Amino Acids (2008) 34:565–572
DOI 10.1007/s00726-007-0010-9

ORIGINAL ARTICLE

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

Shao-Wu Zhang · Yun-Long Zhang · Hui-Fang Yang · Chun-Hui Zhao · Quan Pan

Received: 23 October 2007 / Accepted: 15 November 2007 / Published online: 11 December 2007

Abstract The rapidly increasing number of sequence entering into the genome databank has called for the need for developing automated methods to analyze them. Information on the subcellular localization of new found protein sequences is important for helping to reveal their functions in time and conducting the study of system biology at the cellular level. Based on the concept of Chou's pseudo-amino acid composition, a series of useful information and techniques, such as residue conservation scores, von Neumann entropies, multi-scale energy, and weighted auto-correlation function were utilized to generate the pseudo-amino acid components for representing the protein samples. Based on such an infrastructure, a hybridization predictor was developed for identifying uncharacterized proteins among the following 12 subcellular localizations: chloroplast, cytoplasm, cytoskeleton, endoplasmic reticulum, extracell, Golgi apparatus, lysosome, mitochondria, nucleus, peroxisome, plasma membrane, and vacuole. Compared with the results reported by the previous investigators, higher success rates were obtained, suggesting that the current approach is quite promising, and may become a useful high-throughput tool in the relevant areas.

S.-W. Zhang (✉) · H.-F. Yang · C.-H. Zhao · Q. Pan
College of Automation, Northwestern Polytechnical University,
No. 127 Youyi West Road, Xi'an 710072, China
e-mail: zhangsw@nwpu.edu.cn

Y.-L. Zhang
Department of Computer,
First Aeronautical Institute of Air Force,
Xinyang 464000, China

Keywords Chou's pseudo-amino acid composition · Residue evolutionary conservation · von Neumann entropies · Multi-scale energy · Weighted auto-correlation function

Abbreviations

Chou's PseAA composition	Chou's pseudo-amino acid composition
MSA	Multiple sequence alignments
VNE	von Neumann entropy
IS	Information score
MSE	Multi-scale energy
AAC	Amino acid composition
JACK	Jackknife tests
INDE	Independent dataset tests
MD	Moment descriptors
SVM	Support vector machine

Introduction

One of the fundamental goals in cell biology and proteomics is to identify the functions of proteins in the cellular environment. Determination of protein subcellular localization, purely using experimental approaches, is both time-consuming and expensive. Particularly, the number of new protein sequences yielded by the high-throughput sequencing technology in the post-genomic era has increased explosively. Facing such an avalanche of new protein sequences, it is both challenging and indispensable to develop an automated method for fast and accurate annotation of the subcellular attributes of uncharacterized proteins. The knowledge thus obtained can help us in the

timely utilization of these newly found protein sequences for both basic research and drug discovery (Chou 2004; Lubec et al. 2005).

During the last decade, many theoretical and computational methods were developed in an attempt to predict protein subcellular localization (Cedano et al. 1997; Chou and Elord 1999; Nakai and Horton 1999; Chou 2000, 2001; Chou and Cai 2002, 2003a, b; Cai and Chou 2003, 2004; Cui et al. 2004; Gao et al. 2005; Pan et al. 2003; Zhou and Doctor 2003; Xiao et al. 2005, 2006; Wang et al. 2005; Shi et al. 2006, 2007; Gardy and Brinkman 2006; Chou and Shen 2006a, b, c, 2007a; Xiao and Chou 2007; Liu et al. 2007; Shen and Chou 2005a, b, 2007a; Shen et al. 2007). However, all these prediction methods were established chiefly based on a single classifier, or based on the statistical approach and amino acid physicochemical character to represent protein sequences. Obviously, the prediction quality would be further improved by introducing protein evolutionary information, and multi-classifiers combining. The protein subcellular localization tends to be evolutionary conservation, and the evolution rates of proteins with different subcellular localizations are different (Julenius and Pedersen 2006). Thus, the evolutionary conservation information of protein subcellular localization could be useful for distinguishing the different subcellular localizations.

Here, we introduce a novel method to calculate amino acid evolutionary conservation scores. After the residue conservation scores were obtained, the protein sequence of English symbols can be translated into a number signal sequence. Then we can form a feature vector by wavelet multi-scale energy (MSE) function (Shi et al. 2007) to represent the protein subcellular localization. The samples of proteins can be represented by other different feature vectors formed by weighted auto-correlation function (Zhang et al. 2006a), moment descriptor (Shi et al. 2006). Based on the hybridization representation, a novel ensemble classifier was formed by fusing many individual classifiers through a sum decision rule and a product decision rule (Kitter et al. 1998). The success rates obtained by hybridizing the multi-source information of proteins and fusing multi-classifiers in predicting protein subcellular localization were significantly improved.

Methods

Residue conservation

The residue ranking function assigns a score to each residue, according to which they can be sorted in the order of the presumably decreasing evolutionary pressure they experience. Out of many methods proposed in the literature (Lichtarge et al. 1996; Soyer et al. 2004; Mihalek et al. 2004), Lichtarge research group's hybrid methods (real-valued evolutionary trace method and zoom method) are the two robust methods. These two methods rank the evolutionary importance of residues in a protein family, which is based on the column variation in multiple sequence alignments (MSA) and evolutionary information extracted from the underlying phylogenetic trees (Mihalek et al. 2004). However, the hybrid methods treat the gaps in the multi-sequences alignment as the 21st amino acid. In this paper, we propose an improved algorithm to estimate the residue evolutionary conservation. The processes of calculation are as follows:

First, the initial similarity sequences were created by using three iterations of PsiBlast (Altschul et al. 1997), with the 0.001 E-value cutoff, on the UniProt database (http://www.expasy.org/) of proteins. The PsiBlast resulting sets were aligned by a standard alignment method such as ClustalW 1.83 (Thompson et al. 1994). So, the MSA were obtained.

Second, an MSA is divided into sub-alignments (that is g groups) that correspond to nodes in the tree (Mihalek et al. 2004). This subdivision of an MSA into smaller alignments reflects the tree topology, and therefore the evolutionary variation information within it. Then, the von Neumann Entropy (VNE) (Caffrey et al. 2004; Mintseris and Weng 2005) for a residue belonging to alignment column i in a group g MSA is given by the following equation:

$$\mathrm{VNE}_i^g = -\mathrm{Tr}(\omega_i^g \log_{20} \omega_i^g) \tag{1}$$

Where ω_i^g is a density matrix of group g with trace = 1. Apart from normalization by the trace, the density matrix is given by the product of the relative frequencies of the standard amino acids of group g in each sub-alignment position i and an appropriate similarity matrix, that is, $\omega_i^g = \mathrm{diag}\left[f_{iA}^g, f_{iC}^g, \ldots, f_{i\alpha}^g, \ldots, f_{iY}^g\right] \times$ Similarity matrix; $f_{i\alpha}^g$ is the relative frequency of each amino acid α (α represents one of the 20 standard amino acids, that is, A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y,) of group g in the alignment position i. The base of 20 ensures that all values are bounded between zero and one (assuming that we ignore entities such as "X", "Z", "B", and "–").

The calculation of Eq. (1) is facilitated by first calculating the eigenvalues $\lambda_{i,\alpha}^g$ of ω_i^g, and hence it follows that

$$\mathrm{VNE}_i^g = -\sum_{\alpha=1}^{20} \lambda_{i\alpha}^g \log_{20} \lambda_{i\alpha}^g \tag{2}$$

where $\lambda_{i\alpha}^g$ is the eigenvalues of ω_i^g. Intuitively, the residue is more conservational, the score calculated with Eqs. (1) and (2) is bigger. Then, the von Neumann can be changed into information score (IS).

$$\mathrm{IS}_i^g = 1 - \mathrm{VNE}_i^g = 1 + \sum_{\alpha=1}^{20} \lambda_{i\alpha}^g \log_{20} \lambda_{i\alpha}^g \tag{3}$$

where IS_i^g is the IS of residue belonging to alignment column i in a group g MSA. Considering the effect of alignment gap, Eq. (3) can be rewrote as

$$\mathrm{IS}_i^g = \left(1 + \sum_{\alpha=1}^{20} \lambda_{i\alpha}^g \log_{20} \lambda_{i\alpha}^g\right) \times f_{i,\mathrm{number}}^g \tag{4}$$

Where, $f_{i,\mathrm{number}}^g$ is the number of standard amino acids of group g in the alignment position i, divided by the number of alignment sequences of group g.

The evolutionary score for a residue belong to column i in an MSA is given by the following equations.

$$R_i = 1 + \sum_{n=1}^{N-1} w_{\mathrm{node}}(n) \sum_{g=1}^{n} w_{\mathrm{group}}(g) \times \left[\left(1 + \sum_{\alpha=1}^{20} \lambda_{i\alpha}^g \log_{20} \lambda_{i\alpha}^g\right) \times f_{i,\mathrm{number}}^g\right] \tag{5}$$

where $w_{\mathrm{node}}(n)$, $w_{\mathrm{group}}(g)$ are weights assigned to a node n and a group g, respectively.

$$w_{\mathrm{node}}(n) = \begin{cases} 1 & \text{if } n \text{ on the path to the query protein} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

$$w_{\mathrm{group}}(g) = \begin{cases} 1 & \text{if } g \text{ on the path to the query protein} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Multi-scale energy

Through residue conservation scores calculated, the protein sequence of English letters can be translated into a conservation score sequence. The numerical sequence can be considered as digital signal. Projecting the signal onto a set of wavelet basis functions with various scales, the fine-scale and large-scale conservation information of a protein can be simultaneously investigated. Here, the wavelet basis function used is symlet wavelet (Pittner and Kamarthi 1999). Consequently, the protein can be characterized as the following MSE feature vector (Shi et al. 2007):

$$\mathrm{MSE} = \left[d_1, \ldots, d_j, \ldots, d_m, a_m\right] \tag{8}$$

Here, m is the coarsest scale of decomposition, d_j is the root mean square energy of the wavelet detail coefficients in the corresponding jth scale, and a_m is the root mean square energy of the wavelet approximation coefficients in the scale m. The energy factors d_j and a_m are defined as

$$d_j = \sqrt{\frac{1}{N_j} \sum_{n=0}^{N_j-1} \left[u_j(n)\right]^2} \quad j = 1, 2, \ldots, m \tag{9}$$

$$a_m = \sqrt{\frac{1}{N_m} \sum_{n=0}^{N_m-1} \left[v_m(n)\right]^2} \tag{10}$$

Here, N_j is the number of the wavelet detail coefficients, N_m is the number of the wavelet approximation coefficients, $u_j(n)$ is the nth detail coefficient in the corresponding jth scale, and $v_m(n)$ is the nth approximation coefficient in the scale m. In general, for the protein sequence with length L, m equals INT($\log_2 L$). But, the diverse lengths of protein sequences make that m value must be optimized to select. Here, we select $m = 12$.

In order to represent a protein sequence with a discrete model yet without completely losing its sequence order information, Chou (2001, 2005) introduced the concept of pseudo-amino acid (PseAA) composition. Ever since the concept of Chou's PseAA composition was introduced, various PseAA composition approaches have been developed for improving the prediction quality of protein attributes (see, e.g., Chen et al. 2006a, b; Chen and Li 2007; Diao et al. 2007a, b; Ding et al. 2007; Du and Li 2006; Fang et al. 2007; Gao et al. 2005a; Kurgan et al. 2007; Li and Li 2007; Lin and Li 2007a, b; Mondal et al. 2006; Mundra et al. 2007; Pan et al. 2003; Pu et al. 2007; Shi et al. 2007; Xiao and Chou 2007; Xiao et al. 2006; Zhang et al. 2006a; Zhang and Ding 2007; Zhou et al. 2007; Chou and Shen 2007b). Recently, a web server called PseAAC (Shen and Chou 2007b) was established at http://chou.med.harvard.edu/bioinf/PseAAC/. PseAAC is a flexible web server, by which users can generate 63 different parallel correlation types of PseAA composition and 63 different series correlation types of PseAA composition as well as the dipeptide PseAA composition. Here, we would like to propose a different approach to formulate PseAA composition, combining the MSE with amino acid composition (AAC), which is consisted of the 20-D components of the amino acid frequencies. The protein can be represented by the following (20 + m + 1)-D vector.

$$X = \left[f_1, f_2, \ldots, f_\alpha, \ldots, f_{20}, d_1, d_2, \ldots, d_j, \ldots, d_m, a_m\right]^T \tag{11}$$

Here f_α (α = 1, 2,…, 20) is the occurrence frequencies of 20 standard amino acids in the protein sequence concerned, arranged alphabetically according to their signal letter codes. Conveniently, the feature set based on the residue evolutionary conservation and MSE approach can be written as EMSE.

Multi-classifiers fusion methods

Let us assume that we have R classifiers each representing the given pattern by a distinct measurement vector. Denote the measurement vector used by the ith classifier by x_i. The $p(x_1, x_2,\ldots, x_R|\omega_k)$ represents the joint probability distribution of the measurements extracted by the classifiers. Let us also assume that the representations used are conditionally statistical independent. The use of different representations may be a probable cause of such independence in special cases. We will investigate the consequences of this assumption and write

$$p(x_1, x_2, \ldots, x_R|\omega_k) = \prod_{i=1}^{R} p(x_i|\omega_k) \tag{12}$$

where $p(x_i \mid \omega_k)$ is the measurement process model of the ith representation. According to the Bayes theorem, the product decision rule can be written as (Kittler et al. 1998)

$$\text{assign} \quad \text{protein X} \rightarrow \text{class } \omega_j \quad \text{if}$$
$$P(\omega_j)\prod_{i=1}^{R} P(\omega_j|x_i) = \max_{k=1}^{m} P(\omega_k)\prod_{i=1}^{R} P(\omega_k|x_i) \tag{13}$$

where $p(\omega_k \mid x_i)$ is a posteriori probability yield by the ith classifier. The decision rule (13) quantifies the likelihood of a hypothesis by combining the posteriori probabilities generated by the individual classifier by means of a product rule. It is effectively a severe rule of fusing the classifier outputs as it is sufficient for a single recognition engine to inhibit a particular interpretation by outputting a close to zero probability for it.

In some applications it may be appropriate to assume further that a posterior probability computed by the respective classifier will not deviate dramatically from the prior probabilities. This is a rather strong assumption but it may be readily satisfied when the available observational discriminatory information is highly ambiguous due to high levels of noise. In such a situation, we can obtain a sum decision rule (Kittler et al. 1998).

$$\text{assign} \quad \text{protein X} \rightarrow \text{class } \omega_j \quad \text{if}$$
$$(1-R)P(\omega_j) + \sum_{i=1}^{R} P(\omega_j|x_i) = \max_{i=1}^{m}\left[(1-R)P(\omega_k) + \sum_{i=1}^{R} P(\omega_i|x_i)\right] \tag{14}$$

Results and discussion

Results with different feature extraction methods

The training dataset and independent dataset taken from Chou's paper (Chou 2000) were used to validate the current method. The schematic illustration of the 12 subcellular localizations of proteins is shown in Fig. 1. Among the independent dataset test (INDE), sub-sampling (e.g., five or tenfold cross-validation) test, and jackknife test (JACK), which are often used for examining the accuracy of a statistical prediction method, the jackknife test was deemed the most rigorous and objective (Chou and Zhang 1995) as demonstrated by a penetrating analysis in a recent comprehensive review (Chou and Shen 2007c) and has been increasingly and widely utilized by investigators to test the power of various prediction methods (see, e.g., Cao et al. 2006; Chen et al. 2006a, b, 2007; Chen and Li 2007; Diao et al. 2007a, b; Ding et al. 2007; Du and Li 2006; Fang et al. 2007; Gao and Wang 2006; Gao et al. 2005a, b, 2006a, b; Huang and Li 2004; Jahandideh et al. 2007; Kedarisetti et al. 2006; Li and Li 2007; Lin and Li 2007a, b; Liu et al. 2007; Mondal et al. 2006; Niu et al. 2006; Pan et al. 2003; Shen and Chou 2007c; Shen et al. 2007; Shi et al. 2007; Sun and Huang 2006; Tan et al. 2007; Wang et al. 2005; Wen et al. 2006; Xiao and Chou 2007; Xiao et al. 2005, 2006; Zhang and Ding 2007; Zhang et al. 2003, 2006a, b; Zhou 1998; Zhou and Assa-Munt 2001; Zhou and Doctor 2003; Zhou et al. 2007). Conveniently, the feature set based on the weighted auto-correlation functions approach (Zhang et al. 2006a) can be written as PARJ, which is composed of AAC and the weighted auto-correlation functions of amino acid residue index PARJ860101 (Parker et al. 1986); the feature set based on the Moment

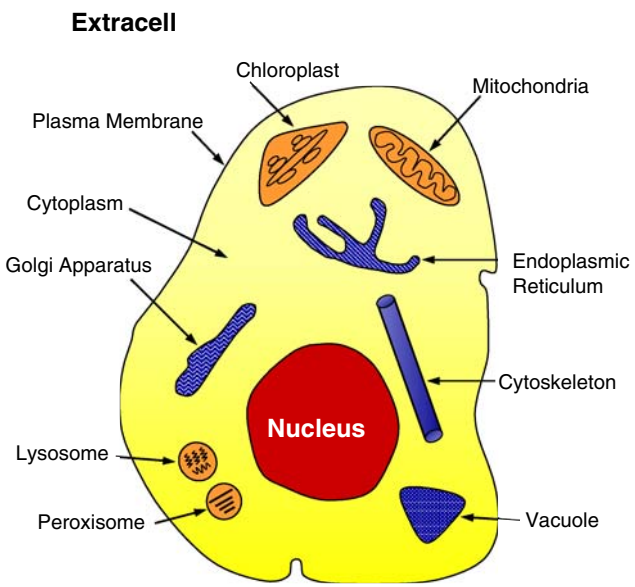

Fig. 1 Schematic illustration to show the 12 subcellular localizations of proteins: chloroplast, cytoplasm, cytoskeleton, endoplasmic reticulum, extracell, Golgi apparatus, lysosome, mitochondria, nucleus, peroxisome, plasma membrane, and vacuole. Note that the vacuole and chloroplast proteins exist only in a plant. Reproduced from Fig. 2 of Chou (2001) with permission

Table 1 Results (in percentage) of four feature extraction methods with SVM and "one-versus-one" classification strategy

	AAC		EMSE		PARJ		MD	
	JACK	INDE	JACK	INDE	JACK	INDE	JACK	INDE
Chloroplast	59.1	60.6	70.8	66.1	59.1	65.1	66.4	77.1
Cytoplasm	85.9	83.9	89	86.7	89	88.8	90.5	89.1
Cytoskeleton	41.2	94.7	44.1	100	47.1	100	50	94.7
Endoplasmic reticulum	32.7	70.8	38.8	85.8	36.7	70.8	34.7	69.8
Extracellular	69.6	84.2	68.3	84.2	73.2	88.4	67.9	87.4
Golgi apparatus	16	0.50	24	25	24	25	16	50
Lysosome	56.8	87.1	51.4	96.8	54.1	96.8	51.4	87.1
Mitochondrial	26.5	12.9	42.2	20.2	41	17.8	38.6	14.1
Nuclear	80.8	76.4	86	80.7	84.1	77.5	81.9	83.1
Peroxisomal	22.2	43.5	18.5	39.1	22.2	47.8	7.4	30.4
Plasma membrane	92.7	96.3	92.8	96.7	96	99	94.3	97.1
Vacuoles	33.3	–	29.2	–	33.3	–	20.8	
Overall accuracy	77.1	80	79.9	83	80.5	83.3	79.4	83.5

descriptors approach (Shi et al. 2006) can be written as MD. The results of four feature extraction methods based on support vector machine (SVM) and "one-versus-one" classification policy (Zhang et al. 2006a) are shown in Table 1.

Table 1 shows that protein evolutionary conservation information can be used to predict subcellular localization. The overall accuracies of EMSE, PARJ and MD are almost equal, but they are all higher than that of AAC in jackknife and independent tests. For EMSE, the predictive accuracy is critically dependent on the input selection of sequences, namely, on the breadth and the depth of the associated sequence similarity tree. That is, how many initial similarity sequences were selected, and how to prune these sequences to form multiple alignment sequences? If the optimal two parameters were selected, we can obtain better results. Considering the computer power, the cutoff of initial similarity sequences was defined as 250, and we did not prune the initial similarity sequences. These results indicate that the performance of predictive system can be improved by using different feature extraction methods. EMSE, PARJ and MD are effective to represent protein sequence and robust for predicting subcellular localization.

Hybrid results of multi-classifiers

Let us assume that the classifiers modeled on ACC, EMSE, PARJ and MD, respectively, are independent. The results of ensemble classifier by fusing four individual classifiers which is modeled on ACC, EMSE, PARJ and MD feature vector set with sum decision rule and product decision rule respectively, are shown in Table 2. From Tables 1 and 2, we can see that the current ensemble hybridization classifier outperforms the individual classifier by 1.3–4.7% in overall accuracy with jackknife test. The performance of sum decision rule is almost equal to that of product decision rule. But the accuracy of some classes such as chloroplast, cytoskeleton, Golgi apparatus and vacuoles has about 2–4% difference between sum decision rule and product decision rule. In addition, the method of multi-classifier fusion assumes that feature vectors used in each classifier should be independent. But the ACC, EMSE, PARJ and MD feature sets used in this paper are not strictly independent. The EMSE, PARJ and MD feature sets

Table 2 Hybrid results (in percentage) of four individual classifier fusion with probability fusion system using sum decision rule and product decision rule

	Sum decision rule		Product decision rule	
	JACK	INDE	JACK	INDE
Chloroplast	67.9	77.1	66.4	74.3
Cytoplasm	91.8	91.2	91.8	91.8
Cytoskeleton	44.1	100	38.2	100
Endoplasmic reticulum	38.8	71.7	38.8	71.7
Extracellular	70.5	92.6	72.3	90.5
Golgi apparatus	28	50	24	50
Lysosome	54.1	96.8	56.8	96.8
Mitochondrial	39.8	14.1	39.8	14.1
Nuclear	87.5	82.3	87.8	82.6
Peroxisomal	14.8	52.2	14.8	47.8
Plasma membrane	96	98.6	95.8	98.7
Vacuoles	25	–	29.2	–
Overall accuracy	81.7	85.2	81.8	85.1

include ACC information. If we delete the redundancy information from EMSE, PARJ and MD feature sets, and use the hybrid method of multi-classifiers, maybe we can then get better fusion results.

Comparison with other prediction methods

The performance of the hybrid method developed in this study was compared with the existing methods such as Pan's (Pan et al. 2003), Gao's (Gao et al. 2005a) and Xiao's (Xiao et al. 2005, 2006), which were also developed from the same dataset. The comparison results of different methods are listed in Table 3. The comparison results demonstrated that the overall prediction accuracies of our hybrid method are higher than that of the other four methods both in the Jackknife and independent tests. For example, the overall accuracy of the hybrid method is 8.1%, 5.4% higher than that of Xiao's method (Xiao et al. 2005) in the Jackknife and independent tests, respectively.

To ensure our methods' validity, we utilize another dataset of Gram-negative bacteria proteins that have been used in previous work (Gardy et al. 2005). Gram-negative bacteria have five major protein subcellular localizations. The dataset consists of 1,444 sequences, 278 of which are cytoplasm, 309 inner membranes, 276 periplasm, 391 outer membranes and 190 extracellular space. The results of different methods are shown in Table 4. From Table 4, we can see that the performances of our methods are better than Gardy's method. The overall accuracy of the fusing EMSE, PARJ and MD is 6.7% higher than that of PSORTb v2.0. These results show that our methods are more effective in predicting subllular localization.

Table 3 Overall accuracy (in percentage) obtained by different methods

	Jackknife test	Independent test
Pan's (Pan et al. 2003)	67.7	73.9
Gao's (Gao et al. 2005a)	69.6	–
Xiao's (Xia et al. 2005)	73.6	79.8
Xiao's (Xiao et al. 2006)	72.6	74.8
Hybrid (AAC + EMSE + PARJ + MD) with sum decision rule	81.7	85.2

Table 4 Results (in percentage) of different methods based on the Gram-negative bacteria dataset

	EMSE	PARJ	MD	Hybrid[a]	PSORTb v2.0 (Gardy et al. 2005)
Cytoplasm	89.2	90.6	90.6	91.7	70.1
Inner membrane	89.6	90.6	90.3	90.3	92.6
Periplasm	81.9	82.6	82.6	83.0	69.2
Outer membrane	91.6	93.9	92.6	94.4	94.9
Extracellular space	79.4	82.6	80	83.1	78.9
Overall accuracy	87.2	88.9	88.2	89.3	82.6

[a] Hybrid method means the fusion of EMSE, PARJ and MD with product decision rule

Conclusion

A new kind of protein evolutionary feature extraction method and a hybrid approach to fuse multi-feature classifiers, were proposed in this paper. The results shows that using residue evolutionary conservation and MSE to represent protein can better reflect protein evolutionary information and predict the subcellular localizations. Weighted auto-correlation function and Moment descriptor methods can optimally reflect the sequence order effect. It is demonstrated that the novel hybrid approach by fusing multi-feature classifiers with sum decision rule and product decision rule is a very intriguing and promising avenue.

Acknowledgments This paper was supported in part by the National Natural Science Foundation of China (No. 60775012 and 60634030) and the Technological Innovation Foundation of Northwestern Polytechnical University (No. KC02), and the Science Technology Research and Development Program of Shaanxi (No. 2006k04-G14).

References

Altschul S, Madden T, Schffer A, Zhang J, Zhang Z, Miller W, Lipman D (1997) Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES (2004) Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci 13:190–202

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

Cai YD, Chou KC (2004) Predicting subcellular localization of proteins in a hybridization space. Bioinformatics 20:1151–1156

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

Cedano J, Aloy P, P'erez-Pons JA, Querol E (1997) Relation between amino acid composition and cellular location of proteins. J Mol Biol 266:594–600

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

Chen J, Liu H, Yang J, Chou KC (2007) Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids 33:423–428
Chen YL, Li QZ (2007) Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition. J Theor Biol 248:377–381
Chou KC (2000) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochem Biophys Res Commun 278:477–483
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 (2004) Review: structural bioinformatics and its impact to biomedical science. Curr Med Chem 11:2105–2134
Chou KC (2005) 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 29:45765–45769
Chou KC, Cai YD (2003a) A new hybrid approach to predict subcellular localization of proteins by incorporating gene oncology composition. Biochem Biophys Res Comm 311:743–747
Chou KC, Cai YD (2003b) Prediction and classification of protein subcellular location: sequence-order effect and pseudo amino acid composition. J Cell Biochem 90:1250–1260
Chou KC, Elrod DW (1999) Protein subcellular location prediction. Protein Eng 12:107–118
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) Cell-PLoc: a package of web-servers for predicting subcellular localization of proteins in various organisms. Nature Protocols. http://chou.med.harvard.edu/bioinf/Cell-PLoc/ (in press)
Chou KC, Shen HB (2007c) Review: recent progresses in protein subcellular location prediction. Anal Biochem 370:1–16
Chou KC, Zhang CT (1995) Review: prediction of protein structural classes. Crit Rev Biochem Mol Biol 30:275–349
Cui Q, Jiang T, Liu B, Ma S (2004) Esub8: a novel tool to predict protein subcellular localizations in eukaryotic organisms. BMC Bioinform 5:66–72
Diao Y, Li M, Feng Z, Yin J, Pan Y (2007a) The community structure of human cellular signaling network. J Theor Biol 247:608–615
Diao Y, Ma D, Wen Z, Yin J, Xiang J, Li M (2007b) Using pseudo amino acid composition to predict transmembrane regions in protein: cellular automata and Lempel-Ziv complexity. Amino Acids. doi:10.1007/s00726-007-0550-z
Ding YS, Zhang TL, Chou KC (2007) Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. Protein Pept Lett 14:811–815
Du P, Li Y (2006) 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 (2007) Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features. Amino Acids. doi:10.1007/s00726-007-0568-2
Gao Y, Shao SH, Xiao X, Ding YS, Huang YS, Huang ZD, Chou KC (2005a) Using pseudo amino acid composition to predict protein subcellular localization: approached with Lyapunov index, Bessel function, and Chebyshev filter. Amino Acids 28:373–376
Gao QB, Wang ZZ, Yan C, Du YH (2005b) Prediction of protein subcellular location using a combined feature of sequence. FEBS Lett 579:3444–3448
Gao QB, Wang ZZ (2006) Classification of G-protein coupled receptors at four levels. Protein Eng Des Sel 19:511–516
Gardy JL, Brinkman FS (2006) Methods for predicting bacterial protein subcellular localization. Nat Rev Microbiol 4:741–751
Gardy JL, Laird MR, Chen F, Rey S, Walsh CJ, Ester M, Brinkman FS (2005) PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics 21:617–623
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 Y, Li Y (2004) Prediction of protein subcellular locations using fuzzy k-NN method. Bioinformatics 20:21–28
Jahandideh S, Abdolmaleki P, Jahandideh M, Asadabadi EB (2007) Novel two-stage hybrid neural discriminant model for predicting proteins structural classes. Biophys Chem 128:87–93
Julenius K, Pedersen AG (2006) Protein evolution is faster outside the cell. Mol Biol Evol 23:2039–2048
Kedarisetti KD, Kurgan LA, Dick S (2006) Classifier ensembles for protein structural class prediction with varying homology. Biochem Biophys Res Commun 348:981–988
Kittler J, Hatef M, Duin RPW, Matas J (1998) On combining classifiers. IEEE Trans Pattern Anal Machine Intell 20:226–239
Kurgan LA, Stach W, Ruan J (2007) Novel scales based on hydrophobicity indices for secondary protein structure. J Theor Biol 248:354–366
Li FM, Li QZ (2007) Using pseudo amino acid composition to predict protein subnuclear location with improved hybrid approach. Amino Acids. doi:10.1007/s00726-007-0545-9
Lichtarge O, Bourne H, Cohen F (1996) An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 257:342–358
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
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 32:493–496
Lubec G, Afjehi-Sadat L, Yang JW, John JP (2005) Searching for hypothetical proteins: theory and practice based upon original data and literature. Prog Neurobiol 77:90–127
Mihalek I, Reš I, Lichtarge O (2004) A family of evolution–entropy hybrid methods for ranking protein residues by importance. J Mol Biol 336:1265–1282
Mintseris J, Weng ZP (2005) Structure function, and evolution of transient and obligate protein-protein interactions. PNAS 102:10930–10935
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 243:252–60
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
Nakai K, Horton P (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci 24:34–36
Niu B, Cai YD, Lu WC, Zheng GY, Chou KC (2006) Predicting protein structural class with AdaBoost learner. Protein Pept Lett 13:489–492
Pan YX, Zhang ZZ, Guo ZM, Feng GY, Huang Z, He L (2003) Application of pseudo amino acid composition for predicting protein subcellular localization: stochastic signal processing approach. J Protein Chem 22:395–402
Parker JM, Guo D, Hodges RS (1986) New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. Biochem 25:5425–5432
Pittner S, Kamarthi SV (1999) Feature extraction from wavelet coefficients for pattern recognition tasks. IEEE Trans Pattern Anal Mach Intell 2:83–88
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
Shen HB, Chou KC (2005a) Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. Biochem Biophys Res Comm 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 (2007a) 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 (2007b) PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition. Anal Biochem. doi:10.10.1016/j.ab.2007.10.012
Shen HB, Chou KC (2007c) Using ensemble classifier to identify membrane protein types. Amino Acids 32:483–488
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, Liang Y, Pan Q (2006) Prediction of protein subcellular localizations using moment descriptors and support vector machine. In: PRIB: 2006. Springer, Berlin, pp 105–114
Shi JY, Zhang SW, Pan Q, Cheng YM, Xie J (2007) SVM-based method for subcellular localization of protein using multi-scale energy and pseudo amino acid composition. Amino Acids 33:69–74
Soyer OS, Goldstein RA (2004) Predicting functional sites in proteins: site-specific evolutionary models and their application to neurotransmitter transporters. J Mol Biol 339:227–242
Sun XD, Huang RB (2006) Prediction of protein structural classes using support vector machines. Amino Acids 30:469–475
Tan F, Feng X, Fang Z, Li M, Guo Y, Jiang L (2007) Prediction of mitochondrial proteins based on genetic algorithm—partial least squares and support vector machine. Amino Acids. doi: 10.1007/s00726-006-0465-0
Thompson J, Higgins D, Gibson T (1994) Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680
Wang M, Yang J, Chou KC (2005) Using string kernel to predict signal peptide cleavage site based on subsite coupling model. Amino Acids (Erratum, ibid. 2005 29:301) 28:395–402
Wen Z, Li M, Li Y, Guo Y, Wang K (2006) 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 SH, Ding YS, Huang ZD, Huang Y, Chou KC (2005) Using complexity measure factor to predict protein subcellular localization. Amino Acids 28:57–61
Xiao X, Shao SH, Ding YS, Huang ZD, Chou KC (2006) Using cellular automata images and pseudo amino acid composition to predict protein subcellular localization. Amino Acids 30:49–54
Xiao X, Chou KC (2007) Digital coding of amino acids based on hydrophobic index. Protein Pept Lett 14:871–875
Zhang SW, Quan Pan, Zhang HC, Zhang YL, Wang HY (2003) Classification of protein quaternary structure with support vector machine. Bioinformatics 19:2390–2396
Zhang SW, Pan Q, Zhang HC, Shao ZC, Shi JY (2006a) Prediction protein homo-oligomer types by pesudo amino acid composition: approached with an improved feature extraction and naive Bayes feature fusion Amino Acids 30:461–468
Zhang ZH, Wang ZH, Zhang ZR, Wang YX (2006b) A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine. FEBS Lett 580:6169–74
Zhang TL, Ding YS (2007) Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes. Amino Acids. doi:10.1007/s00726-007-0496-1
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: Struct Funct Genet 44:57–59
Zhou GP, Doctor K (2003) Subcellular location prediction of apoptosis proteins. Proteins: Struct Funct Genet 50:44–48
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