Predict the Tertiary Structure of Protein with Error-Correcting Output Coding and Flexible Neural Tree

Yiming Chen
School of Information science and Engineering,
University of Jinan
Jinan, PR China
yiming198807@163.com

Yuehui Chen
School of Information science and Engineering,
University of Jinan
Jinan, PR China
yhchen@ujn.edu.cn

***Abstract*—In this paper we intend to apply a new method to predict tertiary structure. A novel hybrid feature adopted is composed of physicochemical composition (PCC), recurrence quantification analysis (RQA) and pseudo amino acid composition (PseAA). We use the Error Correcting Output Coding (ECOC) based on three flexible neural tree models as the classifiers. 640 dataset is selected to our experiment. The predict accuracy with our method on this data set is 60.23%, higher than some other methods on the 640 datasets. So, our method is feasible and effective in some extent.**

Keywords: tertiary structure, feature extraction, ECOC, FNT

I. INTRODUCTION

Protein is the essence of any life; it plays an important role in basic life support, the study of protein tertiary structure is helpful to protein function, and then understands the essence of life phenomenon.

Duo to the gap between the number of protein sequence data and structure data become more and more big, the protein structure prediction is gradually urgent and important. Efforts from many researchers have been done for several decades on this field. We should find more new feature extract methods and new classifiers to improve the predict accuracy of protein tertiary structure. Instead of the traditional classification models，we apply FNT as the base classifier for this field.

We will take several steps in our experiment: 1.Establish a data set; 2. Extract feature to obtain the information of protein sequence; 3. Design a classification model.

II. DATA SET

This paper we select 640 dataset to make the experiment.640 dataset contains 640 protein samples,138 samples are all-α class, 154 samples are all-β class,177 samples are all-α+βclass and 171 samples are all-α/β class, The sequence homology of this dataset is about 25%.It makes our method more persuasive duo to the lower sequence homology.

III. FEATURE EXTRACT METHODS

A. Physicochemical Composion

We divide the 20 amino acids into three groups on the basis of their physicochemical properties, including seven types [7] of hydrophobicity, normalized van der Vaals volume, polarity, polarizibility, charge, secondary structures and solvent accessibility. For instance, we use hydrophobicity attribute to divide amino acids into three groups: polar, neutral and hydrophobic. Then a protein sequence is transformed into a sequence of hydrophobicity attribute. Thus, the composition descriptor contains three values: the global percent compositions of polar, neutral and hydrophobic residues in the new sequence. For seven types of attributes, PCC consists of a total of 3×7=21 descriptor values.

B. Recurrence quantification analysis

Recurrence quantification analysis [2-3] is a powerful nonlinear method in analyzing time series; it has no requirement on the length of time series. Before we extract protein feature in this way, we should get a recurrent plot (RP) [4] of a protein sequence. This is to say, we should make a transition for protein sequence as follows. Firstly, we convert amino acids sequence into nucleotide sequence. But every transition must follow some rules. Here we take the encoding method which is listed in the table 1 as our transition [5]. Secondly, we use Chaos game representation (CGR) [6] to describe a nucleotide sequence on a plot. For a nucleotide sequence, a CGR is defined as a [0, 1] square, the details of RQA please refer to reference [3].

In the experiment, we will use RQA to have an analysis about the RP. DET, ENT, VMAX, LAM, REC and TT are adopted in this paper, so we can obtain twelve features for every protein sequence because of the two time series X and Y.

C. Pseudo Amino Acid composition (PseAA)

According to the concept of Chou's PseAA composition [8], the protein sequence can be formulated as:

$$P = \{p_1, p_2, \ldots . p_{20}, p_{21}, \ldots p_{20+\lambda}\} \lambda < L \qquad (1)$$

$$x_i = \begin{cases} \dfrac{f_i}{\sum_{j=1}^{20} f_j + w\sum_{j=1}^{\lambda} P_j} & 1 \le i \le 20 \\ \dfrac{w\mu_i}{\sum_{j=1}^{20} f_j + w\sum_{j=1}^{\lambda} P_j} & 21 \le i \le 20+\gamma \end{cases} \qquad (2)$$

The first 20 components are the occurrence frequencies of 20 amino acids in sequence. $P_i(21 \le i \le 20+\lambda)$ are the

additional factors that reflect some sort of sequence order information. In this paper the parameter w is set to 5, the parameter λ is set to 20; L is the length of protein sequence.

IV. CLASSIFICATION MODEL.

A. *ECOC framework*

The main idea of ECOC [9] is that decompose multi-class problem into several two-class problems ,the solution of the two-class problems depends on base binary classifiers . This paper we describe a encoding matrix as $M_{R\times n}$ listed in table2. R is the class number of the dataset and n stands for the number of the single classifier. Each row of matrix is called the code word, it corresponds to one class of the dataset, and each column represents the mark of each class in training every base classifier. Every binary classifier outputs a predicted value which forms a vector $H(x) = (h_1(x), h_2(x), \cdots, h_n(x))$. In this paper we use flexible neural tree as base binary classifier and apply Hamming distance function to calculate the distance between the output vector H(x) and each code word of encoding matrix, the corresponding class label of the shortest coding word is the output of the test sample [10].

There are many kinds of forms for the coding [11], like one-to-many matrix, one-to-one matrix, random sparse coding matrix and dense random coding matrix etc.

Table Π. CODING MATRIX

classifier / *structure*	*FNT1*	*FNT2*	*FNT3*
α	0	1	0
α/β	1	0	0
α+β	0	0	1
β	1	1	1

B. *Flexible Neural Tree*

The base prediction model is flexible neural tree (FNT). Probabilistic Incremental Program Evolution (PIPE) and Particle Swarm Optimization algorithms (PSO) are employed for the structure and parameters of FNT. The framework of FNT allows input features selection. It can automatically design the better structure and optimize parameters. The individuals of FNT tend to simplify structure of the similar model due to the evolutionary algorithm. Although the final structure of FNT is usually simpler than that of neural network, it has better generalization ability; FNT can automatically select the input features that contribute more to the final predict result.

The flexible neuron instructor and FNT model are composed of the function set F and terminal instruction set T described as follows:

$$S = F \cup T = \{+2, +3, ..., +N\} \cup \{x_1, x_2 ..... x_n\} \quad (3)$$

The F set +i (i = 2, 3, N) are non-leaf nodes' instructions described as flexible neuron operators which has i inputs. And $\{x_1, x_2, \cdots, x_n\}$ are leaf nodes' instructions i.e.,. i real values which are denoted as W are randomly generated and used for representing the connection strength between the node +i and its children. In addition, two adjustable activation parameters ai and bi are randomly created.

The output of a flexible neuron operator is calculated by the activation function.

$$out_n = f(a_n, b_n, net_n) = e^{-(net_n - a_n/b_n)^2} \quad (4)$$

$$net_n = \sum_{j=1}^{n} w_j \times x_j \quad (5)$$

if a non-terminal instruction, i.e., + i(i = 2, 3, 4,…,N) is selected, i real values defined as W are randomly generated and used for representing the connection strength between the node +i and its children. In addition, two adjustable activation parameters ai and bi are randomly created.

We show a typical flexible neuron operator and a neural tree model in Figure 1. The overall output of flexible neural tree can be computed from left to right by depth-first method, recursively. Duo to the limited space, please see references [12][13] [14]for details of FNT.

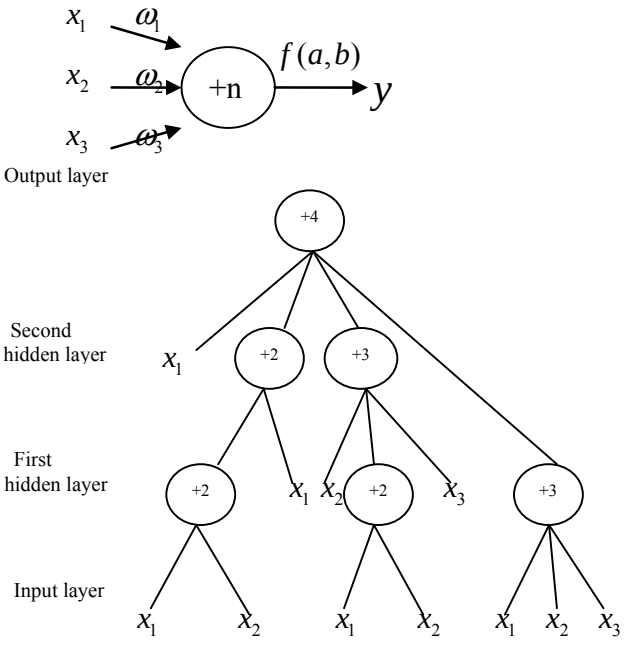

Figure 1 A flexible neuron operator and a typical FNT model

V. EXPERIMENTAL RESULTS

In the classification problems, we generally use the cross-validation method to evaluate the performance of classification method. The 10-jackknife cross-validation was adopted in this paper [16]. We calculate the overall success rate and accuracy of every class. We show the predict result obtained from different algorithms in the table I. From table

III we can conduct that the accuracy of our method is higher than the result of some other experiments.

Table III. THE COMPARISON OF THE RESULTS WITH PRIOR RESEARCH

algorithms	accuracy rate				overall accuracy rate
	α	α/β	β	α+β	
IB1[18]	53.62	46.10	68.93	34.50	50.94
Naïve Bayes[18]	55.07	62.34	80.26	19.88	54.38
Logistic regression[18]	69.57	58.44	61.58	29.82	54.06
This method	55.88	52.63	77.27	57.14	60.73

VI. CONCLUSION

We propose a novel hybrid feature of protein sequence, adopt ECOC as the predict frame. FNT has been successfully applied in this filed, we can select effective feature, and improve the structure and parameters according to it. The results listed in Table III show that our method may make some contribution for protein structure prediction.

ACKNOWLEDGMENT

The authors thank to colleagues in Computational Intelligence Lab for their assistance. This research was partially supported by the Natural Science Foundation of China (61070130), the Key Project of Natural Science Foundation of Shandong Province (ZR2011FZ001), the Key Subject Research Foundation of Shandong Province and the Shandong Provincial Key Laboratory of Network Based Intelligent Computing.

REFERENCES

[1] Shi JY, Zhang SW, Pan Q, Cheng YM, Xie J. "SVM-based method for subcellular localization of protein using multi-scale energy and pseudo amino acid composition Amino Acids",33(1): 69-74 (2007).

[2] Giuliani, A, Sirabella, P., Benigni, R., Colosimo, A, 2000. Mapping protein sequence spaces by recurrence: a case study on chimeric structures. Protein Eng.13,671--678.

[3] Giuliani, A, Tomasi, M., 2002. Recurrence quantification analysis reveals interac-tion partners in paramyxoviridae envelope glycoproteins. Proteins 46, 171-176.

[4] Marwan, N., Romano, M.e., Thiel, M., Kurths, 1, 2007. Recurrenceplots for the analysis of complex systems. Phys.Rep. 438,237-329.

[5] Deschavanne, P, Tuffe ' ry, P., 2008. Exploring an alignment free approach for protein classification and structural class prediction. Biochimie 90, 615-625.

[6] Fiser, A., Tusna 'dy, G.E, Simon, I..Chaos game representation of protein structures. J. Mol. Graphics 12,302-304.

[7] Jianyi Yang, Zhenling Peng, et al. Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation. J. TheoL BioI. 2009, doi: 10.1 OJ6/j.jtbi.2008.12.027.

[8] Chou KC. "Prediction of protein cellular attributes using pseudo-amino acid composition". Proteins: Struct Funct Genet, 43(3): 246-255 (2001).

[9] Huang Y, Li Y D. "Prediction of protein subcellular locations using fuzzy K-NN method". Bioinformatics, 20 (1): 21-28 (2004).

[10] Thomas G. Dietterich G. Bakiri. "Solving multiclass learning problems via Error-Correcting output codes". Artificial Intelligence Research, (2): 263-286 (1995).

[11] LUO D F, JUN, XIONG RONG. "Distance function learning in error-correcting output coding framework" [C]//ICON IP 2006 Proceeding of the 13th International Conference on Neural Information Proceeding LNCS 4233. Berlin: Springer-Berlag: 1-10 (2006).

[12] Chen, Y., Yang, B., Dong, J., Nonlinear systems modelling via optimal design of neural trees.International Journal of Neural systems. 14, (2004) 125-138

[13] Chen, Y., Yang, B., Dong, J., Abraham A.: Time-series forecasting using flexible neural tree model. Information Science, Vol.174, Issues 3/4, pp.219-235, 2005

[14] Chen, Y., Yang, B., Abraham A. "Feature Selection and Classification using Flexible Neural Tree", Neurocomputing, 2006. (In press).

[15] Masulli F, Valentini G. "Effectiveness of error correcting output codes in multiclass learning problems". Lecture Notes in Computer Science 1857, 107-116 (2000).

[16] Chou, K.C., Zhang, C.T., 1995. Review: Prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 30, 275–349.

[17] Chen, C., Chen, L., Zou, X., Cai, P., 2009.Prediction of protein secondary structure content by using the concept of Chou's pseudo-amino acid composition and support vector machine.Protein Pept. Lett.16, 27–31.

[18] Ke Chen, LUKASZ A. KURGAN, Jishou ruan.Prediction of protein structural class using novel evolutionary collocation-based sequence representation. J. Computational Chemistry.2008, 29:1596–1604.

[19] Wang ZX and Yuan Z: How good is the prediction of protein structural class by the component-coupled method? Pattern Recogn 2000, 38:165–175.

[20] Kurgan LA and Homaeian L: Prediction of structural classes for protein sequences and domains-Impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy. Pattern Recogn 2006, 39:2323–2343.

[21] Kedarisetti KD, Kurgan LA and Dick S: Classifier ensembles for protein structural class prediction with varying homology. Biochem Biophys Res Commun 2006, 348:981–988.

TABLE I. THE REVERSE ENCODING FOR AMINO ACIDS

A=GCT	***G=GGT***	***M=ATG***	***S=TCA***	***C=TGC***	***H=CAC***	***N=AAC***	***T=ACT***	***D=GAC***	***I=ATT***
P=CCA	V=GTG	E=GAG	K=AAG	Q=CAG	W=TGG	F=TTC	L=CTA	R=CGA	Y=TAC