Amino Acids (2006) 30: 461–468
DOI 10.1007/s00726-006-0263-8

Amino Acids
Printed in Austria

Prediction of protein homo-oligomer types by pseudo amino acid composition: Approached with an improved feature extraction and Naive Bayes Feature Fusion

S.-W. Zhang[1,2], **Q. Pan**[1], **H.-C. Zhang**[2], **Z.-C. Shao**[1], and **J.-Y. Shi**[1]

[1] College of Automatic Control, Northwestern Polytechnical University, Xi'an, China
[2] College of Life Science, Northwestern Polytechnical University, Xi'an, China

Received August 1, 2005
Accepted September 6, 2005
Published online May 15, 2006; © Springer-Verlag 2006

Summary. The interaction of non-covalently bound monomeric protein subunits forms oligomers. The oligomeric proteins are superior to the monomers within the scope of functional evolution of biomacromolecules. Such complexes are involved in various biological processes, and play an important role. It is highly desirable to predict oligomer types automatically from their sequence. Here, based on the concept of pseudo amino acid composition, an improved feature extraction method of weighted auto-correlation function of amino acid residue index and Naive Bayes multi-feature fusion algorithm is proposed and applied to predict protein homo-oligomer types. We used the support vector machine (SVM) as base classifiers, in order to obtain better results. For example, the total accuracies of A, B, C, D and E sets based on this improved feature extraction method are 77.63, 77.16, 76.46, 76.70 and 75.06% respectively in the jackknife test, which are 6.39, 5.92, 5.22, 5.46 and 3.82% higher than that of G set based on conventional amino acid composition method with the same SVM. Comparing with Chou's feature extraction method of incorporating quasi-sequence-order effect, our method can increase the total accuracy at a level of 3.51 to 1.01%. The total accuracy improves from 79.66 to 80.83% by using the Naive Bayes Feature Fusion algorithm. These results show: 1) The improved feature extraction method is effective and feasible, and the feature vectors based on this method may contain more protein quaternary structure information and appear to capture essential information about the composition and hydrophobicity of residues in the surface patches that buried in the interfaces of associated subunits; 2) Naive Bayes Feature Fusion algorithm and SVM can be referred as a powerful computational tool for predicting protein homo-oligomer types.

Keywords: Naive Bayes Feature Fusion – Support vector machine – Pseudo amino acid composition – Weighted auto-correlation function – Homo-oligomer

Introduction

It is generally accepted that the amino acid sequence of most, not all, proteins contains all the information needed to fold the protein into its correct three-dimension structure (Anfinsen et al., 1961; Anfinsen, 1973). At the next level of protein organization, tertiary structures associate into quaternary structures. 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 non-covalent interactions and, in some cases, disulfide bonds. The association of subunits depends upon the existence of complementary 'patches' on their surfaces. The patches are buried in the interfaces formed by the subunits, thus, play a role in both tertiary and quaternary structure. Jones and Thornton (1997a, b) used a series of parameters to characterize and predict protein–protein interfaces on the basis of patch analysis. This suggests that primary sequences contain quaternary structure information (Garian, 2001).

The results of theoretical computing methods from the primary sequences can be improved not only by adopting powerful algorithms, but also by using an effective feature extraction method. The existing algorithms for predicting protein attributes were mostly based on the amino acid composition (Bahar et al., 1997; Cedano et al., 1997; Chou and Zhang, 1994; Chou, 1995, 2000a; Chou and Elord, 1999; Zhou, 1998; Zhou and Assa-Munt, 2001; Liu and Chou, 1999; Muskal and Kim, 1992; Nakashima et al., 1986; Nakashima and Nishikawa, 1994; Reinhardt and Hubbard, 1998; Zhou and Doctor, 2003). This is because the extremely large numbers of sequence order patterns in proteins and their diverse lengths have made it very difficult to take into account the sequence order effect in both the algorithm formulation and the training data construction. To tackle such a difficult problem, a set of discrete numbers was introduced to approximately reflect the sequence order effect. Recently, several feature extraction methods of taking into account the sequence order effect have been developed and applied successfully for predicting protein attributes, such as incorporating quasi-sequence-order effect (Chou, 2000b; Chou and Cai, 2003a), pseudo-amino acid composition (Chou, 2001, 2005; Chou and Cai, 2003a, b, c, 2004a, b; Gao et al., 2005; Pan et al., 2003; Wang et al., 2004, 2005; Xiao et al., 2005a, b, 2006), or the auto-correlation function (Cornette et al., 1987; Zhang and Zhang, 1998; Feng, 2001; Zhang et al., 2003).

Given a polypeptide chain, will it form a dimmer, trimer, or any other oligomer? This is important, because the functions of proteins are closely related to their quaternary attributes (Chou, 1988, 2004e). The subunit construction of many enzymes provides the structural basis for the regulation of their activities, and indispensable function for many important biological processes. Thus, in the protein universe, there are many different classes of subunit construction, such as monomer, dimmer, trimer, tetramer, and so forth. Some special functions are realized only when protein molecules are formed in oligomers; e.g., GFAT, a molecular therapeutic target for type-2 diabetes, performs its special function when it is a dimer (Chou, 2004a), some ion channels are formed by a tetramer (Chou, 2004b), and some functionally very important membrane proteins are of pentamer (Chou, 2004c, d; Oxenoid and Chou, 2005).

Garian (2001) predicted homodimer and non-homodimer using decision-tree models and a feature extraction method (simple binning function), and found that protein sequences contain quaternary structure information. Chou and Cai (2003b) also researched this question with a pseudo-amino acid composition feature extraction method. In our previous work, we classified homodimers and non-homodomers using the feature extraction method of amino acid index auto-correlation functions (Zhang et al., 2003). In this paper, based on the concept of pseudo amino acid composition (Chou, 2001), we try to develop improved feature extraction method of incorporating sequence order effect, which is that the feature vector representing one protein sequence was composed of 20 amino acid components and a set of sequence weighted auto-correlation functions. This improved feature extraction method is combined felicitously with a support vector machine (Vapnik, 1995, 1998) and Naive Bayes fusion algorithm (Kuncheva, 2002) to predict homo-oligomer types (homodimers, homotrimers, homotetramers and homohexamers).

Materials and methods

Database

The dataset1283 consists of 1283 homo-oligomeric protein sequences, 759 of which are homodimers (2EM), 105 homotrimers (3EM), 327 homotetramers (4EM) and 92 homohexamers (6EM). This dataset was obtained from SWISS-PROT database (Bairoch and Apweiler, 1996) and limited to the prokaryotic, cytosolic subset of homo-oligomers in order to eliminate membrane proteins and other specialized proteins.

Improved feature extraction method

Since the information within the primary sequence is greatly reduced by considering the amino acid composition alone, the sequence orders of amino acids in the query protein have been taken into account. Thus, based on the concept of pseudo amino acid composition (Chou, 2001), an improved feature extraction method has been put forward here, which is the weighted auto-correlation function based on the physicochemical properties of amino acid along the primary sequence of the query protein. In other words, in addition to the 20-D components of the amino acid frequencies, other λ-D components should be added in to form a $(20 + \lambda)$-D vector. Thus the attribute vector will be defined as:

$$\mathbf{x} = [f_1, f_2, \ldots, f_i, \ldots, f_{20}, r_1, r_2, \ldots, r_j, \ldots, r_\lambda]^T \qquad (1)$$

Here f_i $(i = 1, 2, \ldots, 20)$ is the occurrence frequencies of 20 amino acid in the protein concerned, arranged alphabetically according to their signal letter codes. $r_j(j = 1, 2, \ldots, \lambda)$is the weighted auto-correlation function, and λ is an integer to be determined by the optimum prediction. In order to calculate the weighted auto-correlation functions, we replace each residue in the primary sequence by its amino acid index (Shuichi et al., 1999). Here an amino acid index is a set of 20 numerical values representing any of the different physicochemical properties of the 20 amino acids, which may be accessed through the DBGET/LinkDB system at GenomeNet (http://www.genome.ad.jp/dbget) or may be downloaded by anonymous FTP (ftp://genome.ad.jp/db/genoment/aaindex). Consequently, the replacement results in a numerical sequence: $h_1, h_2, \ldots, h_l, \ldots, h_L$.

The weighted auto-correlation functions r_j are defined as:

$$r_j = \frac{w}{L-j}\sum_{l=1}^{L-j} h_l h_{l+j}, \quad j = 1, 2, \ldots, \lambda \qquad (2)$$

Here h_l is the amino acid index for the l-th residue, w is weighted factor and L is the length of protein sequence.

According to the description above, we extract six attribute parameter sets from protein primary sequences, which are clearly shown in Table 1.

Support vector machine (SVM)

The basic idea of applying SVM (Vapnik, 1995, 1998) to pattern classification can be outlined briefly as follows: First, map the input vectors into one feature space (possible with a higher dimension). Then, within this feature space, construct a hyperplane which can separate two classes. The mapping function will involve only the relatively low-dimensional vectors in the input space and dot products in the feature space. These dot products are represented by kernel functions. SVM is of the ability to deal with a large number of features.

Table 1. Seven parameter datasets extracted from protein primary sequences

Symbol	Parameter dataset
A[a]	This set is composed of amino acids compositions and the weighted auto-correlation functions of amino acid residue index of Quan-Sejnowski.
B[b]	This set is composed of amino acids compositions and the weighted auto-correlation functions of amino acid residue index of Quan-Sejnowski.
C[c]	This set is composed of amino acids compositions and the weighted auto-correlation functions of amino acid residue index of Meek-Rossetti.
D[d]	This set is composed of amino acids compositions and the weighted auto-correlation functions of amino acid residue index of Robson-Osguthorpe.
E[e]	This set is composed of amino acids compositions and the weighted auto-correlation functions of amino acid residue index of Sneath.
F	The quasi-sequences-order effect parameter set extracting based on Chou's method (Chou, 2000; Chou and Cai, 2003a).
G	This set is composed of amino acid compositions.

[a] QIAN880132 Weights for coil at the window position of −1 (Qian and Sejnowski, 1988)
[b] QIAN880119 Weights for beta-sheet at the window position of −1 (Qian and Sejnowski, 1988)
[c] MEEJ810101 Retention coefficient in NaC104 (Meek and Rossetti, 1981)
[d] ROBB790101 Hydration free energy (Robson and Osguthorpe, 1979)
[e] SNEP660103 Principal component III (Sneath, 1966)
These index values can be found in the web, http://www.genome.ad.jp/dbget/aaindex.html

The decision function implemented by SVM can be written as:

$$f(x) = \operatorname{sgn}\left(\sum_{\mu \in SV} y_\mu \alpha_\mu k(x, x_\mu) + b \right) \tag{3}$$

Three typical kernel functions are listed below:
Polynomial function

$$k(x_\mu, x_\eta) = (x_\mu \bullet x_\eta + 1)^d \tag{4}$$

Radial basis function (RBF)

$$k(x_\mu, x_\eta) = \exp(-\gamma \|x_\mu - x_\eta\|^2) \tag{5}$$

Sigmoid function

$$k(x_\mu, x_\eta) = \tanh\left[b(x_\mu \bullet x_\eta) + c\right] \tag{6}$$

Naive Bayes Feature Fusion algorithm

According to Kuncheva's multi-classifier fusion idea (Kuncheva, 2002), we introduce Naive Bayes multi-feature fusion algorithm. This scheme assumes that the features are mutually independent; for each feature set, a $c \times c$ confusion matrix CM^k is calculated by applying the classifier output D_k to the training dataset.

$$\mathrm{CM}^k = \begin{bmatrix} cm^k_{1,1} & cm^k_{1,2} & \cdots & cm^k_{1,c} \\ cm^k_{2,1} & cm^k_{2,2} & \cdots & cm^k_{2,c} \\ \vdots & \vdots & \ddots & \vdots \\ cm^k_{c,1} & cm^k_{c,2} & \cdots & cm^k_{c,c} \end{bmatrix} \tag{7}$$

For the feature set $k = 1, 2, \ldots, K$; where each row ϕ corresponds to class w_ϕ and each column φ corresponds to the classifier output $D_k = w_\varphi$. Thus, $cm^k_{\phi\varphi}$ is the number of elements of the k-th feature set whose true class label is w_ϕ, and was assigned by the classifier to class w_φ. By $cm^k_{\bullet,\varphi}$ we denote the total number of elements labeled by the classifier into class w_φ (this is calculated as the sum of the φ-th column of CM^k). Using $cm^k_{\bullet,\varphi}$, a $c \times c$ label matrix LM^k is computed, whose (ϕ, φ)-th entry $lm^k_{\phi,\varphi}$ is an estimate of the probability that the true label is w_ϕ given that the classifier assigns crisp class label w_φ for the k-th feature set.

$$\mathrm{LM}^k = \begin{bmatrix} lm^k_{1,1} & lm^k_{1,2} & \cdots & lm^k_{1,c} \\ lm^k_{2,1} & lm^k_{2,2} & \cdots & lm^k_{2,c} \\ \vdots & \vdots & \ddots & \vdots \\ lm^k_{c,1} & lm^k_{c,2} & \cdots & lm^k_{c,c} \end{bmatrix}, \quad k = 1, 2, \ldots, K \tag{8}$$

$$lm^k_{\phi,\varphi} = P\left(w_\phi \middle| D_k(x) = w_\varphi \right) = \frac{cm^k_{\phi,\varphi}}{cm^k_{\bullet,\varphi}} \tag{9}$$

For every $x \in w_\phi, \phi = 1, 2, \ldots, c$, yields a crisp label vector $D_k(x)$ pointing at one of the classes, say, $w_\varphi, \varphi = 1, 2, \ldots, c$. Let $s_1, \ldots, s_K$ be the crisp class labels assigned to x by the classifier for each feature set. Then, by the independence assumption, the estimate of the probability that the true class label is w_ϕ, is calculated by

$$\Omega_\phi(x) = \prod_{k=1}^{K} P\left(w_\phi \middle| D_k(x) = s_k \right) = \prod_{k=1}^{K} lm^k_{\phi,s_k}$$
$$\phi = 1, 2, \ldots, p, \ldots, c \tag{10}$$

The maximum membership rule will class x in w_p, where $\Omega_p(x)$ is maximal.

Assessment of the prediction system

The prediction quality can be examined using the jackknife test and 10-fold cross-validation (10CV) test. The cross-validation by jackknifing is thought the most objective and rigorous way in comparison with sub-sampling test or independent dataset test (Chou and Zhang, 1995; Zhou and Assa-Munt, 2001). During the process of jackknife analysis, the datasets are actually open, and a protein will in turn move from each to the other. The total prediction accuracy (Q), the prediction accuracy (Q_k) and Matthew's Correlation Coefficient (MCC) (Fasman, 1976) for each class of homo-oligomers calculated for assessment of the prediction system are given by:

$$Q = \sum_{k=1}^{4} p(k) \Big/ N \tag{11}$$

$$Q_k = p(k) \big/ \mathrm{obs}(k) \tag{12}$$

$$\mathrm{MCC}(k) = \frac{p(k)n(k) - u(k)o(k)}{\sqrt{(p(k)+u(k))(p(k)+o(k))(n(k)+u(k))(n(k)+o(k))}} \tag{13}$$

Here, N is the total number of sequences, $\mathrm{obs}(k)$ is the number of sequences observed in k class protein homo-oligomers, $p(k)$ is the number of correctly predicted sequences of k class protein homo-oligomers, $n(k)$ is the number of correctly predicted sequences not of k class protein homo-oligomers, $u(k)$ is the number of under-predicted sequences of k class protein homo-oligomers and $o(k)$ is the number of over-predicted sequences of k class protein homo-oligomers.

Results and discussion

Results of different feature extraction methods and algorithms

The results of the SVM with one-versus-rest approach in jackknife test are shown in Table 2. The total accuracy of G feature set based only on amino acid composition is 71.24%, and the total accuracies of A, B, C, D and E based on the improved feature extraction method – the weighted auto-correlation function are 77.63, 77.16, 76.46, 76.70 and 75.06%, respectively, which are 6.39, 5.92, 5.22, 5.46 and 3.82% respectively higher than that of G feature set. The MCC of each class (2EM, 3EM, 4EM and 6EM) for A, B, C, D and E is bigger than that of the corresponding class for G feature set. These results indicate that the method of the improved feature extraction from the protein sequences is effective and feasible.

From Table 2, we can also see that the results of two different feature extraction methods (the weighted auto-correlation function method and Chou's method (Chou, 2000b; Chou and Cai, 2003a)) which integrate the influence of the sequence orders are always better than that of the method only based on amino acid composition. The results of our method of weighted auto-correlation function are the best, and the total accuracy of A, B, C, D, and E feature sets are 3.51, 3.11, 2.41, 2.65, 1.01% higher respectively than that of F feature set based on Chou's method (Chou, 2000b; Chou and Cai, 2003a). And the MCC of 2EM, 4EM and 6EM class are bigger than that of the corresponding class for F feature set.

Using the Naive Bayes feature fusion algorithm, we try to design some schemes with A, B, C, D, E, F and G feature sets. Some results of these fusion schemes are shown in Table 3, which have better results. From Tables 2 and 3, we can see that the results of different feature sets fusion are better than that of single feature set. The result of A, B, C, D, E and F feature set fusion is the best, its total accuracy is 80.83%, which is 6.78%, 3.2% higher than that of F and A feature set respectively. We can also see that some feature sets fusion can increases prediction accuracy, but some fusion scheme has not marked effect. For example, the accuracy for 'ABCDEF' fusion scheme is 80.83%, but the accuracy for 'ABCDEFG' fusion scheme decreases to 79.81%. The explanation is that strictly speaking, the features are not mutually independent, each set of A, B, C, D, E and F contains information of G feature set, and there may be some redundancy and information conflict between these feature sets.

We have analyzed 402 sets of indices. The total accuracy in 10-fold cross-validation (10 CV) test is used to evaluate the prediction performance of each amino acid

Table 2. Results of one-versus-rest approach and RBF kernel function support vector machine ($C = 1000$) in the jackknife test

	A $\gamma = 0.005$, $w = 100$, $\lambda = 30$		B $\gamma = 0.01$, $w = 100$, $\lambda = 30$		C $\gamma = 0.007$, $w = 1$, $\lambda = 30$		D $\gamma = 0.011$, $w = 10$, $\lambda = 30$		E $\gamma = 0.007$, $w = 1000$, $\lambda = 30$		F $\gamma = 1.3$ $w = 60$, $\lambda = 40$		G $\gamma = 0.04$	
	Q_k%	MCC	Q_k%	MCC	Q_k%	MCC	Q_k%	MCC	Q_k%	MCC	Q_k%	MCC	Q_k%	MCC
2EM	93.68	0.585	92.75	0.548	89.86	0.535	93.02	0.548	89.46	0.520	90.38	0.476	86.96	0.447
3EM	48.57	0.674	43.81	0.646	56.19	0.698	42.86	0.631	43.81	0.638	50.48	0.695	40.95	0.559
4EM	60.55	0.573	62.39	0.595	60.24	0.546	59.63	0.585	60.24	0.516	54.13	0.506	55.05	0.454
6EM	39.13	0.553	39.13	0.584	46.74	0.622	42.39	0.553	44.57	0.605	36.96	0.533	33.40	0.462
Q%	77.63	–	77.16	–	76.46	–	76.70	–	75.06	–	74.05	–	71.24	–

Table 3. Results of several feature fusion schemes

	ACE		ACDE		ABCDE		ABCDEF		ABCDEG		ABCDEFG	
	Q_k%	MCC	Q_k%	MCC	Q_k%	MCC	Q_k%	MCC	Q_k%	MCC	Q_k%	MCC
2EM	94.73	0.614	98.02	0.612	96.84	0.617	98.55	0.631	98.42	0.624	98.68	0.717
3EM	54.29	0.715	54.29	0.722	56.19	0.735	56.19	0.735	56.19	0.735	57.14	0.742
4EM	63.91	0.633	55.66	0.638	60.55	0.659	58.41	0.674	58.10	0.678	54.13	0.653
6EM	43.48	0.620	42.39	0.611	42.39	0.611	42.39	0.611	41.30	0.602	41.3	0.611
Q%	79.89		79.66		80.36		80.83		80.59		79.81	

Table 4. Performance of the prediction system influenced by the sample unbalance between the classes using RBF kernel function support vector machine ($C = 1000$) in a jackknife test

	Database1283				Database368			
	G $\gamma = 0.04$		C $\gamma = 0.007$, $w = 1$, $\lambda = 30$		G $\gamma = 0.04$		C $\gamma = 0.007$, $w = 1$, $\lambda = 30$	
	Q_k%	MCC	Q_k%	MCC	Q_k%	MCC	Q_k%	MCC
2EM	86.96	0.447	89.86	0.535	43.26	0.270	42.61	0.272
3EM	40.95	0.559	56.19	0.698	71.74	0.539	75.43	0.604
4EM	55.05	0.454	60.24	0.546	55.22	0.427	56.30	0.432
6EM	33.40	0.462	46.74	0.622	57.61	0.465	70.00	0.608
Q%	71.24	–	76.46	–	56.96	–	61.09	–

index. Among 402 sets of indices, about 65% can differently improve the prediction results. By the hierarchical clustering (Tomii and Kanehisa, 1996), the 402 indices can be divided into six major classes: 1) α and turn propensities, 2) β propensity, 3) amino acid composition, 4) hydrophobicity, 5) physicochemical properties, 6) other properties. We found that most of hydrophobicity amino acid indices used for predicting have better performance than that of other five classes of amino acid index, which suggests that biologically relevant complex formation is driven predominantly by the hydrophobic effect (Glase, et al., 2001). The results of five typical examples are listed in Tables 2 and 3. The amino acid indices of QIAN880132, QIAN880119, and SNEP660103 belong to the class of α and turn propensities, β propensity, and physicochemical properties respectively, but MEEJ810101 and ROBB790101 belong to the class of hydrophobicity, which have the best results in the class of themselves amino acid index.

Performance of the prediction system influenced by the unbalance of sample numbers among the four classes

To investigate the influence of the sample unbalance among the four classes, we established subset database368. The database368 is randomly selected from the database1283, which consists of 368 homo-oligomeric protein sequences, and each class has 92 protein sequences. The results of C and G sets are shown in the Table 4, and the results of database368 are the mean of five random selections.

From Table 4, it is evident that the database size and the sample unbalance among classes have great influence on the performance of the prediction system. For example, the total accuracy of C set in database368 is 61.09%, which is 15.37% lower than that of in database1283; the prediction accuracy of homodimers is 89.86% in database1283, but decreasing to 42.61% in database368; the prediction accuracy of homohexamers is 46.74% in database1283, but increasing to 70% in database368. Generally, increasing the number of the training set and decreasing the unbalance of the samples among classes can improve the performance of the prediction system, and enhance the system stability. With the increase of the number of protein sequences in the databank, this problem may be solved. In addition, we can see that the performance of C set is always better than that of the G set in both database1238 and database386. These results verify the fact that the feature extraction method of the weighted auto-correlation functions is superior to the method of amino acid composition once again, and the quaternary structure information of feature vectors extracted from primary sequences with this method is more than that of amino acid composition method.

Selection of the weighted factor w and the auto-correlation factor λ

For D set, the classifying results of different weighted factor w in 10 CV test are shown in Fig. 1, which indicates that the classifying results may be influenced by the weighted factor w at a certain extent. Obviously, there is an optimal value of weighted factor w to be selected. For radial basis function, the influence of γ should be taken into account when selecting the weighted factor w. The best results can be obtained by carefully selecting w and γ.

For the C parameter set, the classifying results of the different weighted auto-correlation factor λ in 10CV test are shown in Fig. 2.

From Fig. 2, it is seen that when $\lambda > 20$, the total accuracy, the accuracies of homodimers and homotetramers

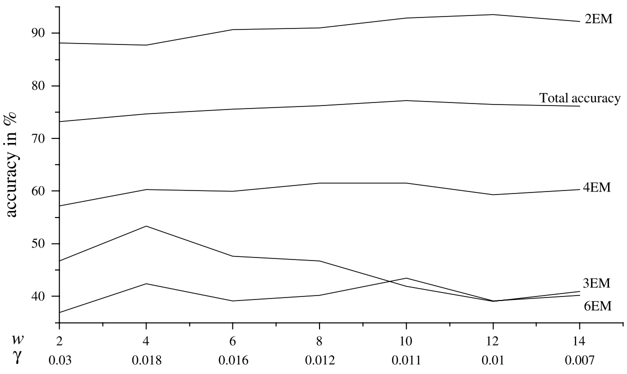

Fig. 1. The relationship between the weighted factor w (x-axis) and the classifying accuracy (y-axis) in the 10 CV test. The classification is performed for D set using RBF kernel function support vector machine ($C = 1000$, $\lambda = 30$)

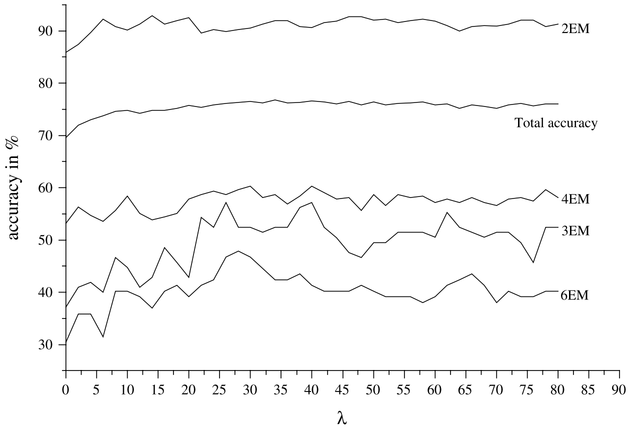

Fig. 2. The relationship between the auto-correlation factor number λ (x-axis) and the classifying accuracy (y-axis) in the 10 CV test. The classification is performed for C set using RBF kernel function support vector machine ($C = 1000$, $w = 1$)

are almost unchanged, especially, the total accuracy keeps nearly at the 75% level. However, the accuracies of homotrimers and homohexamers change more with different λ values. We think these results may be not only related to the sample size, but also to the feature of homo-oligomeric structure, for example, the homotrimer and homohexamer are composed of three subunits or double three subunits. Here, we select $\lambda = 30$.

Assigning a reliability index to the prediction

It is important to know the prediction reliability when using machine learning approaches for predicting protein homo-oligomers. For neural network methods, a Reliability Index (RI) is usually assigned according to the difference between the highest and the second-highest network output score (Rost and Sander, 1993; Reinhardt and Hubbard, 1998; Emanuelsson et al., 2000). The sample idea is easily used to SVM prediction system (Hua and Sun, 2001), i.e. assigning an RI according to the difference (noted as diff(I)) between the highest and the second-highest output value with the one-versus-rest approach in the multi-class prediction. RI is defined as:

$$\mathrm{RI} = \begin{cases} 0 & \text{if} \quad \mathrm{diff}(I) < 0.2 \\ \mathrm{INTEGER}(5\,\mathrm{diff}(I)) & \text{if} \quad 0.2 \le \mathrm{diff}(I) < 1.8 \\ 9 & \text{if} \quad \mathrm{diff}(I) > 1.8 \end{cases} \tag{14}$$

The RI assignment is a useful indication of the level of certainty in the prediction for a particular sequence.

The evaluation of the prediction system for C set in the jackknife test is shown in Figs. 3 and 4.

Figures 3 and 4 show the statistical results for protein homo-oligomeric sequences. The expected prediction accuracy with RI equals to a given value and the fraction of sequences for each given RI were calculated (Fig. 3). For

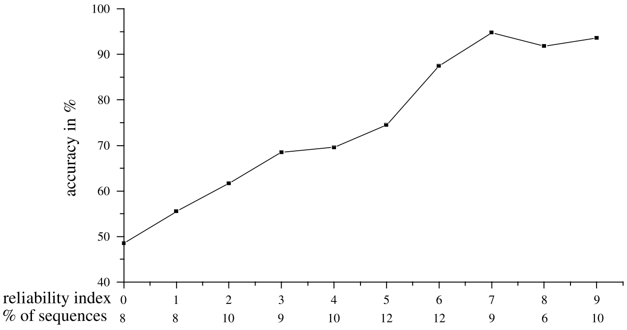

Fig. 3. Expected classifying accuracy with a reliability index equal to a given value. The fractions of sequences that are classified with $\mathrm{RI} = 0, 1, 2, \ldots, 9$ are also given for C attribute set using one-versus-rest policy and RBF kernel function support vector machine ($C = 1000$, $\gamma = 0.007$, $w = 1$, $\lambda = 30$) in jackknife test

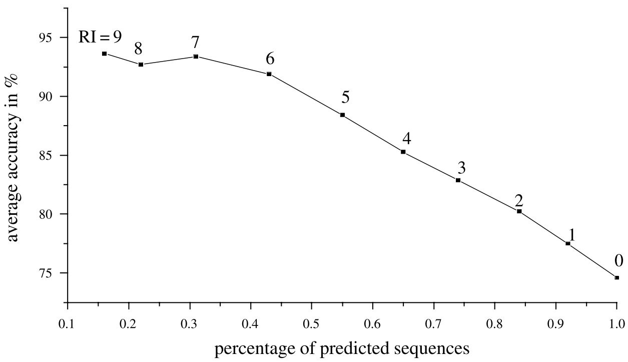

Fig. 4. Average predictive accuracy with a reliability index above a given cut-off

example, the expected accuracy for a sequence with RI = 5 is 74.51% with 12% of all sequences having RI = 5. The average prediction accuracy was also calculated for RI above a given cut-off (Fig. 4). About 74% of all sequences have RI $\geq$ 3, and of these sequences about 82.88% were correctly predicted by SVM prediction system.

Conclusion

The feature vectors composed of amino acid composition and weighted auto-correlation functions of amino acid residue index can reflect the quaternary structure information in a certain extent. With different amino acid indices, the auto-correlation factor λ and the weighted factor w, there are many integrating forms of amino acid composition and the weighted auto-correlation functions. Thus, the best prediction results can be obtained for a given data set by carefully selecting amino acid index, λ and w value.

A remarkable improvement in prediction quality has been observed by using this improved feature extraction method and support vector machine algorithm. The feature vectors based on our improved feature extraction method may contain more protein quaternary structure information, and appear to capture essential information about the composition and hydrophobicity of residues in the surface patches that buried in the interfaces of the associated subunits. Naive Bayes Feature fusion algorithm is effective for predicting homo-oligomer types, but the feature sets should be mutual independent strictly. Only under this condition, the prediction system can get optimal result. The results also indicate that the current approach is quite promising and useful to improve the prediction quality for other protein attributes as well.

Acknowledgements

We thank Professor Runsheng Chen from the Institute of Biophysics Chinese Academy of Science and Professor Zhengxiang Fu from the Northwest Polytechnical University of China for helpful suggestions. This work was supported, in part, by the Technological Innovation Foundation of the Northwestern Polytechnic University (kc02) of China.

References

Anfisen CB (1973) Principles that govern the folding of protein chains. Science 181: 223–230

Anfinsen CB, Haber E, Sela M, White FH (1961) The kinetics of the formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc Natl Acad Sci USA 47: 1309–1314

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

Bairoch A, Apweiler R (1996) The SWISS-PROT protein data bank and its new supplement TrEMBL. Nucleic Acids Res 24: 21–25

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

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

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

Chou KC (2000b) 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 43: 246–255

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

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

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

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

Chou KC (2004e) 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 (2003a) Prediction and classification of protein subcellular location: sequence-order effect and pseudo amino acid composition. J Cell Biochem 90: 1250–1260 (Addendum: J Cell Biochem 91: 1085 (2004))

Chou KC, Cai YD (2003b) Predicting protein quaternary structure by pseudo amino acid composition. Proteins Struct Func Gene 53: 282–289

Chou KC, Cai YD (2003c) A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology. Biochem Biophys Res Commun 311: 743–747

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

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, Elord DW (1999) Prediction of membrane protein types and subcellular locations. Proteins Struct Funct Genet 34: 137–153

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

Cornette JL, Cease KB, Margali H, Spouge JL, Berzofsky JA, Delisi C (1987) Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. J Mol Biol 195: 659–685

Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300: 1005–1016

Fasman GD (ed) (1976) Handbook of biochemistry and molecular biology, 3rd ed. CRC Press, Boca Raton

Feng ZP (2001) Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. Biopolymers 58: 491–509

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

Glase F, Steinberg DM, Vakser IA, Ben-Tal N (2001) Residue frequencies and pairing preferences at protein–protein interfaces. Proteins Struct Funct Genet 43: 89–102

Hua SJ, Sun ZR (2001) Support vector machine approach for protein subcellular localization prediction. Bioinformatics 17: 721–728

Jones S, Thornton JM (1997a) Analysis of protein–protein interaction sites using surface patches. J Mol Biol 272: 121–132

Jones S, Thornton JM (1997b) Prediction of protein–protein interaction sites using patch analysis. J Mol Biol 272: 133–143

Kuncheva LI (2002) Switching between selection and fusion in combining classifiers: an experiment. IEEE Trans 32: 146–156

Liu W, Chou KC (1999) Protein secondary structural content prediction. Protein Eng 12: 1041–1050

Meek JL, Rossetti ZL (1981) Factors affecting retention and resolution of peptides in HPLC. J Chromatogr 211: 15–28

Muskal SM, Kim SH (1992) Predicting protein secondary structure content: a tandem neural network approach. J Mol Biol 225: 713–727

Nakashima H, Nishikawa K (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residuepair frequencies. J Mol Biol 238: 54–61

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

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

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

Qian N, Sejnowski TJ (1988) Predicting the secondary structure of globular proteins using neural network models. J Mal Dial 202: 865–884

Reinhardt A, Hubbard T (1998) Using neural network for prediction of the subcellular location of proteins. Nucleic Acids Res 26: 2230–2236

Robson B, Osguthorpe DJ (1979) Refined models for computer simulation of protein folding. Applications to the study of conserved secondary structure and flexible hinge points during the folding of pancreatic trypsin inhibitor. J Mol Biol 132: 19–51

Rost B, Sander C (1993) Prediction of secondary structure at better than 70% accuracy. J Mol Biol 232: 584–599

Shuichi K, Hiroyuki O, Minoru K (1999) Aaindex: amino acid index database. Nucleic Acids Res 27: 368–369

Sneath PH (1966) Relations between chemical structure and biological activity in peptides. J Theor Biol 12: 157–195

Tomii K, Kanehisa M (1996) Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. Protein Eng 9: 27–36

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

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

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

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

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

Xiao X, Shao S, Ding Y, Huang Z, Huang Y, Chou KC (2005b) 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 sub-cellular location. Amino Acids 30: 49–54

Zhang CT, Zhang R (1998) A new quantitative criterion to distinguish between α/β and $\alpha+\beta$ proteins. FEBS Lett 440: 153–157

Zhang SW, Quan P, Zhang HC, Zhang YL, Wang HY (2003) Classification of protein quaternary structure with support vector machine. Bioinformatics 19: 2390–2396

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 50: 44–48

Authors' address: Shao-Wu Zhang, College of Automatic Control, Northwestern Polytechnical University, 127 YouYi West Rd., Xi'an 710072, Shaanxi, China,
Fax: +86-29-88494352, E-mail: zhangsw@nwpu.edu.cn