimental method of jackknife test is adopted. The consequences indicate that the arithmetic put forwarded here is appropriate for identifying whether the given protein is a homo-oligomer or a hetero-oligomer.

1. Introduction
Protein type prediction is an important aspect of protein research. It can provide useful information. There are four structural types of proteins: Those annotated with “subunit” were extracted typing in “homo-dimer”, “homo-trimer”, “homo-tetramer”, “homo-pentamer”, “homo-hexamer”, “homo-heptamer”, “homo-octamer”, “homo-dodecamer”, “hetero-dimer”, “hetero-trimer”, “hetero-tetramer”, “heteropen-tamers”, “hetero-hexamer”, “hetero-heptamer”, “hetero-octamer” or “hetero-dodecamer”. Array sequential information is often adopted in encoding algorithms. The algorithm is available. The fusion encoding method of PseAA and DC [1,2] was adopted to describe the protein features. DC and PseAA is introduced as follows.

2. Method
2.1. Dataset
4869 proteins are composed the data set [3]: 3468 are homo-oligomers, 1401 are hetero-oligomers.

2.2. Method
There should be an appropriate way to represent protein P. The fusion encoding algorithm of PseAA and DC [1,2] was adopted to describe the protein features. DC and PseAA is introduced as follows.
2.3. Dipeptide Composition
DC [1, 2] is an encoding arithmetic. Protein is showed by a 400-D eigenvector. It could be counted by Eq. 1.

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

2.4. PseAA Composition
PseAA is an encoding method. The protein \( P \) could be counted as

\[
P_{\text{PseAA}} = [p_1, p_2, \ldots, p_{20}, p_{20+1}, \ldots, p_{20+\xi}]^T, \quad (\xi < N) \tag{2}
\]

where the \( 20+\xi \) are generated by

\[
P_k = \begin{cases} 
\frac{f_k}{\sum_{j=1}^{20} f_j + \sum_{j=1}^{\xi} \varphi_j} , & (1 \leq k \leq 20) \\
\frac{w \varphi_{k-20}}{\sum_{j=1}^{20} f_j + \sum_{j=1}^{\xi} \varphi_j} , & (20+1 \leq k \leq 20+\xi)
\end{cases} \tag{3}
\]

The \( w \) value is 0.05. \( \varphi_j \) is the \( j \) th tier relevant coefficient factor. \( f_k \) is the likelihood of 20 amino acids in alignment. The minimum size of protein array in the data pool is \( N = 39 \). In Eqs. 2 and 3, the \( \xi \) is 24. A \( 20+24=44 \)-D vectors is produced.

The fusion encoding algorithm of PseAA and DC was adopted to describe the protein features. A \( 400 + 44 = 444 \)-D vectors are created. We use the algorithm of predigesting the characteristic dimension. A MDM-Isomap [6-8] is adopted.

2.5. MDM-ISOMAP
The calculation process of MDM-Isomap algorithm is divided into three steps [9-11]. Firstly, dissimilar neighbors’ map \( G \)’ is constructed. Secondly, the fully connected graph can represent the points. Thirdly, calculate the Euclidean distances \( d_X(x_i, x_j)' \) between all points \( x_i, x_j \). The minimax distance metric \( M_X(x_i, x_j)' \) is proposed to replace the Euclidean distance metric \( d_X(x_i, x_j)' \). The criteria for judging whether a point is a neighbor are as follows: if \( x_j \) is one of the \( k \)-nn (\( k \)-nearest neighbors) of \( x_i \), or if \( M_X(x_i, x_j)' \leq \varepsilon \) (\( \varepsilon \) is some certain value), they are neighborhoods.

Construction of a diversity neighboring map \( G \). The input vertices in the high-dimensional space \( X \) can be shown as a fully link map. Compute the Euclidean ranges \( d_X(x_i, x_j)' \) among all pairs of input vertices \( x_i, x_j \). The \( d_X(x_i, x_j)' \) reflecting the original vertices among \( x_i \) and \( x_j \). To get the neighboring of the data vertices, the neighboring choice method by the minimax distance metric \( M_X(x_i, x_j)' \) is adopted. The minimax distance metric \( M_X(x_i, x_j)' \) can be gotten by Eq.11. These new diversity relations are shown as a diversity graph over the data points, with edges of weight \( M_{X}(x_i, x_j)' \) among all input vertices. In this new minimax range metric, if \( x_j \) is one of the \( k \) nearest neighboring of \( x_i \), or if \( M_{X}(x_i, x_j)' \leq \varepsilon \), they are neighbors.

Computing the geodesic distance in the given diversity neighboring map. This step is similar to the second step of Isomap method. One difference is that geodesic distance among any two vertices \( x_i, x_j \) is computed on the new dissimilarity neighboring map. Then we get the matrix of shortest pathways distances map \( M_G(x_i, x_j)' \).
Construction of lower dimensional embedding. Use the classical MDS method with the new shortest pathways distances to map the data into a low dimensional space. [12].

3. Experimental Results
The experimental algorithm of jackknife test and classification algorithm of KNN and SVM are adopted.

Table 1. The comparison results of two distinct algorithm

| Algorithm         | The encoding algorithm | Test method (%) |
|-------------------|------------------------|-----------------|
| KNN               | 444-D vectors          | 3416/4869 = 70.16 |
| SVM               | 444-D vectors          | 3403/4869 = 69.89 |
| MDM-Isomap & KNN  | 70-D vectors           | 3918/4869 = 80.47 |
| MDM-Isomap & SVM  | 70-D vectors           | 3999/4869 = 82.13 |

MDM-Isomap algorithm is used to extract significant 70-D feature vectors from the primitive 444-D feature vectors, eigenvectors with high dimensions are changed to eigenvectors with low dimensions.

In Table 1, the analogy of identification outcomes by MDM-Isomap algorithm is illustrated. The analogy of identification outcomes by without MDM-Isomap arithmetic is illustrated in Table 1. The accuracy of the first arithmetic is about 70%. The accuracy of the second arithmetic is 80%.

Table 2. The comparison results of two different algorithm for each type of oligomeric proteins

| Protein          | Jackknife (%)         |
|------------------|-----------------------|
|                  | 444-D vectors | 70-D vectors |
| Homo-oligomer    | 2575/3468 = 74.25 | 2919/3468 = 84.17 |
| Hetero-oligomer  | 841/1401 = 60.03 | 999/1401 = 71.31 |
| Overall          | 3416/4869 = 70.16 | 3918/4869 = 80.47 |

The outcomes for each type of 2 kinds of proteins prediction are illustrated in Table 2. The analogy of identification outcomes by MDM-Isomap arithmetic is illustrated in Table 2. The analogy of identification outcomes by without MDM-Isomap arithmetic is also illustrated in Table 2. The consequences indicate that the arithmetic put forwarded here is very appropriate for protein structure prediction.

4. Conclusions
The composition of proteins nearly correlated with its function. Therefore, it is very important to discuss a method that can automatically forecast protein structure. The fusion encoding method of PseAA and DC was adopted to describe the protein features. Using this encoding method to express protein sequences will produce higher dimensional feature vectors. This paper uses the algorithm of predigesting the characteristic dimension. By extracting significant feature vectors from the primitive feature vectors, eigenvectors with high dimensions are changed to eigenvectors with low dimensions. The experimental
The method of jackknife test is adopted. The consequences indicate that the arithmetic put forwarded here is appropriate for identifying whether the given protein is a homo-oligomer or a hetero-oligomer.

Acknowledgments
This work was supported by National Natural Science Foundation of China (Grant No. 61672022, Grant No. 61502296 and Grant No. 61301249), the Municipal Natural Science Foundation of Shanghai (Grant No. 15ZR1417000) and The Soft Engineering of Key Subjects Construction in Shanghai Second Polytechnic University (Grant No. XXKZD1301). This work was supported by National Natural Science Foundation of China (No. 61620106007). This work was supported by the Collaborative Innovation Platform of Electronic Information Master of Shanghai Polytechnic University.

References
[1] Chou, K.C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins, 43:246-255.
[2] Chou, K.C. (1999) A key driving force in determination of protein structural classes. Biochem. Biophys. Res. Commun. 264:216–224.
[3] Reczeko, M., and Bohr, H. (1994) The DEF data base of sequence based protein fold class predictions. Nucleic Acids Res, 22: 3616-3619.
[4] Tenenbaum, J. B., Silva, V. de, and Langford, J. C. (2000) A global geometric framework for nonlinear dimensionality reduction, Science, 290 (5500):2319-2323.
[5] Choi, H. and Choi, S. (2007) Robust kernel Isomap, Pattern Recognition,40(3):853–862.
[6] Balasubramanian, M., et al. (2002) The Isomap Algorithm and Topological Stability, Science, 295 (5552):7a-7.
[7] Tenenbaum, J.B. (1998) Mapping a manifold of perceptual observations, Advances in Neural Information Processing Systems, 10: 682-687.
[8] Balasubramanian, M., et al. (2002) The Isomap Algorithm and Topological Stability, Science, 295 (5552):7a-7.
[9] Roweis, S. T., Saul, L. K.. (2000) Nonlinear dimensionality reduction by locally linear embedding, Science, 290(5500):2323-2326.
[10] Choi, H., Choi, S.. (2007) Robust kernel Isomap, Pattern Recognition,40(3):853–862.
[11] Saxena, A., Gupta, A., and Mukerjee, A. (2004) Non-linear dimensionality reduction by locally linear isomaps, Lecture Notes in Computer Science, 3316:1038–1043.
[12] Turk, M. Pentland, A. (1991) Eigenfaces for recognition. J. Cogn Neurosci. 3(1):71–86.