Predicting Protein Subcellular Localization Using the Algorithm of Diversity Finite Coefficient Combined with Artificial Neural Network

Zeyue Wu, Yuehui Chen
Department of Information Science and Engineering, University of Jinan
Jinan, 250022, China
Email: fly1024@126.com

***Abstract*—Protein subcellular localization is an important research field of bioinformatics. The subcellular localization of proteins classification problem is transformed into several two classification problems with error-correcting output codes. In this paper, we use the algorithm of the increment of diversity combined with artificial neural network to predict protein in SNL6 which has six subcelluar localizations. The prediction ability was evaluated by 5-jackknife cross-validation. Its predicted result is 81.3%. By com-paring its results with other methods, it indicates the new approach is feasible and effective.**

Keywords-subcellular localization; feature extraction; artificial neural network; ECOC

I. INTRODUCTION

According to the spatial distribution and different functions, cells can be divided into a plurality of cells or cell areas, such as cytoplasm, nucleus, Golgi apparatus, mitochondria, cell membrane, endoplasmic reticulum and so on. These organelles called subcells.

Protein is transported to the specific organelles under protein sorting signals' guidance. If it is transported to the wrong position, it will influence the function of cells, even the whole organisms [1]. Protein is not static in a certain region of the cell. It plays its role through moving in different regions. With the rapid growth of protein quantity in recent years, it is urgent to know proteins' localization because it is closely related to their functions and the role it plays in the biological activities. It is very benefit to basic research and drug design [2].

Various approaches for protein subcellular localization prediction have been developed according to protein sequence information. The earlier approaches in this regard were based on the amino acid composition [3,4,5,6]. However, if the prediction was based on the amino acid composition, it would lose some information, such as sequence length effect of a protein. To solve the problem, Chou introduced the pseudo amino acid composition (PseAA) [7]. With the introduction of the PseAA, it has developed greatly in protein subcellular localization prediction area [8,9,10]. With the introduction of functional domain composition [8], the researchers put gene annotation (GO) into protein subcellular localization prediction area [11]. Zhang [12] developed a new encoding method with grouped weighted for protein sequence. In addition to feature extraction algorithms what were mentioned above, there were position specific scoring matrix (PSSM), auto covariance (AC) and so on.

Moreover, there are a lot of machine methods have been used in protein subcellular localization prediction. In 1998, Reinhardt and Hubbard used BP network in protein subcellular localization prediction. In 2001, Hua and Sun used support vector machine (SVM) firstly to predict protein subcellular localization. Besides, Bias network and Markov model were also introduced into this area. Chen and Li [13,14] had developed two prediction approaches based on increment of diversity (ID) and increment of diversity with support vector machine (ID_SVM). Song and Shi [2] introduced a method that was based on Hilbert-Huang transform. In this paper, a different approach is used for predicting protein subcellular location. We have developed two prediction approaches based on diversity finite coefficients (DFC) and artificial neural network (ANN).

II. MATERIALS AND METHODS

A. Dataset

In this paper, SNL6 dataset was chosen to validate the availability of our classifier. This dataset is founded by Lei and Dai. It is commonly used in subcellular localization. SNL6 contains 504 proteins and thses proteins are localized in 6 subcellular positions. Among the 504 sequences, 61 belonged to chromatin, 55 to nuclear lamina, 56 to nuclear speckle, 219 to nucleolus, 75 to nucleoplasm, and 38 to PML body.

B. Representation of Protein Sequence

Given a protein sequence P with L amino acid resides, it can be formulated as [13,14]

$$P = R_1 R_2 R_3 ... R_L \quad (1)$$

Where, R_1 is the first amino acid of the protein sequence, R_L the L amino acid of the protein sequence.

1) Amino acid composition(AAC)

As mentioned in prior works, the amino acids composition vector of a protein sequence is a simple sequence representation that is widely used in prediction of various structural aspects. Amino acid composition expresses the occurrence frequencies of 20 amino acids in protein P. [13,14] It can be expressed in a formula as follows.

$$P = (p_1, p_2, ... p_{20}) \quad (2)$$

Where, p_i ($i = 1,2,...20$) are the occurrence frequencies of 20 amino acids in sequence P.

2) Dipeptide composition(DC)

Frequencies of amino acid pairs provide more information since they reflect interaction between local (with respect to the sequence) amino acid pairs. The occurrence frequencies of 400 amino acid pairs in protein P can be expressed in a formula as follows.

$$P = (p_1, p_2, ... p_{400}) \tag{3}$$

Where, p_i ($i = 1,2,...400$) are the occurrence frequencies of 400 amino acid pairs in sequence P.

3) N-terminal signal (NTS)

From the second amino acid of N-terminal of a protein, we retain 30 amino acids as this protein. We statistics the frequencies of 20 amino acids of each site [13,14]. It is a 600(20*30=600) dimensional vector and can be expressed in a formula as follows.

$$P = (p_1, p_2, ... p_{600}) \tag{4}$$

C. Ensemble Classifier Prediction System

1) Increment of diversity

FM Li et al. introduced the method which was called increment of diversity [15]. They hypothesized that the state space X consisted of s information symbols and was presented as $X\{x_1, x_2, ... x_s\}$ .Besides, they defined X diversity source. Then X can be expressed as X $[n_1, n_2, ... n_s]$. They used n_i ($i = 1,2,...s$) as the occurrence frequencies of s states, the diversity of X is expressed in a formula as follows.

$$D(X) = D(n_1, n_2, ... n_s) = N \log_b N - \sum_{i=1}^{s} n_i \log_b n_i \tag{5}$$

Here, $N = \sum_{I=1}^{S} n_i$.

If we have two diversities, X $[n_1, n_2, ... n_s]$ and Y $[m_1, m_2, ... m_s]$, the increment diversity of X and Y is,

$$\Delta(X, Y) = D(X + Y) - D(X) - D(Y) \tag{6}$$

According to the above definition, we can prove that the increment of diversity has nonnegative and symmetry properties. That is $\Delta(X, Y) \geq 0$, $\Delta(X, Y) = \Delta(Y, X)$, and it can be written as follows.

$$\Delta(X, Y) = D(M, N) - \sum_{i-1}^{s} D(m_i, n_i) \tag{7}$$

Where, $M = \sum_{i=1}^{s} m_i$, $N = \sum_{i=1}^{s} n_i$.

$$D(M, N) = (M + N)\log_b (M + N) - M \log_b M - N \log_b N$$

$$D(m_i, n_i) = (m_i + n_i)\log_b (m_i + n_i) - m_i \log_b m_i - n_i \log_b n_i$$

If m_i or n_i is zero, then $D(m_i, n_i) = 0$ .In this paper, the value of b is set to 10. The researcher(FM Li) defined diversity finite coefficient in a formula as follows.

$$I(X, Y) = \frac{\Delta(X, Y)}{D(M, N)} = 1 - \sum_{i=1}^{s} \frac{D(m_i, n_i)}{D(M, N)} \tag{8}$$

It is obvious, $0 \leq I(X, Y) \leq 1$.

According to the different occurrence frequencies of 20 amino acids or 400 amino acid pairs or N terminal signal in protein sequences which are localized in 6 subcellular positions, we can form 6 standard diversity sources. Then using the formula (8), we can comstruct 6 diversity finite coefficients of each prediction of protein with 6 standard diversity sources. We take the standard source category of the smallest diversity finite coefficient as the prediction of protein's subcellular location category.

2) Artificial neural network (ANN)

In this paper, the basic classifier is artificial neural network (ANN). ANN can be used in function approximation, pattern recognition, classification and so on. It has strong robustness and toterance and can learn from the uncertain systems [16]. So it has been widely applied to the subcellular location [16]. We adopt the Particle Swarm Optimization (PSO) to optimize the parameters (weights and thresholds) of the ANN.

3) Error-correcting output coding (ECOC)

The motive for using ECOC to solve multiclass problem is that it can solve problem by decomposing a complex multiclass classification into a set of binary classification. In testing process, every single classifier outputs a predicted value which forms a output vector $H(x) = (h_1(x), h_2(x), ... h_n(x))$. It uses Hamming distance function or Euclidean distance function to calculate the distance between the output vector H(x) and each row of encoding matrix, the corresponding class label of the shortest coding is the output of the test sample.

Let us consider a binary ECOC encoding matrix $M \in (1,0)^{N*L}$, where N is class number of the dataset and L stands for the number of the single classifier. Each row in M corresponds to one class, while each column one single classifier [16]. In this paper, we adopt the list of coding method to establish the encoding matrix. In this way, the length of output code is $L = 2^{N-1} - 1$. As the dataset has 6 subcellular positions, L is 31($L = 2^{(6-1)}$).The encoding process is as follows, for the first row of the encoding matrix, all bits are set to 1, the ith row of the encoding matrix, 2^{m-i} 0s and 2^{m-i} 1s appear alternately. The encoding matrix is

{(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),
(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),
(0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1),
(0,0,0,0,1,1,1,1,0,0,0,0,1,1,1,1,0,0,0,0,1,1,1,1,0,0,0,0,1,1,1),
(0,0,1,1,0,0,1,1,0,0,1,1,0,0,1,1,0,0,1,1,0,0,1,1,0,0,1,1,0,0,1),
(0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0)}.

III. EXPERIMENTAL RESULTS

The final feature extraction method which we adopt is N Terminal Signal. We can calculate 6 diversity finite coefficients of each prediction of protein with 6 standard diversity sources. The 6 diversity finite coefficient can form a vector $Z(X) = (I(X,Y_1), I(X,Y_2), ...I(X,Y_6))$ which was put into the artificial neural network. In our study, we adopt 5-jackknife cross-validation to test the prediction quality.

Among the 504 sequences, 61 belonged to chromatin, 55 to nuclear lamina, 56 to nuclear speckle, 219 to nucleolus, 75 to nucleoplasm and38 to PML body. Obviously, the number of the fourth class (Nucleolus) is far greater than that of the other five classes. In the experiment, we found that if we statistics the fourth class samples completely to construct standard diversity source, it will cause classification results biases to the fourth class and the prediction quality of classifier is poor. Therefore, in order to solve this problem, we statistics a part of the fourth class samples to construct the fourth standard diversity source and we statistics the samples of other five classes completely to build other five standard diversity sources. After many experiments, we found that when we randomly selected the fourth class' sample number as 85 to build the fourth class' standard diversity source, the overall classification result is the best.

A. The Comparison with Different Feature Extraction Methods

In our experiment, the feature extraction methods which we adopt are Amino Acid Composition (AAC), Dipeptide Composition and N Terminal Signal respectively. We use diversity finite coefficient combined with artificial neural network to predict protein subcellular localization of the dataset SNL6. The results of classifier which is used diversity finite coefficient combined with artificial neural network with three kinds of feature extraction methods are shown in table I. The results of classifier which is just increment diversity with three kinds of feature extraction methods are shown in table II. By analying the prediction result, it shows that N terminal signal plays an important role in the subcellular localization of proteins. At the same time, the results in table I and table II show that the measure of diversity combined with artificial neural network is better than single method with appropriate feature extraction algorithm.

B. The Comparison With Different Methods

We compared our method with Lei-SVM [17], ESVM [18], Binary tree [16] and IDQD [15]. The comparison of the results is shown in table III. In addition, we compared our method with IDQD in sensitivity, specificity and Markov correlation coefficient. The comparison of the results is shown in table IV.

IV. CONCLUSION

In this paper, we used integrated classifiers to predict the subcellular localization. The overall accuracy rate achieved by this paper was 81.3%, which was better than that by Lei-SVM, ESVM, Binary tree and IDQD. The results indicated that our method was simple and fast. And it did well in subcellular localization results balance. In order to solve the problem of unbalanced data, we first used random sampling principle to build standard diversity sources. And we use the measure of diversity finite coefficient combined with artificial neural network to predict protein subcellular localization. These were innovation of this paper.

ACKNOWLEDGMENT

This research was partially supported by the Natural Science Foundation of China (61070130), the Key Project of Natural Science Foundation of Shandong Province (ZR2011FZ003), the Key Subject Research Foundation of Shandong Province and the Shandong Provincial Key Laboratory of Network Based Intelligent Computing.

REFERENCES

[1] Deng, M., Zhang, K., Mehta, S., Chen, T., Sun, F., "Prediction of Protein Function Using Protein–protein Interaction Data," Journal of computational biology,vol.10, pp.947–960,2003.

[2] C.H. Song, F. Shi, "Prediction of Protein Subcellular Localization Based on Hilbert-Huang Transform. Wuhan university journal of natural sciences," vol.17, pp.048-054,2012.

[3] Cedano J, Aloy P, P'erez-Pons JA, Querol E, "Relation between amino acid composition and cellular location of proteins," J mol biol, vol.266, pp.594–600,1997.

[4] K.C. Chou, "prediction of protein structural classes andsubcellular locations,"Current protein and peptide science," vol.1,pp.171–208, 2000.

[5] K.C. Chou, D.W. Elrod,"Protein subcellular location prediction," Protein engineering,vol.12,pp.107–118,1999.

[6] Nakashima H, Nishikawa K: Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J Mol Biol 238: 54–61(1994)

[7] K.C. Chou,"Prediction of protein cellular attributes using pseudo amino-acid-composition," Erratum: ibid. Proteins: Structure, Function, and Genetics. 2001,pp: 246–255.

[8] K.C. Chou, Y.D. Cai,"Using functional domain composition and support vector machines for prediction of protein subcellular location,"J Biol Chem,vol.277,pp.45765–45769,2002.

[9] K.C. Chou, Y.D. Cai,"A new hybrid approach to predict subcellular localization of proteins by incorporating Gene ontology," Biochem Biophys Res Commun,vol.311,pp.743–747,2003.

[10] K.C. Chou, Y.D. Cai,"Prediction and classification of protein subcellular location: sequence-order effect and pseudo amino acid composition," J Cell Biochem,vol.90,pp.1250–1260,2003.

[11] K.C. Chou, H.B. Shen,"Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization," Biochem Biophys Res Commun, vol.347,pp.150-157,2006.

[12] Z.H. Zhang, Z.H. Wang, Z.R. Zhang, Y.X. Wang ,"A novel methodfor apoptosis protein subcellular localization prediction combin-ing encoding based on grouped weight and support vector machine," FEBS Lett. 2006, pp.6169–6174.

[13] Y.L. Chen, Q.Z. Li,"Prediction of the subcellular location of apoptosis proteins," J Theor Biol,vol.245,pp.775–783,2007.

[14] Y.L. Chen, Q.Z. Li,"Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo-amino acid composition," J Theor Biol,vol.248,pp.377–381,2007.

[15] F.M. Li, Q.Z. Li,"Using pseudo amino acid composition to predict protein subcellular location with improved hybrid approach,"Amino Acid, vol.34,pp.119–125,2008.

[16] Lili GUO, Yuehui Chen,"Predicting protein subcellular localization by fusing binary tree and error-correcting output coding," ICIC2012,LNCS7389.2012,168-173.

[17] Z.D. Lei, Y. Dai,"An SVM-based system for predicting protein subcellular localizations," BMC Bioinf ormatics, vol.6,pp.291-298, 2005.

[18] W.L. Huang, C.W. Tung, H.L. Huang,"ProLoc: Prediction of protein subcellular localization using SVM with automatic selection from physicochemical composition features,"BioSystems,2007.doi: 10. 1016/j.biosystems.2007. 01.001.

TABLE I. THE COMPARISON OF THE RESULTS WITH MY PRIOR RESEARCH

Subset subcelllular location		Different Feature Extraction Algorithms		
		AAC(%)	DC(%)	NTS(%)
1	Chromatin	0/61=0	23/61=37.7	43/61=70.5
2	Nuclear-Lamina	1/55=0.02	18/55=32.7	46/55=83.6
3	Nuclear-speckles	6/56=0.11	30/56=53.6	47/56=83.9
4	Nucleolus	207/208=94.5	189/219=86.3	179/219=81.7
5	Nucleoplasm	12/75=16.0	35/75=46.7	63/75=84.0
6	PML body	0/38=0	25/38=65.8	32/38=84.2
Overall		44.8	63.5	81.3

TABLE II. THE COMPARISON OF THE RESULTS WITH INCREMENT DIVERSITY

Subset subcelllular location		Different Feature Extraction Algorithms		
		AAC(%)	DC(%)	NTS(%)
1	Chromatin	10/61=16.4	5/61=8.2	46/61=75.4
2	Nuclear-Lamina	19/55=34.5	37/55=67.3	42/55=76.4
3	Nuclear-speckles	16/56=28.6	15/56=26.8	40/56=71.4
4	Nucleolus	187/219=85.4	185/219=84.5	180/219=82.2
5	Nucleoplasm	25/75=33.3	43/75=57.3	70/75=93.3
6	PML body	3/38=7.9	0/38=0	18/38=47.4
Overall		51.6	56.5	78.6

TABLE III. THE COMPARISON OF THE RESULTS BETWEEN DIFFERENT METHODS FOR SNL6

Subset Subcelllular Location		Different Methods				
		Lei-SVM[17](%)	ESVM[18](%)	Binary tree+ANN[16](%)	IDQD[15](%)	This paper
1	Chromatin	13/61=21.3	13/61=21.3	37/61=60.7	37/61=60.6	43/61=70.5
2	Nuclear-Lamina	20/55=36.4	20/55=36.4	40/55=72.7	34/55=61.9	46/55=83.6
3	Nuclear-speckles	19/56=33.9	15/56=26.8	37/56=66.0	36/56=64.3	47/56=83.9
4	Nucleolus	182/219=83.1	198/219=90.3	147/219=67.1	205/219=93.6	179/219=81.7
5	Nucleoplasm	21/75=28.0	32/75=42.7	54/75=72.0	51/75=68.0	63/75=84.0
6	PML body	4/38=10.5	7/38=18.4	25/38=65.8	17/38=44.7	32/38=84.2
Overall		259/504=51.4	285/504=56.4	340/504=67.5	380/504=75.4	410/504=81.3

TABLE IV. THE COMPARISON OF THE RESULTS BETWEEN IDQD WITH OUR METHOD IN SENSITIVITY (SEN), SPECIFICITY(SPEC) AND MARKOV CORRELATION COEFFICIENT(MCC)

Different Methods		Subset Subcelllular Location					
		Chromatin	Nuclear-Lamina	Nuclear-speckles	Nucleolus	Nucleoplasm	PML body
IDQD	Sen	0.606	0.619	0.643	0.936	0.680	0.447
	Spec	0.698	0.694	0.720	0.807	0.708	0.654
	MCC	0.607	0.615	0.643	0.758	0.642	0.511
This paper	Sen	0.705	0.836	0.839	0.817	0.840	0.842
	Spec	0.860	0.821	0.922	0.825	0.759	0.681
	MCC	0.752	0.808	0.865	0.685	0.761	0.735