Abstract: The structural class is an important feature widely used to characterize the overall folding type of a protein. How to improve the prediction quality for protein structural classification by effectively incorporating the sequence-order effects is an important and challenging problem. Based on the concept of the pseudo amino acid composition [Chou, K. C. Proteins Struct Funct Genet 2001, 43, 246; Erratum: Proteins Struct Funct Genet 2001, 44, 60], a novel approach for measuring the complexity of a protein sequence was introduced. The advantage by incorporating the complexity measure factor into the pseudo amino acid composition as one of its components is that it can catch the essence of the overall sequence pattern of a protein and hence more effectively reflect its sequence-order effects. It was demonstrated thru the jackknife crossvalidation test that the overall success rate by the new approach was significantly higher than those by the others. It has not escaped our notice that the introduction of the complexity measure factor can also be used to improve the prediction quality for, among many other protein attributes, subcellular localization, enzyme family class, membrane protein type, and G-protein couple receptor type.


Key words: pseudo amino acid composition; complexity measure factor; covariant–discriminant algorithm; invariance theorem

Introduction

One of the critical challenges in science is how to reveal some simple or regular patterns from extremely complicated or highly irregular phenomena, and apply them to predict the desired but still unknown information. The protein structure classification and its prediction are a typical paradigm in this regard. Although the details of the 3D structures of proteins seem extremely complicated and irregular, their overall topological folding patterns are surprisingly simple and regular. Picturized properly, proteins are actually strikingly beautiful from the aesthetical point of view. proteins often have quite similar or identical folding patterns even if they consist of quite different sequences or bear various biological functions. In view of this, about 3 decades ago Levitt and Chothia tried to classify proteins into the following four structural classes: (1) all-α, (2) all-β, (3) α/β, and (4) α + β. The all-α and all-β proteins are essentially formed by α-helices (Fig.1a) and β-strands (Fig.1b), respectively. The α/β class represents those proteins containing both α-helices and β-strands that are largely interspersed in forming mainly parallel β-sheets (Fig. 1c), while the α + β class represents those containing also both α-helices and β-strands but they are largely segregated in forming mainly antiparallel β-sheets (Fig. 1d). Ever since its introduction, the structural class has become an important attribute for characterizing the overall folding type of a protein.

Correspondence to: K.-C. Chou; e-mail: kchou@san.rr.com

Contract/grant sponsor: the Doctoral Foundation from the National Education Committee, China; contract/grant number: 20030255009

This article contains Supplementary Material available at http://www.interscience.wiley.com/jpages/0192-8651/suppmat

© 2006 Wiley Periodicals, Inc.
Prediction of protein structural class is an important topic in protein science (see, e.g., a review, refs. 6 and 7). Many different methods were proposed aimed at such a topic.8–26 Most of these methods were based on the amino acid composition, where the sample of a protein is represented by 20 discrete numbers, with each representing the occurrence frequency of one of the 20 constituent native amino acids (see, e.g., refs. 8, 9, and 12). Obviously, if one uses the conventional amino acid composition to represent the sample of a protein, all its sequence order and length effects are lost. To include these effects, the concept of pseudo amino acid composition was introduced.27 It consists of 20 numbers: the first 20 numbers are none but the 20 components of the conventional amino acid composition; those from 20 + 1 to 20 + λ represent λ factors or functions derived from the sequence of a given protein sequence.27 It is through the additional λ factors that some sequence order and length effects can be incorporated. Ever since the concept of pseudo amino acid composition was introduced, various approaches have been proposed to derive the additional λ components.28–54 Generally speaking, the more the additional λ components, the more the sequence-order effects are incorporated in the pseudo amino acid composition. However, if there are too many additional components, the cluster-tolerant capacity35 will be reduced so as to diminish the success rate of cross-validation. Accordingly, to further improve the prediction quality, the pseudo amino acid composition should be optimized by reducing the number of its additional components and increasing the sequence-order information in the remaining components. But how can we realize this? The present study was initiated in an attempt to approach the problem by introducing the “complexity measure factor” into the pseudo amino acid composition. By doing so, the number of the additional components can be significantly reduced, and yet a considerable amount of information for the sequence order and length can be effectively incorporated.

Method

A protein sequence is actually a symbolic sequence for which the complexity measure factor can be used to reflect its sequence feature or pattern.30 Among known measures of complexity, the Lempel–Ziv (LZ) complexity reflects the order that is retained in the sequence, and hence, was adopted in this study.36 Below, let us first introduce some basic definition about LZ complexity.

Figure 1. Ribbon drawings to show the four structural classes of proteins: (a) all-α, (b) all-β, (c) α/β, and (d) α + β. Reproduced from ref. 6 with permission.

Figure 2. Radar diagrams to show the distinction among the four 21D standard vectors, that is the 21 average components in the pseudo amino acid compositions for the proteins in the following structural classes: (a) all-α, (b) all-β, (c) α/β, and (d) α + β. Here, we use the numerical indexes 1, 2, 3, . . . , 20 to denote the 20 amino acids according to the alphabetical order of their single character codes, and use the index 21 to denote the complexity measure factor.
The LZ complexity of a sequence can be measured by the minimal number of steps required for its synthesis in a certain process. For each step only two operations were allowed in the process: either generating an additional symbol, which ensures the uniqueness of each component or copying the longest fragment from the part of a synthesized sequence. Its substring is expressed by

\[ S[i:j] = \alpha_i \alpha_{i+1} \alpha_{i+2} \cdots \alpha_j \ (1 \leq i < j \leq N). \]  

(1)

The complexity measure factor, \( C_{LS}(S) \), of a nonempty sequence \( S \) synthesized according to the following procedure is defined by the minimal number of steps

\[ H(S) = S[1:i_1]S[i_1+1:i_2] \cdots S[i_{n-1}+1:N]. \]  

(2)

Let us assume that \( S = \alpha_1 \alpha_2 \alpha_3 \cdots \alpha_N \) has been reconstructed by the program up to the digit \( a_{i_1} \) and \( a_{i_2} \) has been newly inserted. The string up to \( a_{i_2} \) will be denoted by \( S[1:i_2] \). where the dot denotes that \( a_{i_2} \) is newly inserted to check whether the rest of the string \( S[i_2+1:N] \) can be reconstructed by a simple copying. First, suppose \( q = a_{i_2-1} \), and see whether \( q \) is reproducible from \( S[1:i_2] \). If the answer is “no,” then we insert \( q = a_{i_2-1} \) into the sequence followed by a dot. Thus, it could not be obtained by the copying operation. If the answer is “yes,” then no new symbol is needed and we can go on to proceed with \( q = a_{i_2} \). and repeat the same procedure. The LZ complexity is the number of dots (plus one if the string is not terminated by a dot). For example, for the string \( S = 0001101001000101 \), the LZ schema of synthesis generates the following components \( H(S) \) and the corresponding complexity \( C_{LS}(S) \):

\[
H(S) = 0001 \cdot 10 \cdot 100 \cdot 1000 \cdot 101
\]

(3)

\[
C_{LS}(S) = 6
\]

implying that the complexity measure factor for the string \( S = 0001101001000101 \) is 6. Listed in Table 1 are the three different codes used to represent the 20 native amino acids. The digit codes adopted there can better reflect the chemical physical properties of an amino acid, as well as its structure and degeneracy. Thus, according to Table 1, a protein sequence can be converted to a series of digital signals by following the above procedure and hence define the value of its \( C_{LS}(S) \). The complexity measure factor thus obtained is used to represent one additional component in formulating the pseudo amino acid composition of that protein.

Now, by following exactly same procedure as described by Chou, a protein \( P \) can be expressed by a vector or a point in a \((20 + \lambda)D = (20 + 1)D = 21D \) space; that is,

\[
P = (p_1, p_2, \cdots, p_{20}, p_{21})^T
\]

(4)

where \( T \) is the transpose operator, and

\[
p_k = \begin{cases} 
\sum_{i=1}^{20} f_i & (1 \leq k \leq 20) \\
\sum_{i=1}^{20} f_i + \sum_{i=1}^{20} w_{fi} & (k = 21)
\end{cases}
\]

(5)

where \( f_i \) is the occurrence frequencies of the 20 native amino acids in a protein, \( w_{fi} = C_{LS}(S) \) is the complexity measure factor for the protein sequence concerned, and \( w \) the...
weight factor. In the present study, the weight factor was set at
\( w = 1/800 \) to make the values of the pseudo amino acid com-
ponents in a region easier to be handled. For readers’ convenience,
the values of the 21 pseudo amino acid components thus obtained
for the 204 proteins investigated here are given in the Online
Supporting Materials A.

Now we can directly use the augmented co-
variant discriminant algorithm to perform the prediction. The covariant
discriminant algorithm\(^{42} \) is a combination of Mahalanobis dis-
tance\(^{43,44} \) and the invariance principle for treating degenerative
space\(^{15} \) that is cited in literature as “Chou’s invariance theo-
rem” (see, e.g., refs. 28 and 45). It is instructive to point out here
that, because of the normalization condition imposed by
eq. (5), the 21 components of the pseudo amino acid composi-
tion are not independent. Therefore, a dimension-reduced op-
eration by leaving out one of the components and making the
rest completely independent is needed when using the aug-
mented covariant discriminant algorithm; that is, a protein
should be defined in a 20D space instead of 21D space. Other-
wise, a divergence difficulty will occur. However, which one of
the 21D components should be removed? The answer is any of
them. The reason is that according to the aforementioned in-
variance theorem,\(^{15} \) the predicted results will remain the same
regardless of which one of the 21 components is left out.

Results and Discussion

As a demonstration, let us use the same dataset studied by
the previous authors.\(^{22,35} \) It consists of 204 proteins, of which 52 all-\( \alpha \),
61 all-\( \beta \), 45 \( \alpha/\beta \), and 46 \( \alpha + \beta \). Their PDB codes are given in
Table 2 of Chou.\(^{35} \)

The power of a statistical prediction method is usually evalu-
ated by the resubstitution test, independent dataset test, and jack-
knife test. Of these three, the jackknife test is deemed the most
rigorous and objective,\(^{16,18,19,45} \) and hence, was adopted for the
current study. The success rates by jackknife test for the afore-
mentioned 204 proteins classified into four structural classes are
given in Table 2, where for facilitating comparison the correspond-
ing rates obtained by the recently developed algorithms, such as the
correlation analysis approach\(^{41} \) and supervised fuzzy clustering
approach,\(^{22} \) are also listed. It can be seen from Table 2 that the
overall success rate by the current approach is 89.7%, which is
remarkably higher than those by the other approaches.

Why could the overall success rate be improved so much by
introducing the complexity measure factor? To address this
problem, let us consider the standard vectors for the four structural
classes, \( P_{\alpha}, P_{\beta}, P_{\alpha/\beta}, \) and \( P_{\alpha+\beta}, \) as defined in ref. 27. Each of the
four standard vectors in the current approach contains 21 com-
ponents (cf. eqs. (4)–(5)), which can be easily derived from the data
in the Online Supporting Materials A. To provide an intuitive
picture, each such 21D standard vector is projected onto a 2D radar
diagram as given in Figure 2, from which we can see that, by
introducing the complexity measure factor into the expression for
protein samples, the standard vectors for the four structural classes
have become remarkably distinct from each other. In contrast to
this, the distinction between \( P_{\alpha/\beta} \) and \( P_{\alpha+\beta} \) is trivial if they are
defined in a 20D space according to the conventional amino acid
composition, as shown in Figure 1 of Du et al.\(^{46} \) That is why the
correlation analysis approach\(^{41} \) could not effectively discriminate
between the two classes, resulting in a poor success rate in pre-
dicting the \( \alpha/\beta \) class (see Table 2). Also, it can be seen from Figure
2 that the 21st components of \( \alpha/\beta \) proteins are larger than those of
\( \alpha + \beta \) proteins, implying that \( \alpha/\beta \) proteins are more complicated
than \( \alpha + \beta \) proteins. This kind of difference cannot be reflected at
all by the conventional 20D amino acid composition, nor effec-
tively reflected by the other pseudo amino acid components, but
they can be distinctly revealed through the complexity measure factor.
That is why the introduction of such a factor as the 21st component
to represent the sample of a protein can significantly enhance the
overall success rate in predicting protein structural class.

Conclusions

It is demonstrated in this study that using the complexity measure
factor as one of the pseudo amino acid components can more
effectively reflect the overall sequence-order feature of a protein.

Table 2. Success Rates of Jackknife Crossvalidation with Different Approaches on the 204 Proteins from
ref. 35.

<table>
<thead>
<tr>
<th>Method</th>
<th>Input</th>
<th>All-( \alpha )</th>
<th>All-( \beta )</th>
<th>( \alpha/\beta )</th>
<th>( \alpha + \beta )</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupervised fuzzy clustering(^{39} )</td>
<td>Amino acid composition</td>
<td>35 ( \frac{28}{32} ) = 67.3%</td>
<td>53 ( \frac{46}{61} ) = 86.9%</td>
<td>21 ( \frac{45}{46} ) = 46.7%</td>
<td>28 ( \frac{46}{46} ) = 60.9%</td>
<td>139 ( \frac{204}{204} ) = 68.1%</td>
</tr>
<tr>
<td>Supervised fuzzy clustering(^{22} )</td>
<td>Amino acid composition</td>
<td>38 ( \frac{28}{32} ) = 73.1%</td>
<td>55 ( \frac{46}{61} ) = 90.2%</td>
<td>28 ( \frac{45}{46} ) = 62.2%</td>
<td>49 ( \frac{46}{46} ) = 63.1%</td>
<td>150 ( \frac{204}{204} ) = 73.5%</td>
</tr>
<tr>
<td>Covariant matrix algorithm(^{33} )</td>
<td>Correlation analysis approach(^{41} )</td>
<td>49 ( \frac{22}{46} ) = 94.2%</td>
<td>53 ( \frac{46}{46} ) = 86.9%</td>
<td>22 ( \frac{45}{46} ) = 48.9%</td>
<td>41 ( \frac{46}{46} ) = 89.1%</td>
<td>165 ( \frac{204}{204} ) = 80.9%</td>
</tr>
<tr>
<td>Augmented covariant discriminant algorithm(^{40} )</td>
<td>Pseudo amino acid composition(^{38} )</td>
<td>43 ( \frac{40}{52} ) = 82.7%</td>
<td>55 ( \frac{46}{61} ) = 90.2%</td>
<td>45 ( \frac{46}{46} ) = 100%</td>
<td>40 ( \frac{46}{46} ) = 87.0%</td>
<td>183 ( \frac{204}{204} ) = 89.7%</td>
</tr>
</tbody>
</table>

*Using the complexity measure factor for the 21st component of pseudoamino acid composition (cf. Online Supporting
Materials A).
leading to higher success rates in predicting the structural class of proteins. It is anticipated that introduction of the complexity measure factor may also have impacts on improving the prediction quality for a series of other protein attributes, such as subcellular localization, membrane types, enzyme family and subfamily classes, enzyme active sites, G-protein coupled receptor classification, and protein quaternary structure types, among many others.

References

43. Mahalanobis, P. C. Proc Natl Sci India 1936, 2, 49.