A New Protein Structure Classification Model

Dong Wang[1], Shiyuan Han[1], Yuehui Chen[1*],Wenzheng Bao[1], Kun Ma[1]
1 School of Information science and Engineering
University of Jinan
Jinan, PR China
baowz55555@126.com

Ajith Abraham[2,3]
2 IT4Innovations,
VSB-Technical University of Ostrava
Ostrava, Czech Republic
3 Machine Intelligence Research Labs (MIR Labs)
Auburn, Washington, USA
ajith.abraham@ieee.org

***Abstract*—Protein structure prediction is an important area of research in bioinformatics .In this paper, we select the features of correlation coefficient sequence and special amino acid composition. The support vector machine and a particular framework of ECOC are employed as classification model. To evaluate the efficiency of the proposed method we choose three benchmark protein sequence datasets (25PDB, 40PDB and ASTRAL) as the test dataset. The final results show that our method is efficient for protein structure prediction.**

Keywords-Prediction structure of protein; Support Vector machine; ECOC

I. INTRODUCTION

During the last several decades, life science studies have gradually become complicated processes involving all sorts of modern technologies. Almost every experiment in today ’ s biology laboratory requires sophisticated devices that can only be manufactured with highly developed modern industry, which is armed with automatic controlling systems, high performance computers and precise manufacturing facilities. Bioinformatics and computational biology are now playing increasingly important roles in life sciences. We are in a new age in which the life sciences have become a precise and quantitative subject; theories and models can be applied to predict a number of experimental results and to guide laboratory practice.

The 3-D (dimensional) structure of a protein is uniquely dictated by its amino acid sequence, the so-called primary structure[1]-[2]. However, owing to the degenerate nature of the sequence-structure relationship, although the number of protein sequences is extremely large, the number of their folding patterns is quite limited. According to their chain folding topologies [3], proteins are usually folded into one of the following four structural classes: all-α, all-β, α/β, and α+β. The all-α and all-β proteins are essentially formed by α-helices and β-sheets respectively. The α/β class represents those proteins in which α-helices and β-strands are largely interspersed with the main sheet consisting mainly of parallel strands, , while the α+β class represents those in which α-helices and β-strands are largely segregated with the β-sheets almost always built up from anti-parallel strands. The four classes of protein structure show in Figure 1.

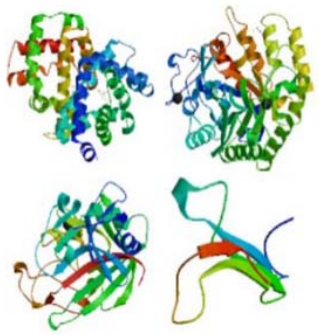

Figure 1. The four classes of protein structure

These class definitions clearly describe the underlying architecture of a protein’s structure, and hence have been generally accepted and are still in common use today. This represents that the degree of degeneracy between protein sequences and structural classes is extremely high. On the other hand, a high degeneracy also exists between protein sequences and amino acid compositions because a same amino acid composition can be derived from many different amino acid sequences.

Although various experimental technologies have been developed for determining prediction the structure of protein, almost every available approach is costly and time consuming. With the development of proteome projects, large amounts of available protein sequences and functional annotations have enabled us to develop computational methods as alternative choices. The computational methods used provide results of low resolution. Only four categories of protein structural were considered when the artificial neuron network was first applied in predicting the structure of protein. However, such methods were in so much demand that many efforts have been made in the last two decades to improve the prediction resolution as well as the prediction accuracy.

Many feature extraction methods have been proposed and are widely used in the protein predictions, such as utilizing amino acid composition(AAC) to predict cellular distribution[4]-[5], utilizing amino acid composition to

predict protein structural class[6]-[7], utilizing dipeptide composition to predict protein subcellular locations[8], utilizing polypeptide to predict protein structural class[9]-[10], utilizing the increment of diversity to predict the subcellular location of apoptosis proteins[11], utilizing the amino acid hydrophobicity to predict protein structural class[12], utilizing grouped weight to predict apoptosis protein subcellular localization[13], and utilizing physicochemical composition features to predict subnuclear localization[14].

II. DATASET

The dataset originally studied by Levitt and Chothia was the first structural dataset consisting of only 31 proteins that were classified completely based on a visual inspection[3]. In order to develop a statistical method for studying protein structural classes, a data-set of much more than 31 proteins must be constructed. Thus, various quantitative classification rules were proposed based on the percentages of α-helices and β-sheets in a protein.

In this research, we apply it to three benchmark datasets: 25PDB, ASTRAL and 40PDB [15]-[17]. 25PDB dataset contains 1673 protein domains, including 443 all-a class, 443 all-b class, 441 a+b class and 346 a/b class. The sequence homology of this dataset is below 40%. The ASTRAL database (including 7 classes) selected has sequence similarity lower than 20% which contains 6424 sequences. In this study, only four major classes that include 2813 sequences were used. The 40PDB dataset includes 40 non-homologous proteins were extracted from the Brookhaven Protein Databank.

It was found that the performance of classification is strongly affected by sequence homology of dataset. So these three datasets were selected just because the sequence homology is different. In this way, the results of classification will be more objective to value the validity of proposed result. The information of the three datasets show in TABEL I.

TABLE I. INFORMATION OF DATASET

Dataset	*All-α*	*All-β*	*α/β*	*α+β*	*total*
25PDB	443	443	441	346	1673
40PDB	165	213	243	169	790
ASTRAL	639	661	749	764	2813

III. FEATURE EXTRACTION

A. Correlation coefficient of sequence

From the viewpoint of molecular biology, traditional amino acid composition methods only 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 [18]. However the autocorrelation coefficient of sequence takes into account both the information of position of protein sequences and the interaction with distance between amino acids sequence. The feature reflects the structure of proteins. Five characteristics are selected to express protein sequence in this research, including hydrophilic, hydrophobicity, volumes of side chains, polarity and polarizability of amino acids. In order to take advantage of the feature, protein sequences are converted into discrete values. The original eigenvalues of the 7 characteristics are showed in TABLE II.

TABLE II. THE ORIGINAL EIGENVALUES OF THE 7 CHARACTERISTICS

AA	*I*	*II*	*III*	*IV*	*V*	*VI*	*VII*
A	-0.4	-0.5	15	8.1	0.046	0.67	1.28
C	0.17	-1.0	47	5.5	0.128	0.38	1.77
D	-1.31	3.0	59	13.0	0.105	-1.20	1.60
E	-1.22	3.0	73	12.3	0.151	-0.76	1.56
F	1.92	-2.5	91	5.2	0.290	2.30	2.94
G	-0.67	0	1	9.0	0	0	0
H	-0.64	-0.5	82	10.4	0.230	0.230	0.64
I	1.25	-1.8	57	5.2	0.186	0.186	1.90
K	-0.67	3.0	73	11.3	0.219	0.219	-0.57
L	1.22	-1.8	57	4.9	0.186	0.186	1.90
M	1.02	3.0	75	5.7	0.221	0.221	-0.57
N	-0.92	0.2	58	11.6	0.134	0.134	1.90
P	-0.49	0	42	8.0	0.131	0.131	1.20
Q	-0.91	0.2	72	10.5	0.180	0.180	-0.22
R	-0.59	3.0	101	10.5	0.291	0.291	-2.10
S	-0.55	0.3	31	9.2	0.062	0.062	0.01
T	-0.28	-0.4	45	8.6	0.108	0.108	0.52
V	0.91	-1.5	43	5.9	0.140	0.140	1.50
W	0.50	-3.4	130	5.4	0.409	0.409	2.60

Y	1.67	-2.3	107	6.2	0.298	0.298	1.60

AA is amino acid. I is hydrophilic. II is hydrophobicity. III is volumes of side chains. IV is Polarity. V is polarizability. VI is the solvent free energy. VII is curve shape index.

As the original eigenvalues differs in some degree, a normalized process should be used. In this research, the process of maximum & minimum standardization is introduced. The expression of this standardization is Eq.1.

$$H(R_i)=\frac{h_0(R_i)-\min(h_0(R_i))}{(\max(h_0(R_i)-\min(h_0(R_i))} \quad (i=1,2,3,4,5,6,7) \tag{1}$$

The normalized eigenvalue is used for the calculation of correlation coefficient of sequence. Based on the amino acid composition principle and the polypeptide composition principle, novel correlation information among *n* amino acid residues is introduced. Eq.2-4 is the expression of the correlation coefficient of dipeptides and tripeptides.

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

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

$$cc-4(\lambda_1,\lambda_2,\lambda_3,k)=\frac{1}{L-\lambda_1-\lambda_2-\lambda_3}\sum_{i=1}^{L-\lambda_1-\lambda_2-\lambda_3}M_i$$
$$M_i=\frac{A_{i,k}\cdot B_{i+\lambda_1,k}\cdot C_{i+\lambda_2,k}\cdot D_{i+\lambda_3,k}}{\sqrt{A_{i,k}\times A_{i,k}^{\ T}}\sqrt{B_{i+\lambda_1,k}\times B_{i+\lambda_1,k}^{\ T}}\sqrt{C_{i+\lambda_1+\lambda_2,k}\times C_{i+\lambda_1+\lambda_2,k}^{\ T}}\sqrt{D_{i+\lambda_1+\lambda_2+\lambda_3,k}\times D_{i+\lambda_1+\lambda_2+\lambda_3,k}^{\ T}}} \tag{4}$$

With limited data processing ability for each classifier, the order of this feature is no more than 4. Firstly, λ, λ_1, λ_2 and λ_3 are the distance between different position amino acids. Secondly, the k is the index of eigenvalues. Thirdly the L is the length of a protein sequence. Finally, the $A_{i,k}$, $B_{i+\lambda,k}$, $B_{i+\lambda1,k}$, $C_{i+\lambda1+\lambda2,k}$ and $D_{i+\lambda1+\lambda2+\lambda3,k}$ is *No.i*, *No.i*+λ, *No.i*+λ*1*, *No.i*+λ*1*+λ*2* and *No.i*+λ*1*+λ*2*+λ*3* position amino acids in a sequence. So we can see that the feature of *cc-2* has 49 elements and the feature of *cc-3* has 441 elements. What's more, the *cc-4* will have 3087 elements.

B. Special Amino Acid Compositions

According to the definition by Levitt and Chothia, there are some major differences between the structure of helix and folding, which is amino acid molecules arrangement, in different structure of the protein sequence [19].

According to the study, we found that the amino acid side chains and water leads to the formation of a protein sequences fold is the interaction between one of the main reasons. Different amino acid combination can appear in different space conformation. According to Lim's[13] research, 6 kinds of hydrophobic combinations (including (*i,i+2*), (*i,i+3*), (*i,i+2,i+4*), (*i,i+5*), (*i,i+3,i+4*), (*i,i+1,i+4*)) frequently exist in the protein. On the one hand ,a class includes a lot of (i,i+2) and (i,i+2,i+4) combinations .On the other hand ,b class contains many (*i,i+3*), (*i,i+5*), (*i,i+3,i+4*) and (i,i+1,i+4) combinations. Based on the theory of Rose's, there are 6 kinds of special amino acid pattern meeting the requirement.

TABLE III. SIX CATEGORIES OF SPECIAL AMINO ACID PATTERNS

No.	Motifs	Occurrence in β strand	Occurrence in α strand
1	*hp*	under represented and not frequent	over represented and frequent
2	*pphhp*	under represented and not frequent	over represented and frequent
3	*hhpp*	under represented and not frequent	over represented and very frequent
4	*pphh*	under represented and not frequent	over represented and very frequent
5	*hphph*	over represented and frequent	under represented and not frequent
6	*phphp*	over represented and frequent	under represented and not frequent

IV. METHODS

A. Support Vector Machine

The support vector machines (SVMs) are developed based on statistical learning theory and are derived from the structural risk minimization hypothesis to minimize both empirical risk and the confidence interval of the learning machine in order to achieve a good generalization capability. SVMs have been proven to be an extremely robust and efficient algorithm for classification [20]. Cortes and Vapnik proposed the current basic SVM algorithm [21].

The preliminary objective of SVM classification is to establish decision boundaries in the feature space, which separate data points belonging to different classes. SVM differs from the other

Classification methods are significantly. Its intent is to create an optimal separating hyper plane between two classes to minimize the generalization error and thereby maximize the margin. If any two classes are separable from among the infinite number of linear classifiers, SVM determines that hyper plane which minimizes the generalization error. and conversely if the two classes are non-separable, SVM tries to search that hyper plane which maximizes the margin and at the same time, minimizes a quantity proportional to the number of Misclassification errors. Thus, the selected hyper plane will have the maximum margin between the two classes, where margin is defined as a summation of the distance between the separating hyper plane and the nearest points on either side of two classes [20]. SVM models were originally developed for the classification of linearly separable classes of objects. Referring to Figure2. Consider a two-dimensional plane consisting linearly separable objects of two separate classes (*class* (+) and *class* (*)).

In real time problems it is not possible to determine an exact separating hyper plane dividing the data within the space and also we might get a curved decision boundary in some cases [22]. Hence SVM can also be used as a classifier for non-separable classes (shows Figure 3). In such cases, the original input space can always be mapped to some higher-dimensional feature space.

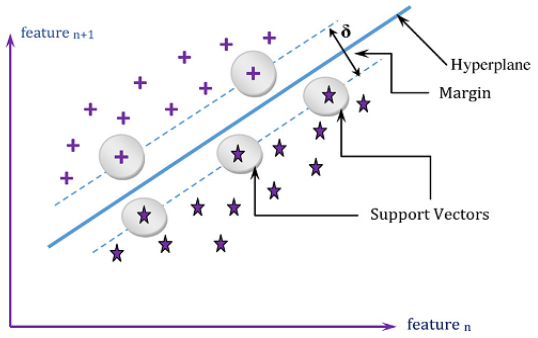

Figure 2. Maximum separation hyperplane

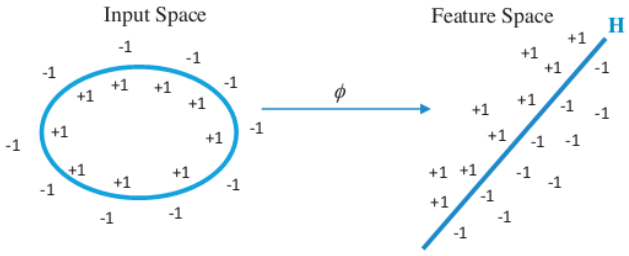

Figure 3. transform from high dimensional to low dimensional

Unique characteristics of SVM classification and kernel methods include: (1) the generalization capability enables trade-off between classifier complexity and error; (2) training is extremely robust and efficient; (3) search space has a unique minimal; (4) their performance is guaranteed as they are purely based on theoretical examples of statistical learning[23].

B. *One vs all framework*

The ECOC framework is to combine binary classifiers (dichotomizers), such as support vector machines (SVMs) and Adaboost[24]-[25]. Dietterich and Bakiri presented the basic ECOC framework represented using a coding matrix of binary symbols [26]. Each column of the coding matrix represents a binary partition of the whole classes in two subsets {0,1}. Alternatively, each row of the matrix is a code word assigned to the corresponding class. The one-vs-all strategy is a special case of the binary symbol based ECOC.

However, all these ECOC coding strategies, either problem-dependent or problem independent, suffer from two shortages: they do not use the combined dichotomizers to extract useful features of the data, and they endow only one codeword for each class. Addressing these two problems may lead to significant improvement of classification accuracy.

V. CLASSIFICATION MODEL

Based on four binary SVM ensemble classification models, we construct the one-vs-all classification mode. The results of each classifier output using hamming distance. Calculating the minimum output to achieve the Hamming distances of the classification. Some cases have more than one category code equals the distance and the minimum distance, at this time will be different weighting the output of the classifier. Finally, the samples of unknown categories based on the specific situation for special treatment. The code of 4 classifiers show in Table IV.

TABLE IV. THE CODING OF EACH CLASSIFIERS

	No.1 classifier	**No.2 classifier**	**No.3 classifier**	**No.4 classifier**
all-α	*1*	*0*	*0*	*0*
all-β	*0*	*1*	*0*	*0*
α+β	*0*	*0*	*1*	*0*
α/β	*0*	*0*	*0*	*1*

VI. RESULT AND ANALYSIS

In this research, we found that the parameter of each classifiers meet the requirement of Table V, the accuracy of prediction of protein structure classes will reach the maximum.

TABLE V. THE INFORMATION OF EACH CLASSIFIERS

Classifier	Feature	Input number
No.1	CC-3,5(I,II,III,V,VI)	125
No.2	CC-3,4(II,III,V,VI,VII)	125
No.3	CC-4,3(I,III,IV)+SAAC	91
No.4	CC-2,5(I,II,III,IV,V)+CC3,4(I,III,IV,VI)	89

Through the computing of each classifier, we found the result of each classifiers shows in TABLE VI.

TABLE VI. THE RESULT OF EACH CLASSIFIERS

Classifier	*Dataset*	*Accuracy (%)*
No.1	25PDB	92.66
No.2	25PDB	90.82
No.3	25PDB	93.57
No.4	25PDB	94.49
No.1	40PDB	90.76
No.2	40PDB	95.64
No.3	40PDB	91.53
No.4	40PDB	95.89
No.1	ASTRAL	89.34
No.2	ASTRAL	92.93
No.3	ASTRAL	91.07
No.4	ASTRAL	94.50

TABLE VII. COMPARISON OF ACCURACIES BETWEEN DIFFERENT METHODS(25PDB, 40PDB, ASTRAL)

Dataset	*Algorithm*	*Overall (%)*
25PDB	Logistic regression [27]	57.1
	Stacking C ensemble [27]	59.9
	SCPRED[28]	78.36
	RKS-PPSC[28]	82.90
	This paper	83.17
40PDB	This paper	84.51
ASTRAL	Bayesian classifier+AA[28]	53.8
	SVM+AA and polypeptide composition, physicochemical properties[28]	54.7
	SCPRED[29]	80.6
	MODAS[29]	83.5
	This paper	82.53

Form above TABLE VII, we can see that, the overall accuracy of method in this paper is more priority than some other methods. There are two reasons for this result. On the one hand, the method is an effective classifier model. On another hand, the methods of ensemble, we constructed here are derived from a large number of experimental results, and highly targeted.

VII. CONCLUSIONS

In this research, the hybrid feature extracted by *cc-n* and SAAC were used to present a protein sequence. The 25PDB, 40PDB and ASTRAL datasets protein sequences were used for conducting all the experiments. Compared with other traditional methods, the method will largely improve the classification accuracy. However, the model exist some drawbacks. Firstly, the number of feature is so large that waste a great many of storage space. Secondly, the one-vs-all model will have the ability to get high accuracy in each classifiers. But the final result will seriously decline. Thirdly, the SVM model can hardly explain the principle of biology. Therefore, in future work should focus on considering the information extracted protein homology.

ACKNOWLEDGMENT

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), the framework of the IT4 Innovations Centre of Excellence project, reg. no. CZ. 1. 05/1. 1. 00/02. 0070 by operational programme Research and Development for Innovations funded by the Structural Funds of the European Union and state budget of the Czech Republic, EU.

REFERENCES

[1] Anfinsen, C.B.. Principles that govern the folding of protein chains. Science(1973) 181, 223 −230 .

[2] Levitt M. Chothia C.: Structural Patterns in Globular Proteins. Nature 261, 552-558 (1976).

[3] Baker, D. A surprising simplicity to protein folding. Nature(2000) 405, 39-42.

[4] M. Levitt, C. Chothia, Structural patterns in globular proteins, Nature 261 (1976) 552-558.

[5] Zhou, G.P., Doctor, K.. Subcellular location prediction of apoptosis proteins. Proteins(2003): Struct. Funct. Genet 50, 44-48.

[6] Du, P.F., Li, Y.D.. Prediction of protein submitochondria locations by hybridiz-ing pseudo-amino acid composition with various physicochemical features of segmented sequence. BMC(2006) Bioinformatics 7, 518.

[7] Feng, K.Y., Cai, Y.D., Chou, K.C.. Boosting classifi er for predicting protein domain structural class. Biochem(2005). Biophys. Res. Commun. 334, 213-217 .

[8] Jahandideh, S., Abdolmaleki, P., Jahandideh, M., Asadabadi, E.B. Novel two-stage hybrid neural discriminant model for predicting proteins structural classes. Biophys(2007). Chem. 128, 87-93.

[9] Hayat, M, Khan, A. Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition. J. Theor(2011). Biol.271 , 10-17.

[10] Gao, Q.B., Zhao, H.Y., Ye, X.F., He, J. Prediction of pattern recognition receptor family using pseudo-amino acid composition. Biochem(2012). Biophys. Res. Commun.417 , 7 3-77.

[11] Jiang, X.Y., Wei, R., Zhang, T.L., Gu, Q. Using the concept of Chou ' s pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy. Protein Peptide(2008) Lett. 15, 392-396

[12] Sun, X.D., Huang, R.B. Prediction of protein structural classes using support vector machines. Amino Acids(2006) 30, 469−475.

[13] Vapnik, V. Statistical Learning Theory. Wiley-Interscience(1998), New York..

[14] Xiao, X., Shao, S.H., Huang, Z.D., Chou, K.C. Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. J. Comput(2006). Chem. 27, 478−482 .

[15] Zhou, G.P., Doctor, K. Subcellular location prediction of apoptosis proteins. Proteins(2003): Struct. Funct. Genet 50, 44-48.

[16] U. Hobohm, C. Sander, Enlarged representative set of protein structures, Protein Science 3 (1994) 522.

[17] W. Li, L. Jaroszewski, A. Godzik, Tolerating some redundancy significantly speeds up clustering of large protein databases, Bioinformatics 18 (1) (2002) 77−82.

[18] Mizianty M, Kurgan LA. Modular Prediction of Protein Structural Classes from Sequences of Twilight-Zone Identity with Predicting Sequences. BMC Bioinformatics(2009), 10:414

[19] Du Xiuquan, Cheng Jiaxing. Inferring protein-protein interactions from sequence using sequence order information[C] Proceedings of the 5th International Conference on ComputerScience and Education (ICCSE 10), Hefei, China, Aug 24-27,2010: 481-486.

[20] Kanaka Durga Kedarisetti, Lukasz Kurgan, Scott Dick.:Classifier ensembles for protein structural class prediction with varying homology. Biochemical and Biophysical Research Communications 348 (2006) 981-988

[21] V. Vapnik.: The Nature of Statistical Learning Theory, Springer, New York,1995.

[22] C. Cortes, V. Vapnik, Support vector networks, Machine Learning 20 (1995) 273−297.

[23] Sujay Raghavendra. N.:Support vector machine applications in the field of hydrology: A review. Applied Soft Computing 19 (2014) 372-386

[24] C.-C. Chang and C.-J. Lin. LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1-27:27, 2011

[25] Guoqiang Zhong, Cheng-Lin Liu.: Error-correcting output codes based ensemble feature extraction. Pattern Recognition 46 (2013) 1091-1100

[26] T.G. Dietterich, Ensemble methods in machine learning, in: Multiple Classi-fier Systems, 2000, pp. 1-15.

[27] N. Nilsson, Learning Machines, McGraw-Hill, 1965.

[28] Kanaka Durga Kedarisetti, Lukasz Kurgan.: Classifier ensembles for protein structural class prediction with varying homology. Biochemical and Biophysical Research Communications 348 (2006) 981-988

[29] Marcin J M izianty and Lukasz Kurgan. Modular prediction of protein s tructural classes from sequences of twilight-zone identity with predicting sequences. BMC Bioinformatics 2009 , 10:414

[30] Shuyan Ding, Shengli Zhang.: A novel protein structural classes prediction method based on predicted secondary structure. Biochimie 94 (2012) 1166-1171