Bio-Medical Materials and Engineering 24 (2014) 3797–3806
DOI 10.3233/BME-141209
IOS Press

Prediction of protein structure classes with flexible neural tree

Wenzheng Bao*, Yuehui Chen* and Dong Wang
School of Information Science and Engineering, University of Jinan, Jinan, P.R. China

Abstract. Prediction of protein structural classes is of great significance to better understand protein folding patterns. An array of methods has been proposed to predict these structures based on sequences. However, the accuracy is strongly affected by the homology of sequences. In the present study, the features based on correlation coefficient of sequence and amino acid composition are extracted. Flexible neutral tree is employed as the classification model. To examine the performance of this method, four benchmark datasets are selected. Altogether, the results show that a higher prediction accuracy of alpha/beta can be achieved by the method compared to others.

Keywords: Protein structural classification, flexible neural tree, correlation coefficient

1. Introduction

The prediction of protein 3D structure is one of the most significant tasks in computational biology. Knowing the protein structure plays an important role in protein functional analysis, binding of protein, designing drug and other related fields and applications. The concept of protein structural classes has been proposed by Levitt and Chothia on a visual inspection of polypeptide chain topologies in a dataset of 31 globular proteins, where these proteins are divided into four mainly structural classes: all-alpha, all-beta, alpha+beta and alpha/beta [1]. The decipherment protein structure helps elucidate protein functions and further helps in drug design [2].

A common way to predict the structure of a protein is to first acquire proteins with known structures (e.g. by crystallography techniques); then from their sequences, the prediction process can be conducted by developing recognition techniques; thereafter, the developed methods can classify unknown protein sequences into one of its classes. Unfortunately, though the sequence length of different proteins (i.e., the number of amino acids within a protein) usually varies, it turns out very often that two proteins with different lengths and low sequential similarities sometimes can be categorized to the same class.

During the past several decades, a large number of algorithms and methods have been proposed to predict protein structural classes from protein primary structure [3–5]. There is an effective method for

*Corresponding authors: Wenzheng Bao, School of Information Science and Engineering, University of Jinan, Xinzhuang West Road No. 336, Jinan, P.R. China. Tel.: +86-18353170677; Fax: +86-531-82767581; E-mail: baowz55555@126.com.
Yuehui Chen, School of Information Science and Engineering, University of Jinan, Xinzhuang West Road No. 336, Jinan, P.R. China. Tel.: +86-531-82767581; Fax: +86-531-82767581; E-mail: yhchen@ujn.edu.cn.

protein structure prediction-fold recognition. General methods for sequence comparison, such as sequence-sequence alignment, sequence-profile alignment and the profile-profile alignment [6], might fail to recognize homologous templates due to the low similarities between the query sequence and the templates. An alternative is to thread the query sequence onto the template structures. Unfortunately, these methods, although widely used, may also fail to achieve satisfactory results at the fold level [7]. For instance, the reported overall accuracy of these methods employing the general widely used dataset 25PDB, whose sequence similarity is only about 25%, are merely about 70%. Recently, enormous efforts have been devoted to improve the prediction accuracy of low-similarity dataset by selecting novel features based on the correlations of amino acids [8,9].

The main goal of increasing protein structure prediction can be broadly divided into two categories: 1) the exploitation of classifiers, such as linear discriminated analysis, Bayesian classifiers, K-Nearest Neighbor, Hidden Markov Model, Artificial Neural Network, Support Vector Machine and ensemble classifiers [10–13]; 2) the development of feature extraction techniques using alphabetical sequence (syntactical-based) and/or using physicochemical properties of the amino acids (attribute-based or physicochemical-based), for instance, pseudo-amino acid composition (AAC), bigrams, autocorrelation and deriving features by considering more physicochemical properties [14–16].

In this study, the feature of information on cross correlation of amino acids is combined with polypeptide composition (PC). A flexible number feature vector is selected and the flexible neural tree (FNT) is chosen as the classifier. Ten-fold cross validation tests on four low-similarity datasets have proved that our method is effective for the prediction.

2. Materials and methods

2.1. Materials

The structural classification of proteins (SCOP) currently include eleven classes [17]: 1) all-alpha proteins; 2) all-beta proteins; 3) alpha/beta proteins; 4) alpha+beta proteins; 5) multi-domain proteins; 6) membrane and cell surface proteins; 7) small proteins; 8) coiled coils proteins; 9) low-resolution proteins; 10) peptides and 11) designed proteins. Our research only focuses on the first four categories, because they include the great majority of protein sequences and are the basis for most comparable approaches [18–20].

Four widely used benchmark datasets, in Table 1, with low similarity are also selected to provide a comprehensive and unbiased comparison with the existing prediction methods. The 1189 dataset with sequence similarity lower than 40% is taken from [10]; the 640 dataset with 25% similarity is taken from [21]; the ASTRAL database (including 7 classes) with 6424 sequences has similarity lower than 20%. In our research, however, only four major classes are used. The C204 dataset includes 204 non-homologous proteins.

Table 1

SCOP class distribution in the datasets used in our research

Dataset	all-alpha	all-beta	alpha/beta	alpha+beta	Total
640	138	154	171	177	640
1189	223	294	334	241	1092
ASTRAL	639	661	749	764	2813
C204	54	61	45	46	204

2.2. Feature extraction

2.2.1. Amino acid composition

The AAC feature is based on statistical property to analyze protein structure. The feature describing a protein sequence is depicted as a vector containing 20 elements, which represents the corresponding frequency of amino acids. The amino acid's fraction is used for AAC prediction. Let ***P*** be a protein sequence, then $P=x_1,x_2, \ldots x_N$, where x_i belongs to A, $i=1,2, \ldots ,N$ and A is the set of 20 amino acids, say, $A=\{a_1, a_2, \ldots , a_{20}\}$. AAC is calculated as the amino acid's fraction of a in the sequence P.

$$AAC(a_i) = \frac{total\ number\ of\ acid\ a_i}{N} (i = 1,2 \cdots 20) \tag{1}$$

where ***N*** is the total number of amino acids in the sequence.

2.2.2. Polypeptide composition

The PC feature depends on AAC. The feature describing protein sequence is represented as a vector containing 20k elements, indicating the corresponding PC frequency. It should be emphasized that PC is associated with AAC and itself. For instance, *PC-2* is the set of 20^2 dipeptides, such as AA, AC, AD… YY, $PC\text{-}2=\{pc\text{-}2_{1,1}, pc\text{-}2_{2,1}, \ldots ,pc\text{-}2_{20,20}\}$. One can see that a_1 is the sum of the numbers from $pc\text{-}2_{1,1}$ to $pc\text{-}2_{1,20}$, so the *PC-k* is made up of *PC-k+1*.

$$PC-2(pc-2_{i,j}) = \frac{total\ number\ of\ dipeptide\ pc-2_{i,j}}{N} (i = 1,2 \cdots 20) \tag{2}$$

$$\begin{aligned} &\sum PC-k+1(pc-k+1_{n1,n2\cdots nk,i})(n1,n2\cdots nk = 1,2\cdots 20) \\ &= PC-k(pc-k+1_{n1,n2\cdots nk}) \qquad (i = 1,2\cdots 20) \end{aligned} \tag{3}$$

2.2.3. Correlation coefficient of sequence

From the viewpoint of molecular biology, traditional AAC methods merely consider the composition of the protein sequence information. In fact, the structure of the protein is folding to various degrees. Some residues have interactions not only with its adjacent residues but also with the residues that are far apart. However, the autocorrelation coefficient of sequence takes into account both the position information of protein sequences and the distant interaction between amino acid sequences. The feature reflects the structure of proteins. Five characteristics are selected to express protein sequence, including hydrophilic, hydrophobicity, volumes of side chains, polarity and polarizability of amino acids. In order to take advantage of the features, protein sequences are converted into discrete values. The eigenvalues of the 5 characteristics are shown in Table 2.

Because the eigenvalues differ in some degree, a normalized process should be used. In this research, the process of maximum & minimum standardization is introduced. This standardization is expressed as

Table 2

The eigenvalues of the 5 characteristics

AA	H1	H2	V	P1	P2	AA	H1	H2	V	P1	P2
A	-0.40	-0.5	15	8.1	0.046	M	1.02	3.0	75	5.7	0.221
C	0.17	-1.0	47	5.5	0.128	N	-0.92	0.2	58	11.6	0.134
D	-1.31	3.0	59	13.0	0.105	P	-0.49	0	42	8.0	0.131
E	-1.22	3.0	73	12.3	0.151	Q	-0.91	0.2	72	10.5	0.180
F	1.92	-2.5	91	5.2	0.290	R	-0.59	3.0	101	10.5	0.291
G	-0.67	0	1	9.0	0	S	-0.55	0.3	31	9.2	0.062
H	-0.64	-0.5	82	10.4	0.230	T	-0.28	-0.4	45	8.6	0.108
I	1.25	-1.8	57	5.2	0.186	V	0.91	-1.5	43	5.9	0.140
K	-0.67	3.0	73	11.3	0.219	W	0.50	-3.4	130	5.4	0.409
L	1.22	-1.8	57	4.9	0.186	Y	1.67	-2.3	107	6.2	0.298

Note: AA represents amino acid; H1 represents Hydrophilic. H2 represents Hydrophobicity; V represents Volumes of side chains; P1 represents Polarity; P2 represents Polarizability.

$$H(R_i)=\frac{h^0(R_i)-\min(h^0(R_i))}{\max(h^0(R_i))-\min(h^0(R_i))} \quad (i=1,2\cdots 5) \tag{4}$$

The normalized eigenvalues are used for calculating correlation coefficient of sequence. Based on the AAC and PC principles, novel correlation information among *n* amino acid residues is introduced. Eqs. (5) and (6) are the expressions of the correlation coefficient of dipeptides and tripeptides, respectively.

$$cc-2(\lambda,k)=\frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda}\frac{A_{i,k}\times B_{i+\lambda,k}}{\sqrt{A_{i,k}\times A_{i,k}^T}\sqrt{B_{i+\lambda,k}\times B_{i+\lambda,k}^T}} \tag{5}$$

$$\begin{aligned} cc-3(\lambda_1,\lambda_2,k)&=\frac{1}{L-\lambda_1-\lambda_2}\sum_{i=1}^{L-\lambda_1-\lambda_2}M_i \\ M_i&=\frac{A_{i,k}\times B_{i+\lambda_1,k}\times C_{i+\lambda_1+\lambda_2,k}}{\sqrt{A_{i,k}\times A_{i,k}^T}\sqrt{B_{i+\lambda_1,k}\times B_{i+\lambda_1,k}^T}\sqrt{C_{i+\lambda_1+\lambda_2,k}\times C_{i+\lambda_1+\lambda_2,k}^T}} \end{aligned} \tag{6}$$

With limited data processing ability for each classifier, the order of this feature is no more than 4. Firstly, λ, λ_1 and λ_2 are the distance between different position amino acids. Secondly, k is the index of eigenvalues. Thirdly, L is the amino acids length number of a protein sequence. Finally, the $A_{i,k}$, $B_{i+\lambda,k}$, $B_{i+\lambda1,k}$, $C_{i+\lambda1+\lambda2,k}$ is *No.i*, *No.i*+λ, *No.i*+λ_1 and *No.i*+λ_1+λ_2 position amino acids in a sequence, respectively. Therefore, one can see that the feature of *cc-2* has 25 elements and the feature of *cc-3* has 125 elements.

Table 3

Feature representation

Feature set ID	Feature set name	Abbr.	Features
1	sequence length	L	1
2	amino acid composition	ACC	20
3	polypeptide composition-2	PC-2	400
4	correlation coefficient-2	cc-2,4	16
5	correlation coefficient-3	cc-3,5	125

2.2.4. Selection of the feature sets

According to the above discussion, the number of these features is so large that it will cost a lot of time. From the view of information efficiency, the classifier hardly makes full use of these features. The entire features in this research are shown in Table 3. In order to reduce the mutual influence between these features, the feature of sequence is normalized. The features of polypeptide composition-2, correlation coefficient-2 and correlation coefficient-3 have done some processes like Eq. (6).

3. Classification algorithm construction

3.1. Flexible neural tree

Flexible neural tree (*FNT)* is a special kind of artificial neural network with flexible tree structures. This kind of structure is put forward by Chen [22,23]. It is relatively easy for a *FNT* model to obtain near-optimal structure using tree structure optimization algorithms. Here, a *FNT* model is employed as the predictor. A tree-structural based encoding method with specific instruction set is selected to represent a FNT model. Flexible Neuron Instructors use function set *F* and terminal instruction set *T* to generate a FNT model, described as follows.

$$S = F \cup T = \{+_2, +_3 \cdots +_N\} \cup \{x_1 \cdots x_n\} \tag{7}$$

where $+_i$ *(i=1, 2 ... N)* denotes non-leaf nodes' instructions and has i arguments, the $x_1, x_2 \ldots x_n$ are leaf nodes' instructions and has no arguments.

The output of a non-leaf node is calculated using a FNT model shown in Figure 1. From this point of view, the instruction +*i* is also called a flexible neuron operator with *i* inputs.

The output of a flexible neuron +n is calculated as follows and the total excitation of $+_n$ is given by

$$net_n = \sum_{j=1}^{n} w_j x_j \tag{8}$$

In Eq. (9), x_j *(j=1,2, ..., n)* are the inputs to node $+_n$. The output of the node +n is then calculated by

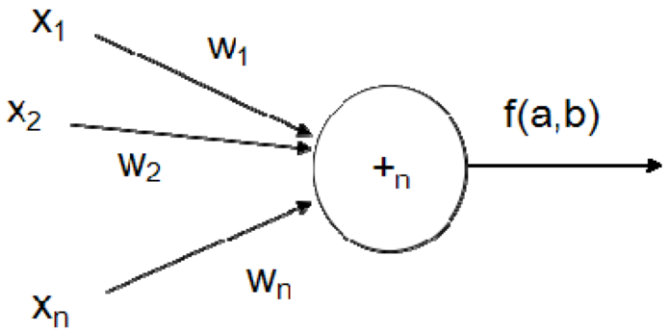

Fig. 1. Non-leaf node of flexible neural tree with a terminal instruction set T={x1,x2,…,xn}.

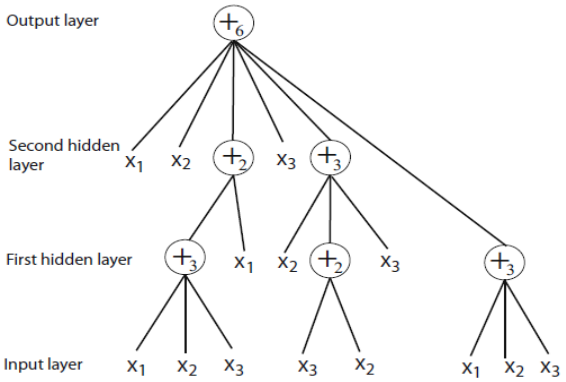

Fig. 2. Typical representation of FNT with function instruction set $\{+_2,+_3,+_4,+_5,+_6\}$ and terminal set $\{x_1,x_2,x_3\}$, which has four layers.

$$out_n = f(a_n, b_n, net_n) = e^{-(\frac{net_n - a_n}{b_n})^2} \tag{9}$$

A typical FNT model is illustrated in Figure 2. Its overall output can be computed from left to right by a depth-first method recursively.

General learning algorithm of FNT

- Step 1. Initialize the values of parameters used in the particle swarm optimization (PSO) algorithms. Set the elitist program as NULL and its fitness value as the biggest positive real number of the computer at hand. Create the initial population.
- Step 2. Construct optimization using PSO algorithm, in which the fitness function is calculated by root mean square error (RMSE).
- Step 3. If the better structure is found, then go to step 4, otherwise go to step 2.
- Step 4. Optimize parameters using PSO algorithm.
- Step 5. If the maximum number of local search is reached, or no better parameter vector is found for a significantly long time (100 steps), then go to step 6; otherwise go to step 4.
- Step 6. If the satisfied solution is found, then stop; otherwise go to step 2.

3.2. Prediction assessment

In statistical analysis, two methods can be used to examine a predictor for its effectiveness in practical application, namely, independent dataset tests and 10-fold cross validation tests. During the process of the 10-fold cross validation tests, 10% of all proteins are treated as test data and others as train data. The overall accuracy (*OA*) is computed for each dataset. In addition, the following two standard performance measures are used to evaluate the prediction accuracy, that is, Sensitivity (*Sens*) and Specificity (*Spec*).

4. Discussion and results

All experiments are performed on a basis of 4 test datasets using 10-fold cross validation tests and report the Sens and Spec for each structural class, as well as the overall accuracy (*OA*). As shown in Table 4, one can find that the accuracies for these datasets reach above 80%. And when comparing the four structural classes, it can also be found that the prediction in alpha/beta class is always the highest (with accuracies of over 92% for these datasets). However, it seems to be less effective to predict the all-alpha class as the prediction accuracy is relatively low (72.49% for 1189 dataset) in comparison with the other classes, which may be due to its non-negligible overlap with other classes.

Throughout the research, one can see that only a portion of features can be preserved. The other features show no obvious effect on classification. The preserved features are shown in Table 5. The feature of sequence length is also normalized.

Table 4

The summary of Sens (%) and Spec (%) in different datasets

Dataset	Class	Sens (%)	Spec (%)
640	all-alpha	76.92	84.62
	all-beta	81.25	89.38
	alpha/beta	94.73	95.27
	alpha+beta	83.87	93.64
	OA	84.51	
1189	all-alpha	72.49	87.52
	all-beta	82.65	92.46
	alpha/beta	93.04	98.45
	alpha+beta	77.24	92.62
	OA	82.56	
ASTRAL	all-alpha	74.31	85.25
	all-beta	77.29	90.23
	alpha/beta	93.82	96.82
	alpha+beta	84.81	89.32
	OA	83.06	
C204	all-alpha	89.46	92.64
	all-beta	93.02	95.35
	alpha/beta	99.27	98.21
	alpha+beta	94.32	98.07
	OA	94.65	

Table 5

The preserved features are summarized

Feature set ID	Feature set name	Abbr.	Features
1	sequence length	L	1
2	correlation coefficient-2 (Hydrophilic, Volumes of side chains, Polarizability)	cc-2,3	9
3	correlation coefficient-3	cc-3,5	125

Table 6

The comparison of the accuracies between all features and selected features

Dataset	Features	Accuracy				
		all-alpha (%)	all-beta (%)	alpha/beta (%)	alpha+beta(%)	Overall (%)
640	all features	77.82	80.42	92.32	82.53	83.62
	selected features	76.92	81.25	94.73	83.87	84.51
1189	all features	73.52	80.53	91.18	75.78	81.31
	selected features	72.49	82.65	93.04	77.24	82.56
ASTRAL	all features	74.31	77.29	93.82	84.81	83.06
	selected features	73.58	77.94	94.01	86.34	83.51
C204	all features	89.46	93.02	99.25	94.32	94.65
	selected features	90.27	94.01	99.01	89.46	94.62

4.1. Analysis of the selected feature vector

Among 135 rational designed features which are used in the vector based on the prediction four protein classes, each of the features plays an essential role in obtaining better prediction accuracy. On the one hand, the features, such as *ACC* and *PC-N*, play an important role in protein sequence. On the other hand, protein sequence homology can achieve a more accurate expression through these selected features in Table 6.

4.2. Comparison with existing methods

There are a large variety of methods for protein structure classification. Thus the prediction efficiency and accuracy of our method are compared with other recently reported prediction methods. Other methods shown in Table 7 are used to test the four experimental datasets, and the comparison results show that our method displays the highest prediction accuracies as revealed in Table 7. However, the accuracies of all-alpha are lower than the methods of second structure.

5. Conclusion

In the current study, newly designed features and a *FNT* model are proposed to predict the protein sequences with low similarity. Compared with existing methods, the proposed method shows higher prediction accuracy in the four experimental datasets. Next, we will focus on the combination between correlation coefficient of sequence and secondary structure sequence. Furthermore, it will be interesting to improve the accuracy of classification model of all-alpha.

Table 7
The comparison between our method and other methods

Dataset	Algorithm	all-alpha (%)	all-beta (%)	alpha/beta (%)	alpha+beta (%)	Overall (%)
1189	Bayes classifier [7]	null	null	null	null	53.8
	Logistic regression [17]	57.0	62.9	25.3	64.6	53.9
	PseAAC [18]	48.9	59.5	26.6	81.7	56.9
	RQA+PCC [19]	63.0	77.5	24	88.5	63.6
	MODAS [20]	92.30	87.10	65.40	87.90	83.50
	SCPERD [21]	89.10	86.70	53.80	89.6	80.60
	RKS-PPSC [24]	89.20	86.70	65.60	82.60	81.30
	PSIPRED [25]	93.72	84.01	66.39	83.53	81.96
	This method	72.49	82.65	77.24	93.04	82.56
640	IB1 [25]	53.62	46.10	68.93	34.50	50.94
	Naive Bayes 5 [25]	55.07	62.3	80.26	19.88	54.38
	Logistic regression [24]	69.57	58.44	61.58	29.82	54.06
	PseAAC [21]	73.91	61.04	81.92	33.92	62.34
	RQA&PCC [24]	59.17	60.83	75	56.67	62.92
	SCPRED [24]	90.60	81.80	66.70	85.90	80.80
	RKS-PPSC [25]	89.10	85.10	71.40	88.10	83.10
	PSIPRED [26]	93.72	84.01	66.39	83.53	83.44
	This method	76.92	81.25	83.87	94.73	84.51
C204	PseAAC [25]	95.92	93.10	72.92	71.43	83.82
	PsePSSM [26]	98.00	96.83	89.13	86.67	93.14
	SerialCombination [25]	100.00	98.36	85.71	83.33	92.65
	Serial fusion [26]	100.00	98.36	85.71	88.88	93.63
	Parallel fusion [26]	100.00	98.39	87.50	91.11	94.61
	This method	89.46	93.02	99.27	94.32	94.65
ASTRAL	SCPRED [24]	93.13	78.33	64.27	83.38	79.14
	RKS-PPSC [25]	94.06	83.38	71.47	85.01	83.01
	PSIPRED [26]	94.53	77.49	71.47	87.28	82.33
	This method	74.31	77.29	93.82	84.81	83.06

Acknowledgement

This research was partially supported by the Youth Project of National Natural Science Fund (61302128), the Key Project of Natural Science Foundation of Shandong Province (ZR2011FZ001), the Natural Science Foundation of Shandong Province (ZR2011FL022), the Key Subject Research Foundation of Shandong Province and the Shandong Provincial Key Laboratory of Network Based Intelligent Computing. This work was also supported by the National Natural Science Foundation of China (Grant No. 61201428, 61203105)

References

[1] M. Levitt and Chothia, Structural patterns in globular proteins, Nature **261** (1976), 552–558.
[2] G.P. Zhou and N. Assa-Munt, Some insights into protein structural class prediction, Proteins **44** (2001), 57–59.
[3] K.C. Chou, Prediction of protein cellular attributes using pseudo-amino acid composition, Proteins **43** (2001), 246–255.
[4] T.L. Zhang and Y.S. Ding, Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes, Amino Acids **33** (2007), 623–629.
[5] T.L. Zhang, Y.S. Ding and K.C. Chou, Prediction protein structural classes with pseudo-amino acid composition: Approximate entropy and hydrophobicity pattern, J. Theor. Biol. **250** (2008), 186–193.

[6] S. Altschul et al., Gapped blast and PSI-blast: A new generation of protein database search programs, Nucleic Acids Res. **25** (1997), 3389–3402.
[7] J. Cheng and P. Baldi, A machine learning information retrieval approach to protein fold recognition. Bioinformatics **22** (2006), 1456–1463.
[8] S. Jahandideh, P. Abdolmaleki, M. Jahandideh and E.B. Asadabadi, Novel two-stage hybrid neural discriminant model for predicting proteins structural classes, Biophys. Chem. **128** (2007), 87–93
[9] S. Costantini and A.M. Facchiano, Prediction of the protein structural class by specific peptide frequencies, Biochimie **91** (2009), 226–229.
[10] X. Xiao, W.Z. Lin and K.C. Chou, Using grey dynamic modeling and pseudo amino acid composition to predict protein structural classes, J. Comput. Chem. **29** (2008), 2018–2024
[11] X. Xiao, P. Wang and K.C. Chou, Predicting protein structural classes with pseudo amino acid composition: An approach using geometric moments of cellular auto maton image, J. T. Heor. Biol. **254** (2008), 691–696.
[12] A. Andreeva, D. Howorth, S. Brenner, T. Hubbard, C. Chothia and A. Murzin, SCOP database in 2004: Refinements integrate structure and sequence family data, Nucleic Acid Research **32** (2004), D226–D229
[13] Y. Cai, X.J. Liu, X.B. Xu and K.C. Chou, Support vector machines for prediction of protein domain structural class, Journal of Theoretical Biology **221** (2003), 115–120.
[14] X.D. Sun and R.B. Huang, Prediction of protein structural classes using support vector machines, Amino Acids **30** (2006), 469–475.
[15] J. Yang, Z. Peng and X. Chen, Prediction of protein structural classes for low-homology sequences based on predicted secondary structure, BMC Bioinform. **11** (Suppl. 1) (2010), S9.
[16] K.C. Chou and G.M. Maggiora, Domain structural class prediction, Protein Eng. **11** (1998), 523–538.
[17] B. Yang, Y.H. Chen and M.Y. Jiang, Reverse engineering of gene regulatory networks using flexible neural tree models, Neurocomputing **99** (2013), 458–466.
[18] L. Kurgan and L. Homaeian, Prediction of structural classes for protein sequences and domains-impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy, pattern recognition, Special Issue on Bioinformatics **39** (2006), 2323–2343.
[19] T.L. Zhang, Y.S. Ding and K.C. Chou, Prediction Protein structural classes with pseudo amino acid composition: Approximate entropy and hydrophobicity pattern, J. Theor. Bi-ol. **250** (2008), 186–193.
[20] G.T. Shao and Y.H. Chen, Predict the tertiary structure of protein with flexible neural tree, Lecture Notes in Computer Science **7390** (2012), 324–331.
[21] L. Kurgan, K. Cios and K. Chen, SCPRED: Accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences, BMC Bioinform. **9** (2008), 226.
[22] Yuehui Chen, Bo Yang, Jiwen Dong and Ajith Abraham, Time-series forecasting using flexible neural tree model, Information Sciences **174** (2005), 219–235.
[23] Bin Yang, Yuehui Chen and Mingyan Jiang, Reverse engineering of gene regulatory networks using flexible neural tree models, Neurocomputing **99** (2013), 458–466.
[24] J. Yang, Z. Peng and X. Chen, Prediction of protein structural classes for low-homology sequences based on predicted secondary structure, BMC Bioinform. **11** (2010), S9.
[25] S. Ding and S. Zhang, A novel protein structural classes prediction method based on predicted secondary structure, Biochimie **94** (2012), 1166–1171.
[26] D. Yu and J. Hu, Learning protein multi-view features in complex space, Amino Acids **44** (2013), 1365–1379.