Engineering, 2013, 5, 409-412
http://dx.doi.org/10.4236/eng.2013.510B083 Published Online October 2013 (http://www.scirp.org/journal/eng)

Prediction Method of Protein Disulfide Bond Based on Pattern Selection*

Pengfei Sun, Yuanquan Cui#, Tiankai Chen, Ying Zhao
College of Computer Science and Technology Harbin Normal University, Harbin, China
Email: sunpengfei2000@yahoo.com.cn, 1512428681 @qq.com

Received 2013

ABSTRACT

The effect of the different training samples is different for the classifier when pattern recognition system is established. The training samples were selected randomly in the past protein disulfide bond prediction methods, therefore the prediction accuracy of protein contact was reduced. In order to improve the influence of training samples, a prediction method of protein disulfide bond on the basis of pattern selection and Radical Basis Function neural network has been brought forward in this paper. The attributes related with protein disulfide bond are extracted and coded in the method and pattern selection is used to select training samples from coded samples in order to improve the precision of protein disulfide bond prediction. 200 proteins with disulfide bond structure from the PDB database are encoded according to the encoding approach and are taken as models of training samples. Then samples are taken on the pattern selection based on the nearest neighbor algorithm and corresponding prediction models are set by using RBF neural network. The simulation experiment result indicates that this method of pattern selection can improve the prediction accuracy of protein disulfide bond.

Keywords: Protein Disulfide Bond; Neural Network; Nearest Neighbor Algorithm; Pattern Selection

1. Introduction

The protein disulfide bond is important component for many proteins; it can maintain the stability and function activity of proteins. The correct orientation of protein disulfide bond is very important to grasp the relationship of the protein structure and its biological function. Therefore, the solution to predict protein disulfide bond has great significance to predict the protein spatial structure and the protein function, but it is still hard to predict protein disulfide bond [1].

Some methods of predicting contact have been developed to solve the problem. Currently, there are many methods spreading internationally, such as artificial neural network, SVM, Genetic Programming, Hidden Markov Model and so on [2-4]. However, generally speaking, the predicting precision of these methods is not high enough. In order to enhance the predicting precision of the protein disulfide bond, a prediction method of protein disulfide bond on the basis of pattern selection and RBF neural network have been brought forward in this paper. In the method proteins with disulfide bond structure from the PDB database are encoded according to the encoding approach and are taken as models of training samples. Then samples are taken on the pattern selection based on the nearest neighbor algorithm and corresponding prediction models are set by using RBF neural network. As a result, the experiment indicates that the method could enhance the predicting accuracy of the disulfide bond effectively.

2. Definition of K-NN Algorithm

In pattern recognition, the k-nearest neighbor algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space. k-NN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification; the k-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms [5].

K-nearest neighbor algorithm(k-NN) is very simple, let $\chi = \{X_1, X_2, \cdots X_n\}$ be N, n-dimensional design samples, it is required to compute the k-nearest neighbors of a test sample X among $\chi = \{X_1, X_2, \cdots X_n\}$, as measured by an appropriate distance function d. The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples. In the classification phase, k

*The project was supported by Scientific Research Fund of Heilongjiang Provincial Education Department under Grant No.11551128.
#Corresponding author.

Copyright © 2013 SciRes.

is a user-defined constant, and an unlabelled vector (a query or test point) is classified by assigning the label which is most frequent among the k training samples nearest to that query point.

3. Selection of Protein Nature and Encoding Method

In the natural world protein has many and varied attributes. It is not realistic to make all the attributes determine the methods as the condition of forming the disulfide bond. In this research, based on the former results, several kinds of important attributes are selected to serve as the input items of the predicting models, and encode these data according to their characteristics.

Attribute 1: hydrophobic of amino acid

In the water medium, globular protein folding always favors in burying the hydrophobic amino acid to the inside of the protein. This phenomenon is called hydrophobic of amino acid. It holds the prominent status in stabling the protein three-dimensional structure. The former researching results indicate that there are differences between the disulfide bond and non disulfide bond in the existence ratio of hydrophobic amino acid. The results reveal that the hydrophobic of amino acid is very important to the forming process of the disulfide bond. Therefore, the property of amino acid can be used as input of prediction model. According to hydrophobic of amino acid, hydrophobic amino acid is coded as 1 and non hydrophobic amino acid is encoded as 0.

Attribute 2: protein secondary structure [6]

Protein secondary structure includes protein folding information which is significant to predict and reconstruct the protein 3D structure. Simultaneously, it is important to the prediction of the protein conjunction which is proceeding in this thesis. The protein secondary structure from the data in the research is chosen from secondary structure database DSSP. DSSP is a database of secondary structure assignments for all proteins in the Protein Data Bank. To any protein in the database of protein 3D structure PDB, the corresponding secondary structure can be derived by its three-dimensional structure. According to the protein secondary structure, the structure code of α is 00, the structure code of β is 01, and other structure code is 10.

Attribute 3: evolution information of protein [7]

In the research of predicting the protein secondary structure, it can be found that using the evolution information of protein will increase the predicting accuracy obviously. It reveals that the evolution information of protein contains the important information of protein structure formation. Thus this research introduces evolution information of protein to predict the protein disulfide bond. In this research, the protein evolution information gains from the HSSP database. The HSSP is a database of protein secondary structures derived by aligning to each protein of known structure all sequences deemed homologous. HSSP contains the sequence information which is based on the sequence's ratio between the protein and its homology in the protein database. According to the protein secondary structure, every amino acid is encoded as *Pi*, in every position, *Pi* is the probability of the presenting some kind of amino acid. The value scope of *i* is 1 to 20.

Because of the consideration of the interaction between the neighboring amino acids, the sliding windows with the length as 7 are chosen as a unit. According to the above encoding method, the predicting amino acid pair (i, j) is encoded and a group of 322-dimensional data are gotten as an input vector $T(i, j)$.

4. Sample Selection Method Based on the K-Nearest Neighbor Algorithm

Artificial neural network (BP), a theoretical classification model, is raised from the simulation of the brain information processing and the learning procedures. It puts forward on the basis of the human science information processing research, contains modern Neurobiology and cognitive science, and has very strong adaptivity and self-learning ability, and the nonlinear mapping ability, robustness and fault-tolerant capability, etc. In recent years, with the development of artificial neural network, it uses in every field of bioinformatics successfully, and the technique of artificial neural network is increasingly becoming an important tool for solving the problems on sequence analysis and pattern recognition of machine learning technology. For neural network, it depends on the quality of the training samples, it may not contain enough information when the performance of the training samples is too small; and if the training samples are too large, it may be too large and make the sample redundancy, increase the training time and is likely to cause overfitting. So the selection process of the training samples has an important meaning on prediction modeling, and how to choose the training sample will be the key to improve the performance of classification. Through the analysis of the working principle of neural network, it is known that neural network is one of the optimization of the nearest neighbor classifier essentially, only its template stored in the network structure by form of weight values, through repeated adjusting relevance weights to the purpose of fitting with the template. And K-Nearest Neighbor Algorithm make representative samples as a template directly, determine the category of the sample according to the distance to the template. So there is no need to iterative process of iterative adjustment. Both compare, neural network training is complex but classification accuracy is higher, and K-Nearest Neighbor Algorithm is low precision but simple and quick. Therefore

Copyright © 2013 SciRes.

we can use the advantage of K-Nearest Neighbor Algorithm which is simply and quickly to structure classifier to choose the sample, and then use these subset samples of this operation as the training sample to build a neural network classifier. So it still maintained the high precision characteristics of neural network classifier.

In the K-Nearest Neighbor Algorithm, the boundary samples play an important role for classification, and all kinds of samples that near the centre are less effective, choosing sample with K-Nearest Neighbor Algorithm is that, trying to delete all kinds of near centre sample and keep boundary samples, only to reduce sample size and improve the accuracy of the classifier. However, in sample selection process, the sequence of subset used for sample selection is a certain order, so that the samples in front can be reserved probability, which will cause boundary samples behind deleted and the class center samples in front of the sequence retained, so the representative of selected samples is unsatisfactory and redundancy, and cannot reach the quality for corresponding sample selection purpose. In order to solve the problem, this study uses an improvement K-Nearest Neighbor Algorithm to improve the quality of selected samples. The basic working principle of this algorithm is generating a sample subset *D* on the basis of the original samples set *T*, so as to make sure the set *T* can still be correctly classified with the condition of decreasing of samples in *D*. When a sample in *T* could not be classified correctly, it will be added to the sample set *D*, until a certain cycle ended and the samples subset *D* does not change. This algorithm through the cycle repeatedly perform the most neighbor algorithm, in order to improve the representation of choosing the sample can reduce redundant sample, to ensure that all the boundary samples in the samples set *T* were chosen to put in selected samples set *D*. The text below is the algorithm description [8]:

Algorithm input:

(1)The initial sample set *T* of the training samples, selected sample set $D = \varnothing$.

(2) Repeated times *n* of the sample choosing procedure.

Algorithm output:

The samples set *D* contains selected samples

Algorithm process:

(1) Choose any one x_1 from the samples set *T*. Store it into the set *D*: $D = x_1$, $T = T - x_1$;

(2) For all samples in the set, execute the following operation: choose any one *x* from *T*, execute nearest neighbor search operation on *x* in the subset *D*, find the sample *s* which is nearest from *x*, $\text{Distance}(x, s) = \min_{s_i \in D} \text{Distance}(x, s_i)$ judge the classes of sample, if $Class(x) \neq class(s)$, then $D = D \cup x$, $T = T - x$;

(3) Repeated execute (2) operation *n* times;

Algorithm end

5. Establishment of the Predicting Model

The artificial neural networks technology, as a kind of important tool of machinery learning technologies, is becoming more and more significant in bioinformatics to solve the problem of sequential analysis and pattern recognition. However, in the training process, the revision of the all network weights and threshold is needed. Therefore, the speed of studying is quite slow. RBF neural network structure is a kind of network based on partial approaches. To each training samples, it only needs to revise few weights and threshold, thus the speed of studying is faster. Consequently, the using of the RBF neural network or not is the classifier to determine whether the contact structure form.

RBF neural network [9] is a kind of back-propagation network, and has two network levels: the hidden layer is radial basis function layer, the output is linear layer. As **Figure 1** shows the network has *Q* group of input vector, the element in every group is *R*, there are S^1 RBF neuron in the intermediate level, S^2 linear neuron in the output level. The output network is:

$$n^1 = \| IW - P \| \bullet * b^1 \tag{1}$$

$$a^1 = radbas(n^1) \tag{2}$$

$$a^2 = purelin(LW^2 a^1 + b^2) \tag{3}$$

Radbas( ) is the radial basis function, generally is the Gauss function, and purelin( ) is the linear output function. The radial basis function network simulates the adjustment of middle part of human brain and the neural network structures which cover the receive territory mutually. Therefore, it is a kind of network based on partial approaches, besides that, its hidden strata node has the mutually independent center and the width, thus has higher classified precision. The simulation experiment indicates that comparing with BP neural network; RBF neural network has the characteristics of faster speed of network training and higher classified precision.

6. Result and Analysis of the Experimental

The experiment protein data used in the experiment is got from PDB, which is a protein structure database; extracting the sequence of amino acids and atomic coordinate information of protein for the as input information of the

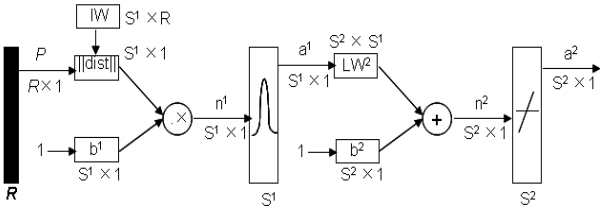

Figure 1. RBF Network Structure.

Copyright © 2013 SciRes.

prediction model. In original PDB data rate file, it usually uses SSBOND records to show that disulfide bonds form information in peptide chain. It has two kinds of bonds, intra-chain and inter-chain disulfide bonds, as a result the data only account intra-chain disulfide bonds. Process the database file from the PDB, remove the protein which contains nonstandard amino acid residues, sequence information deletion, or inaccurate SSBOND sequence information. At last randomly selected 200 proteins as the training sample, and select 50 proteins as the test sample.

The study uses the formula below as the prediction accuracy evaluation criterion to compare the efficiency of the prediction algorithm:

$$A = N_{cp}^{*} / N_{cp} \quad (4)$$

N_{cp}^{*} stands for the correct count of the prediction of disulfide bond structure, N_{cp} is the amount of all. The result is prediction accuracy.

The study calculated accuracy of the prediction of the disulfide bond structure raised in this study, according to this evaluation criterion. And it was compared with other algorithms without K-NN; through the prediction result the prediction algorithm of disulfide bond based on sample selection technology which put forward in the research can improve the precision of prediction (**Table 1**).

7. Conclusion

The study presents the protein disulfide bond structure prediction algorithm based on sample selection technology; it improved the selection of the training samples and classifier performance. Experiment results show that the algorithm can improve the efficiency of the prediction accuracy of disulfide bond structure. Test results show that there are some wrong forecasts, in which non-disulfide bond structure is predicted to be disulfide bond structure. So it affects the prediction accuracy. In the future we will study how to use the structure information of proteins to reduce the error prediction rate of non-disulfide bond structure effectively, so as to increase precision of the disulfide bond structure prediction.

Table 1. Comparison of prediction accuracy.

Method	RBF without K-NN	RBF with K-NN
Accuracy	82.6%	83.5%

REFERENCES

[1] S. M. Muskal, R. S. Holbrook and S. H. Kim, "Prediction of the Disulfide-Bonding State of Cysteine in Proteins," *Protein Engineering*, Vol. 3, 1990, pp. 667-672. http://dx.doi.org/10.1093/protein/3.8.667

[2] P. Fariselli, P. Riccobelli and R. Casadio, "Role of evolutionary Information in Predicting the Disulfide-Bonding State of Cysteine in Proteins," *Proteins*, Vol. 36, 1999, pp. 340-346. http://dx.doi.org/10.1002/(SICI)1097-0134(19990815)36:3<340::AID-PROT8>3.0.CO;2-D

[3] M. H. Mucchielli-Giorgi, S. Hazout and P. Tuffery, "Predicting the Disulfide Bonding State of Cysteines Using Protein Descriptors," *Proteins*, Vol. 46, 2002, pp. 243-249. http://dx.doi.org/10.1002/prot.10047

[4] X.-H. Shi and Y. Wang, "Prediction of Disulfide Bonding in Protein Structure Based on Method of Graph Matching," *Computer Engineering and Applications*, Vol. 43, 2007, pp. 30-32.

[5] P. E. Hart, "The Condensed Nearest Neighbour Rule," *IEEE Transactions on Information Theory*, Vol. 14, 1968, pp. 515-516. http://dx.doi.org/10.1109/TIT.1968.1054155

[6] J.-M. Chandonia and M. Karplus, "New Methods for Accurate Prediction of Protein Secondary Structure," *Proteins*, Vol. 35, 1999, pp. 293-306.

[7] C. Dodge, R. Schneider and C. Sander, "The HSSP Database of Protein Structure-Sequence Alignments and Family Profiles," *Nucleic Acids Research*, Vol. 26, 1998, pp. 313-315.

[8] H.-W. Hao and R.-R. Jiang, "Training Sample Selection Method for Neural Networks Based on Nearest Neighbor Rule," *Acta Automatica Sinica*, Vol. 33, 2007, pp. 1247-1251.

[9] W. Li and Y. Hori, "An Algorithm for Extracting Fuzzy Rules Based on RBF Neural Network," *IEEE Transactions on Industrial Electronics*, Vol. 53, No. 4, 2006.

Copyright © 2013 SciRes.