Using Multi-scale Glide Zoom Window Feature Extraction Approach to Predict Protein Homo-oligomer Types

QiPeng Li, Shao Wu Zhang, and Quan Pan

School of Automation/School of Mechatronics, Northwestern Polytechnical University,
127 YouYi West Rd., Xi'an 710072, Shaanxi, China
liqipeng@nwpu.edu.cn

Abstact. The concept of multi-scale glide zoom window was proposed and a novel approach of multi-scale glide zoom window feature extraction was used for predicting protein homo-oligomers. Based on the concept of multi-scale glide zoom window, we choose two scale glide zoom window: whole protein sequence glide zoom window and kin amino acid glide zoom window, and for every scale glide zoom window, three feature vectors of amino acids distance sum, amino acids mean distance and amino acids distribution, were extracted. A series of feature sets were constructed by combining these feature vectors with amino acids composition to form pseudo amino acid compositions (PseAAC). The support vector machine (SVM) was used as base classifier. The 75.37% total accuracy is arrived in jackknife test in the weighted factor conditions, which is 10.05% higher than that of conventional amino acid composition method in same condition. The results show that multi-scale glide zoom window method of extracting feature vectors from protein sequence is effective and feasible, and the feature vectors of multi-scale glide zoom window may contain more protein structure information.

Keywords: Multi-scale glide zoom window, feature extraction, pseudo amino acid compositions, homo-oligomer.

1 Introduction

In the protein universe, there are many different classes of oligomer, such as monomer, dimer, trimer, tetramer, and so forth. These quaternary structures are closely related to the functions of the proteins [1, 2]. 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 [3], some ion channels are formed by a tetramer [4], and some functionally very important membrane proteins are of pentamer [5,6,7]. 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 structure [8,9]. So, predicting oligomers types from given protein sequences is important.

M. Chetty, A. Ngom, and S. Ahmad (Eds.): PRIB 2008, LNBI 5265, pp. 78–86, 2008.

Garian [10], Chou and Cai [11], Zhang [12] predicted homodimer and non-homodimer using decision-tree models and a feature extraction method (simple binning function), pseudo-amino acid composition feature extraction method, amino acid index auto-correlation functions respectively. Zhang [13] also predicted protein homo-oligomer types by pseudo amino acid composition. They found that protein sequences contain quaternary structure information.

The concept of multi-scale glide zoom window based on the protein sequence was proposed in this paper. Three kinds of feature vector incorporating sequence order effect, that is, amino acids distance sum, amino acids mean distance and amino acids distribution , were extracted from whole protein sequence glide zoom window and kin amino acid glide zoom window of protein sequence. This new feature extraction method is combined felicitously with a support vector machine [14, 15] to predict homodimers, homotrimers, homotetramers and homohexamers.

2 Materials and Methods

2.1 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 [16] and limited to the prokaryotic, cytosolic subset of homo-oligomers in order to eliminate membrane proteins and other specialized proteins.

2.2 The Concept of Multi-scale Glide Zoom Window

Multi-scale glide zoom window of every nature amino acid can be described as multi-scale segment sequence (or, whole sequence) of one protein sequence, that is, the every scale glide zoom window of one nature amino acid can be decided by three factors: constructing rule of *x*th scale glide zoom window, *k*th protein sequence and *i*th amino acid. So, for one protein sequence, we can obtain many glide zoom windows and extract feature vectors from every glide zoom window. This novel multi-scale glide zoom window feature extraction method is very depends on constructing rule of every scale glide zoom window. In this paper, we extract feature vectors of one protein sequence from 2-scale glide zoom window. The first scale glide zoom windows of every nature amino acid are all the whole protein sequence, which provide panorama of a protein sequence. The second scale glide zoom window of every nature amino acid are kin amino acid glide zoom window, which begins from the position where every kin amino acid appears firstly and ends at the position where this kin amino acid appears lastly among the whole protein sequence, which focuses on corresponding local of every nature amino acid in a protein sequence. There are one first scale glide zoom window and twenty second scale glide zoom windows for every protein sequence. For example, for the protein sequence 'MITRMSELFLRTLRDDP', the first scale glide zoom windows of every nature amino acid are all the whole protein sequence itself 'MITRMSELFLRTLRDDP'. The second scale glide zoom window of nature amino acid M is 'MITRM', the second scale glide zoom window of nature amino acid T is 'TRMSELFLRT', the second scale glide zoom window of nature amino acid D is ‘DD’, and so on. If one nature amino acid does not appear in the protein sequence, the second scale glide zoom window of this nature amino acid is empty. The position and the width of every second scale glide zoom window are variable. Apparently, the second scale glide zoom window contains some sequence order information. The width of first scale glide zoom window is equal to the length of the protein sequence.

2.3 The Multi-scale Glide Zoom Window Feature Extraction Methods

Suppose the dataset consists of N homo-oligomeric protein sequences. p^k represents the kth protein sequence. α_i represents the ith amino acid of the nature amino acid set AA, $AA=\{A,R,N,D,C,O,E,G,H,I,L,K,M,F,P,S,T,W,Y,V\}$. Here, We can use $z_i^{x,k}$ to represent the xth scale glide zoom window of α_i in p^k . $f_i^{x,k}$ and $l_i^{x,k}$ represent the first position and last position of $z_i^{x,k}$ in the kth protein sequence p^k , respectively. $L_i^{x,k}$ is defined as length of $z_i^{x,k}$. According to the definition of first scale glide zoom window in section 2.2, every first scale glide zoom window of α_i in p^k is the same whole sequence. Apparently, $z_i^{1,k}$ is p^k . $L_i^{1,k}$ is the length of p^k ,which we can denote as L^k . $f_i^{1,k}$ and $l_i^{1,k}$ are 1 and L^k respectively. According to the definition of second scale glide zoom window in section 2.2, $f_i^{2,k}$ and $l_i^{2,k}$ are first and last position where α_i appear among p^k , respectively. $z_i^{2,k}$ is segment sequence between $f_i^{2,k}$ and $l_i^{2,k}$. $L_i^{2,k}$ is equal to $l_i^{2,k}-f_i^{2,k}$. In order to describe the positions of every nature amino acid in p^k , We first defined a position indicator $o_{i,j}^k$.

$$o_{i,j}^k=\begin{cases}1 \text{ if } \alpha_i \text{ locates in } jth \text{ position of } p^k \\ 0 \text{ if } \alpha_i \text{ does not locate in } jth \text{ position of } p^k\end{cases} \tag{1}$$

Then, we map protein sequence p^k to a position indicator matrix V^k .

$$V^k=\begin{bmatrix} v_1^k \\ \cdots \\ v_i^k \\ \cdots \\ v_{20}^k \end{bmatrix}=\begin{bmatrix} o_{1,1}^k,\cdots,o_{1,j}^k,\cdots,o_{1,L^k}^k \\ \cdots,\cdots,\cdots,\cdots,\cdots \\ o_{i,1}^k,\cdots,o_{i,j}^k,\cdots,o_{i,L^k}^k \\ \cdots,\cdots,\cdots,\cdots,\cdots \\ o_{20,1}^k,\cdots,o_{20,j}^k,\cdots,o_{20,L^k}^k \end{bmatrix}_{20\times L^k} , \quad k=1,\cdots,N \tag{2}$$

Here, position indicator vector v_i^k shows where α_i locates in the p^k.

In order to extract various feature vectors of $z_i^{x,k}$ with v_i^k, we defined a coordinate axis vector $w_i^{x,k}$.

$$w_i^{x,k} = [\xi_{i,1}^{x,k}, \xi_{i,2}^{x,k}, \ldots, \xi_{i,j}^{x,k}, \ldots, \xi_{i,L^k}^{x,k}]_{1\times L^k} \qquad , x = 1,2;\ j = 1,\ldots,L^k \tag{3}$$

Here,

$$\xi_{i,j}^{1,k} = j \qquad , j = 1,\ldots,L^k \tag{4}$$

$$\xi_{i,j}^{2,k} = \begin{cases} j - f_i^{2,k} + 1 & \text{if } f_i^{2,k} \le j \le l_i^{2,k} \\ 0 & \text{if } j < f_i^{2,k} \ or \ j > l_i^{2,k} \end{cases} \tag{5}$$

To integrate more sequence order information, according to the concept of multi-scale glide zoom window, three kinds of feature vector of every scale glide zoom window are extracted to predict homo-oligomers. The three kinds of feature vector of every scale glide zoom window are defined as follows:

1) Amino Acids Distance Sum Feature Vector

The amino acids distance sum feature vector of p^k is expressed as the following 20-D feature vector:

$$S^{x,k} = [\eta_1^{x,k}, \ldots, \eta_i^{x,k}, \ldots, \eta_{20}^{x,k}] \quad k = 1,\cdots,N \tag{6}$$

Here,

$$\eta_i^{x,k} = w_i^{x,k} \times (v_i^k)^T \quad k = 1,\cdots,N \tag{7}$$

Conveniently, $S^1$and S^2 are respectively used to present the amino acids distance sum feature sets of first and second scale glide zoom windows.

2) Amino Acids Mean Distance Feature Vector

The amino acids mean distance feature vector of p^k is expressed as the following 20-D feature vector:

$$M^{x,k} = [\mu_1^{x,k}, \ldots, \mu_i^{x,k}, \ldots, \mu_{20}^{x,k}] \quad k = 1,\cdots,N \tag{8}$$

Here,

$$\mu_i^{x,k} = \begin{cases} \frac{1}{L_i^{x,k}} w_i^{x,k} \times (v_i^k)^T & , \text{ if } L_i^{x,k} \ne 0 \\ 0 & , \text{ if } L_i^{x,k} = 0 \end{cases} \tag{9}$$

Conveniently, M^1 and M^2 are respectively used to present the amino acids mean distance feature sets of first and second scale glide zoom windows.

3) Amino Acids Distribution Feature Vector

The amino Acids distribution feature vector of p^k is expressed as the following 20-D feature vector:

$$D^{x,k} = \left[\rho_1^{x,k}, \cdots, \rho_i^{x,k}, \cdots, \rho_{20}^{x,k}\right], \quad k = 1, \cdots, N \tag{10}$$

Here,

$$\rho_i^{x,k} = \begin{cases} \dfrac{1}{L_i^{x,k}} \displaystyle\sum_{j=f_i^{x,k}}^{l_i^{x,k}} \left(o_{i,j}^k \times j - \dfrac{1}{L_i^{x,k}} w_i^{x,k} \times (v_i^k)^T \right)^2, & \text{if } L_i^{x,k} \neq 0 \\ 0 & , \text{ if } L_i^{x,k} = 0 \end{cases} \tag{11}$$

Conveniently, D^1 and D^2 are respectively used to present the amino acids distribution feature sets of first and second scale glide zoom windows. It is easy to certified that D^1 is equal to D^2, so, we can marked D^1 and D^2 as D.

2.4 Assessment of the Prediction System

The prediction quality can be examined using the jackknife test. The cross-validation by jackknifing is thought the most objective and rigorous way in comparison with sub-sampling test or independent dataset test [17, 18]. 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), Sensitivity (Q(class(*k*))) and Matthew's Correlation Coefficient (MCC) [19] for each class of homo-oligomers calculated for assessment of the prediction system are given by:

$$Q = \sum_{k=1}^{M} p_k \Big/ N \times 100\% \tag{12}$$

$$Q\left(class(k)\right) = p_k / \left(p_k + u_k\right) \tag{13}$$

$$MCC(class(k)) = \frac{p_k n_k - u_k o_k}{\sqrt{\left(p_k + u_k\right)\left(p_k + o_k\right)\left(n_k + u_k\right)\left(n_k + o_k\right)}} \tag{14}$$

Here, M is the total number of classes, p_k is the number of correctly predicted sequences of *k* class protein homo-oligomers, u_k is the number of under-predicted sequences of *k* class protein homo-oligomers, n_k is the number of correctly predicted sequences not of *k* class protein homo-oligomers, o_k is the number of over-predicted sequences of *k* class protein homo-oligomers. According to The dataset1283 used in this paper, M=4, class(1), class(2),class(3) and class(4) are 2,3,4 and 6 respectively. 2, 3, 4 and 6 represent 2EM, 3EM, 4EM and 6EM respectively.

3 Results and Discussion

3.1 The Results of Different Pseudo Amino Acids Composition Feature Sets

C presents the feature set based on the amino acid composition approach [20]. Twenty-seven feature sets of pseudo amino acid composition (PseAAC) are constructed by feature sets D, M^1, M^2, S^1, S^2 of glide zoom window and C. The results of these twenty-seven PseAAC feature sets and feature set C with RBF SVM and one-versus-one strategy in jackknife test are shown in table 1.

From Table 1, we can see that the result of $CDM^1M^2S^2$ is the best in all the feature sets, and the total accuracy is 75.53%, which is 6.71% higher than that of C. The accuracies of feature sets which include M^1, M^2 or both of them are higher than that of other feature sets which do not include M^1, M^2 or both of them. These results suggest that, in every scale glide zoom window, the feature set of amino acids mean distance is more effective and robust than other feature sets. In addition, the accuracies of feature sets which include D, S^1, S^2 except M^1 and M^2 are near that of feature set C. The reasons are that there may be some redundancy and conflict information between these feature sets, or the unbalance of sample numbers among the four classes.

Table 1. Results of 28 Feature sets with RBF SVM and one-versus-one strategy in jackknife test

Feature sets	2EM		3EM		4EM		6EM		Q%
	Q(2) %	MCC(2)	Q(3) %	MCC(3)	Q(4) %	MCC(4)	Q(6) %	MCC(6)	
C	**91.57**	**0.3582**	**42.86**	**0.5726**	**38.53**	**0.3568**	**18.48**	**0.3088**	**68.82**
CD	**95.39**	**0.6630**	**32.38**	**0.5276**	**33.03**	**0.3611**	**1.09**	**0.0992**	**67.58**
CM^1	**92.23**	**0.5152**	**50.48**	**0.6621**	**57.49**	**0.5258**	**29.35**	**0.4412**	**75.45**
CM^2	**91.17**	**0.7497**	**53.33**	**0.6511**	**55.35**	**0.5053**	**30.43**	**0.4373**	**74.59**
CS^1	**95.12**	**0.3341**	**32.38**	**0.5188**	**33.95**	**0.3627**	**2.17**	**0.1403**	**67.73**
CS^2	**94.33**	**0.6813**	**36.19**	**0.5150**	**37.61**	**0.3753**	**3.26**	**0.1439**	**68.59**
CDM^1	**92.89**	**0.5051**	**50.48**	**0.6690**	**55.35**	**0.5155**	**26.09**	**0.4318**	**75.06**
CDM^2	**91.04**	**0.7495**	**53.33**	**0.6511**	**55.05**	**0.4989**	**30.43**	**0.4373**	**74.43**
CDS^1	**94.60**	**0.3325**	**32.38**	**0.5188**	**35.17**	**0.3696**	**3.26**	**0.1720**	**67.81**
CDS^2	**95.92**	**0.6612**	**28.57**	**0.4922**	**32.11**	**0.3569**	**1.09**	**0.0992**	**67.34**
CM^1M^2	**92.36**	**0.5013**	**53.33**	**0.6898**	**55.66**	**0.5178**	**25.00**	**0.3955**	**74.98**
CM^1S^1	**91.44**	**0.5105**	**53.33**	**0.6765**	**57.49**	**0.5183**	**30.43**	**0.4447**	**75.29**
CM^1S^2	**91.96**	**0.5113**	**53.33**	**0.6765**	**56.57**	**0.5201**	**29.35**	**0.4267**	**75.29**
CM^2S^1	**91.30**	**0.5025**	**53.33**	**0.6573**	**55.66**	**0.5065**	**32.61**	**0.4587**	**74.90**
CM^2S^2	**91.17**	**0.7514**	**53.33**	**0.6572**	**55.05**	**0.4973**	**31.52**	**0.4480**	**74.59**
CS^1S^2	**95.65**	**0.3347**	**30.48**	**0.5102**	**33.33**	**0.3641**	**1.01**	**0.0992**	**67.65**
CDM^1S^1	**92.23**	**0.5133**	**53.33**	**0.6765**	**56.27**	**0.5235**	**30.43**	**0.4447**	**75.45**
CDM^2S^2	**90.78**	**0.7481**	**53.33**	**0.6634**	**55.35**	**0.4995**	**31.52**	**0.4480**	**74.43**
CDS^1S^2	**94.07**	**0.3429**	**32.38**	**0.5190**	**37.61**	**0.3679**	**3.26**	**0.1720**	**68.12**
$CM^1M^2S^1$	**92.49**	**0.5085**	**53.33**	**0.6899**	**56.27**	**0.5233**	**26.09**	**0.4151**	**75.29**
$CM^1M^2S^2$	**92.36**	**0.5065**	**53.33**	**0.6899**	**56.27**	**0.5213**	**26.09**	**0.4151**	**75.21**
$CM^1S^1S^2$	**92.89**	**0.5137**	**52.38**	**0.6831**	**56.27**	**0.5235**	**26.09**	**0.4319**	**75.45**
$CM^2S^1S^2$	**91.04**	**0.4985**	**53.33**	**0.6573**	**55.96**	**0.5070**	**32.61**	**0.4657**	**74.82**
$CDM^1M^2S^1$	**92.75**	**0.5125**	**53.33**	**0.6900**	**56.27**	**0.5273**	**26.09**	**0.4152**	**75.45**
$CDM^1M^2S^2$	**92.89**	**0.5145**	**53.33**	**0.6900**	**56.27**	**0.5294**	**26.09**	**0.4152**	**75.53**
$CDM^2S^1S^2$	**91.57**	**0.4965**	**53.33**	**0.6635**	**54.43**	**0.5019**	**32.61**	**0.4657**	**74.75**
$CM^1M^2S^1S^2$	**92.23**	**0.5072**	**53.33**	**0.6831**	**56.57**	**0.5218**	**26.09**	**0.4151**	**75.21**
$CDM^1M^2S^1S^2$	**92.36**	**0.5065**	**53.33**	**0.6831**	**56.27**	**0.5250**	**26.09**	**0.4073**	**75.21**

3.2 The Influence of the Unbalance of Sample Numbers among the Four Classes

We used the weighted factor approach to investigate the influence of the sample unbalance among the four classes. According to the number of four types of protein homo-oligomer, the weighted factor values of 2EM, 3EM, 4EM and 6EM are calculated as follow: 759/759, 759/105, 759/327, 759/92. The results of twenty-eight feature sets using weighted factor approach are shown in table 2.

From table 2, we can see that, in the weighted factor conditions, the total accuracies of all feature sets except CS^1S^2 based on the two scale glide zoom window are higher than that of C. The result of $CDM^1M^2S^1$ is the best, and the total accuracy are 75.37%, which are 10.05 higher than that of feature set C. These results suggest that weighted factor approach can weaken influence of the unbalance of sample numbers among the four classes.

Table 2. Results of 28 feature sets with RBF SVM and one-versus-one strategy in jackknife test using weighted factor approach

Feature sets	2EM		3EM		4EM		6EM		Q%
	Q(2)%	MCC(2)	Q(3) %	MCC(3)	Q(4) %	MCC(4)	Q(6) %	MCC(6)	
C	70.36	0.3577	49.52	0.4772	63.91	0.3859	46.74	0.3752	65.32
CD	76.02	0.4105	53.33	0.5213	64.83	0.4383	42.39	0.4092	68.90
CM^1	78.79	0.4881	59.05	0.5911	69.72	0.5127	51.09	0.4983	72.88
CM^2	78.00	0.4647	59.05	0.5532	67.58	0.5035	53.26	0.5188	72.02
CS^1	74.31	0.4163	57.14	0.5196	65.75	0.4571	48.91	0.4237	68.90
CS^2	76.81	0.4363	55.24	0.5371	66.36	0.4665	45.65	0.4305	70.15
CDM^1	78.92	0.4838	60.00	0.5981	68.50	0.5041	51.09	0.4982	72.72
CDM^2	78.79	0.4723	60.00	0.5677	66.97	0.5039	54.35	0.5356	72.49
CDS^1	75.89	0.4327	58.10	0.5375	65.44	0.4609	47.83	0.4312	69.76
CDS^2	75.76	0.4271	57.14	0.5300	64.83	0.4537	46.74	0.4127	69.37
CM^1M^2	82.35	0.5150	60.95	0.6450	68.50	0.5279	51.09	0.5463	74.82
CM^1S^1	78.52	0.4833	59.05	0.5991	69.42	0.5031	51.09	0.5020	72.64
CM^1S^2	80.24	0.4931	57.14	0.5811	69.72	0.5265	51.09	0.5275	73.58
CM^2S^1	78.52	0.4763	60.00	0.5713	68.20	0.5054	53.26	0.5355	72.56
CM^2S^2	78.39	0.4735	59.05	0.5604	67.89	0.5025	53.26	0.5270	72.33
CS^1S^2	65.88	0.3722	62.86	0.4681	64.53	0.4211	51.09	0.3296	64.22
CDM^1S^1	80.37	0.4797	56.19	0.5736	67.58	0.5117	53.26	0.5533	73.19
CDM^2S^2	80.24	0.4837	60.00	0.5866	66.36	0.5077	54.35	0.5443	73.19
CDS^1S^2	77.47	0.4424	58.10	0.5450	64.83	0.4686	47.83	0.4485	70.54
$CM^1M^2S^1$	82.48	0.5172	61.90	0.6520	68.20	0.5258	52.17	0.5646	74.98
$CM^1M^2S^2$	82.21	0.5085	61.90	0.6519	67.28	0.5164	52.17	0.5595	74.59
$CM^1S^1S^2$	79.18	0.4843	57.14	0.5848	69.42	0.5103	51.09	0.5102	72.88
$CM^2S^1S^2$	76.68	0.4546	62.86	0.5643	66.67	0.4823	53.26	0.5264	71.32
$CDM^1M^2S^1$	83.27	0.5246	61.90	0.6522	67.89	0.5328	52.17	0.5648	75.37
$CDM^1M^2S^2$	83.16	0.5255	61.90	0.6522	68.20	0.5322	51.09	0.5513	75.29
$CDM^2S^1S^2$	80.50	0.4830	60.95	0.5899	65.44	0.5019	54.35	0.5529	73.19
$CM^1M^2S^1S^2$	83.14	0.5176	61.90	0.6521	66.97	0.5236	52.17	0.5646	75.06
$CDM^1M^2S^1S^2$	83.53	0.5223	61.90	0.6568	66.97	0.5269	52.17	0.5647	75.29

4 Conclusion

A novel concept of multi-scale glide zoom window was proposed in this paper. Based on the concept of multi-scale glide zoom window, a protein sequence can be investigated from two scale glide zoom windows (whole protein sequence glide zoom window and kin amino acid glide zoom window). Twenty-seven feature sets were constructed by combining five kinds of feature sets of the two scale glide zoom windows with amino acids composition to form pseudo amino acid compositions (PseAAC). The results show that the twenty-six feature sets based on the two scale glide zoom windows are better than feature set C in the weighted factor conditions, and weighted factor approach can weaken influence of the unbalance of sample numbers among the four classes. In the three kinds of feature sets of the two scale glide zoom window, amino acids mean distance feature set is most effective and robust. It is demonstrated that the concept of multi-scale glide zoom window provide a new scope to investigate primary protein sequence, the feature sets extracted from multi-scale glide zoom window may contain more protein structure information.

Acknowledgements. 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

1. Chou, K.C.: Review: Low-frequency Collective Motion in Biomacromolecules and Its Biological Functions. Biophys. Chem. 30, 3–48 (1988)
2. Chou, K.C.: Review: Structural Bioinformatics and Its Impact to Biomedical Science. Curr. Med. Chem. 11, 2105–2134 (2004e)
3. Chou, K.C.: Molecular Therapeutic Target for Type-2 Diabetes. J. Proteome. Res. 3, 1284–1288 (2004a)
4. Chou, K.C.: Insights from Modelling Three-dimensional Structures of the Human Potassium and Sodium Channels. J. Proteome. Res. 3, 856–861 (2004b)
5. Chou, K.C.: Insights from Modelling the 3D Structure of the Extracellular Domain of Alpha7 Nicotinic Acetylcholine Receptor. Biochem. Biophys. Res. Commun. 319, 433–438 (2004c)
6. Chou, K.C.: Modelling Extracellular Domains of GABA-A Receptors: Subtypes 1, 2, 3, and 5. Biochem. Biophys. Res. Commun. 316, 636–642 (2004d)
7. Oxenoid, K., Chou, J.J.: The Structure of Phospholamban Pentamer Reveals a Channel-Like Architecture in Membranes. Proc. Natl. Acad. Sci. USA 102, 10870–10875 (2005)
8. Anfinsen, C.B., Haber., E., Sela, M., White, F.H.: The Kinetics of the Formation of Native Ribonuclease During Oxidation of the Reduced Polypeptide Chain. Proc. Natl. Acad. Sci. USA 47, 1309–1314 (1961)
9. Anfisen, C.B.: Principles that Govern the Folding of Protein Chains. Science 181, 223–230 (1973)
10. Garian, R.: Prediction of Quaternary Structure from Primary Structure. Bioinformatics 17, 551–556 (2001)
11. Chou, K.C., Cai, Y.D.: Predicting Protein Quaternary Structure by Pseudo Amino Acid Composition. Proteins Struct. Func. Gene. 53, 282–289 (2003b)
12. Zhang, S.W., Quan, P., Zhang, H.C., Zhang, Y.L., Wang, H.Y.: Classification of Protein Quaternary Structure with Support Vector Machine. Bioinformatics 19, 2390–2396 (2003)
13. Zhang, S.W., Pan., Q., Zhang., H.-C., Shao., Z.-C., Shi, J.-Y.: Prediction of 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 (2006)

14. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (1995)
15. Vapnik, V. (ed.): Statistical Learning Theory. Wiley, New York (1998)
16. Bairoch, A., Apweiler, R.: The SWISS-PROT Protein Data Bank and Its New Supplement TrEMBL. Nucleic Acids Res. 24, 21–25 (1996)
17. Chou, K.C., Zhang, C.T.: Prediction of Protein Structural Classes. Crit. Rev. Biochem. Mol. Biol. 30, 275–349 (1995) (review)
18. Zhou, G.P., Assa-Munt, N.: Some Insights Into Protein Structural Class Prediction. Proteins Struct. Funct. Genet. 44, 57–59 (2001)
19. Fasman, G.D. (ed.): Handbook of Biochemistry and Molecular Biology, 3rd edn. CRC Press, Boca Raton (1976)
20. Bahar, I., Atilgan, A.R., Jernigan, R.L., Erman, B.: Understanding the Recognition of Protein Structural Classes by Amino Acid Composition. Proteins 29, 172–185 (1997)