Predicting Protein Subcellular Localization Using the Algorithm of Increment Of Diversity Combined with Weighted K-Nearest Neighbor

Zeyue Wu [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]fly1024@126.com, [b]yhchen@ujn.edu.cn

Keywords: subcellular localization; feature extraction; increment of diversity； Weighted K-Nearest Neighbor

Abstract. Protein subcellular localization is an important research field of bioinformatics. In this paper, we use the algorithm of the increment of diversity combined with weighted K nearest neighbor to predict protein in SNL6 which has six subcelluar localizations and SNL9 which has nine subcelluar localizations. We use the increment of diversity to extract diversity finite coefficient as new features of proteins. And the basic classifier is weighted K-nearest neighbor. The prediction ability was evaluated by 5-jackknife cross-validation. Its predicted result is 83.3% for SNL6 and 87.6 % for SNL9. By comparing its results with other methods, it indicates the new approach is feasible and effective.

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 and so on. 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]. 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]. and pseudo amino acid composition (PseAA) [4]. With the introduction of functional domain composition [4], the researchers put gene annotation (GO) into protein subcellular localization prediction area [5]. Zhang [6] developed a new encoding method with grouped weighted for protein sequence. Chen and Li [7,8] had developed two prediction approaches based on increment of diversity (ID) and increment of diversity with support vector machine (ID_SVM). In this paper, a different approach is used for predicting protein subcellular location. We have developed two prediction approaches based on increment of diversity (ID) and weighted K-nearest neighbor (KNN).

Materials and methods

Dataset

Two datasets are adopted to validate the availability of our classifier. The one is SNL6. This dataset is founded by Lei and Dai. It is commonly used in subcellular localization. SNL6 contains 504 proteins and they are localized in 6 subcellular positions. Among the 504 sequences, 61 belong to chromatin, 55 to nuclear lamina, 56 to nuclear speckle, 219 to nucleolus, 75 to nucleoplasm, 38 to PML body. Another is SNL9 which have 370 proteins localized in 9 subcellular positions: 40 PML body, 15 nuclear speckles, 59 chromation, 65 nuclear diffuse, 115 nucleolus, 31 heterochromatin, 25 nuclear pore, 10 PcG body and 10 Cajal body. SNL9 is founded by Shen and Chou.

Representation of protein sequence

Published by Atlantis Press, Paris, France.

Given a protein sequence P with L amino acid resides, it can be formulated as [7,8]

$$P = R_1 R_2 R_3 ... R_L \tag{1}$$

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

In this paper, N-terminal signal is used to represent protein sequence. From the second amino acid of N-terminal of a protein, we retain 30 amino acids as this protein. Then we statistics the frequencies of 20 amino acids of each site [10,11]. It is a 600(20*30=600) dimensional vector and can be expressed in a formula as follows.

$$P = (p_1, p_2, ... p_{600}) \tag{2}$$

Ensemble classifier prediction system

1) Increment of diversity

FM Li et al. introduced the method which was called increment of diversity [12]. 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.

$$D(X) = D(n_1, n_2, ...n_s) = N \log_b N - \sum_{i=1}^{s} n_i \log_b n_i \tag{3}$$

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,

$$\Delta(X, Y) = D(X + Y) - D(X) - D(Y) \tag{4}$$

According to the above definition, the increment of diversity can be written as follows.

$$\Delta(X, Y) = D(M, N) - \sum_{i-1}^{s} D(m_i, n_i) \tag{5}$$

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.

$$I(X, Y) = \frac{\Delta(X, Y)}{D(M, N)} = 1 - \sum_{i=1}^{s} \frac{D(m_i, n_i)}{D(M, N)} \tag{6}$$

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

For example, the dataset has N subcellular positions. According to the different occurrence frequencies of N terminal signal in protein sequences, we can form N standard diversity sources. Then using the formula (6), we can obtain N diversity finite coefficients of each prediction of protein. And the standard source category of the smallest diversity finite coefficient can be taken as the prediction of protein's subcellular location category.

2) Weighted K-nearest neighbor (WKNN)

In this paper, the basic classifier is weighted K-nearest neighbor algorithm. The thought of the algorithm is to improve the traditional K-nearest neighbor algorithm [13]. Not only the category of nearest neighbor number but also the similarity between nearest and the unknown protein can

influence the result. Researchers choose the Euclidean distance to calculate the similarity of protein sequence.

The process of weighted K-nearest neighbor algorithm has three steps. First, use the traditional K-nearest neighbor algorithm to obtain K-nearest neighbors of the target protein P. Then, calculate the similarity between protein sequence P and K-nearest neighbors by using the formula (7) and (8). At last, we can get the classification of the target protein P. If the highest score is $rank(i)$, then i is decided as the category of the target protein P.

$$S = -\log(D_{pj}) \tag{7}$$

$$rank(i) = \sum_{1 \le j \le k} -\log(D_{Pj}) \tag{8}$$

Where, S is distance weight, $j = 1,2...K$, $i = 1,2...6$. When $D_{Pj} = 0$, then $\log(D_{Pj}) = -300$.

Experimental results

For SNL6, we can calculate 6 diversity finite coefficients of each prediction of protein. The 6 diversity finite coefficients can form a vector $Z(X) = (I(X,Y_1), I(X,Y_2),...I(X,Y_6))$ which was put into the weighted K nearest neighbor algorithm. At the same time, for SNL9, a vector which has 9 diversity finite coefficients of each prediction of protein can be formed. In our study, we adopt 5-jackknife cross-validation to test the prediction quality.

Among the 504 sequences of SNL6, the number of the fourth class (Nucleolus) is far greater than that of the other five classes. In the experiment, we found 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 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. For the same reason, among the 370 sequences of SNL9, we randomly selected the fifth class' sample number as 70 to build the fifth class' standard diversity source, the overall classification result is the best.

The comparison with increment of diversity

In our experiment, we use increment of diversity combined with weighted K-nearest neighbor algorithm to predict protein subcellular localization of the dataset SNL6 and SNL9. For SNL6, the results of classifier which are used increment of diversity combined with weighted K-nearest neighbor algorithm and increment of diversity compared with results which classifier is just increment diversity are shown in table 1 And results of the other dataset SNL9 are shown in table 2. By analying the prediction result, it shows that the measure of diversity combined with weighted K nearest neighbor algorithm is better than single method.

Table 1. The Comparison Of The Results With My Prior Research for SNL6

Subset subcelllular location		Different classifiers	
		increment of diversity(%)	This paper(%)
1	Chromatin	49/61=80.3	39/61=63.9
2	Nuclear-Lamina	43/55=78.2	45/55=81.8
3	Nuclear-speckles	44/56=78.6	44/56=78.6
4	Nucleolus	173/219=80.0	201/219=91.8
5	Nucleoplasm	67/75=89.3	58/75=77.3
6	PML body	19/38=50.0	33/38=86.8
Overall		395/504=78.4	420/504=83.3

Table 2. The Comparison Of The Results With My Prior Research for SNL9

Subset subcelllular location		Different classifiers	
		increment of diversity(%)	This paper(%)
1	PML body	35/40=87.5	35/40=87.5
2	nuclear speckles	14/15=93.3	13/15=86.7

Published by Atlantis Press, Paris, France.

3	chromation	50/59=84.7	45/59=76.3
4	nuclear diffuse	64/65=98.5	61/65=93.8
5	nucleolus	97/115=84.3	103/115=89.6
6	heterochromatin	24/31=77.4	25/31=80.6
7	nuclear pore	17/25=68.0	21/25=84.0
8	PcG body	7/10=70.0	10/10=100.0
9	Cajal body	10/10=100.0	10/10=100.0
Overall		318/370=85.9	324/370=87.6

The Comparison with other different methods

We compared our method for SNL6 with Lei-SVM [11], ESVM [12], Binary tree [10] and IDQD [9]. The comparison of the results is shown in table 3. We compared our method for SNL9 with OET-KNN [13], PSSM [14], AdaBoost [15]. The comparison of the results is shown in table 4.

Table 3. The Comparison Of The Results Between Different Methods For SNL6

Subset Subcelllular Location		Different Methods				
		Lei-SVM[11](%)	ESVM[12](%)	Binary tree+ANN[10](%)	IDQD[9](%)	This paper(%)
1	Chromatin	13/61=21.3	13/61=21.3	37/61=60.7	37/61=60.6	39/61=63.9
2	Nuclear-Lamina	20/55=36.4	20/55=36.4	40/55=72.7	34/55=61.9	45/55=81.8
3	Nuclear-speckles	19/56=33.9	15/56=26.8	37/56=66.0	36/56=64.3	44/56=78.6
4	Nucleolus	182/219=83.1	198/219=90.3	147/219=67.1	205/219=93.6	201/219=91.8
5	Nucleoplasm	21/75=28.0	32/75=42.7	54/75=72.0	51/75=68.0	58/75=77.3
6	PML body	4/38=10.5	7/38=18.4	25/38=65.8	17/38=44.7	33/38=86.8
Overall		259/504=51.4	285/504=56.4	340/504=67.5	380/504=75.4	420/504=83.3

Table 4. The Comparison Of The Results Between Different Methods For SNL9

Subset subcelllular location		Different Methods			
		OET-KNN[13](%)	PSSM[14](%)	AdaBoost[15](%)	This paper(%)
1	PML body	NA	24/40=60.0	38/40=95.0	35/40=87.5
2	nuclear speckles	NA	7/15=50.0	6/15=40.0	13/15=86.7
3	chromation	NA	40/59=66.7	51/59=86.4	45/59=76.3
4	nuclear diffuse	NA	42/65=65.0	50/65=76.9	61/65=93.8
5	nucleolus	NA	108/115=93.9	108/115=93.9	103/115=89.6
6	heterochromatin	NA	23/31=74.2	26/31=83.9	25/31=80.6
7	nuclear pore	NA	14/25=58.3	14/25=56.0	21/25=84.0
8	PcG body	NA	3/10=30.0	6/10=60.0	10/10=100.0
9	Cajal body	NA	2/10=20.0	9/10=90.0	10/10=100.0
Overall		64.32	263/370=71.2	308/370=83.2	324/370=87.6

Conclusion

In this paper, we used integrated classifiers to predict the subcellular localization. The overall accuracy rate of SNL6 achieved by this paper was 83.3%, which was better than that by Lei-SVM, ESVM, Binary tree and IDQD. The overall accuracy rate of SNL9 achieved by this paper was 87.6 %, which was better than that by OET-KNN, PSSM and AdaBoost. 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 increment of diversity combined with weighted K-nearest neighbor algorithm 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,2003, pp.947–960.

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

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

[4] K.C. Chou, Y.D. Cai,Using functional domain composition and support vector machines for predictionof protein subcellular location,J Biol Chem,2002,pp.45765–45769.

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

[6] 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.

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

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

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

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

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

[12] 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.

[13] Shen HB, Chou KC, Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition, Biochem Biophys Rec Commun, pp.752-756.

[14] Mundra P, Kumar KK, Jayaraman VK, Kulkami BD, Using psedo amino acid composition to predict protein subnuclear localization: approached with PSSM, Pattern Recognit Lett28, pp.1610-1615.

[15] Xiaoying Jiang, Rong Wei, Yanjun Zhao, Tongliang Zhang, Using Chou's pseudo amino acid composition based on approximate entropy and ensemble of AdaBoost classifiers to predict protein subnuclear location, Amino Acids, 2008, pp.669-675.