2012 International Conference on Bioinformatics and Computational Biology (ICBCB 2012)
IPCSIT vol.34 (2012) © (2012) IACSIT Press, Singapore

Predicting protein subcellular location using Error-Correcting Output Coding

Lili Guo [+], Yuehui Chen
Computational Intelligence Lab, School of Information Science and Engineering,
University of Jinan, 106 Jiwei Road, 250022 Jinan , P.R.China

ABSTRACT. The knowledge of subcellular localization in cells is important for the proteins function. Developing new methods of predicting the proteins subcellular localization become important research fields in protein science. In this paper, we have proposed a new approach to predict the proteins subcellular localization. The protein is represented by the fusion feature information of three feature extraction methods- the pseudo amino acid composition (PseAA), the physical and chemical composition (PCC), and the amino acids hydration properties composition (HpAA). A novel ensemble classifier is designed using the Error-Correcting Output Coding (ECOC) and seven Artificial Neural Networks (ANN) classifiers. The results of Jackknife cross-validation test are higher than them of some methods on same datasets, which indicate the new approach is feasible and effective.

Keywords: protein subcellular localization, feature extraction, ECOC, ANN, ensemble classifier

1. Introduction

The protein subcellular localization is bound up with the proteins function, and it also guarantees the well working of the complicated cell-system. It is very helpful to learn subcellular localization for understanding proteins' natures, functions, interaction and regulation mechanism between each other [1], which provide reference information for developing new drug .

Biological cell is highly ordered structure. We usually divide intracellular region into different organelles or cell-areas [2], as nucleus, Golgi body, endoplasmic reticulum, chondriosome, endochylema and cytomembrane etc. The proteins are transported to the specific place by sorting signals after proteins synthesis in ribosome, and they can participate in various life activities of cells only they are distributed to the correct positions [3]. The regional distribution in cells affect the process of protein folding, polymerization and modification after translation, also it has profound influence on cells' functions. Knowing the subcellular localization promotes the researching of protein biological functions and protein structures.

With the development of the research on genomics and proteomics, the number of biological data and sequences increase rapidly. It is out of date to study protein localization using experimental method alone which also can't meet the need of the research on life science [4]. In recent years, the protein subcellular localization methods based on machine learning gradually become a hotspot of bioinformatics.

2. Materials and Methods

2.1. Dataset

[+] Corresponding author. Tel.: + 86 13964055381
E-mail address: gfcguoguo@163.com.

One dataset SNL6 (Lei and Dai, 2005) is used to evaluate the proposed prediction system which is Error-Correcting Output Coding (ECOC) [5]. SNL6 has 504 proteins localized in 6 subcellular compartments. Table1 shows the numbers of protein sequences within each subcellular compartment in SNL6. The dataset is obtained from the Nuclear Protein Database (Dellaire et al., 2003), which is a searchable database of information on proteins consisting of more than 2000 vertebrate proteins (mainly from mouse and human) in cell nuclei. The SNL6 proteins are extracted by a Perl script which associate with more than one compartment are eliminated. SNL6 is a non-redundant dataset constructed from PROSET (Brendel, 1992) with low sequence identity (<50%).

Table 1. The numbers of protein sequences within each subcellular compartment in SNL6

Label	Compartment	Number of sequences
1	PML body	38
2	Chromatin	61
3	Nucleoplasm(nuclear diffuse)	75
4	Nucleolus	219
5	Nuclear splicing speckles	56
6	Nuclear lamina	55
Total		504

2.2. Feature extraction methods

2.2.1. Pseudo Amino Acid composition(PseAA)

According to the concept of Chou's PseAA (Chou 2001) composition [6], a sample of protein sequence is a point in (20+λ)-D space.

$$X = \left[x_1, x_2, \cdots, x_{20}, x_{21}, \cdots, x_{20+\lambda}\right]^T \in \Re^{(20+\lambda)}, \tag{1}$$

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

Where, the $f_i (1 \le i \le 20)$ in Ep.（2） is the occurrence frequencies of 20 amino acids in sequence. $P_i (21 \le i \le 20+\lambda)$ is the additional factors that incorporate some sort of sequence order information. The parameter w is weight factors.

2.2.2. Physical and chemical composition (PCC)

The 20 native amino acids are divided into three groups for their physicochemical properties, including seven types [7] of hydrophobicity, normalized van der Vaals volume, polarity, polarizibility, charge, secondary structures and solvent accessibility. For instance, using hydrophobicity attribute all amino acids are divided into three groups: polar, neutral and hydrophobic. A protein sequence is then transformed into a sequence of hydrophobicity attribute. Therefore, the composition descriptor consists of 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.

2.2.3 Amino acids hydration properties composition (HpAA)

In biology, the 20 native amino acids (20 letters in alphabet) are divided into 6 classes. Table2 shows the result of classification. So we make the six kinds of amino acids replace the previous 20 kinds. HpAA is defined as $x(i,j) = N_{i,j}/(N-1)$, where $i, j = 1, 2, \cdots, 6$, and $N_{i,j}$ is the number of dipeptides of amino acid type i and j [8]. So HpAA consists of 6×6=36 descriptor values.

Table 2. Amino acids hydration property classification

Classification	Abbreviation	Amnio 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

2.3. Ensemble classifier prediction system

2.3.1. ECOC framework

The main line of ECOC [9] is : it trains single classifier respectively according to encoding matrix. In testing process, every single classifier outputs a predicted value which forms a output vector $H(x) = (h_1(x), h_2(x), \cdots, h_n(x))$. 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 [10].

The encoding matrix is defined as $M_{K\times n}$ and each element of the matrix is {0, 1}. K refers to class number of the dataset and n stands for the number of the single classifier. Each row in M corresponds to one class , while each column one single classifier. For instance, the all single classifiers are $h_1(x), h_2(x), \cdots, h_n(x)$; if the M(i,j)=0, that means that classifier j regards all the samples of class label i as positive samples, or the samples are regarded as negative ones. The encoding matrix takes many forms [11], as one-to-many matrix, one-to-one matrix, random sparse coding matrix and dense random coding matrix etc.

2.3.2. Artificial neural network(ANN)

The single classifier is ANN. ANN has obtained very good application in many fields of pattern recognition and is such a algorithm which imitates the message processing of people's neurons [12]; it has strong robustness and tolerance and can study uncertain system. So ANN had been applied to the subcellular location very early. Particle swarm optimization (PSO) is adopted to optimize the parameters (weights and thresholds) of the ANN.

3. Experimental Results

The predictions are examined by 3-jackknife test on the 504 proteins classified into 6 subcellular location. The jackknife test is deemed the most objective and rigorous procedure for cross-validation and has been used by more and more investigators to examine the power of various prediction methods [13]. The testing result is in the Table3. In order to compare conveniently, the results of other algorithms which use same dataset are also listed in Table3. The accuracy of different features of our method is 60.395% and 62.886%, and compared with others, our result has certain enhancement, also shows the validity of the ECOC.

In statistic prediction study, it is convenient to introduce an accuracy matrix $[M_{i,j}]$ of size c×c (c is the number of compartments to be predicted). The element $M_{i,j}$ of accuracy matrix is the number of proteins predicted to be in subcellular compartment j, which are actually in the compartment i. Three indexes are applied to evaluate the prediction accuracy [14], i.e., sensitivity (Sn), specificity (Sp), and Matthew's correlation coefficients (CC).

$$S_n = \frac{M_{ii}}{\sum_{j=1}^{c} M_{ij}} \qquad (3)$$

$$S_p = \frac{M_{ii}}{\sum_{j=1}^{c} M_{ji}} \quad (4)$$

$$CC = \frac{M_{ii}(\sum_{k\neq i}\sum_{j\neq i}^{c} M_{jk}) - (\sum_{j\neq i}^{c} M_{ij}) \times (\sum_{j\neq i}^{c} M_{ji})}{[(M_{ii} + \sum_{j\neq i}^{c} M_{ij})(M_{ii} + \sum_{j\neq i}^{c} M_{ji})(\sum_{k\neq i}\sum_{j\neq i}^{c} M_{jk} + \sum_{j\neq i}^{c} M_{ji})(\sum_{k\neq i}\sum_{j\neq i}^{c} M_{jk} + \sum_{j\neq i}^{c} M_{ij})]^{1/2}} \quad (5)$$

$$A_c = (\sum_{i=1}^{c} M_{ii}) / (\sum_{i=1}^{c}\sum_{j=1}^{c} M_{ij}) \quad (6)$$

Table 3. Results of Jackknife test by different algorithms on SNL6

Compartments	Lei-SVM	ESVM	This paper					
			PseAA+PCC+ECOC			PseAA+PCC+HpAA+ECOC		
			Sn(%)	Sp(%)	CC	Sn(%)	Sp(%)	CC
Chromatin	21.3	21.3	85.0	89.6	0.861	70.0	64.2	0.683
Nuclear lamina	36.4	36.4	66.7	75.3	0.705	72.2	75.8	0.749
Nuclear speckles	33.9	26.8	83.3	72.6	0.887	88.9	84.3	0.806
Nucleolus	83.1	90.3	57.5	60.8	0.591	53.4	60.5	0.594
Nuclear diffuse	28.0	42.7	40.0	59.3	0.587	52.0	56.4	0.552
PML body	10.5	18.4	33.3	52.7	0.464	75.0	71.8	0.736
Ac(%)	51.4	56.4	60.241			62.651		

4. Conclusion

A novel approach for protein subcellular localization is proposed. Sample of protein sequence is represented by PseAA, PCC and HpAA. Ensemble classifier-ECOC is used as prediction engine. The ensemble classifier is combined with ECOC frame in which base classification algorithms are ANNs. This paper uses three methods to fuse features to validate the performance of the novel approach. Promising results obtained by jackknife cross-validation test indicate that the proposed approach is effective and practical, and might become a useful tool for prediction protein subcellular localization.

5. Acknowledgments

The authors wish to thank the Nuclear Protein Database for providing the datasets. This work was partially supported by the Natural Science Foundation of China (61070130, 60903176, 60873089), the Program for New Century Excellent Talents in university (NCET-10-0863), the Natural Science Foundation of Shandong Province, China (ZR2010FQ020), the Shandong Distinguished Middle-aged and Young Scientist Encourage and Reward Foundation, China (BS2009SW003), the China Postdoctoral Science Foundation (20100470081), and the Shandong Provincial Key Laboratory of Network Based Intelligent Computing.

6. References

[1]. A.I.Lamond, and W.C.Earnshaw. “Structure and function in the nucleus.Science” , 280, 547-553 (1998).

[2]. Zhang S, Huang B, Xia X, et al. “Bioinformatics research in subcellular localization of protein”. Prog Biochem Biophys, 34(6): 573~579 (2007).

[3]. R.D.Phair, and T.Misteli. “High mobility of proteins in the mammalian cell nucleus”. Nature, 404, 604-609 (2000).

[4]. K.C.Chou, and H.B.Shen. “ Recent progress in protein subcellular location prediction ”. Analytical Biochemistry, 370, 1-16 (2007).

[5]. DIETTERICH T G, BSNIRIG. “Solving multiclass learning problems via error-correcting output codes”. Journal of

Artificial Intelligence Research, 236-286 (1995).

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

[7]. 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).

[8]. Nair R, Rost B. “Inferring subcellular localization through automated lexical analysis”. Bioinformatics, 18 (Suppl): S78-S86 (2002).

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

[13].Breiman L. “Bagging predictors”. Machine Learning. 24 (2): 123-140 (1996).

[14].Reinhardt A, Hubard T. “Using neural networks for prediction of the subcellular location of proteins”. Nucleic Acids Res, 26(9): 2230-2236 (1998).