Available online at www.sciencedirect.com
ScienceDirect
Journal of Molecular Graphics and Modelling 25 (2007) 852–855
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

A simple method for protein structural classification☆

Na Liu [a,b,*], Tianming Wang [b,c]

[a] Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, China
[b] College of Advanced Science and Technology, Dalian University of Technology, Dalian 116024, China
[c] Department of Mathematics, Hainan Normal University, Haikou 571158, China

Received 10 April 2006; received in revised form 15 August 2006; accepted 22 August 2006
Available online 30 August 2006

Abstract

Since the concept of structural classes of proteins was proposed, the problem of protein classification has been tackled by many groups. Most of their classification criteria are based only on the helix/strand contents of proteins. In this paper, we proposed a method for protein structural classification based on their secondary structure sequences. It is a classification scheme that can confirm existing classifications. Here a mathematical model is constructed to describe protein secondary structure sequences, in which each protein secondary structure sequence corresponds to a transition probability matrix that characterizes and differentiates protein structure numerically. Its application to a set of real data has indicated that our method can classify protein structures correctly. The final classification result is shown schematically. So it is visual to observe the structural classifications, which is different from traditional methods.

Keywords: Secondary structure sequence; Structural classes; Content of helix/strand; Transition probability matrix; Structural characteristic vector

1. Introduction

With the development of structural prediction technique, the growth rate at which proteins with known secondary structures are accumulated becomes exponential. It has been publicly accepted that coil (including turn), helix and strand are elementary structural units of protein secondary structures. In the long process of evolution, these units are conserved strongly although the mutation rate of amino acids in proteins is high. So protein secondary structures are important information carriers. It has become urgent and imperative to assign structural classes to these proteins for the reason that it helps to build protein databases and it helps to predict protein function.

The concept of structural classes of proteins was proposed by Levitt and Chothia [1], where the proteins were grouped into four structural classes: all-α class, all-β class, α/β-class and α + β-class. These groups basically characterize the overall structures of proteins even in the up-to-date databases. And this structural classification has been accepted and widely used in protein prediction, in function prediction, etc. Since then, the problem of protein classification has been tackled by many groups, e.g. Nakashima et al. [2] defined the all-α protein as the one in which the content of α-helix is greater than 15% and the content of β-strand is less than 10%; Chou [3] used the amino acid composition of a protein and Mahalanobis distance to assign a protein into one of the four structural classes. These approaches typically use the amino acid composition (AAC) of the protein as the base for classification. More research has been done on the structural classification problem and the related applications, which can be found in [4–18].

According to the definition of the four structural classes, all-α and all-β proteins are defined as to be composed of almost entirely α-helices and β-strands, respectively. The α/β proteins consist of helices and strands that are alternately mixed and the β-strands are often parallel. The α + β proteins consist of the conformation in which α-helices and β-strands are largely separated and the β-strands are often antiparallel. The content of helix/strand is a good index for distinguishing all-α proteins and all-β proteins. It is obvious that the content of helix/strand cannot discriminate the characteristic between α/β proteins and α + β proteins well. So in this paper, we propose a new method to group proteins into appropriate structural classes. To better

☆ The work is supported by the National Natural Science Foundation of China (10571019).
* Corresponding author at: Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, China. Fax: +86 411 84706100.
E-mail address: liunasophia@163.com (N. Liu).

1093-3263/$ – see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2006.08.006

extract structural characteristic, a mathematical model is constructed, from which transition probability matrices can be derived to characterize protein secondary structures. Then structural characteristic vectors are generalized by integrating transition probabilities with the content of elementary structural unit for classifying protein structures. The final classification result is shown schematically. Our test has proved that our method can distinguish different structural classes well. In other words, given a newly determined protein with known secondary structure sequence, our method can help it find its structural class.

2. Methodology

2.1. Building mathematical model and computing transition probability matrices for proteins

A secondary structure sequence [19] is a linear sequence defined over alphabet $\Lambda = \{C, H, E\}$, where H represents helix, E represents strand and the rest are represented by C (mainly coil and turn). Consider the definition of stochastic process: a stochastic process is a collection $\{X(t)|t \in T\}$ of random variables $X(t)$ defined on the probability space (Ω, Γ, P), where T is called index set, Ω represents the sample space that is constituted by all the basic events, Γ represents the event set that is constituted by all possible events and P represents the probability, which is a function defined over Γ. Given any t, the possible values of $X(t)$ are called the states of the process at t. So the secondary structure sequence may be regarded as a realization of a stochastic process. In this stochastic process, the states of the stochastic process are $\{C, H, E\}$ and the index set is a finite ordered sequence of non-negative integer numbers. Then transition probability matrix may be employed to describe a realization of a stochastic process. It records the overall situation that certain state transfers to another state in a realization. The transition probability matrix for proteins is as follows, in the form of Table 1.

Generally different realizations have different transition probability matrices. Obviously, there are $3 \times 3 = 9$ transition probabilities. They are computed by the following formula:

$$p_{a_i a_j} = \frac{n_{a_i a_j}}{\sum_{k=1}^{3} n_{a_i a_k}}, \quad \text{if } \sum_{k=1}^{3} n_{a_i a_k} \neq 0;$$

$$p_{a_i a_j} = 0, \quad \text{if } \sum_{k=1}^{3} n_{a_i a_k} = 0$$

where a_i represents the ith element of alphabet $\{H, E, C\}$; $n_{a_i a_j}$ enumerates the frequency of the incident that letter a_i is followed by letter a_j in a secondary structure sequence.

Table 1
Transition probability matrix for protein secondary structure sequence

Bases	H	E	C
H	p_{HH}	p_{HE}	p_{HC}
E	p_{EH}	p_{EE}	p_{EC}
C	p_{CH}	p_{CE}	p_{CC}

Table 2
Transition probability matrix for the former secondary structure sequence

Bases	H	E	C
H	0.75	0.16667	0.08333
E	0.16667	0.83333	0
C	0	0.4	0.6

Take the following two secondary structure sequences for example. The contents of helix/strand of them are the same: the content of H is 0.5, the content of E is 1/3 and the content of C is 1/6.

Sequence 1: *CCEEEEEHHHHEEEHHHHHEEEEHHHCCCEEEEEEE*

Sequence 2: *HHHHCCHHHHHHCCCEEEEEEEEEEEEEEEEE*

That means they cannot be discriminated by only using the content of helix/strand. In fact they belong to different structural classes. Sequence 1 belongs to α/β-class and Sequence 2 belongs to α + β-class. However, their transition probability matrices are different, as shown in Tables 2 and 3, respectively.

2.2. Constructing structural characteristic vectors for proteins

Our aim to construct structural characteristic vectors is to define an index that can characterize protein structures numerically for classifying proteins into appropriate structural classes. The more information the vector extracts, the better the classification result will be. Since both transition probability matrix and the content of elementary structural units are indices from different perspectives, we integrate them together to define the structural characteristic vector:

$$\mathrm{SCV} = (\ p_{HH} \quad p_{HE} \quad p_{HC} \quad p_{EH} \quad p_{EE} \quad p_{EC} \quad p_{CH} \quad p_{CE} \quad p_{CC} \quad m_H \quad m_E\ )$$

where m_H represents the content of H and m_E represents the content of E.

The structural characteristic vector characterize secondary structure sequences/proteins numerically. They generalize the distribution patterns of elementary structure units in secondary structure sequence and are indices for their corresponding protein structures. For the above-mentioned two secondary structure sequences, their structural characteristic vectors are: SCV = (0.75 0.16667 0.08333 0.16667 0.83333 0 0 0.4 0.6 0.33333 0.52778) and SCV = (0.8 0 0.2 0 1 0 0.2 0.2 0.6 0.33333 0.5).

Table 3
Transition probability matrix for the latter secondary structure sequence

Bases	H	E	C
H	0.8	0	0.2
E	0	1	0
C	0.2	0.2	0.6

2.3. Classifying proteins into appropriate structural classes by using SCVs

Now each secondary structure sequence corresponds to a SCV. SCV is a vector in 11-dimension space. SCV carries protein structural information. Hence the variation between SCVs also indicates the difference between their secondary structures/proteins. In mathematics, the Euclidean distance is widely used to measure the variation degree between vectors. So we compute Euclidean distance between SCVs to measure the structural variation between proteins. For proteins A and B, the variation between them is measured by the following formula:

$$d_{AB} = \sqrt{\sum_{k=1}^{11} (SCV_{Ak} - SCV_{Bk})^2} \quad (1)$$

where SCV_{Ak} is the kth component of structural characteristic vector for protein A.

According to the formula, the smaller d_{AB} is, the less deviation there is between SCV_A and SCV_B and hence the more similar protein A is to protein B in terms of structure.

Given n proteins, their structural characteristic vectors can be calculated easily. Then, using formula (1), the distance between any pair of protein structures can be obtained.

By arranging all the values into a matrix, we will get a pair-wise distance matrix $D = (d_{i,j})$, $i, j = 1, 2, \ldots, n$, which provides the structural variation information on each pair of proteins. By inputting this distance matrix into XLminer programme that is developed by Cytel Software Corp that offers a set of data mining tools (http://www.xlminer.com/RSS), these proteins will be classified into clusters. The proteins that are in the same cluster may be regarded to have the same structural class.

3. Results and discussion

To test the validity of our method, in this section, we apply our method to a set of real data. All the secondary structure sequences in this data are retrieved from PALI database [20]. They are from all-α class, we randomly choose 1dly, 1dlw, 1idra, 1mwb and 5ngka. In all-β class, we choose 1ee8, 1k3a$_1$, 1k82a$_1$, 1liz and 1nnja$_1$. In α/β class, we choose 1l9g, 1mug and 1nio. In α + β class, we choose 1e3pa$_6$, 1e3pa$_7$, 1oysa$_2$, and 1r61. Fig. 1 shows the classification result of these 17 proteins by using our method. We observe that all the proteins that belong to α-class, β-class, α/β-class and α + β-class have been separated well and grouped into respective structural classes accurately. This verifies the validity of our method.

In general, given a newly determined secondary structure sequence, it can be put into appropriate structural class by our method: first, retrieve several representative proteins in α-class, β-class, α/β-class and α + β-class, respectively, from protein database. Then mix these proteins with the newly determined structural sequence to form a new set of data. Now apply our method to this set of data according to the above-introduced rule. The cluster into which the newly determined secondary structure sequence is grouped can tell us the structural class of this newly determined sequence.

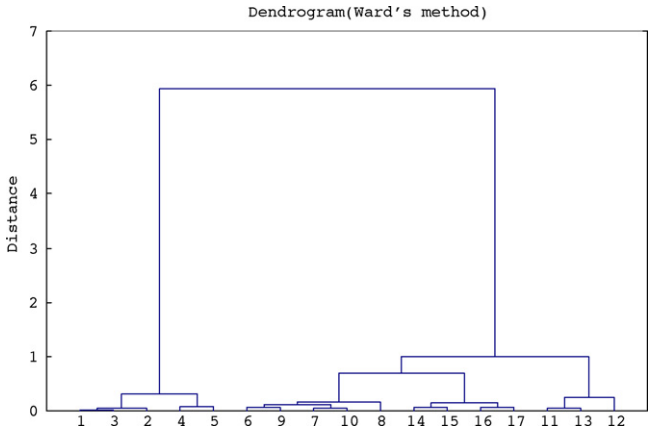

Fig. 1. The dendrogram of the hierarchical clustering for the 17 proteins in terms of structure: 1, 1dlwa; 2, 1dlya; 3, 1idra; 4, 1mwba; 5, 1ngka; 6, 1ee8a$_2$; 7, 1k3xa$_2$; 8, 1k82a$_2$; 9, 1l1za$_2$; 10, 1nnja$_2$; 11, 1l9ga; 12, 1muga; 13, 1ui0a; 14, 1e3pa$_6$; 15, 1e3pa$_7$; 16, 1oysa$_2$; 17, 1r6la$_2$.

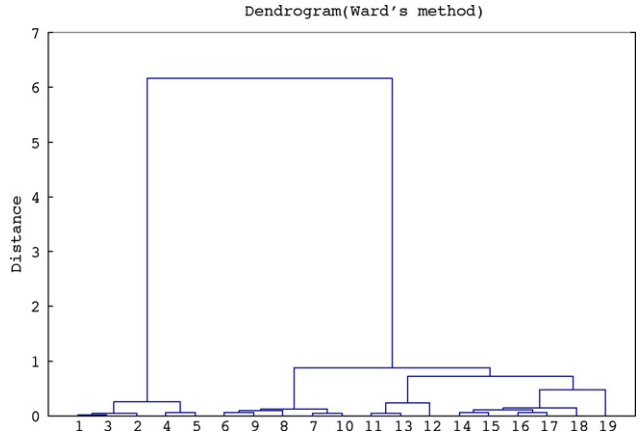

Fig. 2. The dendrogram of the hierarchical clustering for the 19 proteins in terms of structure: 1, 1dlwa; 2, 1dlya; 3, 1idra; 4, 1mwba; 5, 1ngka; 6, 1ee8a$_2$; 7, 1k3xa$_2$; 8, 1k82a$_2$; 9, 1l1za$_2$; 10, 1nnja$_2$; 11, 1l9ga; 12, 1muga; 13, 1ui0a; 14, 1e3pa$_6$; 15, 1e3pa$_7$; 16, 1oysa$_2$; 17, 1r6la$_2$; 18, 1ptf; 19, 1ctf.

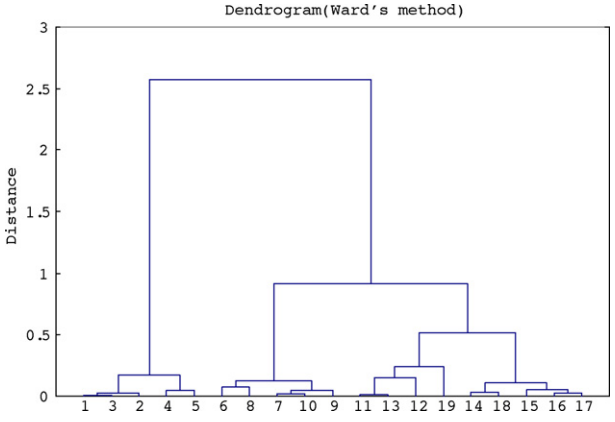

Fig. 3. The dendrogram of the hierarchical clustering for the 19 proteins in terms of structure: 1, 1dlwa; 2, 1dlya; 3, 1idra; 4, 1mwba; 5, 1ngka6, 1ee8a$_2$; 7, 1k3xa$_2$; 8, 1k82a$_2$; 9, 1l1za$_2$; 10, 1nnja$_2$; 11, 1l9ga; 12, 1muga; 13, 1ui0a; 14, 1e3pa$_6$; 15, 1e3pa$_7$; 16, 1oysa$_2$; 17, 1r6la$_2$; 18, 1ptf; 19, 1ctf.

Taking the 17 proteins as representative proteins, we now add another 2 proteins to this data set: lptf and lctf. They belong to α + β-class. Following the steps generalized above, we have correctly assigned them into the cluster that represents α + β-class, see Fig. 2. For comparison, we show the result obtained by the content method, see Fig. 3. We note that this method cannot assign ldtf to α + β-class. Instead it is assigned to the α/β-class.

4. Conclusions

Despite years of research and the wide variety of approaches that have been utilized, the protein folding problem still remains an open problem. Structural class assignment is one of the task. The structural class of a protein has been used in some secondary structure prediction algorithms. Once, the structural class of a protein is known, it can be used to reduce the search space of the structure prediction problem: most of the structure alternatives will be eliminated and the structure prediction task will become easier and faster. Most of the previous approaches use the content of helix/strand as classification criteria. To better distinguish the difference of distribution patterns covered in α/β-class and α + β-class, in this paper, we propose a simple new method for protein structural class assignment. In this method, we construct a mathematical model to describe protein secondary structure sequences, from which transition probability matrices are derived to reveal the patterns covered in secondary structure sequences. Then we define a structural characteristic vector by integrating the information from transition probability matrix with that from the content of helix/strand. Based on the structural characteristic vectors, proteins can be classified into appropriate structural classes. We have chosen a set of proteins from each class, whose structural classes have been known, to see how well our method will classify them into structural classes. The result indicates that our method can classify them into accurate structural classes. Its advantage is that it can distinguish α/β-class and α + β-class better than the approaches that are based only on the content of helix/strand. This is due to the fact that the transition probabilities contain the information on alternately mixed/largely separately patterns of helices-strands. Different from traditional methods, the final classification result obtained by our method is shown schematically which is visual to observe the structural classifications.

There is one point that deserves our attention: our method has not considered the characteristics resulted from the parallel/anti-parallel arrangement of β-strands, which is also the factor that can affect the discrimination between α + β-class and α/β-class. This shortage of course has impact on the accuracy of our method though it performs better than the content method. Thus a question has been left to us: to find appropriate parameters that can explain the parallel/anti-parallel arrangement of β-strands.

Acknowledgement

The authors thank all the anonymous for their valuable suggestions and support.

References

[1] M. Levitt, C. Chothia, Structural patterns in globular proteins, Nature 261 (1976) 552–558.
[2] H. Nakashima, K. Nishikawa, T. Ooi, The folding type of a protein is relevant to the amino acid composition., J. Biochem. 99 (1986) 153–162.
[3] K.C. Chou, A novel approach to predicting protein structural classes in a (20-L)-D amino acid composition space, Proteins 21 (1995) 319–344.
[4] P. Klein, C. Delisi, Prediction of protein structural class from the amino acid sequence, Biopolymers 25 (1986) 1659–1672.
[5] P.Y. Chou, Prediction of protein structural classes from amino acid composition, in: G.D. Fasman (Ed.), Prediction of Protein Structure and the Principles of Protein Conformation, Plenum Press, New York, 1989, pp. 549–586.
[6] C.T. Zhang, K.C. Chou, An optimization approach to predicting protein structural class from amino acid composition, Protein Sci. 1 (1992) 401–408.
[7] B.A. Metfessel, P.N. Saurugger, D.P. Connelly, S.S. Rich, Cross-validation of protein structural class prediction using statistical clustering and neural networks, Protein Sci. 2 (1993) 1171–1182.
[8] J.M. Chandonia, M. Karplus, Neural networks for secondary structure and structural class predictions, Protein Sci. 4 (1995) 275–285.
[9] I. Bahar, A.R. Atilgan, R.L. Jernigan, B. Erman, Understanding the recognition of protein structural classes by amino acid composition, Proteins 29 (1997) 172–185.
[10] K.C. Chou, A key driving force in determination of protein structural classes, Biochem. Biophys. Res. Commun. 264 (1999) 216–224.
[11] Y.D. Cai, G.P. Zhou, Prediction of protein structural classes by neural network, Biochimie 82 (2000) 783–787.
[12] Y.D. Cai, X.J. Liu, X. Xu, K.C. Chou, Prediction of protein structural classes by support vector machines, Comput. Chem. 26 (2002) 293–296.
[13] C.H. Ding, I. Dubchak, Multi-class protein fold recognition using support vector machines and neural networks, Bioinformatics 17 (2001) 349–358.
[14] A.C. Tan, D. Gilbert, Y. Deville, Multi-class protein fold classification using a new ensemble machine learning approach, Genome Inf. 14 (2003) 206–217.
[15] P. Robert, J. Sheridan, R. Scott Dixon, I.D. Venkataraghavan, K.P.S. Kuntz, Amino acid composition and hydrophobicity patterns of protein domains correlate with their structures, Biopolymers 24 (10) (1985) 1995–2023.
[16] D.G. Kneller, F.E. Cohen, R. Langridge, Improvements in protein secondary structure prediction by an enhanced neural network, J. Mol. Biol. 214 (1990) 171–182.
[17] D.M. Alex, A.O. Christine, M.T. Janet, Analysis of domain structural class using an automated class assignment protocol, J. Mol. Biol. 262 (1996) 168–185.
[18] G. Deleage, J. Dixon, Use of class prediction to improve protein secondary structure prediction, in: C.D. Fasman (Ed.), Prediction of Protein Structure and the Principles of Protein Conformation, Plenum Press, New York, 1989, pp. 587–597.
[19] C.T. Zhang, R. Zhang, S curve, a graphic representation of protein secondary structure sequence and its application, Biopolymers 53 (2000) 539–549.
[20] V.S. Gowri, S.B. Pandit, P.S. Karthik, N. Srinivasan, S. Balaji, Integration of related sequences with protein three-dimensional structural families in an updated Version of PALI database, Nucl. Acids Res. 31 (2003) 486–488.