Amino Acids (2008) 35:591–598
DOI 10.1007/s00726-008-0086-x

ORIGINAL ARTICLE

Using Chou's pseudo amino acid composition to predict protein quaternary structure: a sequence-segmented PseAAC approach

Shao-Wu Zhang · Wei Chen · Feng Yang · Quan Pan

Received: 19 February 2008 / Accepted: 28 February 2008 / Published online: 22 April 2008

Abstract In the protein universe, many proteins are composed of two or more polypeptide chains, generally referred to as subunits, which associate through noncovalent interactions and, occasionally, disulfide bonds to form protein quaternary structures. It has long been known that the functions of proteins are closely related to their quaternary structures; some examples include enzymes, hemoglobin, DNA polymerase, and ion channels. However, it is extremely labor-expensive and even impossible to quickly determine the structures of hundreds of thousands of protein sequences solely from experiments. Since the number of protein sequences entering databanks is increasing rapidly, it is highly desirable to develop computational methods for classifying the quaternary structures of proteins from their primary sequences. Since the concept of Chou's pseudo amino acid composition (PseAAC) was introduced, a variety of approaches, such as residue conservation scores, von Neumann entropy, multiscale energy, autocorrelation function, moment descriptors, and cellular automata, have been utilized to formulate the PseAAC for predicting different attributes of proteins. Here, in a different approach, a sequence-segmented PseAAC is introduced to represent protein samples. Meanwhile, multiclass SVM classifier modules were adopted to classify protein quaternary structures. As a demonstration, the dataset constructed by Chou and Cai [(2003) Proteins 53:282–289] was adopted as a benchmark dataset. The overall jackknife success rates thus obtained were 88.2–89.1%, indicating that the new approach is quite promising for predicting protein quaternary structure.

Keywords Sequence-segmented PseAAC · Residue conservation · Von Neumann entropy · Multiscale energy · Moment descriptor · Support vector machine

Introduction

The "protein quaternary structure" refers to the number of polypeptide chains (subunits) involved in forming a protein and the spatial arrangement of its subunits. The concept of quaternary structure is derived from the fact that many proteins are composed of two or more subunits that associate through noncovalent interactions and, in some cases, disulfide bonds to form oligomers. In the protein universe there are many different classes of subunit construction, such as monomer, dimer, trimer, tetramer, and so forth. The oligomers may be homo-oligomers or hetero-oligomers; the former consist of identical polypeptide chains, whereas the latter are nonidentical. Such complexes are involved in various biological processes, including metabolism, signal transduction and chromosome replicating, etc., and play very important roles in protein functions (Terry and Richard 1998). Some examples include enzymes, hemoglobin, DNA polymerase, and ion channels. The oligomeric proteins have more advantages than the monomers from a functional evolution point of view, and contribute significantly to evolutionary stability in that changes in the quaternary structure can occur through each individual chain or through their reorientation relative to each other (Klotz et al. 1975; Einstein and Schachman 1989; Price 1994).

S.-W. Zhang (✉) · W. Chen · F. Yang · Q. Pan
College of Automation,
Northwestern Polytechnical University,
710072 Xi'an, China
e-mail: zhangsw@nwpu.edu.cn

Q. Pan
e-mail: quanpan@nwpu.edu.cn

Recent research into the utilization of computational methods to determine quaternary structures appears to be heading in three main directions. One direction is the study of domain–domain docking or the type of interaction in the protein complexes (Kim and Ison 2005; Chen and Zhou 2005; Zhu et al. 2006). In this approach, the docking or interaction type is examined based on the protein structures deposited in the PDB. The methodology involves generalizing the association mechanisms of multiple proteins in the complexes to the quaternary structures in general. It has been observed that the overall prediction success rate across a genome-wide study is poor. However, the performance can be improved significantly if only those proteins that have informative (or related) proteins in the training set are considered. The second direction seeks out geometric regularities and constraints to reduce the huge search spaces of quaternary structures (Inbar et al. 2005; Chen and Skolnick 2007; Liu et al. 2007a). The third direction involves the classification of quaternary attributes: given a protein primary sequence, determining whether it takes a tertiary structure of a single chain or a quaternary structure with other proteins (Garian 2001; Zhang et al. 2003, 2006a; Chou and Cai 2003; Yu et al. 2006). This is important, because the functions of proteins are closely related to their quaternary attributes. For example, some critical ligands only bind to dimers (Chou 2004a, 2004b) but not to monomers; some marvelous allosteric transitions only occur in tetramers (Chou 1988, 1989, 2004c; Doyle et al. 1998), not other oligomers; and some ion channels are formed by dimers (Call et al. 2006) or tetramers (Schnell and Chou 2008), whereas others are formed by pentamers (Chou 2004d, 2004e; Oxenoid and Chou 2005). The association of subunits depends upon the existence of complementary “patches” on their surface structures.

This suggests that primary sequences contain quaternary structure information (Garian 2001; Zhang et al. 2003, 2006a; Chou and Cai 2003; Yu et al. 2006). Therefore, we can develop an automated method to predict protein quaternary structure from protein primary sequences. To explore this problem, Garian (2001) developed a method which used decision-tree models and a feature extraction approach (simple binning function) to successfully predict homodimers and nonhomodimers. Chou and Cai (2003) also researched this question using a pseudo-amino acid composition (PseAAC) feature extraction method to predict monomers, homodimers, homotrimers, homotetramers, homopentamers, homohexamers and homooctamers. In our previous work, we successfully predicted homodimers and nonhomodimers, homodimers, homotrimers, homotetramers and homohexamers using a weighted autocorrelation function feature extraction approach (Zhang et al. 2003; 2006a). In this paper, we try to develop another approach, the sequence-segmented PseAAC method, to predict monomers (1EM), homodimers (2EM), homotrimers (3EM), homotetramers (4EM), homopentamers (5EM), homohexamers (6EM) and homooctamers (8EM).

Materials and methods

Datasets

The training data used here was constructed by Chou and Cai (2003), and it consists of 3,174 protein sequences, of which 382 are classified monomers, 817 homodimers, 593 homotrimers, 884 homotetramers, 54 homopentamers, 287 homohexamers, and 157 homooctamers. They each contain more than 50 sequences.

Dataset construction was governed by the following criteria:

Clearness: the collected samples were only those protein sequences that had their quaternary attributes marked

Nonredundancy: if several proteins had high sequence similarity, only one was kept in order to avoid redundancy

Statistical significance: subsets were dropped from further consideration if they contained too few entries to be of statistical significance

Methods of representing the protein sequence

Without loss of generality, we assume that there are N protein sequences in the dataset. Let L^k be the length of the kth sequence p^k and α_i be the ith element of 20 natural amino acids represented by the letters A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y, respectively.

Sequence-segmented amino acid composition

Suppose the kth protein sequence p^k is segmented into M segmentations of the same length, and the amino acid composition (AAC) of each segment is calculated. Therefore, the protein sequence p^k can be represented using the following formula:

$$\mathrm{AACS}^k = \left\{ \begin{matrix} c^k_{1,1} & \cdots & c^k_{1,m} & \cdots & c^k_{1,M} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c^k_{i,1} & \cdots & c^k_{i,m} & \cdots & c^k_{i,M} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c^k_{20,1} & \cdots & c^k_{20,m} & \cdots & c^k_{20,M} \end{matrix} \right\}_{20\times M}, \quad k = 1, \ldots, N \tag{1}$$

where M is the total number of segments, $\left[c^k_{1,m}, \ldots, c^k_{i,m}, \ldots, c^k_{20,m}\right]^{\mathrm{T}}$ is the AAC of the mth segment of p^k, and $c^k_{i,m}$ is defined as

$$c^k_{i,m} = M \cdot t^k_{i,m} \Big/ L^k, \quad m = 1, \ldots, M, \; i = 1, \ldots, 20 \qquad (2)$$

where $t^k_{i,m}$ is the count of α_i that appears in the mth segment of the protein sequence p^k.

Conveniently, the feature set based on this sequence-segmented amino acid composition approach can be denoted AACS$_m$.

Sequence-segmented pseudo amino acid composition

Note that the use of the amino acid composition to represent a protein segment as described in the above section would result in the loss of all of its sequence-order information. To avoid losing the sequence-order information, a logical approach is to use the entire sequence to represent the protein segment. However, this kind of approach fails to work when the query protein does not have significant homology to proteins with known characteristics (Chou and Shen 2007). In order to avoid the complete loss of sequence-order information and also enable more effective prediction for those proteins that do not have significant homology to characterized proteins, a feasible approach is to use the pseudo amino acid composition (PseAAC) to represent the protein sample. The PseAAC (Chou 2001) was originally proposed for predicting protein subcellular localization and membrane protein type (Chou 2001), while the amphiphilic PseAAC (Chou 2005) was proposed for predicting the enzyme functional classification. The essence of PseAAC is to use a discrete model to represent a protein sample without complete losing its sequence-order information. According to its definition, the PseAAC for a given protein sample is expressed by a set of 20 + λ discrete numbers, where the first 20 represent the 20 components of the classical amino acid composition while the additional λ numbers incorporate some of its sequence-order information via various different kinds of coupling modes. Ever since the concept of PseAAC was introduced, various PseAAC approaches have been proposed to deal with different problems in proteins and protein-related systems (Chen et al. 2006a, 2006b; Chen and Li 2007a, 2007b; Diao et al. 2008; Du and Li 2006; Fang et al. 2008; Gao et al. 2005; Kurgan et al. 2007; Li and Li 2008; Lin and Li 2007a, 2007b; Liu et al. 2005; Mondal et al. 2006; Mundra et al. 2007; Nanni and Lumini 2008a, 2008b; Pu et al. 2007; Shi et al. 2007a; Shi et al. 2007b; Wang et al. 2004; Xiao et al. 2006; Zhang et al. 2006a, 2006b, 2007a; Zhang and Ding 2007; Zhou et al. 2007a, 2007b). Owing to its wide usage, a very flexible PseAA composition generator called "PseAAC" (Shen and Chou 2008) was recently established at the website http://chou.med.harvard.edu/bioinf/PseAA/, which users can use to generate 63 different kinds of PseAAC. Here, we shall use a different approach to formulate the PseAAC, the so-called sequence-segmented PseAAC.

The sequence-segmented PseAAC method can be described as follows. First, the kth protein sequence p^k is segmented into M same-length segmentations. Second, the PseAAC of each segment is calculated using the PseAAC method. Then, the protein sequence p^k is characterized as the following matrix:

$$\mathrm{PseAAS}^k = \left\{ \begin{array}{ccccc} c^k_{1,1} & \cdots & c^k_{1,m} & \cdots & c^k_{1,M} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c^k_{i,1} & \cdots & c^k_{i,m} & \cdots & c^k_{i,M} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c^k_{20,1} & \cdots & c^k_{20,m} & \cdots & c^k_{20,M} \\ \theta^k_{1,1} & \cdots & \theta^k_{1,m} & \cdots & \theta^k_{1,M} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \theta^k_{j,1} & \cdots & \theta^k_{j,m} & \cdots & \theta^k_{j,M} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \theta^k_{\lambda,1} & \cdots & \theta^k_{\lambda,m} & \cdots & \theta^k_{\lambda,M} \end{array} \right\}, \qquad (3)$$
$$k = 1, 2, \ldots, N$$
$$m = 1, 2, \ldots, M$$

where $\left[c^k_{1,m}, \ldots, c^k_{i,m}, \ldots c^k_{20,m}, \theta^k_{1,m}, \ldots \theta^k_{j,m}, \ldots \theta^k_{\lambda,m}\right]^{\mathrm{T}}$ is the PseAAC feature vector of the mth segment of p^k. The first 20 elements represent the AAC, and the following λ elements represent the PseAAC, which can be calculated using different PseAAC approaches.

Sequence-segmented moment descriptor PseAAC According to our previous moment descriptor approach (Shi et al. 2006), the λ elements (λ = 40) can be calculated by the following formula:

$$\theta^k_{j,m} = \begin{cases} \frac{1}{L^k_m} \sum\limits_{l=1}^{L^k_m} s^k_{j,l} \cdot l\,, & (\;1 \leq \mathrm{j} \leq 20) \\ \frac{1}{L^k_m} \sum\limits_{l=1}^{L^k_m} \left[s^k_{(j-20),l} \cdot l - \theta^k_{(j-20),m} \right]^2, & (21 \leq \mathrm{j} \leq 40) \end{cases},$$
$$l = 1, 2, \ldots, L^k_m \qquad (4)$$

where L^k_m is the length of the mth subsequence of protein sequence p^k, and $s^k_{j,l}$is the position indicator of natural amino acid α_j of the subsequence p^k_m, which is defined as

$$s^k_{j,l} = \begin{cases} 1 & \text{if amino acid } \alpha_j \text{ appears at position } \\ & l \text{ in the sub sequence } p^k_m \\ 0 & \text{if amino acid } \alpha_j \text{ does not appear at position } \\ & l \text{ in the sub sequence } p^k_m \end{cases}$$

For convenience, the feature set based on this sequence-segmented moment descriptor PseAAC approach can be denoted MDS$_m$.

Sequence-segmented multiscale energy PseAAC of the amino acid evolutionary conservation scores According to our previous work (Shi and Zhang et al. 2007a; Zhang et al. 2007a, 2007b), the residue conservation scores are calculated with the von Neumann entropy, and then the protein sequence of letters can be translated into a conservation score sequence. The numerical sequence can be segmented into M same-length segmentations. Using the Symlet wavelet basis function (Pittner and Kamarthi 1999), the λ elements of mth subsequence p_m^k can be calculated by the following formulae:

$$\theta_{j,m}^k = d_{j,m}^k = \sqrt{\frac{1}{\Omega_{j,m}^k}\sum_{\omega=0}^{\Omega_{j,m}^k-1}\left[u_{j,m}^k(\omega)\right]^2}, \quad 1 \leq j \leq \lambda - 1 \tag{5}$$

$$\theta_{\lambda,m}^k = a_{(\lambda-1),m}^k = \sqrt{\frac{1}{\Omega_{(\lambda-1),m}^k}\sum_{\omega=0}^{\Omega_{(\lambda-1),m}^k-1}\left[v_{(\lambda-1),m}^k(\omega)\right]^2} \tag{6}$$

where $(\lambda - 1)$ is the coarsest scale of decomposition, $d_{j,m}^k$ is the root mean square energy of the wavelet detail coefficients at the corresponding jth scale, $a_{(\lambda-1)}^k$ is the root mean square energy of the wavelet approximation coefficients at the scale $(\lambda - 1)$, $\Omega_{j,m}^k$is the number of the wavelet detail coefficients, $\Omega_{(\lambda-1),m}^k$is the number of the wavelet approximation coefficients, $u_{j,m}^k(\omega)$is the ωth detail coefficient at the corresponding jth scale, and $v_{(\lambda-1),m}^k(\omega)$ is the ωth approximation coefficient at the scale $(\lambda - 1)$. In general, for the mth protein subsequence with length L_m^k, $(\lambda - 1)$ equals $\mathrm{INT}(\log_2 L_m^k)$.

For convenience, the feature set based on this sequence-segmented PseAAC approach can be denoted MSES$_m$.

Sequence-segmented autocorrelation function PseAAC According to Chou's approach (Chou and Cai 2003), the λ elements can be calculated by the following formulae:

$$\theta_{j,m}^k = \frac{1}{L_m^k - j}\sum_{l=1}^{L_m^k - j} J_{l,l+j}^k, \qquad j = 1, 2, \ldots, \lambda \tag{7}$$

$$J_{l,l+j}^k = \frac{1}{\Lambda}\sum_{g=1}^{\Lambda}\left[\Phi_g(R_{l+j}) - \Phi_g(R_l)\right]^2 \tag{8}$$

Here L_m^k is the length of the mth subsequence of p^k, $\Phi_g(R)$ is the gth function of the amino acid R, and Λ is the total number of functions considered. In this current study, three different functions $(\Lambda = 3)$ are used to reflect the characters of an amino acid: $\Phi_1(R_i)$ refers to the hydrophobicity of amino acid R_i, taken from Tanford (1962), $\Phi_2(R_i)$ refers to the hydrophilicity of amino acid R_i, taken from Hopp and Woods (Hopp and Woods 1981), and $\Phi_3(R_i)$ is the side-chain mass of R_i, which can be obtained from any biochemistry textbook.

For convenience, the feature set based on this sequence-segmented PseAAC approach can be denoted ACFS$_m$.

Multiclass support vector machine

A support vector machine (SVM) is a learning machine based on statistical learning theory (Vapnik 1998). Due to its powerful discrimination, it has been successfully applied in medicine, bioinformatics, computational biology, etc. SVM was originally designed for binary classification, whereas the prediction of protein quaternary structures is a multiclass prediction problem. We can decompose the multiple classes into a series of binary classes, and construct multi-binary-class SVM classifiers to solve such a problem. Normally, the "one-versus-one (OVO)" or "one-versus-all" approach is employed for a multi-class SVM classifier (Hsu and Lin 2002). In this current study, the "OVO" approach was used. This method involves constructing an individual binary SVM classifier for each pair of classes. Hence, if there are η classes, a total of $\eta(\eta - 1)/2$ classifiers will be constructed. Unseen test instance prediction follows the voting strategy. Predictions are made with each binary classifier and a label is assigned to the class with the maximum number of votes. When a tie occurs (i.e., the two classes have identical votes), class label assignment is made on the basis of the largest index.

All of the computations were performed using the LIBSVM standard package, which can be freely downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/ for academic research (Hsu and Lin 2002). The various user-defined parameters, e.g., the radial basis kernel function (RBF) parameter γ and the regularization parameter C, were optimized on the training dataset.

Assessment of the prediction system

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical applications: independent dataset test, subsampling test, and jackknife test (Chou and Zhang 1995; Zhou 1998). However, as elucidated by Chou and Shen (2008) and demonstrated in Chou and Shen (2007), among the three cross-validation methods, the jackknife test is deemed to be the most objective method that will always yield a unique result for a give benchmark dataset, and hence it has recently been used by many investigators to examine the accuracies of various predictors (Chen et al. 2007; Diao et al. 2008; Ding et al. 2007; Fang et al. 2008; Gao et al. 2005; Guo et al. 2006; Li and Li 2008; Liu et al. 2007b; Nanni and Lumini 2008a, 2008b; Niu et al. 2006; Shen and Chou 2007; Shen et al. 2007; Shi et al. 2007a, 2007b; Sun and Huang 2006; Tan et al. 2007; Wang et al. 2005; Wen et al. 2007; Xiao et al. 2005, 2006; Zhang et al.

2006a, 2007a; Zhang and Ding 2007; Zhou and Assa-Munt 2001; Zhou and Cai 2006; Zhou and Doctor 2003; Zhou et al. 2007a, 2007b). During the process of jackknife analysis, the datasets are actually open, and a protein will move from one to another. The total prediction accuracy (Q) and the prediction accuracy for each class of protein quaternary structure (Q_η) calculated when assessing the the prediction system are given by:

$$Q = \sum_{\eta=1}^{7} p(\eta) \bigg/ N \qquad (9)$$

$$Q_\eta = p(\eta)/\mathrm{obs}(\eta) \qquad (10)$$

Here, N is the total number of sequences, obs(η) is the number of sequences observed in the η class protein quaternary structure, and $p(\eta)$ is the number of correctly predicted sequences of the η class protein quaternary structure.

Results and discussion

Results from different sequence-segmented PseAAC methods

Different feature vector sets (e.g., AACS$_m$, MDS$_m$, MSES$_m$ and ACFS$_m$) were employed as input feature vectors for RBF SVM. The performance of each trained module was evaluated with a jackknife cross-validation test. The classification performances of the different sequence-segmented PseAAC methods with the “OVO” approach are summarized in Table 1, which shows that the overall success rates of AACS$_4$, MDS$_5$, MSES$_3$ and ACFS$_5$ are 87.59, 89.07, 88.28 and 88.15%, respectively, which are 3.19, 2.21, 1.95 and 2.58% higher than those from their corresponding nonsegmentation methods; that is, AACS$_1$, MDS$_1$, MSES$_1$ and ACFS$_1$. The feature vector sets MDS$_m$, MSES$_m$ and ACFS$_m$ extracted from the current sequence-segmented PseAAC methods involve some information about long-distance interactions between residues, MSES$_m$ also involves protein evolutionary conservation information, and ACFS$_m$ also involves the physicochemical properties of the residues. The results indicate that the segmentation (that is, the subsequence) may be related to the protein function domain, and that it contains more protein quaternary structure information.

Table 1 The results (in percentages) from using different segmented PseAA methods with RBF SVM and the OVO approach in jackknife tests

	AACS$_m$		MDS$_m$		MSES$_m$		ACFS$_m$	
	m = 1	m = 4	m = 1	m = 5	m = 1	m = 3	m = 1	m = 5
1EM	89.01	89.27	89.79	87.96	88.22	90.05	84.82	89.27
2EM	89.60	91.55	91.80	92.41	91.80	91.31	92.78	92.29
3EM	81.28	83.98	83.64	86.00	83.14	84.82	81.96	84.65
4EM	89.14	92.65	91.29	93.78	90.84	92.99	91.06	92.87
5EM	75.93	74.07	74.07	79.63	75.39	77.78	68.52	79.63
6EM	62.37	73.52	68.99	78.05	71.43	76.66	65.16	73.52
8EM	74.52	78.34	78.34	82.80	77.71	79.62	75.80	80.25
Q%	84.40	87.59	86.86	89.07	86.33	88.28	85.57	88.15

Performance of the prediction system influenced by the number of segments

The performance of the prediction system can be affected by m, the segmentations of a protein sequence. The results obtained using the fivefold cross-validation test (5CV) are shown in Fig. 1. From Fig. 1, it is clear that compared with the case of m = 1, the overall success rates are significantly enhanced by segmenting the protein sequence into m segmentations. However, the overall success rate does not always monotonously increase with m. Actually, different datasets may have different optimal values for m that yield the highest overall success rate. For example, the optimal m values for the feature sets of AACS$_m$, MDS$_m$, MSES$_m$ and ACFS$_m$ are 4, 5, 3 and 5, respectively, and their corresponding overall success rates are 86.11, 87.72, 86.7 and 86.39%, respectively.

Comparison with Chou’s results

The corresponding comparison with Chou’s method (Chou and Cai 2003) is shown in Table 2. Our prediction performance is superior to that of Chou’s method. The results

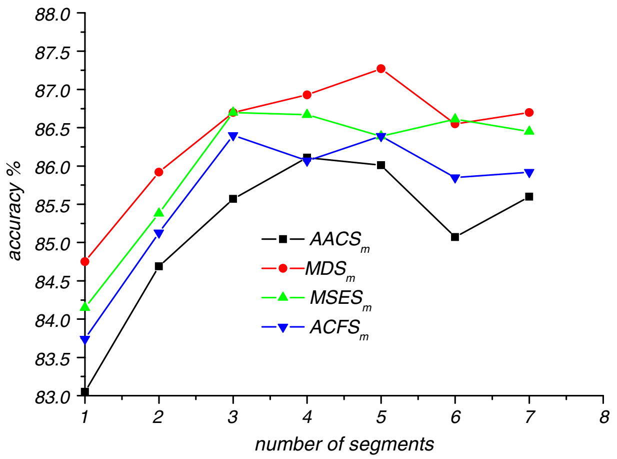

Fig. 1 The relationship between the number of segments (x-axis) and the prediction accuracy (y-axis) in the 5CV test. Prediction is performed using the RBF kernel function support vector machine

Table 2 Comparison with Chou's method (Chou and Cai 2003)

	1EM	2EM	3EM	4EM	5EM	6EM	8EM	Q%
Chou's results	80.9	85.7	77.9	85.4	1.9	62.7	54.1	78.5
MDS_5	87.96	92.41	86.00	93.78	79.63	78.05	82.80	89.07
$MSES_3$	90.05	91.31	84.82	92.99	77.78	76.66	79.62	88.28
$ACFS_5$	89.27	92.29	84.65	92.87	79.63	73.52	80.25	88.15

show that the current sequence-segmented PseAAC methods can successfully predict protein quaternary structures. It may also be very applicable to similar prediction tasks.

Conclusion

In the current study, a novel approach involving a sequence-segmented PseAAC was introduced in order to predict protein quaternary structures. The rates of correct identification suggest that the subsequences of an oligomeric protein do contain more information about its quaternary structure. Feature vectors based on the sequence-segmented PseAAC approach appear to capture essential information about the compositions and hydrophobicities of residues in the surface patches buried in the interfaces of associated subunits. The results also indicate that the current sequence-segmented PseAAC approach is quite promising and may at least provide a complimentary approach to existing methods.

Acknowledgments The authors would like to thank Prof. Kuo-Chen Chou (Gordon Life Science Institute, San Diego, California, USA) for providing the datasets. 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).

References

Call ME, Schnell JR, Xu C, Lutz RA, Chou JJ, Wucherpfennig KW (2006) The structure of the zetazeta transmembrane dimer reveals features essential for its assembly with the T cell receptor. Cell 127:355–368

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

Chen HL, Skolnick J (2007) M-TASSER: an algorithm for protein quaternary structure prediction. Biophys J BioFAST. doi: 10.1529/biophysj.107.114280

Chen H, Zhou HX (2005) Prediction of interface residues in protein–protein complexes by a consensus neural network method: test against NMR data. Proteins 61:21–35

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

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

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

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

Chou KC (2004a) Molecular therapeutic target for type-2 diabetes. J Proteome Res 3:1284–1288

Chou KC (2004b) Review: Structural bioinformatics and its impact to biomedical science. Curr Med Chem 11:2105–2134

Chou KC (2004c) Insights from modelling three-dimensional structures of the human potassium and sodium channels. J Proteome Res 3:856–861

Chou KC (2004d) Insights from modelling the 3D structure of the extracellular domain of alpha7 nicotinic acetylcholine receptor. Biochem Biophys Res Commun 319:433–438

Chou KC (2004e) Modelling extracellular domains of GABA-A receptors: subtypes 1, 2, 3, and 5. Biochem Biophys Res Commun 316:636–642

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

Chou KC, Cai YD (2003) Predicting protein quaternary structure by pseudo amino acid composition. Protein Struct Funct Genet 53:282–289

Chou KC, Shen HB (2007) Review: recent progresses in protein subcellular location prediction. Anal Biochem 370:1–16

Chou KC, Shen HB (2008) Cell-PLoc: a package of web-servers for predicting subcellular localization of proteins in various organisms. Nat Protoc 3:153–162

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

Diao Y, Ma D, Wen Z, Yin J, Xiang J, Li M (2008) Using pseudo amino acid composition to predict transmembrane regions in protein: cellular automata and Lempel–Ziv complexity. Amino Acids 34:111–117

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

Doyle DA, Morais CJ, Pfuetzner RA, Kuo A, Gulbis JM, Cohen SL, Chait BT, MacKinnon R (1998) The structure of the potassium channel: molecular basis of K+ conduction and selectivity. Science 280:69–77

Du P, Li Y (2006) Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. BMC Bioinformatics 7:518

Einstein E, Schachman HK (1989) Determining the roles of subunits in protein function. In: Creighton TE (ed) Protein function: a practical approach. IRL, London, pp 135–176

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

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

Garian R (2001) Prediction of quaternary structure from primary structure. Bioinformatics 17:551–556

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

Hopp TP, Woods KR (1981) Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci USA 78:3824–3828

Hsu C, Lin CJ (2002) A comparison of methods for multi-class support vector machines. IEEE Trans Neural Netw 13:415–425

Inbar Y, Benyamini H, Nussinov R, Wolfson HJ (2005) Prediction of multimolecular assemblies by multiple docking. J Mol Biol 349(2):435–447

Kim WK, Ison JC (2005) Survey of the geometric association of domain–domain interfaces. Proteins 61:1075–1088

Klotz IM, Darnell DW, Langerman NR (1975) Quaternary structure of proteins. In: Neurath H, Hill RL (eds) The proteins, vol 1, 3rd edn. Academic, New York, pp 226–241

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 (2008) Using pseudo amino acid composition to predict protein subnuclear location with improved hybrid approach. Amino Acids 34:119–125

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

Liu Y, Carbonell J, Gopalakrishnan V, Weigele P (2007a) Discriminative graphical models for protein quaternary structure motif detection. In: ICML2007 Workshop on Constrained Optimization and Structured Output Spaces, Corvallis, OR, 24 June 2007

Liu DQ, Liu H, Shen HB, Yang J, Chou KC (2007b) Predicting secretory protein signal sequence cleavage sites by fusing the marks of global alignments. Amino Acids 32:493–496

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

Nanni L, Lumini A (2008a) Combing ontologies and dipeptide composition for predicting DNA-binding proteins. Amino Acids. doi:10.1007/s00726-007-0018-1

Nanni L, Lumini A (2008b) Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. Amino Acids. doi:10.1007/s00726-007-0016-3

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

Oxenoid K, Chou JJ (2005) The structure of phospholamban pentamer reveals a channel-like architecture in membranes. Proc Natl Acad Sci USA 102:10870–10875

Pittner S, Kamarthi SV (1999) Feature extraction from wavelet coefficients for pattern recognition tasks. IEEE Trans Pattern Anal Mach Intell 21:83–88

Price NC (1994) Assembly of multi-subunit structure. In: Pain RH (ed) Mechanisms of protein folding. Oxford University Press, New York, pp 160–193

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

Schnell JR, Chou JJ (2008) Structure and mechanism of the M2 proton channel of influenza A virus. Nature 451:591–595

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

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 (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: Rajapakse JC et al. (eds) PRIB 2006, LNBI 4146. Springer, Berlin, pp 105–114

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

Shi JY, Zhang SW, Pan Q, Zhou GP (2007b) Using pseudo amino acid composition to predict protein subcellular location: approached with amino acid composition distribution. Amino Acids. doi:10.1007/s00726–007-0623-z

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 33:669–675

Tanford C (1962) Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. J Am Chem Soc 84:4240–4274

Terry BF, Richard MC (1998) Determination of protein–protein interactions by matrix-assisted laser desorption/ionization mass spectrometry. J Mass Spectrom 33:697–704

Vapnik V (1998) Statistical learning theory. Wiley, New York

Wang M, Yang J, Chou KC (2005) Using string kernel to predict signal peptide cleavage site based on subsite coupling model. Amino Acids 28:395–402 (Erratum, ibid. 2005, 29:301)

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

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, Huang Y, Chou KC (2005) Using complexity measure factor to predict protein subcellular location. 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 location. Amino Acids 30:49–54

Yu XJ, Wang C, Li YX (2006) Classification of protein quaternary structure by function domain composition. BMC Bioinformatics 7:187

Zhang TL, Ding YS (2007) Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes. Amino Acids 33:623–629

Zhang SW, Pan Q, 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 pseudo amino acid composition:

approached with an improved feature extraction and naive Bayes feature fusion. Amino Acids 30:461–468

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

Zhang SW, Zhang YL, Yang HF, Zhao CH, Pan Q (2007a) 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. doi:10.1007/s00726-007-0010-9

Zhang SW, Zhang YL, Pan Q, Cheng YM, Chou KC (2007b) Estimating residue evolutionary conservation by introducing von Neumann entropy and a novel gap-treating approach. Amino Acids. doi:10.1007/s00726-007-0586-0

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, Cai YD (2006) Predicting protease types by hybridizing gene ontology and pseudo amino acid composition. Proteins Struct Funct Genet 63:681–684

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 (2007a) Improved prediction of subcellular location for apoptosis proteins by the dual-layer support vector machine. Amino Acids. doi:10.1007/s00726-007-0608-y

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

Zhu HB, Domingues FS, Sommer I, Lengauer T (2006) NOXclass: prediction of protein–protein interaction types. BMC Bioinformatics 7:27