Using Pseudo Amino Acid Composition to Predict Protein Structural Classes: Approached with Complexity Measure Factor

XUAN XIAO,[1,2] SHI-HUANG SHAO,[1] ZHENG-DE HUANG,[1] KUO-CHEN CHOU[1,3]
[1]*Institute of Information, Donghua University, Shanghai 200051, People's Republic of China*
[2]*Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 33300, People's Republic of China*
[3]*Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, California 92130*

Received 4 August 2005; Accepted 20 September 2005
DOI 10.1002/jcc.20354
Published online in Wiley InterScience (www.interscience.wiley.com).

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.

 J Comput Chem 27: 478–482, 2006

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.[1–5] 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) $\alpha + \beta$. 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 $\alpha + \beta$ class represents those containg 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

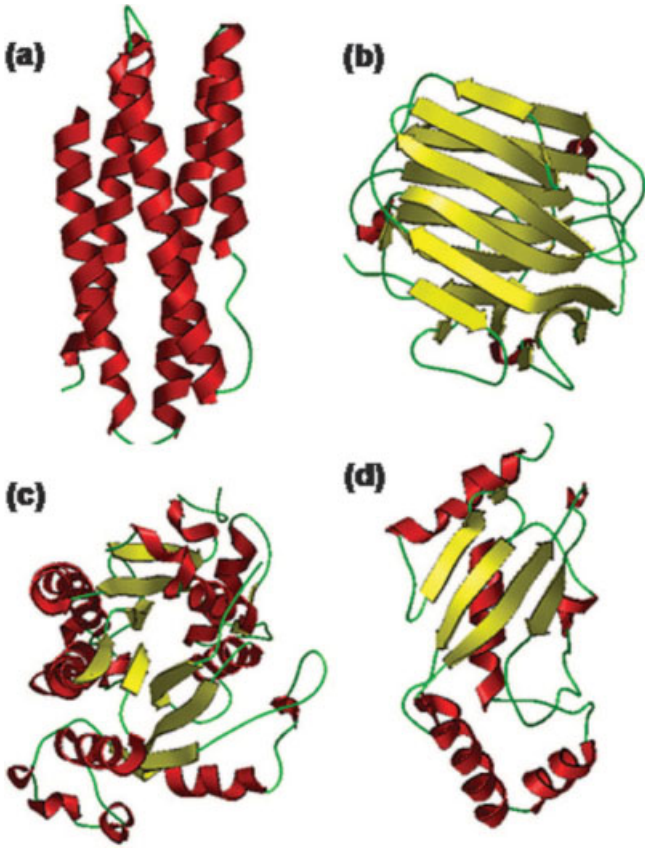

Figure 1. Ribbon drawings to show the four structural classes of proteins: (a) all-α, (b) all-β, (c) α/β, and (d) $\alpha + \beta$. Reproduced from ref. 6 with permission.

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 + λ 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.[28–34] 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 capacity[35] 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.

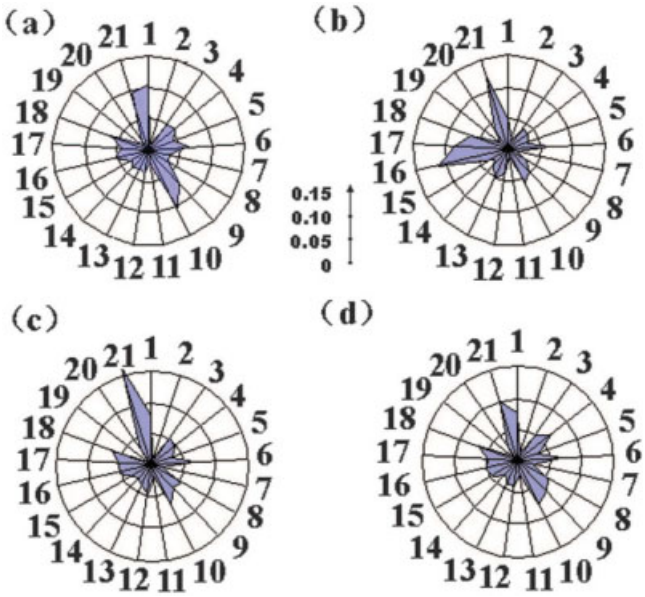

Figure 2. Radar diagrams to show the distinction among the four 21D standard vectors, that is the 21 average components in the pseudo amino acid compositions for the proteins in the following structural classes: (a) all-α, (b) all-β, (c) α/β, and (d) $\alpha + \beta$. Here, we use the numerical indexes 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.

Table 1. Three Different Types for Coding Amino Acids.

Type	Code									
Character	K	N	D	E	P	Q	R	S	T	G
Decimal	6	8	9	10	11	12	13	14	15	16
Binary	00110	01000	01001	01010	01011	01100	01101	01110	01111	10000
Character	A	H	W	Y	F	L	M	I	V	C
Decimal	17	18	20	21	23	24	26	27	28	30
Binary	10001	10010	10100	10101	10111	11000	11010	11011	11100	11110

$A \Leftrightarrow$ Alphabet of symbols (for a binary sequence we have two symbols, namely 0 and 1)
$S \Leftrightarrow$ Finite-length sequences formed by A, for example, $S = \alpha_1\alpha_2\alpha_3 \cdots \alpha_N$
v(S) $\Leftrightarrow$ Vocabulary of sequence S; it is the set of all substrings of S
S$\pi \Leftrightarrow$ Number of the elements in the set S minus one; that is, if $A = \{0,1\}$ and $S = 010$ then $v(S) = \{0,1,01,10,010\}$ and $v(S\pi) = \{0,1,01\}$.

The LZ complexity of a sequence can be measured by the minimal number of steps required for its synthesis in a certain process.[37] For each step only two operations were allowed in the process: either generating an additional symbol, which ensures the uniqueness of each component $S(i_{k-1} + 1: i_k]$ or copying the longest fragment from the part of a synthesized sequence. Its substring is expressed by

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

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

$$H(S) = S[1:i_1]\bullet S[i_1 + 1:i_2]\bullet \cdots S[i_{k-1} + 1:i_k]\bullet \cdots S[i_{m-1} + 1:N]. \tag{2}$$

Let us assume that S $= \alpha_1,\alpha_2,\alpha_3, \ldots \alpha_N$ has been reconstructed by the program up to the digit a_r, and a_r has been newly inserted. The string up to a_r will be denoted by $S[1:r]\cdot$, where the dot denotes that a_r is newly inserted to check whether the rest of the string $S[r + 1: N]$ can be reconstructed by a simple copying. First, suppose $q = a_{r+1}$, and see whether q is reproducible from $S[1:r]q\pi$. If the answer is "no," then we insert $q = a_{r+1}$ 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 $= a_{r+1}a_{r+2}$ 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 $C_{\mathrm{IS}}(S)$:

$$\begin{cases} H(S) = 0\bullet001\bullet10\bullet100\bullet1000\bullet101 \\ C_{LS}(S) = 6 \end{cases} \tag{3}$$

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.[38] 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 $C_{\mathrm{LS}}(S)$. 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,[27] a protein **P** can be expressed by a vector or a point in a $(20 + \lambda)$D $= (20 + 1)$D $= 21$D space; that is,

$$\mathbf{P} = (p_1, p_2, \cdots, p_{20}, p_{21})^{\mathbf{T}} \tag{4}$$

where **T** is the transpose operator, and

$$p_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{20} f_i + wf_{21}}, & (1 \le k \le 20) \\ \dfrac{wf_{21}}{\sum_{i=1}^{20} f_i + wf_{21}}, & (k = 21) \end{cases} \tag{5}$$

where f_i $(i = 1, 2, \ldots 20)$ are the occurrence frequencies of the 20 native amino acids in a protein,[13] $f_{21} = C_{\mathrm{LS}}(S)$ is the complexity measure factor for the protein sequence concerned, and w the

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

Method	Input	All-α	All-β	α/β	$\alpha+\beta$	Overall
Unsupervised fuzzy clustering[39]	Amino acid composition	$\frac{35}{52} = 67.3\%$	$\frac{53}{61} = 86.9\%$	$\frac{21}{45} = 46.7\%$	$\frac{28}{46} = 60.9\%$	$\frac{139}{204} = 68.1\%$
Supervised fuzzy clustering[22]	Amino acid composition	$\frac{38}{52} = 73.1\%$	$\frac{55}{61} = 90.2\%$	$\frac{28}{45} = 62.2\%$	$\frac{29}{46} = 63.1\%$	$\frac{150}{204} = 73.5\%$
Covariant matrix algorithm[13]	Correlation analysis approach[41]	$\frac{49}{52} = 94.2\%$	$\frac{53}{61} = 86.9\%$	$\frac{22}{45} = 48.9\%$	$\frac{41}{46} = 89.1\%$	$\frac{165}{204} = 80.9\%$
Augmented covariant discriminant algorithm[40]	Pseudo amino acid composition[a,25]	$\frac{43}{52} = 82.7\%$	$\frac{55}{61} = 90.2\%$	$\frac{45}{45} = 100\%$	$\frac{40}{46} = 87.0\%$	$\frac{183}{204} = 89.7\%$

[a]Using the complexity measure factor for the 21st component of pseudoamino acid composition (cf. Online Supporting Materials A).

weight factor. In the present study, the weight factor was set at w = 1/800 to make the values of the pseudo amino acid components 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[40] covariant discriminant algorithm to perform the prediction. The covariant discriminant algorithm[42] is a combination of Mahalanobis distance[43,44] and the invariance principle for treating degenerative space[15] that is cited in literature as "Chou's invariance theorem" (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 composition are not independent. Therefore, a dimension-reduced operation by leaving out one of the components and making the rest completely independent is needed when using the augmented covariant discriminant algorithm; that is, a protein should be defined in a 20D space instead of 21D space. Otherwise, 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 invariance 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-α, 61 all-β, 45 α/β, and 46 $\alpha + \beta$. Their PDB codes are given in Table 2 of Chou.[35]

The power of a statistical prediction method is usually evaluated by the resubstitution test, independent dataset test, and jackknife 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 aforementioned 204 proteins classified into four structural classes are given in Table 2, where for facilitating comparison the corresponding 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, $\bar{P}^{\alpha}$, $\bar{P}^{\beta}$, $\bar{P}^{\alpha/\beta}$, and $\bar{P}^{\alpha+\beta}$, as defined in ref. 27. Each of the four standard vectors in the current approach contains 21 components [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 $\bar{P}^{\alpha/\beta}$ and $\bar{P}^{\alpha+\beta}$ 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.[41] That is why the correlation analysis approach[41] could not effectively discriminate between the two classes, resulting in a poor success rate in predicting the α/β class (see Table 2). Also, it can be seen from Figure 2 that the 21st components of α/β proteins are larger than those of $\alpha + \beta$ proteins, implying that α/β 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 effectively reflected by the other pseudo amino acid components, but 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,

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[28,30,42,45,46,47] membrane types,[29,33,48–51] enzyme family and subfamily classes,[34,52–54] enzyme active sites,[55,56] G-protein coupled receptor classification,[57–59] and protein quaternary structure types,[60] among many others.

References

1. Finkelstein, A. V.; Ptitsyn, O. B. Prog Biophys Mol Biol 1987, 50, 171.
2. Chou, K. C.; Carlacci, L. Proteins Struct Funct Genet 1991, 9, 280.
3. Chou, K. C. Biochem Biophys Res Commun 2004, 316, 636.
4. Chou, K. C. Biochem Biophys Res Commun 2004, 319, 433.
5. Chou, K. C. Current Medicinal Chemistry 2004, 11, 2105.
6. Chou, K. C. Curr Protein Peptide Sci 2000, 1, 171.
7. Chou, K. C. Current Protein and Peptide Science 2005, 6, 423.
8. Chou, P. Y. In Prediction of Protein Structure and the Principles of Protein Conformation; Fasman, G. D., Ed.; Plenum Press: New York, 1989, p. 549.
9. Nakashima, H.; Nishikawa, K.; Ooi, T. J Biochem 1986, 99, 152.
10. Chou, K. C. Amino Acids 1993, 6, 231.
11. Klein, P.; Delisi, C. Biopolymers 1986, 25, 1659.
12. Chou, J. J.; Zhang, C. T. J Theoret Biol 1993, 161, 251.
13. Chou, K. C.; Zhang, C. T. J Biol Chem 1994, 269, 22014.
14. Mao, B.; Chou, K. C.; Zhang, C. T. Protein Eng 1994, 7, 319.
15. Chou, K. C. Proteins Struct Funct Genet 1995, 21, 319.
16. Chou, K. C.; Zhang, C. T. Crit Rev Biochem Mol Biol 1995, 30, 275.
17. Bahar, I.; Atilgan, A. R.; Jernigan, R. L.; Erman, B. Proteins Struct Funct Genet 1997, 29, 172.
18. Zhou, G. P.; Assa-Munt, N. Proteins Struct Funct Genet 2001, 44, 57.
19. Zhou, G. P. J Protein Chem 1998, 17, 729.
20. Cai, Y. D.; Li, Y. X.; Chou, K. C. BBA 2000, 1476, 1.
21. Cai, Y. D.; Zhou, G. P. Biochimie 2000, 82, 783.
22. Shen, H. P.; Yang, J.; Liu, X. J.; Chou, K. C. Biochem Biophys Res Commun 2005, 334, 577.
23. Feng, K. Y.; Cai, Y. D.; Chou, K. C. Biochem Biophys Res Commun 2005, 334, 213.
24. Liu, W.; Chou, K. C. J Protein Chem 1998, 17, 209.
25. Zhang, C. T.; Chou, K. C. Protein Sci 1992, 1, 401.
26. Metfessel, B. A.; Saurugger, P. N.; Connelly, D. P.; Rich, S. T. Protein Sci 1993, 2, 1171.
27. Chou, K. C. Proteins Struct Funct Genet 2001, 43, 246 (Erratum: Proteins Struct Funct Genet 2001, 44, 60).
28. Pan, Y. X.; Zhang, Z. Z.; Guo, Z. M.; Feng, G. Y.; Huang, Z. D.; He, L. J Protein Chem 2003, 22, 395.
29. Wang, M.; Yang, J.; Xu, Z. J.; Chou, K. C. J Theoret Biol 2005, 232, 7.
30. Xiao, X.; Shao, S.; Ding, Y.; Huang, Z.; Huang, Y.; Chou, K. C. Amino Acids 2005, 28, 57.
31. Xiao, X.; Shao, S.; Ding, Y.; Huang, Z.; Chen, X.; Chou, K. C. Amino Acids 2005, 28, 29.
32. Gao, Y.; Shao, S. H.; Xiao, X.; Ding, Y. S.; Huang, Y. S.; Huang, Z. D.; Chou, K. C. Amino Acids 2005, 28, 373.
33. Chou, K. C.; Cai, Y. D. J Chem Inform Modeling 2005, 45, 407.
34. Chou, K. C. Bioinformatics 2005, 21, 10.
35. Chou, K. C. Biochem Biophys Res Commun 1999, 264, 216.
36. Gusev, V. D.; Nemytikova, L. A.; Chuzhanova, N. A. Bioinformatics 1999, 15, 994.
37. Ziv, J.; Lempel, A. IEEE Trans Inf Theory 1976, IT-22, 75.
38. Xiao, X.; Shao, S. H.; Ding, Y. S.; Huang, Z. D.; Chou, K. C. Amino Acids 2005, DOI: 10.1007/s00726-005-0225-6.
39. Zhang, C. T.; Chou, K. C.; Maggiora, G. M. Protein Eng 1995, 8, 425.
40. Chou, K. C. Biochem Biophys Res Commun 2000, 278, 477.
41. Du, Q. S.; Wei, D. Q.; Chou, K. C. Peptides 2003, 24, 1863.
42. Chou, K. C.; Elrod, D. W. Protein Eng 1999, 12, 107.
43. Mahalanobis, P. C. Proc Natl Inst Sci India 1936, 2, 49.
44. Pillai, K. C. S. In Encyclopedia of Statistical Sciences; Kotz, S.; Johnson, N. L., Eds.; John Wiley & Sons: New York. 1985, p. 176.
45. Zhou, G. P.; Doctor, K. Proteins Struct Funct Genet 2003, 50, 44.
46. Chou, K. C.; Cai, Y. D. J Cell Biochem 2003, 90, 1250 (Addendum, J Cell Biochem 2004, 91, 1085).
47. Chou, K. C.; Cai, Y. D. Bioinformatics 2005, 21, 944.
48. Chou, K. C.; Elrod, D. W. Proteins Struct Funct Genet 1999, 34, 137.
49. Cai, Y. D.; Zhou, G. P.; Chou, K. C. Biophys J 2003, 84, 3257.
50. Wang, M.; Yang, J.; Liu, G. P.; Xu, Z. J.; Chou, K. C. Protein Eng Design Select 2004, 17, 509.
51. Shen, H.; Chou, K. C. Biochem Biophys Res Commun 2005, 334, 288.
52. Chou, K. C.; Cai, Y. D. Protein Sci 2004, 13, 2857.
53. Chou, K. C.;Elrod, D. W. J Proteome Res 2003, 2, 183.
54. Chou, K. C.; Cai, Y. D. Biochem Biophys Res Commun 2004, 325, 506.
55. Chou, K. C.; Cai, Y. D. Proteins Struct Funct Genet 2004, 55, 77.
56. Cai, Y. D.; Zhou, G. P.; Jen, C. H.; Lin, S. L.; Chou, K. C. J Theoret Biol 2004, 228, 551.
57. Elrod, D. W.; Chou, K. C. Protein Eng 2002, 15, 713.
58. Chou, K. C.; Elrod, D. W. J Proteome Res 2002, 1, 429.
59. Chou, K. C. J Proteome Res 2005, 4, 1413.
60. Chou, K. C.;Cai, Y. D. Proteins Struct Funct Genet 2003, 53, 282.