Amino Acids (2007) 33: 623–629
DOI 10.1007/s00726-007-0496-1
Printed in The Netherlands

Amino Acids

Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes

T.-L. Zhang[1] and **Y.-S. Ding**[1,2]

[1] College of Information Sciences and Technology, Donghua University, Shanghai, China
[2] Engineering Research Center of Digitized Textile and Fashion Technology, Ministry of Education, Donghua University, Shanghai, China

Received December 22, 2006
Accepted January 15, 2007
Published online February 19, 2007;

Summary. Compared with the conventional amino acid composition (AA), the pseudo amino acid composition (PseAA) as originally introduced by Chou can incorporate much more information of a protein sequence; this remarkably enhances the power to use a discrete model for predicting various attributes of a protein. In this study, based on the concept of Chou's PseAA, a 46-D (dimensional) PseAA was formulated to represent the sample of a protein and a new approach based on binary-tree support vector machines (BTSVMs) was proposed to predict the protein structural class. BTSVMs algorithm has the capability in solving the problem of unclassifiable data points in multi-class SVMs. The results by both the 10-fold cross-validation and jackknife tests demonstrate that the predictive performance using the new PseAA (46-D) is better than that of AA (20-D), which is widely used in many algorithms for protein structural class prediction. The results obtained by the new approach are quite encouraging, indicating that it can at least play a complimentary role to many of the existing methods and is a useful tool for predicting many other protein attributes as well.

Keywords: Protein structure classes – Pseudo amino acid composition – Correlation of amino acid – Hydrophobic amino acid couple – Binary tree support vector machines

Abbreviations: BTSVM, binary-tree support vector machine; PseAA, pseudo amino acid composition; SVM, support vector machines

1. Introduction

Prediction of protein structural class is still a hotly researched field in bioinformatics. Protein 3-D (dimensional) structure is related to its function. However, it is difficult to predict 3-D structure from protein sequence directly. The proteins or domains in structural classification of proteins databank (SCOP) (Murzin et al., 1995; Lo et al., 2002; Andreeva et al., 2004) are classified into seven structural classes: all-α, all-β, $\alpha+\beta$, α/β, multi-domain, small protein, and peptide. More than 80% proteins are deposited into the former four classes. Many efforts were focused on the four structural classes, i.e., all-α, all-β, $\alpha+\beta$, and α/β.

The results of Nishikawa et al. (1982, 1983a, b) indicated that the structural class of a protein is correlated with its AA. Several approaches have been developed to predict protein structural classes based on the AA alone. Different clustering and classification algorithms have been used in those prediction approaches, such as the discriminate analysis (Klein and Delisi, 1986; Klein, 1986), the least Euclidian distance (Nakashima et al., 1986), the Mahalanobis distance (Chou and Zhang, 1994; Chou, 1995), covariant discriminant (Chou et al., 1998; Chou and Maggiora, 1998; Zhou, 1998; Chou, 2000), fuzzy clustering (Zhang et al., 1995; Shen et al., 2005), support vector machines (SVMs) (Cai et al., 2001), amino acid principal component analysis (AAPCA) (Du et al., 2006), and rough sets algorithm (Cao et al., 2006).

Although promising results have been achieved by the improvement of algorithms, the representation of protein with AA lacks the sequence order information. The concept of pseudo amino acid composition (PseAA) as originally introduced by Chou (2001) can incorporate much more information of a protein sequence so as to remarkably enhance the power of using a discrete model to predict various attributes of a protein (Chen et al., 2006; Chou, 2002, 2005a; Chou and Cai, 2005; Gao et al., 2005a, b; Liu et al., 2005a, c; Mondal et al., 2006; Pan et al., 2003; Shen and Chou, 2005a, b; 2006a; Shen et al., 2006; Wang et al., 2004, 2005, 2006; Xiao et al., 2005b, c, 2006b, c; Zhang et al., 2006a, b).

Meanwhile, Chou and Cai (2004a) have reported that an overwhelming high success rates could be obtained by using the functional domain composition for predicting protein structural class on a benchmark dataset in which none of protein has greater than 25% sequence identity to any other in a same class. The functional domain composition of proteins has also been used in membrane protein type prediction (Cai et al., 2003) and subcellular location prediction (Chou and Cai, 2002).

Support vector machines (SVM) algorithm is a powerful tool in solving classification problems. SVM has been used to predict various attributes of proteins such as subcellular location (Chou and Cai, 2002), and membrane type (Cai et al., 2003; Wang et al., 2004). Toward protein structural classes, the prediction approaches based on SVM algorithms have achieved promising predictive accuracy on different datasets (Cai et al., 2001; Isik et al., 2004; Markowetz et al., 2003). The prediction of protein structural classes is a multi-classes classification task. The conventional multi-classes SVMs include one against one, one against others, and DAG. However, there are some unclassifiable data points existing in conventional multi-classes SVMs. The methods solving unclassifiable data points in multi-classes SVMs include fuzzy SVM (Lin and Wang, 2002; Huang and Liu, 2002; Abe, 2004), SVM-tree (Pang, 2004; Pang et al., 2005). Here, we introduce a new method to predict protein structural class, which is based on BTSVM and PseAA of protein, which is a 46-D vector, including amino acid composition, correlation of amino acids with different distances, and occurrence frequencies of hydrophobic amino acid couples in protein sequence. BTSVMs algorithm has the capability to solve the problem of unclassifiable data points. The new approach provides encouraging predictive result on self-consistency and jackknife test methods. Compared with the reported methods, the test results show that the new method is promising and will become potential tool for protein function prediction.

2. PseAA representation

In this study, the protein sequence is represented by PseAA (46-D), where the first 20-components are the 20 occurrence frequencies of the native amino acids in a protein, the next 20-components are the correlation factors of amino acids, and the last 6 components are the occurrence frequencies of hydrophobic amino acid couple in protein sequence. The correlation factors and hydrophobic amino acid couple in a protein sequence are given below.

2.1 Correlation factors of amino acid

The sequence of a protein with N residues can be written as $R = R_1R_2 \ldots R_N$. According to Chou (2001, 2005b), the sequence order effect can be approximately reflected through the following equation:

$$\tau_\lambda = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda} J_{i,i+\lambda} \tag{1}$$

where $J_{i,i+\lambda}$ is the correlation factor of residue i and $i+\lambda$. In this study, $J_{i,i+\lambda}$ is defined as the product of hydrophobic values of two residues

$$J_{i,i+\lambda} = h(R_i)h(R_{i+\lambda}) \tag{2}$$

Similar to Liu et al. (2005a, c), Fourier transforms are used to get the frequency information of protein sequence order effect.

$$F(k) = \frac{1}{L}\sum_{i=1}^{L} \tau(i)\exp(-j2\pi ki/L), \quad k = 0, 1, 2, \ldots, L-1 \tag{3}$$

We get the power spectral density of protein sequence with Eq. (4) and the energy spectral density with Eq. (5)

$$P(k) = \frac{F(k)}{L}, \quad k = 0, 1, 2, \ldots, L-1 \tag{4}$$

$$E(k) = |P(k)|^2, \quad k = 0, 1, 2, \ldots, L-1 \tag{5}$$

Based on the theory of digital signal processing, the high-frequency components are more noisy and hence only the low-frequency components are more important. This is just like the case of protein internal motions where the low-frequency components are functionally more important (Chou, 1988, 1989a). In this study, we pick the 20 lower frequencies of energy spectral density to represent the correlation of amino acid with different distances in protein sequence.

2.2 Hydrophobic amino acid couples

In the process of folding from sequence to 3-D structure, the hydrophobic force between polar amino acid side chains and water is the main driving force. Each amino acid has hydrophobic value based on the polar of side chain. Hydrophobic amino acid couple at the different position of protein sequence determines the conformation of protein. Different hydrophobic amino acid couple occurs in different protein secondary structure. According to the theories of Lim (1974), hydrophobic amino acid couples have six kinds of classes: $i, i+2$; $i, i+3$; $i, i+2, i+4$;

$i, i+5$; $i, i+3, i+4$; $i, i+1, i+4$. Hydrophobic amino acid couple $i, i+2$ and $i, i+2, i+4$ occur more frequently in beta-sheets while the patterns $i, i+3$, $i, i+3, i+4$ and $i, i+1, i+4$ tend to occur more often in alpha-helices. The couple $i, i+5$ is an extension of the concept of the ''helical wheel'' or ''amphipathic'' alpha-helix.

The selection of hydrophobic amino acids is based on the theory of Rose et al. (1985). The amino acids that have a mean fractional area loss (on transfer from the standard state to the folded protein) of greater than 80% are included. These amino acids tend to be buried in the hydrophobic core of a protein. Seven amino acids meet this criterion: Val (V), Leu (L), Ile (I), Met (M), Cys (C), Phe (F), and Trp (W). The method of calculation occurrence frequencies of six patterns in protein is described as below:

Step 1: A protein sequence R is expressed by $R = R_1R_2R_3 \ldots R_N$, where R_i is the *i-th* residue in protein sequence.

Step 2: If the conditions $R_i \in \Psi$ and $R_{(i+2)} \in \Psi$ are satisfied simultaneously,

$$C_{(i,i+2)} = C_{(i,i+2)} + 1, \tag{6}$$

where $\Psi\{V,L,I,M,C,F,W\}$, $F(i,i+2) = \frac{1}{N}\sum_{i=1}^{N-2} C_{(i,i+2)}$, $F(i,i+2,i+4)$, $F(i,i+3)$, $F(i,i+3,i+4)$, $F(i,i+1,i+4)$, and $F(i,i+5)$ are computed according to the rule mentioned above.

3. BTSVMs

SVM is one kind of machine learning algorithm based on statistical learning theory. The theory of SVM for pattern recognition has been described in details in the book of Vapnik (1995). The process of SVM is demonstrated simply as follows: first, the samples are transformed into a high-demonstrational feature space through kernel function. Then, optimal separating hyperplane is decided in this space.

SVM is regarded as a typical binary classifier. The methods of applying SVM to solve multi-class classification problems have one-against-one, one-against-others and DAGSVM. The methods mentioned above have the problems of existing unclassifiable data points. Several approaches are developed on this issue. Here, we adopt the approach of SVM multi-class classification based on binary tree (Tang et al., 2005).

The idea of SVM multi-class classification based on binary tree can be stated briefly as follows. First, all classes are separated into two subclasses. Then, each subclass is separated into two sub-subclasses until all subclasses comprise single class. In this way, the multi-classes problem can be transformed into many binary-classes problems. Toward k classes, we need just $k-1$ SVM classifiers when we construct SVM classifier between two subclasses. This approach can avoid the problem of unclassifiable zones. All classes are ranged from small to large, based on the volume of distribution of input vectors. When the volume of one class equals to another, we arrange the class with smaller tab number in the front. In the way, we can get the array of all classes: $n_1, n_2, \ldots, n_k$. Making the n_1 as positive training sample and others as negative training sample, the first SVM sub-classifier is obtained. After deleting n_1 from the array, the secondary SVM sub-classifier is obtained by the method mentioned above. The positive training sample of the $(k-1)$-*th* SVM sub-classifier is n_{k-1} and negative is n_k.

Here, we consider the type of distribution of input vectors as just super sphere for the input vector being multi-dimension. The volume of super sphere is calculated as below.

Step 1, getting the center of gravity of samples, see Eq. (7),

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{7}$$

Step 2, the semi-diameter of super sphere is defined as the maximum value of subtraction of samples and center of gravity of samples,

$$R = \max\{\|\overline{x} - x_i\|\} \tag{8}$$

where $\|\bullet\|$ represent Euclidian distance.

Step 3, the volume of super sphere is the product of d power of semi-diameter of super sphere and π,

$$v = \pi R^d \tag{9}$$

4. Results and discussion

We chose the same dataset as constructed by Chou (2000), which has been used in several previous studies (Zhang et al., 1995; Shen et al., 2005; Du et al., 2003; Xiao et al., 2006c), to validate the performance of the current approach. The dataset contains 204 proteins classified into four structural classes: 52 all-α, 61 all-β, 45 α/β, and 46 $\alpha+\beta$. In the process of SVM training and classification, the kernel function is crucial. We compare the predictive accuracy rates for four structural classes using linear kernel, RBF kernel, and poly kernel functions under 10-fold cross-validation test. The test results are listed in Table 1. RBF kernel function behaves the best performance. We

Table 1. 10-Fold cross validating results of four structural classes with different kernel functions

Kernel function	Accuracy rates (%)				
	all-α	all-β	α/β	α + β	Total
Linear kernel	$\frac{35}{52} = 67.3$	$\frac{57}{61} = 93.4$	$\frac{41}{45} = 91.1$	$\frac{25}{46} = 54.3$	$\frac{158}{204} = 77.4$
RBF kernel	$\frac{48}{52} = 92.3$	$\frac{61}{61} = 100.0$	$\frac{44}{45} = 97.8$	$\frac{35}{46} = 76.1$	$\frac{188}{204} = 92.2$
Poly kernel	$\frac{40}{52} = 76.9$	$\frac{59}{61} = 96.7$	$\frac{42}{45} = 93.3$	$\frac{27}{46} = 58.7$	$\frac{168}{204} = 82.4$

select it as kernel function of SVM in the following experiments.

The success rates thus obtained for the all-α, all-β, α/β, and α + β classes by the 10-fold cross-validation test are 90.4, 95.1, 73.3, and 63.0%, respectively. The overall accuracy is 81.9%. The accuracy rates of the new PseAA approach are higher than those of the AA approach for whichever structural class and overall accuracy rate. The new PseAA method is a kind of combined one, which integrates the concepts of amino acid composition, correlation of amino acids, and hydrophobic amino acids couples. The function of a protein is determined directly by its amino acid sequence and spatial structure. When a protein is folded into a 3D (dimensional) structure, the amino acid with farther distance in linear sequence will become neighbor residues in the 3D structure. So, the correlation of amino acids could represent the features of the protein sequence. Hydrophobic amino acid couples at different positions often occur in different protein secondary structure states. Lim (1974) has obtained the frequent occurrence of six kinds of hydrophobic patterns in helices and strand. The frequencies of hydrophobic amino acid couples reflect the secondary structure distribution in protein sequence. Thus, the new PseAA could reflect essential characteristics of protein sequence in different structural classes. The test results are also validating the effectiveness of the feature expression of protein (see Table 1).

Three indexes are applied to evaluate the prediction accuracy, that is, sensitivity (S_n), personality (S_p), and correlation coefficient (CC).

$$S_n = \frac{TP}{TP + FN} \tag{10}$$

$$S_p = \frac{TP}{TP + FP} \tag{11}$$

$$CC = \frac{TP \times TN - FP \times FN}{[(TP + FN)(TP + FP)(TN + FN)(TN + FP)]^{1/2}} \tag{12}$$

where *TP* (true positives) is the protein number of right prediction in a structure class, *FN* (false negatives) is the protein number of wrong prediction in a structure class, and *FP* (false positives) is the number of the proteins in other classes to be predicted in this class. *TN* (true negatives) is the number of proteins observed in other classes that are not predicted in this class. S_n represents the accuracy, and S_p represents the reliability in procedure of prediction. The *CC* is a single parameter characterizing the matching extent between the observed and predicted structural classes.

In statistical prediction, the following three cross-validation tests are often used to examine the power of a predictor: independent dataset test, sub-sampling test, and jackknife test. Of these three, the jackknife test is thought the most rigorous and objective one [see (Chou and Zhang, 1995) for a comprehensive review in this regard], and hence has been used by more and more investigators (Chou and Shen, 2006a, c, d; Feng, 2001, 2002; Guo et al., 2006b; Liu et al., 2005a, b; Luo et al., 2002; Sun and Huang, 2006; Wang et al., 2005; Wen et al., 2007; Xiao et al., 2005a, c, 2006b, c; Zhang et al., 2006a; Zhou, 1998) (Cao et al., 2006; Chen et al., 2006; Chou and Shen, 2006a, b, c, d, e; Gao and Wang, 2006; Gao et al., 2005b; Guo et al., 2006a, b; Liu et al., 2007; Mondal et al., 2006; Shen and Chou, 2005a, b; Shen and Chou, 2006a, b; Shen et al., 2006; Sun and Huang, 2006; Wen et al., 2007; Xiao et al., 2005a, c, 2006a; Zhang et al., 2006a) in examining the power of

Table 2. Predicted results by different test methods

Structural classes	Self-consistency test (%)			Jackknife test (%)		
	S_n	S_p	*CC*	S_n	S_p	*CC*
all-α	100.0	100.0	100.0	90.4	85.4	83.6
all-β	100.0	100.0	100.0	100	92.3	94.4
α/β	100.0	100.0	100.0	97.8	97.8	97.1
α + β	100.0	100.0	100.0	73.9	89.5	76.6
Total	204/204 = 100.0%			186/204 = 91.2%		

Table 3. Comparison of test results of different predictive algorithms by the jackknife test

Algorithms	Protein features	Predictive accuracy rate (%)				
		all-α	all-β	α/β	α+β	Overall
Unsupervised fuzzy clustering (Zhang et al., 1995)	AA	$\frac{35}{52}=67.3$	$\frac{55}{61}=90.2$	$\frac{21}{45}=46.7$	$\frac{28}{46}=60.9$	$\frac{139}{204}=68.1$
Supervised fuzzy clustering (Shen et al., 2005)	AA	$\frac{38}{52}=73.1$	$\frac{55}{61}=90.2$	$\frac{28}{45}=62.2$	$\frac{29}{46}=63.1$	$\frac{150}{204}=73.5$
Covariant matrix algorithm (Du et al., 2003)	Correlation analysis approach	$\frac{49}{52}=94.2$	$\frac{53}{61}=86.9$	$\frac{22}{45}=48.9$	$\frac{41}{46}=89.1$	$\frac{165}{204}=80.9$
Complexity measure factor (Xiao et al., 2006c)	PseAA (Complexity)	$\frac{43}{52}=82.7$	$\frac{55}{61}=90.2$	$\frac{45}{45}=100$	$\frac{40}{46}=87.0$	$\frac{183}{204}=89.7$
This paper	PseAA (46-D)	$\frac{47}{52}=90.4$	$\frac{61}{61}=100$	$\frac{44}{45}=97.8$	$\frac{34}{46}=73.9$	$\frac{186}{204}=91.2$

various prediction methods. Three indexes as defined in Eqs. (10–12) for four structural classes are calculated under self-consistency and jackknife tests, respectively. The results are illuminated in Table 2. In self-consistency test, the predicted results of S_n, S_p, and *CC* are 100%, respectively. The same results have been obtained in prior works (Cao et al., 2006; Xiao et al., 2006c). The overall accuracy rate of jackknife test is 91.2%. In the four structural classes, the S_n values of all-β, α/β and *all*-α classes are more than 90%, which are 100, 97.8, and 90.4%, respectively. The S_n values of α + β class is the smallest, i.e. 73.9%. The S_p and *CC* values of the four structural classes are promising and satisfying.

Several approaches (Zhang et al., 1995; Shen et al., 2005; Du et al., 2003; Xiao et al., 2006c) have been reported on prediction of protein structural classes using the same test dataset constructed by Chou (2000). To compare the test results of the new method with those of prior works, the test results of different algorithms by jackknife test are shown in Table 3. As we can see, the new approach achieves the highest overall accuracy (91.2%) by jackknife test.

Comparing with the reported prediction methods on the same dataset, our approach behaves encouraging predictive performance. The new PseAA representation in this study could better reflect the essential of a protein sequence.

As mentioned above, Chou and Cai (2004a) have used functional domain composition to predict protein structural class on a benchmark dataset classified into 7 classes where none of protein has greater than 20% sequence identity to any other in a same class, and obtained an overall success rate of 98%. This is indeed a huge improvement in this area. In the future work, we are to improve the current approach by combining the functional domain composition as well.

5. Conclusions

A new approach based on PseAA and BTSVM is reported to predict protein structure classes. BTSVM eliminates the unclassifiable data points happed in multi-class SVM. The protein sequence is represented by a new kind of PseAA: 46-D vectors, including amino acid composition, correlation of amino acid with different distances, and hydrophobic amino acid couples. The predictive performance on PseAA of protein expression is better than that of amino acid composition by 10-fold cross-validation. In order to compare predictive performance objectively, the same dataset (Chou, 2000) is applied to test the new approach. The new approach achieves quite encouraging results. BTSVM has the capability to solve the problem of unclassifiable data points. We believe that the new approach might be further improved by combining the function domain composition (Chou and Cai, 2004a). It is reasonable to believe that the new method could also be used to predict other protein attributes, such as subcellular location and enzyme family class.

Acknowledgments

This work was supported in part by Program for New Century Excellent Talents in University from Ministry of Education of China (No. NCET-04-415), the Cultivation Fund of the Key Scientific and Technical Innovation Project from Ministry of Education of China (No. 706024), International Science Cooperation Foundation of Shanghai (No. 061307041), and Specialized Research Fund for the Doctoral Program of Higher Education from Ministry of Education of China (No. 20060255006).

References

Abe Shigeo (2004) Fuzzy LP-SVM for multiclass problems. ESANN 2004 proceedings- European symposium on artificial neural networks Bruges (Belgium), 28–30 April 2004 d-side public., pp 429–434

Andreeva A, Howorth D, Brenner SE, Hubbard TJP, Chothia C, Murzin AG (2004) SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 32: D226–D229

Cai YD, Liu XJ, Xu X, Zhou GP (2001) Support vector machines for predicting protein structural class. BMC Bioinformatics 2: 3–7

Cai YD, Zhou GP, Chou KC (2003) Support vector machines for predicting membrane protein types by using functional domain composition. Biophys J 84: 3257–3263

Cao Y, Liu S, Zhang L, Qin J, Wang J, Tang K (2006) Prediction of protein structural class with Rough Sets. BMC Bioinformatics 7: 20

Chen C, Zhou X, Tian Y, Zou X, Cai P (2006) Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. Anal Biochem 357: 116–121

Chou KC (1988) Review: Low-frequency collective motion in biomacromolecules and its biological functions. Biophys Chem 30: 3–48

Chou KC (1989a) Low-frequency resonance and cooperativity of hemoglobin. Trends Biochem Sci 14: 212

Chou KC (1993) Mini Review: Prediction of protein folding types from amino acid composition by correlation angles. Amino Acids 6: 231–246

Chou KC (1995) A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins 21: 319–344

Chou KC (2000) Prediction of protein structural classes and subcellular locations. Curr Protein Peptide Sci 1: 171–208

Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. Proteins Struct Funct Genet (Erratum: ibid., 2001, Vol. 44, 60) 43: 246–255

Chou KC (2002) A new branch of proteomics: prediction of protein cellular attributes. In: Weinrer, PW, Lu Q (eds) Gene cloning and expression technologies, chapter 4, pp 57–70. Eaton Publishing, Westborough, MA

Chou KC (2005a) Review: Progress in protein structural class prediction and its impact to bioinformatics and proteomics. Curr Prot Pept Sci 6: 423–436

Chou KC (2005b) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21: 10–19

Chou KC, Cai YD (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem 277: 45765–45769

Chou KC, Cai YD (2004a) Predicting protein structural class by functional domain composition. Biochem Biophys Res Commun (Corrigendum: ibid., 2005, Vol. 329, 1362) 321: 1007–1009

Chou KC, Cai YD (2004b) Predicting enzyme family class in a hybridization space. Protein Science 13: 2857–2863

Chou KC, Cai YD (2005) Prediction of membrane protein types by incorporating amphipathic effects. J Chem Inform Model 45: 407–413

Chou KC, Maggiora GM (1998) Domain structural class prediction. Protein Eng 11: 523–538

Chou KC, Shen HB (2006a) Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochem Biophys Res Commun 347: 150–157

Chou KC, Shen HB (2006b) Large-scale plant protein subcellular location prediction. J Cell Biochem (DOI 10.1002/jcb.21096)

Chou KC, Shen HB (2006c) Large-scale predictions of Gram-negative bacterial protein subcellular locations. J Proteome Res 5: 3420–3428

Chou KC, Shen HB (2006d) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. J Proteome Res 5: 1888–1897

Chou KC, Shen HB (2006e) Predicting protein subcellular location by fusing multiple classifiers. J Cell Biochem 99: 517–527

Chou KC, Zhang CT (1992) A correlation coefficient method to predicting protein structural classes from amino acid compositions. Eur J Biochem 207: 429–433

Chou KC, Zhang CT (1994) Predicting protein folding types by distance functions that make allowances for amino acid interactions. J Biol Chem 269: 22014–22020

Chou KC, Zhang CT (1995) Review: Prediction of protein structural classes. Crit Rev Biochem Mol Biol 30: 275–349

Chou KC, Liu W, Maggiora GM, Zhang CT (1998) Prediction and classification of domain structural classes. Proteins Struct Funct Genet 31: 97–103

Chou PY (1980) Amino acid composition of four classes of proteins. In: Abstracts of Papers, Part I, Second Chemical Congress of the North American Continent, Las Vegas

Chou PY (1989b) Prediction of protein structural classes from amino acid composition. In: Fasman GD (ed) Prediction of protein structure and the principles of protein conformation. Plenum Press, New York, pp 549–586

Du QS, Wei DQ, Chou KC (2003) Correlation of amino acids in proteins. Peptides 24: 1863–1869

Du QS, Jiang ZQ, He WZ, Li DP, Chou KC (2006) Amino acid principal component analysis (AAPCA) and its applications in protein structural class prediction. J Biomol Struct Dyn 23: 635–640

Feng KY, Cai YD, Chou KC (2005) Boosting classifier for predicting protein domain structural class. Biochem Biophys Res Commun 334: 213–217

Feng ZP (2001) Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. Biopolymers 58: 491–499

Feng ZP (2002) An overview on predicting the subcellular location of a protein. In Silico Biol 2: 291–303

Gao QB, Wang ZZ (2006) Classification of G-protein coupled receptors at four levels. Protein Eng Des Select 19: 511–516

Gao QB, Wang ZZ, Yan C, Du YH (2005a) Prediction of protein subcellular location using a combined feature of sequence. FEBS Lett 579: 3444–3448

Gao Y, Shao SH, Xiao X, Ding YS, Huang YS, Huang ZD, Chou KC (2005b) Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter. Amino Acids 28: 373–376

Guo J, Lin Y, Liu X (2006a) GNBSL: A new integrative system to predict the subcellular location for Gram-negative bacteria proteins. Proteomics 6: 5099–5105

Guo YZ, Li M, Lu M, Wen Z, Wang K, Li G, Wu J (2006b) Classifying G protein-coupled receptors and nuclear receptors based on protein power spectrum from fast Fourier transform. Amino Acids 30: 397–402

Huang HP, Liu YH (2002) Fuzzy support vector machines for pattern recognition and data mining. Int J Fuzzy Syst 4: 826–835

Isik Z, Yanikoglu B, Sezerman U (2004) Protein structural class determination using support vector machines. In: Cevdet A, Tuğrul D, Ibrahim K (eds) 19th International Symposium, Kemer-Antalya, Turkey, October 27–29, pp 82–89

Klein P (1986) Prediction of protein structural class by discriminant analysis. Biochim Biophys Acta 874: 205–215

Klein P, Delisi C (1986) Prediction of protein structural class from the amino acid sequence. Biopolymers 25: 1659–1672

Levitt M, Chothia C (1976) Structural patterns in globular proteins. Nature 261: 552–557

Lim VI (1974) Algorithms for the prediction of a-helical and 0-structural regions in globular proteins. J Mol Biol 88: 873–894

Lin CF, Wang SD (2002) Fuzzy support Machines. IEEE Trans Neural Networks 13: 464–471

Liu DQ, Liu H, Shen HB, Yang J, Chou KC (2007) Predicting secretory protein signal sequence cleavage sites by fusing the marks of global alignments. Amino Acids 32 (in press) (DOI: 10.1007/s00726-006-0466-z)

Liu H, Wang M, Chou KC (2005a) Low-frequency Fourier spectrum for predicting membrane protein types. Biochem Biophys Res Commun 336: 737–739

Liu H, Yang J, Ling JG, Chou KC (2005b) Prediction of protein signal sequences and their cleavage sites by statistical rulers. Biochem Biophys Res Commun 338: 1005–1011

Liu H, Yang J, Wang M, Xue L, Chou KC (2005c) Using Fourier spectrum analysis and pseudo amino acid composition for prediction of membrane protein types. Protein J 24: 385–389

Liu W, Chou KC (1998) Prediction of protein structural classes by modified Mahalanobis discriminant algorithm. J Protein Chem 17: 209–217

Lo Conte L, Brenner SE, Hubbard TJP, Chothia C, Murzin A (2002) SCOP database in 2002: refinements accommodate structural genomics. Nucleic Acids Res 30: 264–267

Luo RY, Feng ZP, Liu JK (2002) Prediction of protein strctural class by amino acid and polypeptide composition. Eur J Biochem 269: 4219–4225

Markowetz F, Edler L, Vingron M (2003) Support vector machines for protein fold class prediction. Biometr J 45: 377–389

Metfessel BA, Saurugger PN, Connelly DP, Rich SS (1993) Cross-validation of protein structural class prediction using statistical clustering and neural networks. Protein Sci 2: 1171–1182

Mondal S, Bhavna R, Mohan Babu R, Ramakumar S (2006) Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. J Theor Biol 243: 252–260

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536–540

Nakashima H, Nishikawa K, Ooi T (1986) The folding type of a protein is relevant to the amino acid composition. J Biochem 99: 152–162

Nishikawa K, Ooi T (1982) Correlation of amino acid composition of a protein to its structural and biological characters. J Biochem 91: 1821–1824

Nishikawa K, Kubota Y, Ooi T (1983a) Classification of proteins into groups based on amino acid composition and other characters. I. Angular distribution. J Biochem 94: 981–995

Nishikawa K, Kubota Y, Ooi T (1983b) Classification of the proteins into groups based on amino acid composition and other characters. II. Grouping into four types. J Biochem 94: 997–1007

Niu B, Cai YD, Lu WC, Zheng GY, Chou KC (2006) Predicting protein structural class with AdaBoost learner. Protein Peptide Lett 13: 489–492

Pan YX, Zhang ZZ, Guo ZM, Feng GY, Huang ZD, He L (2003) Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach. J Protein Chem 22: 395–402

Pang SN (2004) Constructing SVM multiple tree for face membership authentication. ICBA 2004, Lecture Notes in Computer Science 3072, pp 37–43

Pang SN, Kim D, Bang SY (2005) Face membership authentication using SVM classification tree generated by membership-based LLE data partition. IEEE Trans Neural Network 16: 436–446

Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH (1985) Hydrophobic of amino acid residue in globular proteins. Science 229: 834–838

Shen HB, Chou KC (2005a) Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. Biochem Biophys Res Commun 337: 752–756

Shen HB, Chou KC (2005b) Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types. Biochem Biophys Res Commun 334: 288–292

Shen HB, Chou KC (2006a) Ensemble classifier for protein fold pattern recognition. Bioinformatics 22: 1717–1722

Shen HB, Chou KC (2006b) Using ensemble classifier to identify membrane protein types. Amino Acids 32 (in press) (DOI: 10.1007/s00726-006-0439-2)

Shen HB, Yang J, Liu XJ, Chou KC (2005) Using supervised fuzzy clustering to predict protein structural classes. Biochem Biophys Res Commun 334: 577–581

Shen HB, Yang J, Chou KC (2006) Fuzzy KNN for predicting membrane protein types from pseudo amino acid composition. J Theor Biol 240: 9–13

Sun XD, Huang RB (2006) Prediction of protein structural classes using support vector machines. Amino Acids 30: 469–475

Tang FM, Wang ZD, Chen MY (2005) One multiclass classification methods for support vector machine. Control Decision 20: 746–750

Vapnik VN (1995) The nature of statistical learning theory. Springer, New York

Wang M, Yang J, Liu GP, Xu ZJ, Chou KC (2004) Weighted-support vector machines for predicting membrane protein types based on pseudo amino acid composition. Protein Eng Des Select 17: 509–516

Wang M, Yang J, Xu ZJ, Chou KC (2005) SLLE for predicting membrane protein types. J Theor Biol 232: 7–15

Wang SQ, Yang J, Chou KC (2006) Using stacked generalization to predict membrane protein types based on pseudo amino acid composition. J Theor Biol 242: 941–946

Wen Z, Li M, Li Y, Guo Y, Wang K (2007) Delaunay triangulation with partial least squares projection to latent structures: a model for G-protein coupled receptors classification and fast structure recognition. Amino Acids 32: 277–283

Xiao X, Shao S, Ding Y, Huang Z, Chen X, Chou KC (2005a) An application of gene comparative image for predicting the effect on replication ratio by HBV virus gene missense mutation. J Theor Biol 235: 555–565

Xiao X, Shao S, Ding Y, Huang Z, Chen X, Chou KC (2005b) Using cellular automata to generate Image representation for biological sequences. Amino Acids 28: 29–35

Xiao X, Shao S, Ding Y, Huang Z, Huang Y, Chou KC (2005c) Using complexity measure factor to predict protein subcellular location. Amino Acids 28: 57–61

Xiao X, Shao SH, Chou KC (2006a) A probability cellular automaton model for hepatitis B viral infections. Biochem Biophys Res Commun 342: 605–610

Xiao X, Shao SH, Ding YS, Huang ZD, Chou KC (2006b) Using cellular automata images and pseudo amino acid composition to predict protein sub-cellular location. Amino Acids 30: 49–54

Xiao X, Shao SH, Huang ZD, Chou KC (2006c) Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. J Comput Chem 27: 478–482

Zhang CT, Chou KC (1992) An optimization approach to predicting protein structural class from amino acid composition. Protein Sci 1: 401–408

Zhang CT, Chou KC, Maggiora GM (1995) Predicting protein structural classes from amino acid composition: application of fuzzy clustering. Protein Eng 8: 425–435

Zhang SW, Pan Q, Zhang HC, Shao ZC, Shi JY (2006a) Prediction protein homo-oligomer types by pseudo amino acid composition: Approached with an improved feature extraction and naive Bayes feature fusion. Amino Acids 30: 461–468

Zhang TL, Ding Y, Chou KC (2006b) Prediction of protein subcellular location using hydrophobic patterns of amino acid sequence. Comput Biol Chem 30: 367–371

Zhou GP (1998) An intriguing controversy over protein structural class prediction. J Protein Chem 17: 729–738

Zhou GP, Assa-Munt N (2001) Some insight into protein structural class prediction. Protein Struct Funct Genet 50: 44–48

Authors' address: Yong-Sheng Ding, College of Information Sciences and Technology, Donghua University, Shanghai 201620, China,
Fax: +86-21-67792325, E-mail: ysding@dhu.edu.cn