A New Encoding Scheme to Improve the Performance of Protein Structural Class Prediction

Zhen-Hui Zhang[1], Zheng-Hua Wang[2], and Yong-Xian Wang[2]

[1] Institute of Science, National University of Defense Technology,
410073 Changsha, China
zhangzhenhui2000@163.com

[2] Institute of Computer, National University of Defense Technology,
410073 Changsha, China
yongxian_wang@yahoo.com

Abstract. Based on the concept of coarse-grained description, a new encoding scheme with grouped weight for protein sequence is presented in this paper. By integrating the new scheme with the component-coupled algorithm, the overall prediction accuracy of protein structural class is significantly improved. For the same training dataset consisting of 359 proteins, the overall prediction accuracy achieved by the new method is 7% higher than that based solely on the amino-acid composition for the jackknife test. Especially for $\alpha+\beta$ the increase of prediction accuracy can achieve 15%. For the jackknife test, the overall prediction accuracy by the proposed scheme can reach 91.09%, which implies that a significant improvement has been achieved by making full use of the information contained in the protein sequence. Furthermore, the experimental analysis shows that the improvement depends on the size of the training dataset and the number of groups.

1 Introduction

It is generally accepted that protein structure is determined by its amino acid sequence [1] and that the knowledge of protein structures plays an important role in understanding their functions. Understanding the relation between amino acid sequence and three-dimensional protein structure is one of the major goals of contemporary molecular biology. A priori knowledge of protein structural classes has become quite useful from both an experimental and theoretical point of view. The concept of protein structural classes was proposed by Levitt and Chothia more than 20 years ago [2]. According to this concept, a protein is usually classified into one of the following structural classes: all-α, all-β, α/β and $\alpha + \beta$. The structural class of a protein presents an intuitive description of its overall folding and the restrictions of the structural class have a high impact on its secondary and tertiary structure prediction [3]. Some researchers have claimed that the knowledge of structural classes might be used to decrease the complexity of searching conformational space during energy optimization, and

L. Wang, K. Chen, and Y.S. Ong (Eds.): ICNC 2005, LNCS 3611, pp. 1164–1173, 2005.

provide useful information for a heuristic approach to find the tertiary structure of a protein. Owing to the importance and the relative simplicity of structural class prediction, considerable attention has been focused on this problem during the past years [3-11].

Historically, Nishikawa's found [4] that structural classes of proteins correlate strongly with amino acid composition. It marked the onset of algorithm developments aimed at predicting the structural class of a protein from its amino acid composition solely. There have been a number of algorithms about this topic, such as the least Hamming distance, the least Euclidian distance, the discriminate analysis, the vector decomposition, the component-coupled algorithm, and fuzzy structural vectors. Although the amino-acid composition is very convenient to calculate, the full information contained in the protein sequence is reduced considerably. The prediction accuracy is limited by the amino-acid composition-based approach. It is the aim of this study to overcome this drawback.

Based on the concept of coarse-grained description, a protein sequence was reduced to a few of binary sequences, which we named characteristic sequences. For each characteristic sequence, a canonical weight function was introduced to realize the grouped weight encoding of protein sequence. We name this new encoding approach of protein sequence as EBGW (Encoding Based on Grouped Weight) approach. Integrating the new scheme (EBGW) with the component-coupled algorithm, it shows that the overall prediction accuracy of protein structural class is significantly improved. Furthermore, the methodology presented here might be useful for other studies of protein structure.

2 Methods

For many quite different things, we can treat them as one if they have some same characters. This is the main idea of coarse-grained and was applied to DNA sequence analysis in [12]. It is well known that the three-dimensional structure of protein is more conservative than its protein sequence. In the process of folding, the insertion, deletion or permutation of single amino acid residue may not destroy the three-dimensional structure. The most important influencing factor of protein folding is the unique character of amino acid residue. Thus ,in the following, we present a new encdoing scheme (named EBGW) of amino acid sequence based on the different character of amino acid residue and coarse-grained idea.

2.1 EBGW of Protein Sequence

Considering the hydrophobicity and charged character, we can divide the 20 amino acid residues into four different classes as follows [13]:

neuter and non-polarity residue	$C1=\{G,A,V,L,I,M,P,F,W\}$
neuter and polarity residue	$C2=\{Q,N,S,T,Y,C\}$
acidic residue	$C3=\{D,E\}$
alkalescence residue	$C4=\{H,K,R\}$

Thus, we can get three combinations, each of which can partition the 20 amino acid residues into two disjoint group: C1+C2 vs C3+C4, or C1+C3 vs C2+C4, and C1+C4 vs C2+C3.

Definition 1. *(Characteristic Sequence) Let $A(n) = a_1a_2\cdots a_n$ be a protein sequence, we can transform it into three binary sequences by three homomorphic maps $\Phi_i(A(n)) = \Phi_i(a_1)\Phi_i(a_2)\cdots\Phi_i(a_n) \quad (i = 1, 2, 3)$ which are defined as follows:*

$$\Phi_1(a_j) = \begin{cases} 1 & if\ a_j \in C1 \cup C2 \\ 0 & if\ a_j \in C3 \cup C4 \end{cases} \quad (j = 1, 2, \cdots, n) \tag{1}$$

$$\Phi_2(a_j) = \begin{cases} 1 & if\ a_j \in C1 \cup C3 \\ 0 & if\ a_j \in C2 \cup C4 \end{cases} \quad (j = 1, 2, \cdots, n) \tag{2}$$

$$\Phi_3(a_j) = \begin{cases} 1 & if\ a_j \in C1 \cup C4 \\ 0 & if\ a_j \in C2 \cup C3 \end{cases} \quad (j = 1, 2, \cdots, n) \tag{3}$$

Denote $H(n)^i = \Phi_i(A(n)) = h_1^ih_2^i\cdots h_n^i$ $(i = 1, 2, 3)$, we call $H(n)^1, H(n)^2$, $H(n)^3$ as 1-,2- and 3-characteristic sequences of the protein sequence, respectively.

For simplicity, in the following text we denote $H(n) = h_1h_2\cdots h_n$ as any characteristic sequence of three defined above.

Definition 2. *(Weight) Let $H(n) = h_1h_2\cdots h_n$ be a characteristic sequence, the weight of $H(n)$ is defined as the enumeration of digit 1 in $H(n)$.*

We can see that the weight of characteristic sequence is dependent on the sequence length. So it could not be applied to the comparison or analysis of sequences with different lengths.

Definition 3. *(Canonical Weight) Let $H(n) = h_1h_2\cdots h_n$ be a characteristic sequence, the canonical weight $w(n)$ is defined as the frequency of digit 1 occurs in $H(n)$, that is $w(n) = p/n$, where p is the weight of $H(n)$.*

Definition 4. *(Encoding Based on Grouped Weight) Let $H(n) = h_1h_2\cdots h_n$ be a characteristic sequence, assume L be a positive integer, we can partition $H(n)$ into L pieces of subsequence. The process of subsequence partitioning can refer to Figure 1. From Figure 1 we know that the length of each subsequence is progressive increase. Let $H(\lfloor kn/L \rfloor)$ $(k = 1, 2, \cdots, L)$ be subsequences of $H(n)$ whose length are $\lfloor kn/L \rfloor$ $(k = 1, 2, \cdots, L)$, where$\lfloor\bullet\rfloor$ is the operation returning a number down to the nearest integer, and $w(\lfloor kn/L \rfloor)$ $(k = 1, 2, \cdots, L)$ be the canonical weight of $H(\lfloor kn/L \rfloor)$ $(k = 1, 2, \cdots, L)$, we can get $W = [w(\lfloor n/L \rfloor), w(\lfloor 2n/L \rfloor), \cdots w(\lfloor Ln/L \rfloor)]$ which we call as the EBGW string of characteristic sequence $H(n)$.*

Thus, given a protein sequence $A(n) = a_1a_2\cdots a_n$, we can transform it into three characteristic sequences$H(n)^1, H(n)^2, H(n)^3$ by using definition 1.

$H(n)$	10101100101111001101001000010001010011011010111000010101011101111
$H(10)$	1010110010
$H(21)$	101011001011110011010
$H(32)$	10101100101111001101001000010001
$H(43)$	1010110010111100110100100001000101001101101
$H(54)$	101011001011110011010010000100010100110110101110000101
$H(65)$	10101100101111001101001000010001010011011010111000010101011101111

n=65 L=6 length of subsequence is 10,21,32,43,54,65 respectively

Fig. 1. Partitioning subsequence of characteristic sequence

For each characteristic sequence $H(n)^i (i = 1, 2, 3)$, it can be encoded into a L-dimension vector $W^i (i = 1, 2, 3)$ with definition 4. That is, we can transform a protein sequence into a 3L-dimension vector $X = [W^1, W^2, W^3] = [x_1, x_2, \cdots x_{3L}]$, we call x as the EBGW string of protein sequence A.

In EBGW approach, characteristic sequence is introduced based on the concept of coarse-grained. It reflects the distribution of residues with the same unique characteristic and portrays the essence of protein sequence. Although the amino-acid composition is very convenient to calculate, the information contained in the protein sequence is reduced considerably. In EBGW approach, grouping presented can contain more information in the protein sequence. If grouping based on the amino acid composition, a protein sequence can be transformed into a 20L-dimension vector, where L is the number of groups. However, grouping based on characteristic sequence, a protein sequence can be transformed into a 3L-dimension vector. The computational complexity is largely decreased. From definition 4, we know that the larger the value of L used, the more information of EBGW approach contained, and the higher accuracy of test reached. On the other hand, information may be less when L equals the length of protein sequence. So the optimal value of L should be carefully chosen for different dataset.

2.2 Component-Coupled Algorithm

Suppose there are N proteins forming a set S, i.e.

$$S = S^{\alpha} \cup S^{\beta} \cup S^{\alpha+\beta} \cup S^{\alpha/\beta} \tag{4}$$

where the subset S^{α} consists of only all-α proteins, the subset S^{β} consists of only all-β proteins, and so forth. According to the EBGW approach, any protein in the set S corresponds to a vector (or a point) in the 3L-dimension space, i.e.

$$X_k^{\xi} = [x_{k,\,1}^{\xi}, x_{k,\,2}^{\xi}, \cdots, x_{k,\,3L}^{\xi}] \quad (k = 1, 2, \cdots, N_{\xi}) \tag{5}$$

where $\xi = \alpha, \beta, \alpha + \beta, \alpha/\beta$ denotes one of the four different structural classes and N_{ξ} is the number of proteins in the subset ξ.

The standard vector for the subset $S^{\xi}$is defined by

$$X^{\xi} = [x_1^{\xi}, x_2^{\xi}, \cdots, x_{3L}^{\xi}] \tag{6}$$

where

$$x_i^{\xi} = \frac{1}{N_{\xi}} \sum_{k=1}^{N_{\xi}} x_{k,i}^{\xi} \quad i = 1, 2, \cdots 3L \tag{7}$$

Suppose X is a protein whose structural class is to be predicted. It can be either one of the N proteins in the set S or a protein outside it. It also corresponds to a point $[x_1, x_2, \cdots x_{3L}]$ in the 3L-dimension space with EBGW approach.

The component-coupled algorithm is based on the squared Mahalanobis distance, defined by

$$F_M^2(X, X^{\xi}) = (X - X^{\xi}) C_{\xi}^{-1} (X - X^{\xi})^T + \ln \mathrm{K}_{\xi} \tag{8}$$

where $C_{\xi} = (c_{i,j}^{\xi})_{3L \times 3L}$ is a covariance matrix given by

$$C_{\xi} = \begin{bmatrix} c_{1,1}^{\xi} & c_{1,2}^{\xi} & \cdots & c_{1,3L}^{\xi} \\ c_{2,1}^{\xi} & c_{2,2}^{\xi} & \cdots & c_{2,3L}^{\xi} \\ \vdots & \vdots & \ddots & \vdots \\ c_{3L,1}^{\xi} & c_{3L,2}^{\xi} & \cdots & c_{3L,3L}^{\xi} \end{bmatrix}$$

and the superscript T is the transposition operator; C_{ξ}^{-1} is the inverse matrix of C_{ξ}. The matrix elements $C_{i,j}^{\xi}$ are given by

$$c_{i,j}^{\xi} = \frac{1}{N_{\xi} - 1} \sum_{k=1}^{N_{\xi}} [x_{k,i}^{\xi} - x_i^{\xi}][x_{k,j}^{\xi} - x_j^{\xi}] \; (i, j = 1, 2, \cdots 3L) \tag{9}$$

K_{ξ} is the product of all positive eigenvalues of C_{ξ}. The target protein X is predicted to be the structural class for which the corresponding Mahalanobis distance has the least value, as can be formulated as follow

$$F_M^2(X, X^{\lambda}) = \min \left\{ F_M^2(X, X^{\alpha}), F_M^2(X, X^{\beta}), F_M^2(X, X^{\alpha+\beta}), F_M^2(X, X^{\alpha/\beta}) \right\} \tag{10}$$

where λ can be α,β,α/β or $\alpha+\beta$ and the superscript λ in Equation (10) will give the subset (or structural class) to which the predicted protein X should belong.

2.3 Evaluation of the Prediction Results

In order to assess the accuracy of a prediction algorithm, the sensitivity for each type is calculated according to Baldi et al [14]. Evaluating a given prediction method is a common but quite subtle problem. Usually, a prediction method is evaluated by the prediction results for a training data set and testing data set, respectively. According to the statistical terminology, the former is called a test of resubstitution reflecting the self-consistency, and the latter is a test of cross-validation reflecting the extrapolating effectiveness of the algorithm studied. As is well known, the single-test-set analysis, sub-sampling and jackknife analysis are the three methods often used for cross-validation examination[5]. In the single-test-set examination, the selection of a testing dataset is arbitrary, and the accuracy thus obtained lacks an objective criterion unless the training database is an ideal one and the testing dataset is sufficiently large. Another approach for cross-validation is sub-sampling analysis, according to which a given dataset is divided into a training set and a testing set. However, how to divide the whole dataset into a training set and a testing set is a serious problem. The number of possible divisions might be extremely large. In comparison with the single-set-test examination and the sub-sampling analysis, the jackknife test, also called the leave-one-out test seems to be most effective. In the jackknife test, each domain in the dataset is singled out in turn as a test domain and all the rule-parameters are determined from the remaining domains. Hence the memorization effects that are included in the resubstitution tests can be completely removed. During the process of jackknife analysis, both the training and testing datasets are actually open, and a domain will in turn move from each to the other. Both tests of resubstitution and jackknife are used to evaluate the new prediction method proposed here.

3 Results and Discussion

3.1 Dataset

To facilitate the comparison between our approach and the amino-acid composi tion- based approach, the same datasets and algorithm used by Chou and Maggiora [5] are used here. In their work several datasets were selected from structural classification of proteins (SCOP) [15] for the study of a four-class prediction. These datasets consist of 253, 359, 225 and 510 proteins respectively. The datasets T359 and T253 are mainly used here for comparison. Furthermore, the datasets T225 and T510 are also used as a practical application. The Protein DataBank codes of these proteins are referred to [5], and the constructions of all the datasets are listed in Table 1.

Table 1. Datasets used in this paper

Dataset	The number of sequences in different classes				Total
	$all-\alpha$	$all-\beta$	α/β	$\alpha+\beta$	
T253	63	58	61	71	253
T225	61	45	56	63	225
T359	82	85	99	93	359
T510	109	130	135	136	510

3.2 Prediction Results

The prediction results for dataset T359 are listed in Table 2. For convenience, we abbreviate the two approaches as follows: AAC, the amino acid composition-based approach; EBGW, encode based on grouped weight approach. In the following tables, we also abbreviate Resb and Jack as the Resubstitution test and the jackknife test respectively. As seen from Table 2, the overall prediction accuracy achieved by EBGW is 5% higher than AAC for the Resubstitution test. Meanwhile, the overall prediction accuracy achieved by EBGW is about 7% higher than AAC for the jackknife test. As the jackknife test is thought of a rigorous cross-validation, the improvement of the overall prediction accuracy for the jackknife test is considered remarkable. Carefully analysis the data in Table 2, we find that the prediction accuracy for each class is improved. Especially for the protein structure class of $\alpha + \beta$ the increase of prediction accuracy can achieve about 15%. Note that the above results for EBGW approach is dependent on the datasets and the number of groups L adopted (L=13 here). We will discuss this point as well as several other points below.

Table 2. Prediction results for dataset T359 using EBGW (L=13)

Method/test	Prediction accuracy for each class				Overall accuracy
	$all - \alpha$	$all - \beta$	α/β	$\alpha + \beta$	
EBGW/Resb	100%	100%	100%	98.92%	**99.72**%
AAC[5]/Resb	93.90%	94.12%	95.96%	93.55%	94.43%
EBGW/Jack	95.12%	85.88%	89.90%	93.55%	**91.09**%
AAC[5]/Jack	89.02%	83.53%	85.86%	78.49%	84.12%

3.3 The Optimal Choice of the Number of Groups

The number of groups is denoted as L in definition 4. Usually, the larger the value of L is used, the higher the accuracy of the resubstitution test can get. However, our study shows that a great number of groups do not always lead to a better prediction result for the jackknife test. For the dataset T359 we find that L=13 leads to the highest prediction accuracy of jackknife test, i.e. 327/359=91.09%, while for the dataset T510 we find that L= 14 is the best choice. We should point out that the optimal L value is dependent on the dataset. For the different datasets discuss here, the optimal L value are found to vary from 7 to 14 (see Table 2-6).

3.4 The Impact of the Size of Dataset to the Prediction Accuracy

We also try to discover how the size of dataset can affect the prediction accuracy. A smaller dataset consisting of 253 proteins was used early in Chou et al [5], which has less overlap with those of dataset T359. The same prediction was performed for dataset T253, the results are shown in Table 4(L=8). The overall prediction accuracy of the resubstitution test for EBGW is about 0.79% higher

Table 3. Optimal choice of the number of groups for dataset T359

L	Prediction accuracy for each class in jackknife test				Overall accuracy
	all − α	*all* − β	α/β	$\alpha+\beta$	
10	87.80%	82.35%	86.87%	89.25%	86.63%
11	91.46%	85.88%	80.81%	92.47%	87.47%
12	90.24%	83.53%	88.89%	91.40%	88.58%
13	95.12%	85.88%	89.90%	93.55%	**91.09%**
14	98.78%	83.53%	87.88%	88.17%	89.42%
15	98.78%	72.94%	87.88%	88.17%	86.91%

than that for AAC, whereas the overall prediction accuracy of the jackknife test for EBGW is about 2.77% higher than that for AAC. Because the component-coupled algorithm needs more training data to make its prediction mechanism work properly, the decrease in the improvement (from 7% to about 3%) may be caused by the smaller size of the dataset T253. We should point out that in the case of the smaller dataset T253, the optimal number of groups used is changed, here we find that L=8 leads to the highest overall prediction accuracy.

Table 4. Prediction results for dataset T253 using EBGW (L=8)

Method/test	Prediction accuracy for each class				Overall accuracy
	all − α	*all* − β	α/β	$\alpha+\beta$	
EBGW/Resb	90.48%	98.28%	96.72%	98.59%	**96.05**%
AAC[5]/Resb	95.24%	93.10%	98.36%	94.37%	95.26%
EBGW/Jack	82.54%	75.86%	75.41%	91.55%	**81.82**%
AAC[5]/Jack	84.13%	79.31%	70.49%	81.69%	79.05%

To test the new approach for a larger dataset, another dataset, which was used in Chou [5], consisting of 510 proteins extracted from SCOP is used here. Performing exactly the same prediction as for the dataset T359, the detailed prediction results are shown in Table 5 (L=14). The overall prediction accuracy of the jackknife test for EBGW is 91.96%, indicating that a higher overall prediction accuracy is achieved with EBGW approach. This prediction confirms again the point of view that in order to work properly, the new method needs a much larger training dataset.

Table 5. Prediction results for dataset T510 using EBGW (L=14)

Method/test	Prediction accuracy for each class				Overall accuracy
	all − α	*all* − β	α/β	$\alpha+\beta$	
EBGW/Resb	100%	100%	99.26%	100%	**99.80**%
EBGW/Jack	91.74%	88.46%	91.11%	96.32%	**91.96**%

3.5 Application

The prediction quality of EBGW approach can be improved has been demonstrated above through both resubstitution and jackknife tests. Here, a practical application is presented to indicate the consistency of this kind of improvement.

The procedure consists of the following two steps: (1) constructing a training dataset from which the prediction-rule-parameters are derived; (2) constructing an independent testing dataset for which the prediction is performed using the parameters derived from the training dataset. Another two datasets, which used in Chou [5], consisting of 225 and 510 proteins extracted from SCOP are used here as training dataset and testing dataset respectively. By following the same prediction procedure, the structural classes of the 510 proteins in the testing dataset can be predicted based on the parameters derived from the 225 proteins in the training dataset. The prediction results are summarized in Table 6, from which it can be seen that the overall prediction rate of correct prediction for the independent testing dataset by EBGW approach is about 3% higher than those by AAC approach. This is consistent with the jackknife test results demonstrated in the previous section.

Table 6. Prediction results for dataset T510 as testing dataset using EBGW (L=11)

Method	Prediction accuracy for each class				Overall accuracy
	$all-\alpha$	$all-\beta$	α/β	$\alpha+\beta$	
EBGW	82.57%	93.08%	85.93%	94.85%	89.41%
AAC[5]	74.31%	90.00%	91.85%	87.50%	86.47%

4 Conclusions

Instead of the approach based on amino acid composition solely, a new encoding scheme named EBGW is presented in this paper. Applying EBGW approach to some non-redundant datasets with component-coupled algorithm, considerable improvements in the overall prediction accuracy are achieved compared with the AAC approach. The experiment results show that EBGW approach is convenient to calculate and provides an effective tool to extract valuable information from protein sequences, which may be a useful tool in other assignment problems in proteomics and genome research.

References

1. Anfinsen C.B: Principles that govern the folding of protein chains. Science. **181** (1973) 223-230
2. Levitt M, Chothia C: Structure patterns in globular proteins. Nature. **262** (1976) 552-557
3. Chou K.C, Zhang C.T: Prediction of protein structural classes. Crit.Rev.Biochem.Mol.Biol. **30** (1995) 275-349

4. Nakashima H, Nishikawa K, Ooi T: The folding type of a protein is relevant to the amino acid composition. J.Biochem. **99** (1986) 152-162
5. Chou K.C, Maggiora G.M: Domain structural class prediction. Protein Engineering.**11**(1998) 523-538
6. Bu W.S, Feng Z.P, Zhang Z.D, Zhang C.T: Prediction of protein (domain) structural classes based on amino-acid index. Eur. J. Biochem.**266** (1999) 1043-1049
7. Li X.Q, Luo L.F: The definition and recognition of protein structural class. Progress in Biochemistry and Biophysics. **29**(2002) 124-127
8. Li X.Q, Luo L.F: The recognition of protein structural class. Progress in Biochemistry and Biophysics.**29** (2002) 938-941
9. Wang Z.X, Yuan Z: How good is prediction of protein structural class by the component-coupled method?. Proteins.**38** (2000) 165-175
10. Cai Y.D, Liu X.J, Xu X.B, Zhou G.P: Support Vector Machines for predicting protein structural class. BMC Bioinformatics. **2**(2001) 3
11. Luo R.Y, Feng Z.P, Liu J.K: Prediction of protein structural class by amino acid and ploypeptide composition. Eur.J.Biochem. **269**(2002) 4219-4225
12. He P.A, Wang J: Numerical characterization of DNA primary sequence. Internet Electronic Journal of Molecular Design. **1** (2002) 668-674
13. Lin J.C, Yang K.C: Biochemistry. Shenyang: liaoning science and technology press. (1996) 6-7
14. Baldi P, Brunak S, Chauvin Y, Andersen C.A, Nielsen H: Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics.**16** (2000) 412-424
15. Loredana L.C, Steven E.B, Tim J.P.H, Cyrus C, Alexey G.M: SCOP dataset in 2002: refinements accommodate structural genomics. Nucleic Acids Research. **30**(2002) 264-267