Amino Acids (2009) 37:415–425
DOI 10.1007/s00726-008-0170-2

ORIGINAL ARTICLE

Prediction of protein structural classes by Chou's pseudo amino acid composition: approached using continuous wavelet transform and principal component analysis

Zhan-Chao Li · Xi-Bin Zhou · Zong Dai · Xiao-Yong Zou

Received: 8 July 2008 / Accepted: 3 August 2008 / Published online: 23 August 2008

Abstract A prior knowledge of protein structural classes can provide useful information about its overall structure, so it is very important for quick and accurate determination of protein structural class with computation method in protein science. One of the key for computation method is accurate protein sample representation. Here, based on the concept of Chou's pseudo-amino acid composition (AAC, Chou, Proteins: structure, function, and genetics, 43:246–255, 2001), a novel method of feature extraction that combined continuous wavelet transform (CWT) with principal component analysis (PCA) was introduced for the prediction of protein structural classes. Firstly, the digital signal was obtained by mapping each amino acid according to various physicochemical properties. Secondly, CWT was utilized to extract new feature vector based on wavelet power spectrum (WPS), which contains more abundant information of sequence order in frequency domain and time domain, and PCA was then used to reorganize the feature vector to decrease information redundancy and computational complexity. Finally, a pseudo-amino acid composition feature vector was further formed to represent primary sequence by coupling AAC vector with a set of new feature vector of WPS in an orthogonal space by PCA. As a showcase, the rigorous jackknife cross-validation test was performed on the working datasets. The results indicated that prediction quality has been improved, and the current approach of protein representation may serve as a useful complementary vehicle in classifying other attributes of proteins, such as enzyme family class, subcellular localization, membrane protein types and protein secondary structure, etc.

Keywords Pseudo-amino acid composition · Support vector machine · Wavelet power spectrum

Z.-C. Li · X.-B. Zhou · Z. Dai · X.-Y. Zou (✉)
School of Chemistry and Chemical Engineering,
Sun Yat-Sen University, Guangzhou,
People's Republic of China
e-mail: ceszxy@mail.sysu.edu.cn

Introduction

The structural class has become one of the important features of a protein, and has played an important role in both experimental and theoretical studies in protein science (Chen et al. 2006a), because a prior knowledge of protein structural classes is helpful to improve the prediction accuracy of the protein secondary and tertiary structure (Chou 1992; Deleage and Roux 1987). It becomes more and more interesting and challenging to predict protein structural classes, because the gap between the number of known protein sequences and known protein structures is widening rapidly, and the experiments of X-ray crystallographic and NMR are expensive and time-consuming.

According Levitt and Chothia's definition (1976), a protein of known structure can generally be categorized into one of four structural classes: all-α, all-β, α/β, and $\alpha+\beta$. The results observed by Muska and Kim (1992) suggested that protein structure class correlates strongly with its amino acid composition (AAC). Actually, most classifiers were based on their AAC for predicting protein structural classes (Bahar et al. 1997; Cai et al. 2001; Cai and Zhou 2000; Chou 1995, 1999a; Chou and Zhang 1994; Zhou 1998; Zhou and Assa-Munt 2001) (for a systematic description in this area, see comprehensive reviews by Chou (2000, 2005b). However, the successful prediction rate may be declined due to the complete lack of sequence-order effects in primary sequence. To take into account the

effects, a diverse set of descriptors were proposed for enhancing the prediction quality; these include pair-coupled AAC (Chou 1999b), various auto-correlation descriptors (Feng and Zhang 2000; Horne 1988; Lin and Pan 2001), the polypeptide composition (Luo et al. 2002), other composition (Du et al. 2006) and pseudo-amino acid composition (PseAAC; Chou 2001). Another important progress in this area is the introduction of function domain composition by Chou and Cai (2004a) to incorporate the information of various function types.

The PseAAC was originally introduced by Chou to improve the prediction quality for protein subcellular localization and membrane protein type. It can be used to represent a protein sequence with a discrete model yet without completely losing its sequence-order information. Since the concept of Chou's PseAAC was introduced, various PseAAC approaches have been stimulated to deal with varieties of problems in proteins and protein-related systems (Aguero-Chapin et al. 2006; Caballero et al. 2007; Cai and Chou 2006; Chen et al. 2006a, b; Chen and Li 2007a, b; Chou 2005a; Chou and Cai 2004b; Du and Li 2006a, b; Fang et al. 2008; Gao et al. 2005a, b; Gonzalez-Diaz et al. 2006, 2007a, b, c; Jiang et al. 2008; Kurgan et al. 2007; Li and Li 2008; Lin 2008; Lin et al. 2008; Lin and Li 2007a, b; Liu et al. 2005a, b; Mondal et al. 2006; Mundra et al. 2007; Nanni and Lumini 2008; Pan et al. 2003; Pu et al. 2007; Shen and Chou 2005a, b, 2006, 2007f; Shen et al. 2006, 2007; Wang et al. 2004, 2006; Xiao et al. 2006a, b; Zhang and Fang 2008; Zhang et al. 2008a; Zhang et al. 2006; Zhou et al. 2007). Because of its wide usage, recently a very flexible pse-AAC generator, called "PseAAC" (Shen and Chou 2008), was established at the website http://chou.med.harvard.edu/bioinf/PseAAC/, by which users can generate 63 different kinds of PseAA composition. Stimulated by these success, diverse PseAAC based on different digital signal processing approach was developed and utilized to predict various attributes of proteins. Liu et al. (2005a) construct PseAAC by using low-frequency Fourier spectrum analysis. Xiao et al. (2006b) introduced a kind of PseAAC by measuring the complexity of a protein digital signal sequence. Then, Lin and Li (2007b) proposed a novel method to generate PseAAC, which was based on the diversity of the amino acid and dipeptide composition. More recently, Zhang et al. (2008b) used the approximate entropy and hydrophobicity pattern of a protein sequence to construct PseAAC, and Xiao et al. (2008) introduced the grey dynamic modeling to construct PseAAC. However, all these existing PseAAC methods did not consider interaction of long-range in the primary structure (long-range interaction is a major driving force for the protein folding process). Consequently, classification accuracy would further be enhanced by considering the interaction of long-range in the primary structure with more effective digital signal processing approaches.

In view of the above facts, the work presents a new kind of PseAAC based on the continuous wavelet transform (CWT) and principal component analysis (PCA). The procedures of the method are as followings. Firstly, the digital signal is obtained by mapping each amino acid according to various physicochemical properties. Secondary, CWT is utilized to extract new feature vector from the digital signal based on wavelet power spectrum (WPS), and PCA is utilized to reorganize the feature vector to decrease information redundancy and computation complexity. Finally, the reorganized feature vector and AAC are used to generate PseAAC. Evaluated by the jackknife cross-validation test with support vector machines (SVM), the performance of current method exhibited improvement compared with several published results.

Materials and methods

Protein dataset

As is well-known, protein sequence homology in dataset has an effect on the prediction accuracy, i.e. prediction accuracy will be overestimated when using highly homologous protein sequences. Thus, in order to test current method strictly and facilitate the comparison, the dataset constructed by Kurgan and Homaeian (2006) and other two dataset constructed by Zhou (1998) were used as the working dataset. The dataset constructed by Kurgan and Homaeian contains 1673 proteins and domains, of which 443 all-α, 443 all-β, 346 α/β, and 441 $\alpha + \beta$. One of the two dataset constructed by Zhou, consists of 277 domains: 70 all-α domains, 61 all-β domains, 81 α/β domains, and 65 $\alpha + \beta$ domains; the other consists of 498 domains: 107 all-α domains, 126 all-β domains, 136 α/β domains, and 129 $\alpha + \beta$ domains.

Continuous wavelet transform and its power spectrum

Since introduced in the early 1980s, wavelets transform (WT) has become a popular signal analysis tool due to their ability to elucidate simultaneously both spectral and temporal information within the signal (Zhou et al. 2003). WT overcomes the shortcoming of Fourier analysis, which is based on functions that are localized in frequency domain, not in time domain, thus leading to location-specific features in the signal being lost (Subramani et al. 2006). A digital signal can be decomposed into many groups of coefficients in different scales with CWT, and these coefficient vectors can exhibit characteristics in time domain and frequency domain. WPS is a graphical representation

of cumulative information variations at each scale of decomposition of a data and can be a powerful tool for analyzing the WT of a dataset at different decomposition level (Prabakaran et al. 2007). CWT and WPS can be described as follows (Yang 1999):

$$W_f(a,b) = 1/\sqrt{a} \int f(t)\psi\left(t - b/a\right)\mathrm{d}t \quad (1)$$

$$\mathrm{WPS}[j] = \sum_{k=1}^{k} C_{j,k}^2 \quad (2)$$

where a is scale factor and b is the shift factor, and all of them belong to the set of real numbers, and a > 0. $f(t)$ is digital signal sequences. $\psi(t)$ is wavelet core. $W_f(a,b)$ is the result of inner product operation between $f(t)$ and $\psi(t)$. $C_{j,k}$ is coefficient vectors in different scales, where j is level of decomposition and k is the order of the coefficient. WPS[j] is the values of WPS at jth level of decomposition.

Principal component analysis

One of the key matters in classification is to find ways to reduce dimensionality and eliminate information redundancy in data without sacrificing accuracy. Principal component analysis (PCA) is a prominent method, which can be used to deal with this kind of problem. The most common derivation of PCA is in terms of a standardized linear projection which maximizes the variance in the projected space. i.e. PCA generates a new orthogonal space, in this space, a set of variable called principal components. Each principal component is a linear combination of the original variables. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible (Mazzatorta et al. 2006). PCA can be described as follows (Wang and Paliwal 2003; Polat and Güneş 2008a, b; Widodo and Yang 2007): for a given p-dimensional dataset X. The m principal axes $T_1,T_2,\ldots,T_m$, where $1 \le m \le p$, are orthonormal axes onto which the retained variance is maximum in the projected space. Generally, $T_1, T_2,\ldots,T_m$ can be given by the m leading eigenvectors of the sample covariance matrix:

$$S = 1/N \sum_{i=1}^{N} (xi - \mu)^T (xi - \mu) \quad (3)$$

where $x_i \in X$, μ is the sample mean and N is the number of samples, so that:

$$STi = \lambda iTi \quad i \in 1,\ldots,m \quad (4)$$

where λ_i is the ith largest eigenvalue of S. The m principal components of a given observation vector $x_i \in X$ are given by

$$y = [y_1, y_2, \ldots, y_m] = \left[T_1^T x, T_2^T x, \ldots, T_m^T x\right] = T^T x \quad (5)$$

The m principal components of X are decorrelation in the projected space. More details of PCA can be referred to the literature (Jolliffe 1986).

Protein representation based on CWT and PCA

An important issue in the prediction of protein structure class is to represent the primary sequence of proteins with a certain encoding scheme. As is well known, multifarious physicochemical properties of amino acids are important factors in protein folding. Consequently, we choose 9 physicochemical properties which can be referred to previous research (Du and Li 2006a; Gao et al. 2005a, Gao et al. 2005b; Lio and Vannucci, 2000) to map the amino acid sequence to corresponding numerical sequence, and all the properties were listed in Table 1. All amino acid indexes were extracted from AAIndex database (Kawashima et al. 2000).

Here, the protein of 1bea was chosen from the dataset constructed by Kurgan and Homaeian (2006) as an example to describe protein primary structure representation by CWT and PCA. The protein is a kind of serine protease inhibitor and contains 127 residues. The hydrophobic

Table 1 The physicochemical properties used in current research

Properties description	Entry	Reference
Hydrophobicity	JURD980101	Juretic et al. (1998)
Hydrophilicity	HOPT810101	Hopp-Woods (1981)
Refractivity	MCMT640101	McMeekin et al. (1964)
Flexibility	BHAR880101	Bhaskaran-Ponnuswamy (1988)
Normalized van der Waals vloume	FAUJ880103	Fauchere et al. (1988)
Transfer free energy to surface	BULH740101	Bull-Breese (1974)
Electron-ion interaction potential values	COSI940101	Cosic (1994)
Mean polarity	RADA880101	Radzicka-Wolfenden (1988)
Isoelectric point	ZIMJ680104	Zimmerman et al. (1968)

Table 2 The hydrophobic values of 20 amino acids

Amino acid	Hydrophobicity	Amino acid	Hydrophobicity
A	1.1	M	1.9
C	2.5	N	−3.5
D	−3.6	P	−1.9
E	−3.2	Q	−3.68
F	2.8	R	−5.1
G	−0.64	S	−0.5
H	−3.2	T	−0.7
I	4.5	V	4.2
K	−4.11	W	−0.46
L	3.8	Y	−1.3

values of the 20 amino acids were listed in Table 2. Firstly, the amino acid sequence of 1bea was mapped to numerical sequence based on the hydrophobic values and shown in Fig. 1. Secondly, Meyer wavelet was chosen, and scale vector was selected in the range of 1–100 for the step of 1 to perform CWT for the numerical sequence, and the result were shown in Fig. 2.

In Fig. 2, the abscissa is the residue position of the amino acids, the ordinate is the decomposition scales, and the coefficient in different scales are coded from black (relative minimum) to white (relative maximum). Finally, WPS was calculated according to the wavelet coefficients vector obtained from every scale, and WPS values were illustrated in Fig. 3. Further, PCA was used to reorganize the WPS to minimize the random errors and redundant information.

According to Chou's PseAAC (2001), a protein sample can be represented by a vector or a point in (20+λ)-D space. In other words, protein can be expressed as following formula:

$$\mathrm{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{20+1} \\ \vdots \\ X_{20+\lambda} \end{bmatrix} \tag{6}$$

Where

$$x_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + \omega \sum_{j=1}^{\lambda} \theta_j} & (1 \le u \le 20) \\ \dfrac{\omega\, \theta_{u-20}}{\sum_{i=1}^{20} f_i + \omega \sum_{j=1}^{\lambda} \theta_j} & (20+1 \le u \le 20+\lambda) \end{cases} \tag{7}$$

where, f_i ($i = 1, 2,\ldots,20$) are the normalized occurrence frequencies of the 20 native amino acids in the protein X, and ω is a weight factor that adjusts the latter λ-D components to be in similar scales with the first 20 components. In current study, θ_j are the j-th principal component of

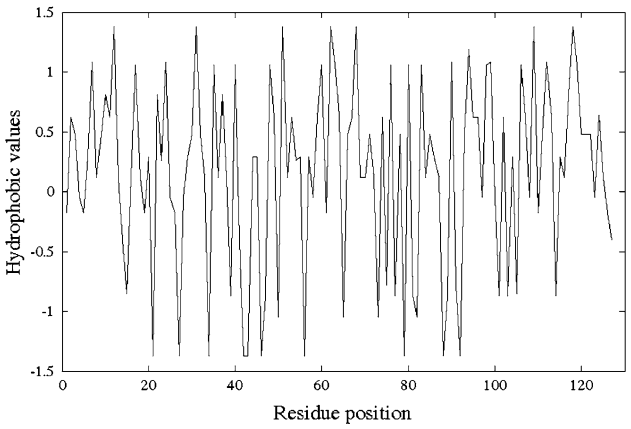

Fig. 1 Amino acid numerical sequence of 1bea

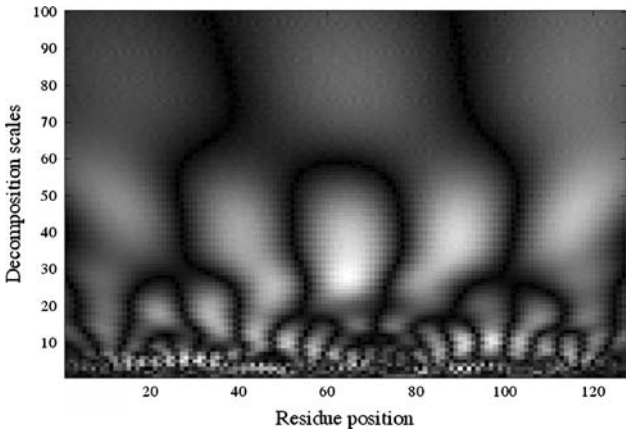

Fig. 2 The CWT plot for amino acid numerical sequence of 1bea

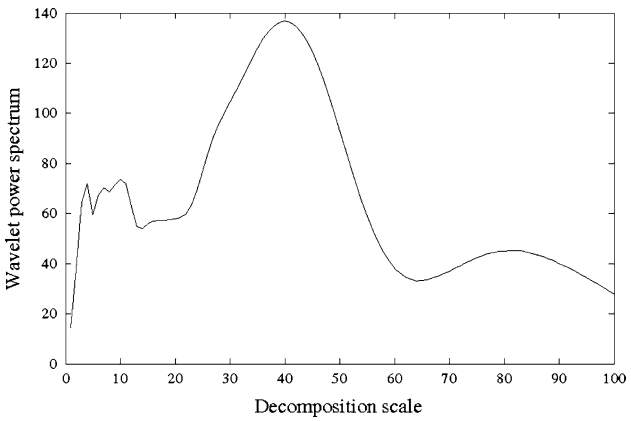

Fig. 3 WPS values in different decomposition scale

WPS. In order to get better prediction accuracy and simplify the method in finding best ω, Shi et al. (2007) parameters normalization approach was applied to (20 + λ)-D components. In this situation, we need not to find best ω, and it was set at 1.

Prediction algorithm

The support vector machine (SVM) (Cortes and Vapnik 1995) is a machine learning system based on statistical learning theory. Compared with other machine learning systems, SVM has many attractive features, including the absence of local minima, the speed and scalability, and the ability to condense information contained in the training set (Chen et al. 2006b). In this research, the publicly available LIBSVM (Chang and Lin 2001) software was used to process the classification. The radial basis function (RBF) was selected as the kernel function and one-versus-one strategy was utilized to predict protein structure class. The whole procedure of current method was illustrated in Fig. 4.

Assessment of predictive performances

For the current study, some assessments of predictive performances were given by Eqs. (8)–(10), including the overall prediction accuracy, the prediction accuracy and Matthew's correlation coefficient (MCC) (Matthews 1975).

$$\text{overall\_accuracy} = \left(\sum_{i=1}^{k} p(i)\right) \Big/ N \tag{8}$$

$$\text{accuracy}(i) = p(i)/(p(i) + o(i)) \tag{9}$$

$$\text{MCC}(i) = \frac{p(i)n(i) - u(i)o(i)}{\sqrt{(p(i) + u(i))(p(i) + o(i))(n(i) + u(i))(n(i) + o(i))}} \tag{10}$$

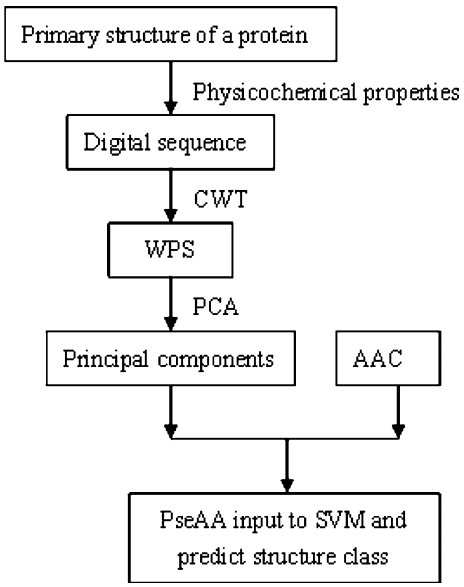

Fig. 4 The chart of the current method

As described by Du and Li (2006a), N is the total number of sequences in the dataset, k is the class number, $p(i)$, $n(i)$, $u(i)$, $o(i)$ are the numbers of true positives, true negatives, false positives and false negatives of the ith class.

Results and discussion

Selection of optimal decomposition scale

Continuous wavelet transform has different resolution at the different decomposition scale. CWT has higher resolution to the lower frequency region of the digital signal when the decomposition scale is larger, so it can be used to study the interaction of long-range in the primary structure and the relation between the structural classes and the entire protein sequence. At the same time, CWT has higher resolution to the high frequency region when the decomposition scale is smaller, it also can be used to analyze local character of the primary structure. It can be observed from Fig. 2 that there are 5 bright regions at the scale of 25–50, which may character the interaction of long-range. There are also many bright regions at the scale of 1–20, which may represent the local character of the primary structure. However, there is no obvious bright region at the scale 50–100, indicating that there are no the interaction of long-range and the local character in the range of these scale. Consequently, decomposition scale was selected in the range of 1–48 for the step of 1 to perform CWT for each physicochemical properties sequence, for investigating the interaction of long-range and the local character simultaneously.

Comparison with different physicochemical properties and wavelet

The effects of wavelet functions and physicochemical properties on prediction accuracy were also investigated. We chose 29 kinds of wavelets function and 9 physicochemical properties to test the maximal overall prediction accuracy of 277 proteins based on fivefold cross-validation.

The overall prediction accuracies were illustrated in Fig. 5 according to 9 physicochemical properties and 29 kinds of wavelet function, including meyr, haar, mexh, morl, db1, db2, db3, db4, db5, db6, db7, db8, db9, db10, bior1.1, bior1.3, bior1.5, bior2.2, bior2.4, bior2.6, bior2.8, bior3.1, bior3.3, bior3.5, bior3.7, bior3.9, bior4.4, bior5.5, bior6.8.

It is very clear from Fig. 5 that the overall prediction accuracies are in the range of 0.7–0.81, and the highest success rates are not obtained for any kind of the wavelets function and each physicochemical property. However, the overall prediction accuracy are in the range of 0.81–0.85

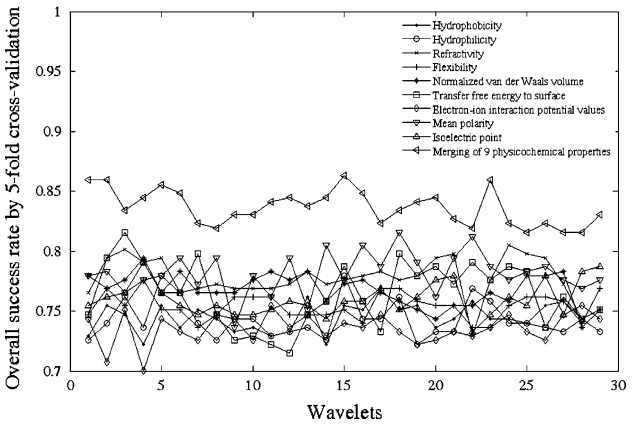

Fig. 5 Overall prediction accuracy based on 29 kinds of wavelet function and 9 physicochemical properties

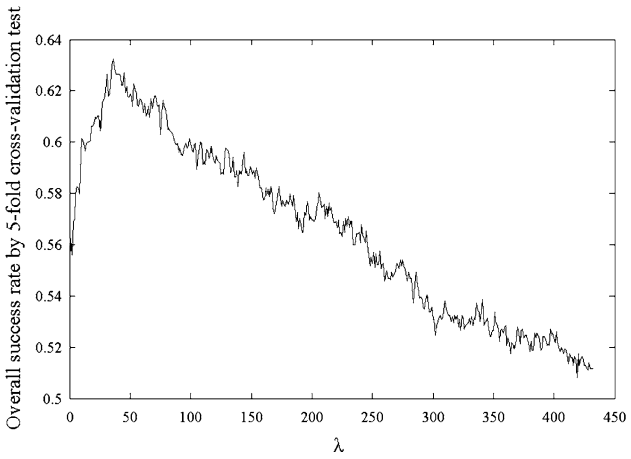

Fig. 6 The effect of λ on overall success rate for the working dataset constructed by Kurgan and Homaeian

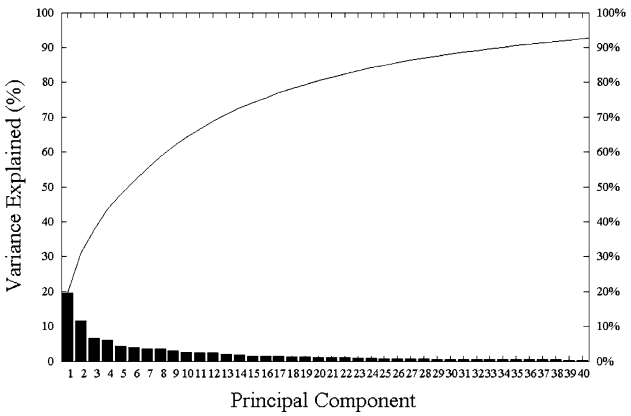

Fig. 7 The variance and the cumulative variance calculated by first 40 principal components

for each wavelet functions by merging nine physicochemical properties (i.e. we combined WPS of nine physicochemical properties to become a $9 \times 48 = 432$ dimension feature vector, PCA was used to reorganize the feature vector and form PseAAC). The results from merging of physicochemical properties are higher than any kind of physicochemical properties, and the highest overall prediction accuracy is obtained when meyr, haar, bior1.1 or bior3.3 is used, indicating the synergistic effects of many kinds of physicochemical properties are the main factor to the protein structure classes. Consequently, meyr wavelet and merging of physicochemical properties were utilized to predict structural classes.

Selection of optimal value for λ in PseAAC

Previous investigations indicated that λ may have different optimal values for different datasets and prediction models. In current study, the value of λ is from 1 to 432 for any dataset. Here, we only carries on discusses to the database constructed by Kurgan and Homaeian, in order to save length and avoid tedious. Figure 6 illustrates the effect of λ on overall success rate by fivefold cross-validation test for the working dataset. Figure 7 showed the variance and the cumulative variance calculated by only first 40 principal components for the convenience of plot. In that, the curve indicates the cumulative variance and bars indicates the variance of the every principal components.

We can see from Fig. 6 that the overall success rate by the fivefold cross-validation test can be improved as the λ value increases, and the best prediction result is achieved when λ is equal to 36. However, overall accuracy of the prediction is decreased with the continue increase of λ. The results from Fig. 7 indicated that the variance enhances with the increase of λ, and 36 significant principal components successfully described over 90% of the variance of the original 432 descriptors. In theory, the overall prediction rate should be enhanced, when variance is explained by more principal components. However, the result showed that the overall prediction rate is not continuously improved with the increase of λ. The reason may be caused by the fact when the λ is large, the noises in the principal components will augment. And the noises will lead to more information redundancy and more instable for the prediction model. Accordingly, for the dataset, the optimal value for λ is 36 (i.e. given a protein in the dataset, we can use $20 + 36 = 56$ dimension PseAAC feature vector to represent it). Similarly, we can use $20 + 90 = 110$ and $20 + 127 = 147$ dimension PseAAC to represent a protein primary structure in the other two datasets constructed by Zhou, respectively.

Results and comparison with different methods

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its

effectiveness in practical application: independent dataset test, subsampling test, and jackknife test (Chou and Zhang 1995). However, as elucidated by (Chou and Shen 2008) and demonstrated in (Chou and Shen 2007c), among the three cross-validation methods, the jackknife test is deemed the most objective that can always yield a unique result for a given benchmark dataset, and hence has been increasingly and widely used by investigators to examine the accuracy of various predictors (Cai and Chou 2003, 2005, 2006; Cai et al. 2005; Chen and Li 2007a, b; Chou and Shen 2006a, b, c, 2007a, 2007b, 2007d, Chou and Shen 2008; Diao et al. 2007; Ding et al. 2007; Du et al. 2003, 2006; Gao et al. 2005a; Jiang et al. 2008; Jin et al. 2008; Li and Li 2008; Lin 2008; Lin et al. 2008; Liu et al. 2005a, b; Niu et al. 2006, 2008; Shen and Chou 2005a, b, 2006, 2007a, b, c, d, e, g; Shen et al. 2006, 2007; Wang et al. 2004, 2006, 2008; Xiao and Chou 2007; Xiao et al. 2006a, b; Zhang and Fang 2008; Zhang et al. 2006, 2008b; Zhou et al. 2007). In the jackknife test, each protein in the dataset is in turn singled out as an independent test sample and all the rule-parameters are calculated without using this protein. For the prediction method proposed here, the optimized parameter combination for C, γ and λ based on fivefold cross-validation tested were utilized to perform jackknife cross-validation test, and the success rates were listed in Tables 3, 4 and 5, compared with several published results for the same dataset.

From Table 3 we can see that the overall success rate by the current approach is 64.0%, about 1.3% higher than the SVM (Kurgan and Chen 2007) based on custom feature. Most notably, the success rates for the current dataset are increased by 29.8% in comparison with the SVM (Kurgan and Homaeian 2006). In additional, our method has best prediction accuracy for the most difficult cases of α/β and $\alpha + \beta$ class. The MCC of the current method is in the range of 0.28–0.62, which is also comparable to the results from the previous literatures (Kurgan and Homaeian 2006; Kurgan and Chen 2007; Kedarisetti et al. 2006). Meanwhile, it is worth noting that StackingC ensemble, LLSC-PRED algorithms and SVM (Kurgan and Chen 2007) based on custom feature achieved higher classification success rate than the SVM (Kurgan and Homaeian 2006; Kurgan and Chen 2007), and these results even comparable to the present method.

As shown in Table 4, the overall accuracy of the current approach is 85.9, which is 11% higher than that of neural network, about 6% higher than component coupled, SVM as well as rough sets. The MCC is in the range of 0.77–0.86 for different structure classes, indicating the present method is a reliable tool for the dataset contained 277 proteins.

It demonstrates from Table 5 that overall success rates of our method for 498 proteins reach 95.2%, which is about 5% higher than component coupled, neural network and rough sets, respectively. To the best of our knowledge, our method is superior to the previous approaches in terms of predictive accuracy. The MCC is from 0.92 to 0.96 for different structure classes, indicating that our method

Table 3 Comparison of different methods by the jackknife test for 1,673 proteins

Method	Success rate (%)/MCC								
	All-α		All-β		α/β		$\alpha + \beta$		Overall
SVM (Kurgan and Homaeian, 2006; Kurgan and Chen 2007)	50.1	0.16	49.4	0.16	28.8	0.05	29.5	0.05	34.2
StackingC ensemble (Kurgan and Chen 2007; Kedarisetti et al. 2006)	74.6	0.62	67.9	0.53	70.2	0.55	32.4	0.22	61.3
SVM (Kurgan and Chen 2007)	77.4	0.65	66.4	0.54	61.3	0.55	45.4	0.27	62.7
LLSC-PRED (Kurgan and Chen 2007)	75.2	0.63	67.5	0.54	62.1	0.54	44.0	0.27	62.2
Our method	76.5	0.62	67.3	0.51	66.8	0.50	45.8	0.28	64.0

Table 4 Comparison of different methods by the jackknife test for 277 proteins

Method	Success rate (%)/MCC								
	All-α		All-β		α/β		$\alpha + \beta$		Overall
Component coupled (Zhou 1998)	84.3	N/A	82.0	N/A	81.5	N/A	67.7	N/A	79.1
Neural network (Cai and Zhou 2000)	68.6	N/A	85.2	N/A	86.4	N/A	56.9	N/A	74.7
SVM (Cai et al. 2001)	74.3	N/A	82.0	N/A	87.7	N/A	72.3	N/A	79.4
LogitBoost (Feng et al. 2005)	81.4	N/A	88.5	N/A	92.6	N/A	72.3	N/A	84.1
Rough sets (Cao et al. 2006)	77.1	N/A	77.0	N/A	93.8	N/A	66.2	N/A	79.4
SVM fusion (Chen et al. 2006b)	85.7	N/A	90.2	N/A	93.8	N/A	80.0	N/A	87.7
Current method	85.7	0.77	90.2	0.86	87.7	0.82	80.1	0.78	85.9

Table 5 Comparison of different methods by the jackknife test for 498 proteins

Method	Success rate (%) / MCC								
	All-α		All-β		α/β		$\alpha + \beta$		Overall
Component coupled (Zhou 1998)	93.5	N/A	88.9	N/A	90.4	N/A	84.5	N/A	89.2
Neural network (Cai and Zhou 2000)	86.0	N/A	96.0	N/A	88.2	N/A	86.0	N/A	89.2
SVM (Cai et al. 2001)	88.8	N/A	95.2	N/A	96.3	N/A	91.5	N/A	93.2
LogitBoost (Feng et al. 2005)	92.6	N/A	96.0	N/A	97.1	N/A	93.0	N/A	94.8
Rough sets (Cao et al. 2006)	87.9	N/A	91.3	N/A	97.1	N/A	86.0	N/A	90.8
SVM fusion (Chen et al. 2006b)	99.0	N/A	96.0	N/A	80.9	N/A	91.5	N/A	91.4
Hybrid neural discriminant (Jahandideh et al. 2007a)	95.3	N/A	88.9	N/A	94.1	N/A	93.0	N/A	92.8
Hybrid model (Jahandideh et al. 2007b)	96.3	N/A	92.1	N/A	95.6	N/A	93.8	N/A	94.4
Current method	94.4	0.92	96.8	0.96	97.0	0.93	92.3	0.92	95.2

exhibited good performances. In additional, to the dataset of 498 proteins, the overall success rate and success rate to α/β class by the present method is about 4 and 16% higher than our former study (Chen et al. 2006b), respectively. However, we should point out that the current method obtained higher overall success rate and success rate of α/β than SVM fusion network to 498 proteins, however, the present method achieved lower overall success rate and success rate of α/β than the SVM fusion network to 277 proteins.

In one word, these results suggested that our method is superior or comparable to other existing methods. And from both the rationality of testing procedure and the success rates of test results, the current method can improve the prediction quality of protein structure class, and can serve as a useful complementary tool. And how to use more effective digital signal processing method is the major task in our feature work.

Conclusion

In this article, a new feature extraction method was presented based on CWT and PCA to take into account the sequence-order effects and long distance interaction in primary sequence. The results of jackknife cross-validation test from working dataset showed that the current method is helpful for the prediction of protein structure class. Our study also indicated that success rates can be improved significantly if many kinds of physicochemical properties are considered in protein representation. Moreover, it can be anticipated that the current method may also have impacts on improving the success rates for many other protein attributes, such as subcellular localization, membrane types, enzymes family and subfamily classes, and G-protein coupled receptor classification.

Acknowledgments The authors acknowledge financial support from the National Nature Science Foundation of China (no. 20575082), the Ph.D. Programs Foundation of Ministry of Education of China (no. 20070558010), the Natural Science Foundation of Guangdong Province (no. 7003714), the Scientific Technology Project of Guangdong Province (no. 2005B30101003) and the Scientific Technology Project of Guangzhou City (no. 2007Z3-E0441).

References

Aguero-Chapin G, Gonzalez-Diaz H, Molina R, Varona-Santos J, Uriarte E, Gonzalez-Diaz Y (2006) Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonases; isolation and prediction of a novel sequence from Psidium guajava L. FEBS Lett 580:723–730

Bahar I, Atilgan AR, Jernigan RL, Erman B (1997) Understanding the recognition of protein structural classes by amino acid composition. Proteins 29:172–185

Caballero J, Fernandez L, Garriga M, Abreu JI, Collina S, Fernandez M (2007) Proteometric study of ghrelin receptor function variations upon mutations using amino acid sequence autocorrelation vectors and genetic algorithm-based least square support vector machines. J Mol Graph Model 26:166–178

Cai YD, Chou KC (2003) Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition. Biochem Biophys Res Comm 305:407–411

Cai YD, Chou KC (2005) Predicting enzyme subclass by function domain composition and pseudo amino acid composition. J Proteome Res 4:967–971

Cai YD, Chou KC (2006) Predicting membrane protein type by function domain composition and pseudo amino acid composition. J Theor Biol 238:395–400

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

Cai YD, Zhou GP, Chou KC (2005) Predicting enzyme family classes by hybridizing gene product composition and pseudo-amino acid composition. J Theor Biol 234:145–149

Cai YD, Zhou GP (2000) Prediction of protein structural classes by neural network. Biochimie 82:783–785

Cao YF, Liu S, Zhang L, Qin J, Wang J, Tang KX (2006) Prediction of protein structural class with rough sets. BMC Bioinform 7:20

Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Chen C, Tian YX, Zou XY, Cai PX, Mo JY (2006a) Using pseudo-amino acid composition and support vector machine to predict protein structural class. J Theor Biol 243:444–448
Chen C, Zhou XB, Tian YX, Zou XY, Cai PX (2006b) Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. Anal Biochem 357:116–121
Chen YL, Li QZ (2007a) Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition. J Theor Biol 248:377–381
Chen YL, Li QZ (2007b) Prediction of the subcellular location of apoptosis proteins. J Theor Biol 245:775–783
Chou KC (1992) Energy-optimized structure of antifreeze protein and its binding mechanism. J Mol Biol 223:509–517
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 (1999a) A key driving force in determination of protein structural classes. Biochem Biophys Res Commun 264:216–224
Chou KC (1999b) Using pair-coupled amino acid composition to predict protein secondary structure content. J Protein Chem 18:473–480
Chou KC (2000) Review: Prediction of protein structural classes and subcellular locations. Curr Protein Pept Sc 1:171–208
Chou KC (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins (Erratum: ibid., 2001, Vol. 44, 60) 43:246–255
Chou KC (2005a) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21:10–19
Chou KC (2005b) Review: Progress in protein structural class prediction and its impact to bioinformatics and proteomics. Curr Protein Pept Sci 6:423–436
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 subcellular localization of proteins by hybridizing functional domain composition and pseudo-amino acid composition. J Cell Biochem 91:1197–1203
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) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. J Proteome Res 5:1888–1897
Chou KC, Shen HB (2006c) Predicting protein subcellular location by fusing multiple classifiers. J Cell Biochem 99:517–527
Chou KC, Shen HB (2007a) Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. J Proteome Res 6:1728–1734
Chou KC, Shen HB (2007b) MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem Biophys Res Comm 360:339–345
Chou KC, Shen HB (2007c) Review: Recent progresses in protein subcellular location prediction. Anal Biochem 370:1–16
Chou KC, Shen HB (2007d) Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochem Biophys Res Comm 357:633–640
Chou KC, Shen HB (2008) Cell-PLoc: A package of web-servers for predicting subcellular localization of proteins in various organisms. Nat Protocols 3:153–162
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
Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297
Deleage G, Roux B (1987) An algorithm for protein secondary structure prediction based on class prediction. Protien Eng 1:289–294
Diao Y, Li M, Feng Z, Yin J, Pan Y (2007) The community structure of human cellular signaling network. J Theor Biol 247:608–615
Ding YS, Zhang TL, Chou KC (2007) Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. Protein Peptide Lett 14:811–815
Du P, Li Y (2006a) Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. BMC Bioinform 7:518
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
Du QS, Wei DQ, Chou KC (2003) Correlation of amino acids in proteins. Peptides 24:1863–1869
Du PF, Li YD (2006b) Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. BMC Bioinform 7:518
Fang Y, Guo Y, Feng Y, Li M (2008) Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features. Amino Acids 34: 103–109
Feng KY, Cai YD, Chou KC (2005) Boosting classifier for predicting protein domain structural class. Biochem Biophys Res Commun 334:213–217
Feng ZP, Zhang CT (2000) Prediction of membrane protein types based on the hydrophobic index of amino acids. J Protein Chem 19:269–275
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
Gonzalez-Diaz H, Aguero-Chapin G, Varona J, Molina R, Delogu G, Santana L, Uriarte E, Podda G (2007a) 2D-RNA-coupling numbers: a new computational chemistry approach to link secondary structure topology with biological function. J Comput Chem 28:1049–1056
Gonzalez-Diaz H, Perez-Bello A, Uriarte E, Gonzalez-Diaz Y (2006) QSAR study for mycobacterial promoters with low sequence homology. Bioorg Med Chem Lett 16:547–553
Gonzalez-Diaz H, Perez-Castillo Y, Podda G, Uriarte E (2007b) Computational chemistry comparison of stable/nonstable protein mutants classification models based on 3D and topological indices. J Comput Chem 28:1990–1995
Gonzalez-Diaz H, Vilar S, Santana L, Uriarte E (2007c) Medicinal chemistry and bioinformatics - current trends in drugs discovery with networks topological indices. Curr Top Med Chem 10:1015–1029
Horne DS (1988) Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities. Biopolymers 27:451–477
Jahandideh S, Abdolmaleki P, Jahandideh M, Asadabadi EB (2007a) Novel two-stage hybrid neural discriminant model for predicting proteins structural classes. Biophys Chem 128:87–93
Jahandideh S, Abdolmaleki P, Jahandideh M, Hayatshahi SHS (2007b) Novel hybrid method for the evaluation of parameters contributing in determination of protein structural classes. J Theor Biol 244:275–281

Jiang X, Wei R, Zhang TL, Gu Q (2008) Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy. Protein Peptide Lett 15:392–396

Jin Y, Niu B, Feng KY, Lu WC, Cai YD, Li GZ (2008) Predicting subcellular localization with AdaBoost learner. Protein Peptide Lett 15:286–289

Jolliffe IT (1986) Principal component analysis. Springer-Verlag, New York

Kawashima S, Ogata H, Kanehisa M (2000) AAindex: amino acid index database. Nucleic Acids Res 28:374

Kedarisetti KD, Kurgan L, Dick S (2006) Classifier ensembles for protein structural class prediction with varying homology. Biochem Biophys Res Commun 348:981–988

Kurgan L, Chen K (2007) Prediction of protein structural class for the twilight zone sequences. Biochem Biophys Res Commun 357:453–460

Kurgan L, Homaeian L (2006) Prediction of structural classes for protein sequences and domains-impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy. Pattern Recogn 39:2323–2343

Kurgan L, Stach W, Ruan J (2007) Novel scales based on hydrophobicity indices for secondary protein structure. J Theor Biol 248:354–366

Levitt M, Chothia C (1976) Structural patterns in globular proteins. Nature 261:552–558

Li FM, Li QZ (2008) Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. Protein Peptide Lett 15:612–616

Lin H (2008) The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. J Theor Biol 252:350–356

Lin H, Ding H, Feng B, Guo FB, Zhang AY, Huang J (2008) Predicting subcellular localization of mycobacterial proteins by using chou's pseudo amino acid composition. Protein Peptide Lett 15:739–744

Lin H, Li QZ (2007a) Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant. Biochem Biophys Res Commun 354: 548–551

Lin H, Li QZ (2007b) Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components. J Comput Chem 28:1463–1466

Lin Z, Pan XM (2001) Accurate prediction of protein secondary structural content. J Protein Chem 20:217–220

Lio P, Vannucci M (2000) Wavelet change-point prediction of transmembrane proteins. Bioinformatics 16:376–382

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

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

Matthews BW (1975) Comparison of predicted and observed secondary structure of T4 phage Iysozyme. Biochim Biophys Acta 405:442–451

Mazzatorta P, Cronin MTD, Benfenati E (2006) A QSAR study of avian oral toxicity using support vector machines and genetic algorithms. QSAR Comb Sci 7:616–628

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

Mundra P, Kumar M, Kumar KK, Jayaraman VK, Kulkarni BD (2007) Using pseudo amino acid composition to predict protein subnuclear localization: approached with PSSM. Pattern Recogn Lett 28:1610–1615

Muska SM, Kim SH (1992) Predicting protein secondary structure contenta tandem neural network approach. J Mol Biol 255: 713–727

Nanni L, Lumini A (2008) Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. Amino Acids 34:653–660

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

Niu B, Jin YH, Feng KY, Liu L, Lu WC, Cai YD, Li GZ (2008) Predicting membrane protein types with bagging learner. Protein Peptide Lett 15:590–594

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

Pu X, Guo J, Leung H, Lin Y (2007) Prediction of membrane protein types from sequences and position-specific scoring matrices. J Theor Biol 247:259–265

Polat K, Güneş S (2008a) Principles components analysis, fuzzy weighting pre-processing and artificial immune recognition system based diagnostic system for diagnosis of lung cancer. Expert Syst Appl 34:214–221

Polat K, Güneş S (2008b) Computer aided medical diagnosis system based on principal component analysis and artificial immune recognition system classifier algorithm. Expert Syst Appl 34:773–779

Prabakaran S, Sahu R, Verma S (2007) Classification of multi class dataset using wavelet power spectrum. Data Min Knowl Disc 15:297–319

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 (2006) Ensemble classifier for protein fold pattern recognition. Bioinformatics 22:1717–1722

Shen HB, Chou KC (2007a) EzyPred: a top–down approach for predicting enzyme functional classes and subclasses. Biochem Biophys Res Comm 364:53–59

Shen HB, Chou KC (2007b) Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. Protein Eng Des Sel 20:39–46

Shen HB, Chou KC (2007c) Hum-mPLoc: An ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. Biochem Biophys Res Commun 355:1006–1011

Shen HB, Chou KC (2007d) Nuc-PLoc: A new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. Protein Eng Des Sel 20: 561–567

Shen HB, Chou KC (2007e) Signal–3L: a 3-layer approach for predicting signal peptide. Biochem Biophys Res Comm 363:297–303

Shen HB, Chou KC (2007f) Using ensemble classifier to identify membrane protein types. Amino Acids 32:483–488

Shen HB, Chou KC (2007g) Virus-PLoc: A fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. Biopolymers 85:233–240

Shen HB, Chou KC (2008) PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition. Anal Biochem 373:386–388

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

Shen HB, Yang J, Chou KC (2007) Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. Amino Acids 33:57–67

Shi JY, Zhang SW, Pan Q, Cheng YM, Xie J (2007) Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. Amino Acids 33:69–74

Subramani P, Sahu R, Verma S (2006) Feature selection using haar wavelet power spectrum. BMC Bioinform 7:432

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 Sel 17:509–516

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

Wang T, Yang J, Shen HB, Chou KC (2008) Predicting membrane protein types by the LLDA algorithm. Protein Peptide Lett 15: no. 9

Wang XC, Paliwal KK (2003) Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition. Pattern Recogn 36:2429–2439

Widodo A, Yang BS (2007) Application of nonlinear feature extraction and support vector machines for fault diagnosis of induction motors. Expert Syst Appl 33:241–250

Xiao X, Chou KC (2007) Digital coding of amino acids based on hydrophobic index. Protein Peptide Lett 14:871–875

Xiao X, Lin WZ, Chou KC (2008) Using grey dynamic modeling and pseudo amino acid composition to predict protein structural classes. J Comput Chem 29:2018–2024

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

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

Yang FS (1999) The engineering analysis and application of wavelet transform. Science Press, Beijing

Zhang GY, Fang BS (2008) Predicting the cofactors of oxidoreductases based on amino acid composition distribution and Chou's amphiphilic pseudo amino acid composition. J Theor Biol 253:310–315

Zhang SW, Zhang YL, Yang HF, Zhao CH, Pan Q (2008a) Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: an approach by incorporating evolutionary information and von Neumann entropies. Amino Acids 34:565–572

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

Zhang TL, Ding YS, Chou KC (2008b) Prediction protein structural classes with pseudo amino acid composition: approximate entropy and hydrophobicity pattern. J Theor Biol 250:186–193

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

Zhou GP, Assa-Munt N (2001) Some insights into protein structural class prediction. Proteins 44:57–59

Zhou X, Wang X, Dougherty ER (2003) Binarization of microarray data based on a mixture model. J Mol Cancer Therapy 2: 679–684

Zhou XB, Chen C, Li ZC, Zou XY (2007) Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. J Theor Biol 248: 546–551