ELSEVIER

Biophysical Chemistry 128 (2007) 87–93

Biophysical Chemistry

http://www.elsevier.com/locate/biophyschem

Novel two-stage hybrid neural discriminant model for predicting proteins structural classes

Samad Jahandideh [a], Parviz Abdolmaleki [a,*], Mina Jahandideh [b], Ebrahim Barzegari Asadabadi [a]

[a] Department of Biophysics, Faculty of Science, Tarbiat Modares University, Tehran, Iran
[b] Department of Mathematics, Faculty of Science, Vali-E-Asr University, Rafsanjan, Iran

Received 17 December 2006; received in revised form 5 March 2007; accepted 6 March 2007
Available online 13 March 2007

Abstract

In order to establish novel hybrid neural discriminant model, linear discriminant analysis (LDA) was used at the first stage to evaluate the contribution of sequence parameters in determining the protein structural class. An in-house program generated parameters including single amino acid and all dipeptide composition frequencies for 498 proteins came from Zhou [An intriguing controversy over protein structural class prediction, J. Protein Chem. 17(8) (1998) 729–738]. Then, 127 statistically effective parameters were selected by stepwise LDA and were used as inputs of the artificial neural networks (ANNs) to build a two-stage hybrid predictor. In this study, self-consistency and jackknife tests were used to verify the performance of this hybrid model, and were compared with some of prior works. The results showed that our two-stage hybrid neural discriminant model approach is very promising and may play a complementary role to the existing powerful approaches.

Keywords: Linear discriminant analysis (LDA); Artificial neural networks (ANNs); Sequence parameters; Amino acid composition; Protein structural class

1. Introduction

Nowadays, it is generally accepted that all information regarding the structure of a protein is coded in the amino acid sequence [2]. The functional properties of proteins are determined by their three-dimensional (3D) structure which, in turn, depends on amino acid sequence. To understand the function of proteins, scientists are trying to predict the 3D structure of a protein from its amino acid sequence. Finding the rules relating the amino acid sequence to 3D protein structure is one of the major goals of contemporary molecular biology.

Structural protein classes were defined over 20 years ago as being general ways of describing folds that reflected content of the secondary-structure elements and their arrangement in folded proteins [16]. Protein folds can be classified into four main classes: all-α, all-β, α/β and α+β. Since then, various quantitative classification rules have been proposed based on the percentages of α-helices and β-sheets in a protein.

As two golden standards for exact determination of protein structure, NMR and X-ray crystallography are presently restricted to limited number of proteins due to the inherent complications in their interpretation as well as their technical limitations. These practical limitations encouraged the development of theoretical methods to predict the structures of proteins based on the available data. In addition, as a result of large-scale sequencing projects, the gap between the number of known protein sequences and number of known structures is increasing. This gap highlights the demand to theoretical methods of protein structure determination.

From historical point of view, Nishikawa's findings [20–22] revealed the strong correlation between the structural classes of proteins and amino acid composition. Since then, many different theoretical methods have been proposed for predicting the structural class of proteins such as statistical analysis which uses parameters obtained from known protein sequences and tertiary structure [6], information theory [8], nearest neighbor methods [26], multiple alignment [14,24], neural networks [18], component coupled [5], combination of multiple alignment and

* Corresponding author. Department of Biophysics, Faculty of Science, Tarbiat Modares University, Tehran, Iran. P.O. Box 14115/175. Tel.: +98 21 88950325.
E-mail address: Parviz@modares.ac.ir (P. Abdolmaleki).

0301-4622/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.bpc.2007.03.006

neural networks [23], 3D–1D compatibility [10], support vector machines [3], rough sets [4] and hybrid model [11]. In addition, many more studies have applied various methods to predict protein structural classes [28–41].

Generally, linear discriminant analysis (LDA) as an essential linear statistical model has been applied in modeling tasks. The utilization of LDA has often been criticized because of linear relationship between dependent and independent variables. Basically, LDA is designed for the case when the underlying relationships between variables are linear.

Artificial neural networks (ANNs) provide a new alternative to LDA in modeling tasks, particularly in situations where the dependent and independent variables exhibit complex non-linear relationships. Even ANNs have shown to have better classification capability than LDA. It is, however, also being criticized for its long training process in designing the optimal networks topology and hence has limited its applicability in handling modeling problems.

In our previous work [11] we established a hybrid model using multinomial logistic regression and ANNs. Results of that work showed that combination of ANNs as a non-algorithmic model and multinomial logistic regression as an algorithmic model provide better results than either one alone. At the present study, LDA as another algorithmic model were used in the first stage of hybrid neural discriminant model. Using the LDA as the first stage of this model, we could increase the reliability of our model in predicting protein structural classes. The rationale underlying this analysis is using the LDA to build the most influent set of parameters which then are fed into well established ANNs. In addition, the predicting capability of hybrid neural discriminant model was compared to the predicting results from previous models on the same database.

2. Materials and methods

2.1. Database

The database comprising of 498 protein domain sequences as described by Zhou [27] which was collected from SCOP database [19] was used in the present study. These protein domain sequences have been categorized as shown in Fig. 1. We used the database to test our model through self-consistency test and jackknife test and to compare the prediction accuracy and individual accuracies of each structural class with other models.

Our parameters including amino acid and all dipeptide composition frequencies were generated using in-house programs in MATLAB language. In order to check the fidelity of these programs, results were compared with the outputs of COMPSEQ program (http://bioweb.pasteur.fr/seqanal/interfaces/compseq.html) on the same database.

2.2. Model development

In the first stage of this hybrid model, LDA acts on the database to select parameters and predict protein structural class. Then the ANNs, which act non-linearly in the second stage, were fed by the outputs of LDA to predict protein

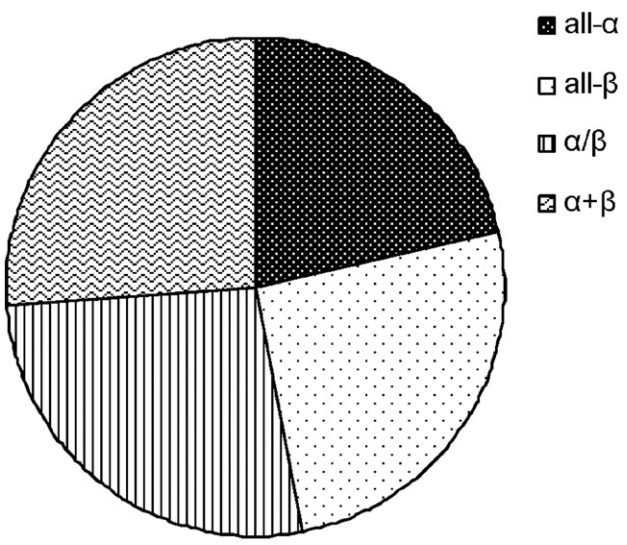

Fig. 1. Distribution of different protein structural classes in the database.

structural class. The jackknife technique, in which all cases were used in both the training and testing processes, was applied to train and test models on the database. The procedure is as follows: given a training set of N proteins, the first protein in the training set, t_1, is set aside (left out). Then the model is trained on the remaining $N-1$ proteins and tested on the left out sample. Then sample t_1 is inserted back into the database and the next protein, t_2, is left out. This procedure is repeated until every protein in the database had the opportunity to be a left out sample. It therefore provided as many simulations as the number of samples in each database. Although this method is time-consuming, it is especially useful for the small databases such as ours. In addition to jackknife test we used self-consistency test on the database.

2.2.1. Linear discriminant analysis (LDA)

LDA was first proposed by Fisher in the 1930s as a discrimination and classification tool. These days, LDA has been reported as the most commonly discussed and used statistical technique in modeling tasks [15].

LDA can be used to build a predictive model of the group membership based on observed characteristics of each case. This procedure generates a discriminant function (or, for more than two groups, a set of discriminant functions) based on linear combinations of the predictor variables that provide the best discrimination among the groups. The functions are generated from the samples with known membership; the functions can then be applied to new cases with measurements for the predictor variables but with unknown group membership. The LDA can be expressed as

$$D = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n,$$

where D represents the discriminant score, β_0 is the intercept term, and β_i (i=,…,n) represents the β coefficient associated with the corresponding explanatory variable X_i (i=1,…,n).

The variables used to compute the linear discriminant function are chosen in stepwise manner using the Wilk's lambda

Table 1
Optimized structure of ANNs used in this study

Learning rate	0.2
Error goal	0.02
Trans function of hidden layer	Logsig
No. of input nodes	127
Iterations	Over 1500
No. of hidden nodes	18
No. of output nodes	2
Training algorithm	Conjugate gradient method

method: at each step, the variable that minimizes the overall Wilk's lambda is entered. F value, which can be calculated using Wilk's lambda, allows the assessment of relative importance among the candidate variables. This criterion was used for entering and removing variables: a variable is entered into the model if its F value is greater than the entry value (F_{min}), thereafter F values of the rest variables in the model are re-calculated and those with F values less than the removal value (F_{max}) are removed. This procedure is continued until F values of the rest variables are all less than the defined F_{min} [12]. The Wilk's lambda for the overall discrimination can take values in the range from 0 (perfect discrimination) to −1 (no discrimination).

As a matter of fact, LDA has been widely devoted to considerably wide range of application areas, such as medicine, biology, marketing research, chemistry, finance, business, engineering and archaeology [1,7,13,25].

2.2.2. Artificial neural networks (ANNs)

As a powerful non-linear predictor in hybrids with the LDA, the ANNs were used in this study. In this way, the selected 127 variables from LDA were used as input nodes of ANNs. This is supposed to reduce the number of input nodes, simplify the network structure, and shorten the model building time. The networks which were employed were classical feed-forward ones to associate protein sequential parameters to the structural classes. Using this algorithm, parameters of training cases are fed into the networks. The final outputs estimated by the network are compared with the real class of the cases, producing a mean of the sum-of-squares error (MSE). MSE is propagated back into the networks to adjust the randomly chosen weights. The training cases are tested with new weights and the process is repeated. Through such process, the MSE is minimized.

After optimizing the preliminary structure, we found that the best architecture in terms of computational experience was that characterized by two output neurons representing the protein structural classes ([1 1] for all-α, [0 1] for all-β, [0 0] for α+β and [1 0] for α/β), one hidden layer containing eighteen neurons and another input one, containing 127 neurons, each of which corresponded to a selected parameter proposed by the LDA. In order to determine the well optimized structure for the network, we constructed a larger number of networks, varying the number of hidden neurons, iteration and learning rates. This optimal network produced the least MSE when trained. We used several training algorithms such as gradient descent, resilient back-propagation, conjugate gradient and quasi-Newton. The best result was obtained using conjugate gradient training algorithm.

Our networks were trained perfectly over 1500 iterations. Also the optimal learning rate was found to be 0.2. The parameters of the optimized neural network are listed in Table 1. The software used to construct the neural networks was in-house written in the MATLAB programming language.

2.2.3. Performance evaluation

Threshold-dependent measures were used to assess the performance of LDA and ANNs. These measures can be derived from the four scalar quantities; TP_c (true positives: number of correctly predicted proteins in class c), TN_c (true negatives: number of correctly predicted proteins not in class c), FP_c (false positives: number of underpredicted proteins) and FN_c (false negatives: number of overpredicted proteins).

The following three measures including the prediction accuracy (PA), individual accuracy (IA) and Mathews correlation coefficient (MCC) [17] were calculated for the output of the models using the following formulas:

$$PA = \frac{(TP_c + TN_c)}{t} \times 100$$

$$IA = \left(\frac{TP_c}{Ind_c}\right) \times 100$$

$$MCC_c = \frac{(TP_c)(TN_c) - (FP_c)(FN_c)}{\sqrt{(TP_c + FP_c)(TP_c + FN_c)(TN_c + FP_c)(TN_c + FN_c)}}$$

where t is the total number of examples and Ind_c is the number of proteins reside in class c.

Statistical analysis was performed using SPSS 13 for Windows (SPSS Inc., Chicago, USA).

3. Results

3.1. Results of LDA

In LDA, the minimum F entry value (F_{min}) is set to 2.71 and the maximum F removal value (F_{max}) is set to 3.84. These settings are based on default values in software SPSS. We ran

Table 2
Performance comparison between two stages of the hybrid neural discriminant model

Test	Performance measures	First stage (Linear discriminant analysis)				Second stage (neural network)			
		All-α	All-β	α/β	α+β	All-α	All-β	α/β	α+β
Self-consistency	Individual accuracy (%)	100	100	100	100	100	100	100	100
	MCC	1	1	1	1	1	1	1	1
Jackknife	Individual accuracy (%)	94.4	89.7	92.6	92.2	95.3	88.9	94.1	93.0
	MCC	0.52	0.46	0.60	0.53	0.66	0.74	0.72	0.66

Table 3
Results of self-consistency and jackknife tests

Test	Algorithm	Individual accuracy for each class				Prediction accuracy (%)
		All-α (%)	All-β (%)	α/β (%)	α+β (%)	
Self-consistency	Component coupled	95.80	95.20	94.90	95.40	95.80
	Neural network	100	98.40	96.30	84.50	94.60
	SVM	100	100	100	100	100
	Rough sets	100	100	100	100	100
	Multinomial logistic regression[a]	100	100	100	100	100
	Hybrid model[a]	100	100	100	100	100
	LDA	100	100	100	100	100
	Hybrid neural discriminant model	100	100	100	100	100
Jackknife	Component coupled	93.50	88.90	90.40	84.50	89.20
	Neural network	86.00	96.00	88.20	86.00	89.20
	SVM	88.80	95.20	96.30	91.50	93.20
	Rough Sets	87.90	91.30	97.10	86.00	90.80
	Multinomial logistic regression[a]	92.50	88.10	90.50	89.90	90.40
	Hybrid model[a]	96.30	92.10	95.60	93.80	94.40
	LDA	94.39	89.68	92.64	92.24	92.17
	Hybrid neural discriminant model	95.32	88.88	94.11	93.02	92.77

[a] The results of these models were reported from our previous work [11].

LDA on the database using jackknife and self-consistency tests. As an indicator of the optimized step in LDA, we used the Wilk's lambda through jackknife test. The average Wilk's lambda was 0.000 in the 155th step, suggesting that 100% of the variance associated with results of NMR and X-ray crystallography was accounted for in the model. LDA selected 127 sequence parameters among 420 sequence parameters at the 155th step. All of sixteen sequence parameters selected by multinomial logistic regression at the first stage of our hybrid model in previous work are among these 127 selected sequence parameters using LDA in the present work.

The results of jackknife and self-consistency tests were evaluated by the performance evaluative measures. The results shown in Table 2 are obtained according to the output of the model. Results show that IA has the highest value in all-α protein structural class. These results are in agreement to our results from multinomial logistic regression in previous work. In addition, LDA as the first stage of hybrid neural discriminamt model in the present work showed higher PA in comparison with multinomial logistic regression as the first stage of hybrid model in the previous work.

3.2. Results of ANNs

The selected parameters in jackknife LDA procedure were fed into the ANNs to build two-stage hybrid neural discriminant model. As the second stage of the hybrid modeling procedure, and using conjugate gradient training algorithm method on the database, the three-layer neural networks with optimized architecture including two output neurons, one hidden layer containing eighteen neurons and an input layer containing another set of 127 neurons, all performance measures were high. The results of self-consistency and jackknife tests have been shown in Table 2. The second stage same as the first stage showed the highest value of IA in all-α protein structural class.

PA and IA in each class for the two-stage hybrid neural discriminant model in comparison with our previous hybrid model and some of other methods on the same database are shown in Table 3. The results of the hybrid neural discriminant model showed all the percentages of correct prediction on the database reach 100% in self-consistency test, which is the same as the results of our previous hybrid model, SVM and rough sets based methods [11,3,4]. The highest PA value of 94.6% has been obtained in a previous study using jackknife test [32]. However, our results indicated that the hybrid neural discriminant model same as our previously proposed hybrid model captured the characteristics between sequences and their classes through single amino acid and all dipeptide composition frequencies. Jackknife test was performed on the database. The comparison should be focused on the jackknife test because it is more rigorous and objective method. From the result of jackknife test, it is obvious that the PA in comparison with our previous hybrid model is a little lower in second stage of the hybrid neural discriminant model.

4. Discussion and conclusion

Determining the 3D fold of a protein is still remained a difficult task, despite great achievements in protein science. Indeed, golden standard techniques such as NMR and X-ray crystallography are expensive and time-consuming. Consequently, there is a large gap between the number of known protein sequences and the number of known 3D protein structures. The computational prediction of structures from amino acid sequence has therefore come to play a key role in narrowing the gap. The previous reports indicated that these computational methods have been successful in providing useful information for the biological research community.

In order to establish powerful hybrid model using combination of algorithmic and non-algorithmic models, we used LDA

in the present work and multinomial logistic regression in the previous work, at the first stage of hybrid modeling procedures as algorithmic models. LDA and multinomial logistic regression were used to select the effective sequence parameters that are applied for prediction of protein structural classes. Then ANNs as the non-algorithmic model was used in hybrid modeling procedures at the second stage.

Regarding the fact that almost all previously used models detected all-α cases better than other classes [9], it is revealed that IA values in the results of two stages of hybrid discriminant model are in agreement with many other previous works done A plausible reason for this tendency of predictors is the predominant role of short and medium range interactions in all-α proteins. Similarly, uniformly lower accuracy in the prediction of the other classes implies the dominance of long-range interactions [9].

As recommendation for further research we proposed using new defined protein database for prediction of protein structural class in seven structural classes including 2230 protein domains which consists of protein and domains with only less than 20% sequence identity to each other [42]. Using such database is much more rigorous and interesting. In addition, another aspect of mentioned research is applying functional domain composition that is also an extremely important and promising direction.

As it is obvious from the results of jackknife test, α/β class has the higher IA than α+β class. This may be related to the proportion of α/β class in the training sets in which α/β class occupied the bigger part. As a supervised learning method, it makes it easier to capture characteristics that feed more training objects to neural networks.

In statistical prediction, the following three cross-validation tests are often used to examine the power of a predictor: independent database test, sub-sampling test, and jackknife test. Of these three the jackknife test is thought the most rigorous and objective one (see [43] for comprehensive review in this regard), and hence has been used by more and more investigators [36,39,44–67] in examining the power of various prediction methods.

Hybrid models can predict the structural class of proteins, using restricted number of sequence parameters among 420 sequence parameters this is one of the most important advantages of these models. In general, the results showed that by use of hybrid neural discriminant model, one can provide adequate information for an accurate prediction applying a few sequence parameters, only including single and dipeptide compositions.

Using dipeptide composition strengthens the approach in comparison with using amino acid composition [68,69]. It is anticipated that the prediction quality of the current approach can be further improved if the amino acid composition is substituted by the pseudo amino acid composition [70] because the latter can incorporate a considerable amount of the sequence order effects as demonstrated in predicting other attributes of proteins [39,40,56,61,62,67,71–74].

This study showed, LDA as an algorithmic model could predict protein structural classes with high PA that is comparable with other algorithmic and non-algorithmic models. ANNs at the second stage of hybrid model did not show obvious difference accuracy in PA value comparing with LDA at the first. A plausible reason for this result is the enormous number of selected sequence parameters in the first stage of hybrid model that caused overfitting in training of ANNs. In fact, with the presence of more selected sequence parameters in comparison with our previous hybrid model that selected only sixteen parameters, needed time for neural network training procedure in the second stage and the probability of overfitting occurrence is increased and therefore lower precision and reliability obtained are in this way.

In the procedure of applying hybrid models for prediction of protein structural class, we concluded that these models can give more precise, more reliable and more accurate results than other commonly used models. In this regard, using such hybrid models would be recommended in other fields of structural bioinformatics.

References

[1] E.I. Altman, Financial ratios discriminant analysis and prediction of corporate bankruptcy, Journal of Finance 23 (1968) 589–609.
[2] C.B. Anfinsen, Principles that govern the folding of protein chains, Science 181 (1973) 223–230.
[3] Y.D. Cai, X.J. Liu, X.U. Xu, G.P. Zhou, Support vector machines for predicting protein structural class, BMC Bioinformatics 2 (1) (2001) 3–9.
[4] Y. Cao, S. Liu, L. Zhang, et al., Predicting of protein structural class with rough sets, BMC Bioinformatics 7 (1) (2006) 3–8.
[5] K.C. Chou, G.M. Maggiora, Domain structural class prediction, Protein Engineering 11 (7) (1998) 523–538.
[6] P.Y. Chou, G.D. Fasman, Prediction of protein conformation, Biochemistry 13 (1974) 222–245.
[7] E.B. Deakin, A discriminant analysis of predictors of business failure, Journal of Accounting Research 10 (1972) 167–179.
[8] J. Garnier, D. Osguthorpe, B. Robson, Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins, Journal of Molecular Biology 120 (1978) 97–120.
[9] M.M. Gromiha, S. Selvaraj, Protein secondary structure prediction in different structural classes, Protein Engineering 11 (4) (1998) 249–251.
[10] M. Ito, Y. Matsuo, K. Nishikawa, Prediction of protein secondary structure using the 3D–1D compatibility algorithm, CABIOS 13 (1997) 415–423.
[11] S. Jahandideh, P. Abdolmaleki, M. Jahandideh, S.H. Sadat Hayatshahi, Novel hybrid method for the evaluation of parameters contributing in determination of protein structural classes, Journal of Theoretical Biology 244 (2007) 275–281.
[12] R. Jennrich, P. Sampson, Stepwise discriminant analysis, in: W.J. Dixon (Ed.), BMD Biomedical Computer Programs, University of California Press, Berkley, 1979.
[13] J.C. Kim, D.H. Kim, J.J. Kim, J.S. Ye, H.S. Lee, Segmenting the Korean housing market using multiple discriminant analysis, Construction Management & Economics 18 (2000) 45–54.
[14] R.D. King, M.J.E. Sternberg, Protein secondary structure prediction based on position-specific scoring matrices, Protein Science 5 (1996) 2298–2310.
[15] G. Lee, T.K. Sung, N. Chang, Dynamics of modeling in data mining: interpretive approach to bankruptcy predication, Journal of Management Information Systems 16 (1999) 63–85.
[16] M. Levitt, C. Chothia, Structural patterns in globular proteins, Nature 261 (1976) 552–557.
[17] B.W. Mathew, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochimica et Biophysica Acta 405 (1975) 442–451.
[18] B.A. Metfessel, P.N. Saurugger, D.P. Connelly, S.S. Rich, Cross-validation of protein structural class prediction using statistical clustering and neural networks, Protein Science 2 (7) (1993) 1171–1182.
[19] A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures, Journal of Molecular Biology 247 (4) (1995) 536–540.

[20] K. Nishikawa, T. Ooi, Correlation of amino acid composition of a protein to its structural and biological characters, Journal of Biochemistry 91 (1982) 1821–1824.
[21] K. Nishikawa, Y. Kubota, T. Ooi, Classification of proteins into groups based on amino acid composition and other characters, Journal of Biochemistry 94 (1983) 981–995.
[22] K. Nishikawa, Y. Kubota, T. Ooi, Classification of the proteins into groups based on amino acid composition and other characters grouping into four types, Journal of Biochemistry 94 (1983) 997–1007.
[23] B. Rost, C. Sander, Combining evolutionary information and neural networks to predict protein secondary structure, Proteins 19 (1994) 55–72.
[24] R.B. Russell, G.J. Barton, The limits of protein secondary structure prediction accuracy from multiple sequence alignment, Journal of Molecular Biology 234 (1993) 951–957.
[25] L.J. Trevino, J.D. Daniels, FDI theory and foreign direct investment in the United States: a comparison of investors and non-investors, International Business Review 4 (1995) 177–194.
[26] T.M. Yi, E.S. Lander, Protein secondary structure prediction using nearest-neighbor methods, Journal of Molecular Biology 232 (1993) 1117–1129.
[27] G.P. Zhou, An intriguing controversy over protein structural class prediction, Journal of Protein Chemistry 17 (8) (1998) 729–738.
[28] K.C. Chou, Review: progress in protein structural class prediction and its impact to bioinformatics and proteomics, Current Protein and Peptide Science 6 (2005) 423–436.
[29] K.C. Chou, A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space, Proteins, Structure, Function, and Genetics 21 (1995) 319–344.
[30] K.C. Chou, C.T. Zhang, Predicting protein folding types by distance functions that make allowances for amino acid interactions, Journal of Biological Chemistry 269 (1994) 22014–22020.
[31] G.P. Zhou, N. Assa-Munt, Some insights into protein structural class prediction, Proteins, Structure, Function, and Genetics 44 (2001) 57–59.
[32] K.Y. Feng, Y.D. Cai, K.C. Chou, Boosting classifier for predicting protein domain structural class, Biochemical and Biophysical Research Communications 334 (2005) 213–217.
[33] H.B. Shen, J. Yang, X.J. Liu, K.C. Chou, Using supervised fuzzy clustering to predict protein structural classes, Biochemical and Biophysical Research Communications 334 (2005) 577–581.
[34] K.C. Chou, Review: Prediction of protein structural classes and subcellular locations, Current Protein and Peptide Science 1 (2000) 171–208.
[35] K.C. Chou, A key driving force in determination of protein structural classes, Biochemical and Biophysical Research Communications 264 (1999) 216–224.
[36] X. Xiao, S.H. Shao, Z.D. Huang, K.C. Chou, Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor, Journal of Computational Chemistry 27 (2006) 478–482.
[37] B. Niu, Y.D. Cai, W.C. Lu, G.Y. Zheng, K.C. Chou, Predicting protein structural class with AdaBoost learner, Protein and Peptide Letters 13 (2006) 489–492.
[38] Q.S. Du, Z.Q. Jiang, W.Z. He, D.P. Li, K.C. Chou, Amino acid principal component analysis (AAPCA) and its applications in protein structural class prediction, Journal of Biomolecular Structure and Dynamics 23 (2006) 635–640.
[39] C. Chen, X. Zhou, Y. Tian, X. Zou, P. Cai, Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network, Analytical Biochemistry 357 (2006) 116–121.
[40] C. Chen, Y.X. Tian, X.Y. Zou, P.X. Cai, J.Y. Mo, Using pseudo-amino acid composition and support vector machine to predict protein structural class, Journal of Theoretical Biology 243 (2006) 444–448.
[41] Y. Cao, S. Liu, L. Zhang, J. Qin, J. Wang, K. Tang, Prediction of protein structural class with rough sets, BMC Bioinformatics 7 (20) (2006) 10.1186/1471-2105-1187-1120.
[42] K.C. Chou, Y.D. Cai, Predicting protein structural class by functional domain composition, Biochemical and Biophysical Research Communications 321 (2004) 1007–1009 (Corrigendum: ibid., 2005, Vol. 329, 1362).
[43] K.C. Chou, C.T. Zhang, Review: prediction of protein structural classes, Critical Reviews in Biochemistry and Molecular Biology 30 (1995) 275–349.
[44] G.P. Zhou, An intriguing controversy over protein structural class prediction, Journal of Protein Chemistry 17 (1998) 729–738.
[45] Z.P. Feng, Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition, Biopolymers 58 (2001) 491–499.
[46] Z.P. Feng, An overview on predicting the subcellular location of a protein, In. Silico. Biol. 2 (2002) 291–303.
[47] R.Y. Luo, Z.P. Feng, J.K. Liu, Prediction of protein structural class by amino acid and polypeptide composition, European Journal of Biochemistry 269 (2002) 4219–4225.
[48] M. Wang, J. Yang, Z.J. Xu, K.C. Chou, SLLE for predicting membrane protein types, Journal of Theoretical Biology 232 (2005) 7–15.
[49] X. Xiao, S. Shao, Y. Ding, Z. Huang, Y. Huang, K.C. Chou, Using complexity measure factor to predict protein subcellular location, Amino Acids 28 (2005) 57–61.
[50] X. Xiao, S. Shao, Y. Ding, Z. Huang, X. Chen, K.C. Chou, An application of gene comparative image for predicting the effect on replication ratio by HBV virus gene missense mutation, Journal of Theoretical Biology 235 (2005) 555–565.
[51] K.C. Chou, H.B. Shen, Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization, Biochemical and Biophysical Research Communications 347 (2006) 150–157.
[52] H. Liu, J. Yang, J.G. Ling, K.C. Chou, Prediction of protein signal sequences and their cleavage sites by statistical rulers, Biochemical and Biophysical Research Communications 338 (2005) 1005–1011.
[53] H. Liu, M. Wang, K.C. Chou, Low-frequency Fourier spectrum for predicting membrane protein types, Biochemical and Biophysical Research Communications 336 (2005) 737–739.
[54] X.D. Sun, R.B. Huang, Prediction of protein structural classes using support vector machines, Amino Acids 30 (2006) 469–475.
[55] K.C. Chou, H.B. Shen, Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers, Journal of Proteome Research 5 (2006) 1888–1897.
[56] S.W. Zhang, Q. Pan, H.C. Zhang, Z.C. Shao, J.Y. Shi, Prediction protein homo-oligomer types by pseudo amino acid composition: Approached with an improved feature extraction and naive Bayes feature fusion, Amino Acids 30 (2006) 461–468.
[57] Y.Z. Guo, M. Li, M. Lu, Z. Wen, K. Wang, G. Li, J. Wu, Classifying G protein-coupled receptors and nuclear receptors based on protein power spectrum from fast Fourier transform, Amino Acids 30 (2006) 397–402.
[58] Z. Wen, M. Li, Y. Li, Y. Guo, K. Wang, 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 (2) (2007) 277–283.
[59] X. Xiao, S.H. Shao, Y.S. Ding, Z.D. Huang, K.C. Chou, Using cellular automata images and pseudo amino acid composition to predict protein sub-cellular location, Amino Acids 30 (2006) 49–54.
[60] K.C. Chou, H.B. Shen, Large-scale predictions of Gram-negative bacterial protein subcellular locations, Journal of Proteome Research 5 (2006) 3420–3428.
[61] H.B. Shen, K.C. Chou, Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types, Biochemical and Biophysical Research Communications 334 (2005) 288–292.
[62] Y. Gao, S.H. Shao, X. Xiao, Y.S. Ding, Y.S. Huang, Z.D. Huang, K.C. Chou, Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, bessel function, and chebyshev filter, Amino Acids 28 (2005) 373–376.
[63] H.B. Shen, K.C. Chou, Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition, Biochemical and Biophysical Research Communications 337 (2005) 752–756.
[64] H.B. Shen, K.C. Chou, Ensemble classifier for protein fold pattern recognition, Bioinformatics 22 (2006) 1717–1722.
[65] Q.B. Gao, Z.Z. Wang, Classification of G-protein coupled receptors at four levels, Protein Engineering Design and Selection 19 (2006) 511–516.
[66] J. Guo, Y. Lin, X. Liu, GNBSL: a new integrative system to predict the subcellular location for Gram-negative bacteria proteins, Proteomics 6 (2006) 5099–5105.

[67] S. Mondal, R. Bhavna, R. Mohan Babu, S. Ramakumar, Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification, Journal of Theoretical Biology 243 (2006) 252–260.
[68] K.C. Chou, Using pair-coupled amino acid composition to predict protein secondary structure content, Journal of Protein Chemistry 18 (1999) 473–480.
[69] W. Liu, K.C. Chou, Protein secondary structural content prediction, Protein Engineering 12 (1999) 1041–1050.
[70] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition, Proteins, Structure, Function, and Genetics 43 (2001) 246–255 (Erratum: ibid., 2001, Vol. 44, 60).
[71] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (2005) 10–19.
[72] K.C. Chou, Y.D. Cai, Prediction of membrane protein types by incorporating amphipathic effects, Journal of Chemical Information and Modeling 45 (2005) 407–413.
[73] Z.M. Guo, Prediction of Membrane protein types by using pattern recognition method based on pseudo amino acid composition, Master Thesis, Bio-X Life Science Research Center, Shanghai Jiaotong University (2002).
[74] H. Liu, J. Yang, M. Wang, L. Xue, K.C. Chou, Using Fourier spectrum analysis and pseudo amino acid composition for prediction of membrane protein types, The Protein Journal 24 (2005) 385–389.