Predict the Tertiary Structure of Protein with Binary Tree and Ensemble Strategy

Yiming Chen[1, a], Yuehui Chen [2,b]

[1]Department of Information Science and Engineering, University of Jinan, Jinan, 250022, China

[2]Department of Information Science and Engineering, University of Jinan, Jinan, 250022, China

[a]yiming198807@163.com,[b]yhchen@ujn.edu.cn

Keywords: tertiary structure;binary tree; selective ensemble;FNT

Abstract. In this paper we intend to apply a new method to predict tertiary structure. Several feature extraction methods adopted are physicochemical composition, recurrence quantification analysis (RQA) , pseudo amino acid composition (PseAA) and Distance frequency. We construct the binary tree Classification model, and adopt flexible neural tree models as the classifiers. We will train a number of based classifiers through different features extraction methods for every node of binary tree, then employ the selective ensemble method to ensemble them. 640 dataset is selected to our experiment. The predict accuracy with our method on this data set is 63.58%, higher than some other methods on the 640 datasets. So, our method is feasible and effective in some extent.

Introduction

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

In recent years, 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 many years on this field. We should find some new feature extract methods and new classifiers to improve the predict accuracy of protein tertiary structure. We apply FNT as the base classifier in this field instead of the traditional classification models.

We 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.4. Ensemble these based classifiers.

Dataset

There are four benchmark datasets in the field of protein tertiary structure; they are C204 datasets, 1189 datasets, 640 datasets and 25PDB datasets. This paper we select 640 dataset to make the experiment. The sequence homology of this dataset is about 25%.It make our method more persuasive duo to the lower sequence homology.

Feature extract methods

Physicochemical Composion

The 20 amino acids are divided 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. PCC consists of a total of 3×7=21 descriptor values because of seven types of attributes.

Recurrence quantification analysis

Published by Atlantis Press, Paris, France.

Recurrence quantification analysis [2-6] is a powerful nonlinear method in analyzing time series; before we extract protein feature, a recurrent plot (RP) [4] of a protein sequence should be obtained. We should make a transition for protein sequence. Firstly, we convert amino acids sequence into nucleotide sequence. The encoding method is listed in the table 1 [5]. Secondly, we use Chaos game representation (CGR) [6] to describe a nucleotide sequence on a plot. CGR is defined as a [0, 1] square for a nucleotide sequence; please refer to reference [3] for the details of RQA.

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.

Subsection distance frequency

The 20 native amino acids are divided into 6 classes [22] according to the properties. We show the result of classification in table2.

After the partition, protein sequence can be represented by combination of the six letters. For every kind of amino acids, we will separately calculate the distance-value's occurrence number of two letters which belong to same type. One sequence thus gets one vector Vi [23] on the basis of distance frequency:

$$V_i = \left[v_1^L, v_2^L, \cdots, v_s^L, v_1^B, v_2^B, \cdots, v_s^B, \cdots, v_1^C, v_2^C, \cdots, v_s^C \right].$$

This paper the value of s is set to 11. In order to obtain the partial information we try to divide the protein sequence into three parts.

Pseudo Amino Acid composition (PseAA)

According to PseAA composition [8], the protein sequence can be described as:

$$P = \{p_1, p_2, \ldots\ldots p_{20}, p_{21}, \ldots p_{20+\lambda}\} \lambda < L \tag{1}$$

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

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.

Classification model

Binary tree model

We design the Binary tree Classification model based on FNT classifier as follows

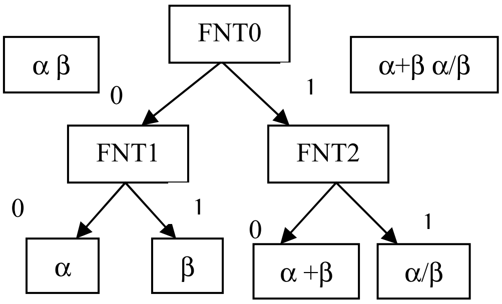

Figure 1 Binary Tree classification model

For the first node (FNT0), four classes of protein tertiary structure are divided into two groups αβ class and α+β α/β class. The second node is used to classify α and β from αβ class, finally. We can distinguish α+β and α/β class through FNT2 node.

Flexible Neural Tree

The base classifier is flexible neural tree (FNT). We use Probabilistic Incremental Program Evolution (PIPE) and Particle Swarm Optimization algorithms (PSO) to optimize the structure and parameters of FNT. FNT allows input features selection and the individuals of FNT tend to simplify structure of the similar model due to the evolutionary algorithm. The flexible neuron instructor and FNT model are composed of the function set F and terminal instruction set T described as follows:

$$S = F \cup T = \{+2, +3, \ldots\ldots, +N\} \cup \{x_1, x_2 \ldots\ldots x_n\} \tag{3}$$

The F set +i (i = 2, 3, N) are non-leaf nodes' instructions which has i inputs. $\{x_1, x_2, \cdots, x_n\}$ are leaf nodes' instructions i.e.,. The output of a flexible neuron operator is calculated by the activation function.

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

$$net_n = \sum_{j=1}^{n} w_j \times x_j \tag{5}$$

Duo to the limited space, please see references [12][13][14] for details of FNT.

Integration

Selective ensemble method[24] is a learning algorithm, it trains different kinds of based classifier and selects some of them to ensemble. Selecting a part of based classifier is effective than that select all based classifier.

This paper we will use five kinds of feature extraction methods to construct five different based classifier for every node of Binary tree Classification model, they are Physical and chemical composition, the fusion of Recurrence quantification analysis and Physical and chemical composition, the fusion of Pseudo Amino Acid composition and Recurrence quantification analysis, subsection Distance frequency, the fusion of Pseudo Amino Acid composition and subsection Distance frequency. Then, we apply the selective ensemble method to ensemble these based classifier.

Experimental results

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 result obtained from different method in the table 3. From the table 3 we can conduct that the accuracy of our method is better than the result of some other experiments.

Conclusion

We construct a Binary tree Classification model based on flexible neural tree models as the classifiers. We adopt five different to construct five different based classifier and use the selective ensemble strategy to ensemble them. The results listed in Table Ш show that our method may make some contribution for protein structure prediction.

Acknowledgment

This research was 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.

[22] Pa’nek J，Eidhammer I，Aasland R．A new method for identification of protein (Sub) families in a set of proteins based on hydropathy di stribution in proteins．Proteins：Struct Funct Bioinformatics，2005，58：923—934.

[23] Zhang Li，Liao Bo，Li Dachao，Zhu Wen．A novel representation for apoptosis protein subcellular localization prediction using support Vector machine．J Theor Bi01．2009，259：361-365．

[24] Zhihua, Z., Jianxin, W., Wei, T.: Ensembling neural networks: Many could be better than all. Artif. Intell. 137, 239–263 (2002)

Table 1 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

Table 2. Amino acids hydration property classification

Classification	Abbreviation	Amino acids
hydrophily	L	R,D,E,N,Q,K,H
hydrophobicity	B	L,I,V,A,M,F
neutral	W	S,T,Y,W
proline	P	P
glycocoll	G	G
cysteine	C	C

Table 3 The comparison of the results with prior research

algorithms	accuracy rate				overall accuracy rate
	α	β	α+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
SVM[18]	73.91	61.04	81.92	33.92	62.34
This method	74.07	66.67	44.12	69.44	63.58

Published by Atlantis Press, Paris, France.