I.J. Engineering and Manufacturing, **2011, 6, 8-13**
Published Online December 2011 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijem.2011.06.02

Available online at http://www.mecs-press.net/ijem

Prediction of Protein Subcellular Localization Using EDA based Ensemble Classifiers

Ying Li

School of Information Technology, Shandong Women's University, Jinan, Shandong Province, China

Abstract

The function of protein is closely correlated with its subcellular locations. New composed proteins can perform normal biological function only after they are translocated to correct subcellular locations. In this paper, a new selective ensemble classifiers based on EDA algorithm has been proposed. In the method, pseudo amino acid composition was firstly applied to form the protein feature sets, then 10 neural networks is generated to learn the subsets which are re-sampling from feature subsets with PSO algorithm. At last, appropriate classifiers are selected to construct the prediction committee with EDA algorithm. Experiment shows that the proposed method produces the best prediction accuracy than the other methods on SNL6 database.

Index Terms: Protein subcellular location；Estimation of Distribution Algorithm (EDA)；selective ensemble；Pseudo amino acid composition

1. Introduction

With the rapid development of contemporary biology, a great number of original data of protein sequence have been gained, but it is a great challenge for all the researchers to determine the functions of protein according to protein sequences. Researchers show that the function of protein is closely correlated with its subcellular locations. New composed proteins can perform normal biological function only after they are translocated to correct subcellular location. So to determine the subcelluar location has a great significance for the studying of the function of proteins.

Traditional methods to determine protein subcellular location are biology experimental method. Generally there are three methods, namely the cell fractionation, the fluorescence microscopy and the electronic microscopy. But using these methods need highly cost and too much time. With the rapid expansion of today's database of protein sequences, to determine the proteins one by one is not practical. So, in recent years, to develop a mathematical method which predicting the protein subcellular location from the primary sequence of proteins directly become popular. Nakia uses the regulation of "if —then" to create an expert system which lays on mathematics for the prediction. Yuan uses second-order Markov model to predict the protein subcellular location. Reinhardt proposed a neural network method and Hua uses amino acid composition as input, using

* Corresponding author.
E-mail address: cherry_jn@126.com

support vector machine to predict the subcellular location. Neural network ensemble is a way of increasing the neural network's generalization ability. Neural network ensembles use a finite number of member networks to study the problem, and then make the final decision according to the output of the member networks in the ensemble.

The neural network ensemble is easy to be realized and can acquire stronger generalization ability than a single neural network, so it is widely used in increasing the generalization accuracy of neural networks. But in practice, we can hardly acquire the best performance if the member networks integrated indiscriminately. We can acquire better generalization ability only by selecting networks with good performance. Therefore the selective ensemble is proposed to enhance the performance of neural network. The method of selective ensemble has been widely used in the study of bioinformatics, and has achieved good effects.

In this paper, a new selective classifier ensemble based on EDA algorithm has been proposed. In this method, pseudo acid composition is used as network input to train every neural network. Then the EDA (Evaluation Distribution Algorithm) is used to evaluate the performance of training network and automatically select the best neural networks to take part in the ensemble. The testing of SNL6 database shows that the proposed method produces the best prediction accuracy.

2. Theoretical Methods

A. Pseudo Amino Acid Composition

We cannot predict the protein subcellular location directly from the protein sequence, so it is of great significance for people to develop an appropriate representation method for protein sequence in order to increase the prediction accuracy. The representation method is to transform the alphabetical representation of protein sequence with different length into a finite dimension of discrete vector with same length.

Nakashima and Nishikawa [7] find that the protein subcellular location is closely correlated with the amino acid composition of protein sequence and they are the first to use the method to represent protein based on amino acid composition. There are totally twenty kinds of amino acids which are made of protein, so using this method we can form a vector consisted of twenty kinds of amino acid frequencies. However, this method only considers the amino acid composition, neglects the mutual locations between the amino acids. So this method can not convey the important information of protein completely.

In order to overcome the shortcomings of the amino acid composition, Chou [8] proposed pseudo Amino Acid (PseAA) for protein in order to contain ranging order information of amino acids completely. In this method, the protein P is defined as follows (1):

$$P = [p_1, p_2, \cdots, p_{20}, p_{20+1}, \cdots, p_{20+\lambda}]^T \ (\lambda < L) \tag{1}$$

L is the length of protein sequence; λ is a variable parameter, ranging from 20 to 40. pi is calculated as follows (2):

$$p_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w\sum_{k=1}^{\lambda} \tau_k} & (1 \le u \le 20) \\ \dfrac{w\tau_{u-20}}{\sum_{i=1}^{20} f_i + w\sum_{k=1}^{\lambda} \tau_k} & (20+1 \le u \le 20+\lambda) \end{cases} \tag{2}$$

Where w is a weighting factor, generally set at 0.5, τk is k-tier correlation factor, expressing the sequence correlation among k amino acid residues, it is defined as follows (3):

$$\tau_k = \frac{1}{L-k}\sum_{i=1}^{L-k} J_{i,i+k}, (k < L) \tag{3}$$

Ji, j+k are coupling factors in the domain of amino acids from i to i+k. It is defined as follows (4):

$$J_{i,i+k} = \frac{1}{3}\sum_{m=1}^{3}(H_m(R_i) - H_m(R_{i+k}))^2 \tag{4}$$

Where H1(Ri), H2(Ri), H3(Ri) are the value of hydro phobic property, the value of hydrophilic qualitative and side chain atomic weight of amino acid residues. These three numerical values are all standardized. In the protein representation of PseAA, (P1,P2,...,P20)is the traditional amino acid composition ; (P20+1,P20+2,... ,P20+λ) are λ-tier correlation factors, namely the first tier, the second tier , until the λ-tier mode of sequence order correlation. PseAA adds the correlation factors related to sequence order; therefore it can be more scientific to represent protein than simply using the information of amino acid composition.

B. Integrated Neural Network Based on EDA

Selecting partial neural networks to take part in the ensemble can acquire better network generalization properties than using all the neural networks. So we must select the better networks subsets to integrate. There are many methods which can help us to complete this selecting, such as Greedy Hill Climbing. These methods test whether the performance of the ensemble may change through increasing or decreasing a neural network. Each time we should select the best neural networks which can make the subset of neural network acquire best performance. But this method is easy to fail into local optimum because we can not change the chosen network. In order to avoid the shortcoming, this paper adopted the theory of evolutionary computation method to select the best neural networks.

a). Evaluation Distribution Algorithm (EDA)

Evaluation Distribution Algorithm (EDA) is an evolutionary computation algorithm firstly proposed by Carranaga P and Loiaw J in 2002. This method emphasizes the mutual relations of the trajectory of individual genes and abandons the trivial crossover and mutation operation of genetic algorithm. This method selects the best N individuals from each generation to generate the probability distribution of best individual chromosome, then generating the individuals of next generation. Partial individual chromosome in each generation contains the information of the best individuals, so the best chromosomes of good individuals will keep by maximum probability in the individuals in next generation. Through iteration, the probability distribution of individual chromosome will be transformed into the probability distribution which can generate more excellent individuals in next generation. The algorithm of EDA is described as follows:

1) Randomly generate a set of λ individuals (t = 0).
2) Evaluate the λ individuals.
3) Select μ individuals (where μ ≤ λ) to be parents. Develop a probability distribution/density function Pt based on the parents
4) Create λ offspring using Pt
5) Evaluate the offspring
6) The λ offspring replace the μ parents (t = t +1)
7) If not achieve the goal, go to while

b). Selective Neural network Ensemble Algorithm Based on EDA

The pseudo amino acid composition which regarded as the features of protein sequence is extracted out to

form the training data. The sampling from the data set, k different data subsets can be used to train k neural networks with different nodes in hidden layers. Assuming that k neural networks are C1,C2,C3，…，Ck, respectively and S is a subset of set {C1，C2，C3，…，Ck}. We can use k (meaning length) and s (meaning binary vector) to represent the feature subset. If neural network Ci is selected, the assignment for bit i of the vectors is one. If not, the assignment for bit i is zero. The vector s is used to represent the individual chromosomes of EDA. It is defined as follows (5):

$$S = [c_1, c_2, \cdots, c_k]^T \quad c_i \in \{0,1\} \quad (1 \le i \le k) \tag{5}$$

We must choose suitable fitness function in order to measure the properties of each individual; we create a verification set v in the data set and calculate the training error Ev of each individual in verification data set v. Then 1/Ev is the fitness function. It is defined as follows (6):

$$E_{vi} = \sum_{j=1}^{K} p_{ij} \times classifier_j \tag{6}$$

Where Evi is the error of individual i of EDA population; K is the number of neural networks taking part in the ensemble. Pij is the value for bit j of individual i of the population. Classifier j is the training error in verification data set v of the neural network j.

3. Experiment

In this paper, SNL6 data set are used to test the performance of the proposed method. SNL6 data set was created by Lei and Dai in 2005. In this data set, there are five hundred and four protein sequences whose homology is less than 50 percent. This information belongs to six different protein subcellular location respectively, PML body, Chromatin, Nucleoplam, Nucleolus, Nuclear splicing speckles and nuclear lamina. All these sequences in SNL6 are from the world famous biology database —Nuclear Protein Database which contains over 2000 proteins known subcellular location. In order to avoid the data redundancy and increase the generality, the protein with more than one subcellular location and the two proteins whose similarities are over fifty percent are discarded. The details of the data set are as follows:

Table 1 The number of sequence of six kinds of subcellalar location in data set SNL6.

Subcellular Location	Number of Sequence
PML body	38
Chromatin	61
Nucleoplasm	75
Nucleolus	219
Nuclear splicing speckles	56
Nuclear lamina	55
Total number	504

Firstly we use PseAA to encode the protein sequences whose data are concentrated. In the dealing process, we assign λ as five, ten, fifteen, twenty, twenty-five and thirty respectively to extract the features of the sequences. The experiment shows that the classification accuracy is comparatively low when λ is less than 10 and the classification accuracy will gradually increase with the increasing of the value of λ when λ is over fifteen. This

shows that the mutual actions between the amino acids whose sequence distance is over twenty only have small effects to the protein subcellular locations. In this paper, we choose λ=20 to extract the features from the protein sequence, so we can acquire a forty-dimension feature vector from each protein sequence. Encoding all the five hundred and four protein in SNL6, we can acquire a feature data set containing five hundred and four feature vectors.

In order to increase the diversity of training samples, we use the method of Adaboost to take samples from the feature data sets to form ten different training subsets. Then we use them to train the neural networks which contain different numbers of neural nodes in hidden layer. The PSO (particle swarm optimization) algorithm is used to adjust the weights of the neural networks. Then we encode each neural network as a binary string to form an individual chromosome of EDA. At last, we can acquire best neural network ensemble classifiers using Evaluation Distribution Algorithm (EDA).

In the process of classification, the method of five-cross-validation is used to test the data sets. The five-cross-validation is described as follows: First, we divide the experimental data sets into five non-intersecting subsets. Then we choose four groups of them to form a training set and the other one is used as testing set. Five different experiments are performed in order to evaluate the classification performance. In this paper, for each group of data sets, we perform twenty experiments. The experimental results are as follows:

Table 2 The Prediction Accuracy In SNL6

Sub cellular Location	Lei-SVM	ESVM	Our Method
PML body	10.5	18.42	44.74
Chromatin	21.3	21.31	49.18
Nucleoplasm	28.0	42.67	66.67
Nucleolus	83.1	90.32	76.26
Nuclear Splicing Speckas	33.9	26.79	51.79
Nuclear Lamina	36.4	36.37	52.73
Total number	51.4	56.37	63.89

The above experimental results are described as follows: Using the method proposed by this paper, the protein subcellular location accuracy can reach 64.89 percent, 7.52 percent higher than that of ESVM and 12.49 percent higher than that of Lei-SVM [11]. More important, the accuracy of predicting different protein subcellular location is more stable because of the increasing of network generalization. Its accuracy is between 44.74 percent and 76.26 percent. Even for PML body which is the hardest to predict, its predicting accuracy can reach 44.17 percent, much higher than the other methods.

4. Conclusion

The prediction of protein subcellular location is a hotpot of today's study of bioinformatics and also one of the problems hard to be solved. In this paper, a novel integrated neural network classifier based on EDA is proposed. It firstly uses many single neural networks to learn the pseudo composition which represents the characteristic information of protein sequence, then to determine the best composition of neural network ensembles using EDA algorithm. At last, the network ensembles are used to predict the protein subcellular location. The experiment shows that this method has achieved a quite good performance, compared with the other method.

Reference

[1] Fujiwara Y, Asogawa M. Prediction of subcellular localization using amino acid composition and order[J]. Genome Informatics, 2001, 12 : 103-112

[2] Mullar R H, Karsten M, Sven G. Solid lip id nanoparticles for controlled drug delivery2a review of the state or the art[J]. Eur J Pharm B iopharm, 2000, 50(1) : 161-177.

[3] Nakaia K, Kanehisa M. A knowledge base for predicting protein localization sites in eukaryotic cells[J]. Genomics, 1992, 14 : 897-911.

[4] Yuan Z. Prediction of protein subcellar location using Markov chain models[J]. FEB S Letter, 1999, 451 : 23-26.

[5] Renhardt A, Hubbard T. Using neural networks for prediction of the subcellular location of proteins[J]. Nucleic Acids Research, 1998, 26, 2230-2236.

[6] Hua S, Sun Z. Support vector machine approach for protein subcellular localization prediction[J]. Bioinformatics, 2001, 17 : 721-728

[7] Nakashima H, Nishikawa K. Discrimination of intracellular and ex-tracellular proteins using amino acid composition and residue – pair frequencies[J]. J Mol Boil, 1994, 238 : 54-61.

[8] Chou K C. Prediction of protein cellular attributes using pseudo amino acid composition[J]. Protein Struct. Funct. Genet, 2001, 43 : 246-255.

[9] Chou Z H, Wu J, Tang W. Ensembling neural networks: many could be better than all. Artificial Intelligence, 2002, 137(1-2) : 239-263

[10] Larranaga P, Lozano J A. Estimation of distribution algorithms : A new tool for Evolutionary Computation[M]. Berlin : Kluwer Academic Publishers, 2002

[11] Lei Z, Dai Y. An SVM-based system for predicting protein subnuclear localization[J]. BMC Bioinformatics, 2005, 6 : 291-298.

[12] Huang W L, Tung C W, Huang H L, et al. ProLoc : prediction of protein subnuclear localization using SVM with automatic selection from physico-chemical composition features[J]. Biosystems, 2007, 90(2) : 573-581.