Predicting Protein Interaction Sites Based on a New Integrated Radial Basis Functional Neural Network

Xiaoli Shen\textsuperscript{a} and Yuehui Chen\textsuperscript{b}

School of Information Science and Engineering, University of Jinan, Jiwei road 106, Jinan 250022, P.R.China

\textsuperscript{a}xiaoli\_shen@sina.com, \textsuperscript{b}yhchen@ujn.edu.cn

Keywords: protein interaction sites, features, radial basis functional neural network, integrated, genetic algorithm based selected ensemble.

Abstract. Interactions among proteins are the basis of various life events. So, it is important to recognize and research protein interaction sites. A control set that contains 149 protein molecules were used here. Then 10 features were extracted and 4 sample sets that contained 9 sliding windows were made according to features. These 4 sample sets were calculated by Radial Basis Functional neural networks which were optimized by Particle Swarm Optimization respectively. Then 4 groups of results were obtained. Finally, these 4 groups of results were integrated by Genetic Algorithm based Selected Ensemble (GASEN) and better accuracy was got. So, the integrated method was proved to be effective.

Introduction

Proteins are polymers that are made up of amino acids. They all have their unique three-dimensional structures and implement specific functions respectively. But they don’t function alone. They complete a particular function via their interactions. Protein interactions control the various processes of life [1].

If we want to understand the principle of protein interactions, we must make clear that which parts of a protein participate in protein interactions firstly. So, this leads to the concept of interaction sites. An interaction site is an amino acid residue in a chain. If an amino acid residue is involved in an interaction, then it is defined as an interaction site. Otherwise, this amino acid residue is defined as a non-interaction site[2]. It is a two-type classification problem to predict protein interaction sites based on bioinformatics. Many authors had done a lot of work: Mile Sikic etal[3] listed 17 different features and used random forest to predict. Man Lan etal[4] adopted SVM to predict, but they mainly focused on feature generation and representations.

In this paper, we extracted 10 important features and used Radial Basis Functional (RBF)[5] neural network as the classifier. Then, Genetic Algorithm based Selected Ensemble (GASEN)[6] was used to integrate the results generated by the single classifier and better accuracy was got.

Materials and Methods

Data Set. We selected a non-redundant control data set that contains 149 protein molecules. It includes 92 hetero-complexes and 57 homo-complexes. The data set can be available on the SPIDER web site (http://spider.cchmc.org)[7].

Features. Feature is the first key to predict successfully. We extracted 10 different features:

1) sequence profiles (SP)[8,9]: it represents the relative frequency of an amino acid type at each position. It can be generated by multiple sequence alignment.
2) entropy (E)[9,10]: it is the measurement of sequence variability at one position. Here, it expressed the order among elements (amino acid residues).
3) relative entropy (RE)[9]: it is the normalization of entropy. It is changed between 0 and 100.
4) conservation weight (CW)[9]: it is the measurement of sequence conservation at one position. It is changed between 0 and 1.
5) complex accessible surface area (CASA)[2,9]: it expresses the total solvent exposure in a bound complex.

6) sequence variability (SV)[9]: it is on a scale of 0-100 and can be derived from the NALIGN alignments.

7) back-bone ASA (b-ASA)[3,11]: it was calculated by PSAIA.

8) side-chain ASA (s-ASA)[3,11];

9) polar ASA (p-ASA)[3,11];

10) non-polar ASA (n-ASA)[3,11];

**Definition of Protein Interaction Sites.** Usually, there are two methods can be used to define an interaction site[8]. We chose the first. Before protein complex forming, a residue exists in a monomer. So we call ASA of it as MASA. After complex formed, we call ASA of it as CASA. Then a residue in the complex was defined as a surface residue, if MASA / total ASA of a free amino acid ≥ 20%[12]. Total ASA of 20 amino acids were calculated by Huanxiang Zhou[13]. Finally, a residue was defined as an interaction site in the surface residues, if MASA-CASA ≥1(Å²)[2]. The others were non-interaction sites.

**Creation of Sample Sets.** We made 4 sample sets according to the above 10 features:

1) : SP;

2) : 1) + E + RE + CW;

3) : 2) + CASA + SV;

4) : 3) + b-ASA + s-ASA + p-ASA + n-ASA;

Here, the 4 sets were all made into 9 sliding windows, and they contained 20 * 9, 23 * 9, 25 * 9 and 29 * 9 values in each residue respectively.

**RBF Neural Network and Integration.** Classifier is the second key of this subject. We adopted RBF neural network as the classifier. It only contains 3 layers. The RBF that we adopted was:

\[
H_i(x) = \exp\left(-\frac{1}{2} \| x - c_i \|^2 / 2\sigma_i^2 \right), \quad i = 1, 2, ..., I. \tag{1}
\]

\(H_i(x)\) are the results of the hidden layer. \(x\) are the input vectors. \(c_i\) are the centers of the function. \(\sigma_i\) are the widths around one center. \(I\) represent the number of nodes in the layer. The results of the third layer can be calculated by:

\[
f_j = \sum_{i=1}^{I} \omega_{ij} H_i(x), \quad j = 1, 2, ..., J. \tag{2}
\]

\(f_j\) are the final results of the entire neural network. \(\omega_{ij}\) are the weights which contact the second layer and the third layer. \(J\) represent the number of nodes in this layer.

We used PSO[14] to optimize the parameters. In our RBF neural networks, the center and width of the function needed to be optimized. The weights that linked the second layer and the third layer needed to be optimized too. So, these parameters should be included in each particle of PSO.

Finally, GASEN was used to integrate the results of RBF neural networks. It is the extension of generalized ensemble method (GEM). GEM is calculated by the following formula:

\[
f_{GEM} = \sum_{i=1}^{n} \omega_i f_j(x). \tag{3}
\]
\( \omega_i \) is the weight and it is changed between 0 and 1. The sum of \( n \omega_i \) is 1. In GASEN, the weights were optimized by Genetic Algorithm. It simultaneously optimized 4 weights each cycle.

**Experiments and Results**

**Evaluation of The Results.** First, we defined the followings:
- TP: the number of interaction sites that were predicted correctly;
- TN: the number of non-interaction sites that were predicted correctly;
- FP: the number of non-interaction sites that were predicted as interaction sites;
- FN: the number of interaction sites that were predicted as non-interaction sites;
- N: the number of all sites in a protein molecule.

The following metrics were used to evaluate the prediction results[2, 7]:

1. Sensitivity of the positive data:
   
   \[
   \text{sen}^+ = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{4}
   \]

2. False alarm rate:
   
   \[
   \text{FAR} = \frac{\text{FP}}{\text{FP} + \text{TN}}. \tag{5}
   \]

3. Accuracy of prediction:
   
   \[
   \text{acc} = \frac{\text{TP} + \text{TN}}{N}. \tag{6}
   \]

4. Matthews Correlation Coefficient:
   
   \[
   \text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FN} \cdot \text{FP}}{\sqrt{\text{(TP+FN)(TN+FP)(TP+FP)(TN+FN)}}}. \tag{7}
   \]

**Experimental Results.** We adopted 10 cross-validation. The 149 protein molecules were divided into 10 groups (the last group contained 14 molecules). Each time, one group was selected as test set and the remaining groups were train set. Thus every sample set was carried out 10 times.(see Fig. 1)

![Fig. 1. The accuracy of RBF neural networks and their integration](image)

Every time, PSO and GASEN iterated 1000 times to generate the final result. The final experimental results were the average of 10 times. (see Table 1)
Table 1. The final experimental results of RBF neural networks and their integration
methods sen [%] FAR[%] acc[%] MCC[%]
Sample Set 1 93.3791 53.9768 65.7926 47.5746
Sample Set 2 94.1524 42.5145 67.7865 58.7108
Sample Set 3 94.3806 14.8154 75.2054 82.4262
Sample Set 4 95.2763 18.5355 77.3842 79.1662
GASEN 96.3019 18.466 80.3705 80.7351

On the SPPIDER web site, their final accuracy were 72.48% and 74.18%. Our posterior 4 results were all better than their’s.

In the end, we used 1qz8 to validate our method. It is a fragment of SARS corona virus NSP9. There are two chains (A, B) in this molecule and we used the first. The chain contains 111 amino acid residues and we predicted 96 ones by GASEN correctly. (see Fig. 2-3)

Fig.2. Visualization of interaction sites (blue) and non-interaction sites (red) of 1qz8 that we defined in ”Definition of Protein Interaction Sites”.

Fig.3. Visualization of the result of 1qz8 in GASEN. Blue showed the interaction sites that were predicted correctly. Red showed the non-interaction sites that were predicted correctly. Yellow (15 ones) showed the residues that were not predicted correctly.

Conclusion

In this paper, we extracted 10 features and some were new. The 4 sample sets that we created were also new and unique. Finally, we demonstrated our new ideas. We hope that more and more different and new methods about computational intelligence and biology can be applied to this subject in the future.
Acknowledgement

This research was supported by the Natural Science Foundation of China (NSFC) (60573065) and the Key Subject Research Foundation of Shandong Province (Y2007G33).

References

[1] W. Xiangyu, C. Shouliang and G. Mingde: General Biology (Higher Education Press, Beijing 2005).
[2] Y. Changhui, H. Vasant and D. Drena in: Predicting Protein-Protein Interaction Sites From Amino Acid Sequence. Technical report, Iowa State University (2002).
[3] S. Mile, T. Sanja and V. Kristian: Prediction of Protein-Protein Interaction Sites in Sequences and 3D Structures by Random Forests. PloS. Comput. Biol. Forum Vol. 5 (2009), p. 1-9
[4] L. Man, L. T. Chew and S. Jian: Feature generation and representations for protein-protein interaction classification. J. Biomed. Inform. Forum Vol. 42 (2009), p. 866-872
[5] P. B. John and R. B. Lauren: Sensitivity of RBF interpolation on an otherwise uniform grid with a point omitted or slightly shifted. Appl. Numer. Math. Forum Vol. 60 (2010), p. 659-672
[6] Z. Zhihua, W. Jianxin and T. Wei: Ensembling neural networks: Many could be better than all. Artif. Intell. Forum Vol. 137 (2002), p. 239-263
[7] E. Iakes, B. Lisa, F. Piero, C. Rita, V. Alfonso and L. T. Michael: Progress and challenges in predicting protein-protein interaction sites. Brief. Bioinform. Forum Vol. 10 (2009), p. 233-246
[8] L. Yang, T. Zhengquan and W. Yifei in: SVM-based protein interaction sites prediction. Technical report, Shanghai University (2006).
[9] Information on ftp://ftp.ebi.ac.uk/pub/databases/hssp/
[10] L. Chun and Q. Wei: Mathematical description of biological macromolecules and its application. (Dalian University of Technology Press, Liaoning 2009).
[11] J. Mihel, M. Sikic, S. Tomic, B. Jeren and K. Vlahovicke: PSAIA-Protein Structure and Interaction Analyzer. BMC. Struct. Biol. Forum Vol. 8 (2008), p. 21
[12] M. Wei, W. Feifei and P. Xinjun: Prediction of Protein-Protein Interaction Sites Using Support Vector Machine. J. App. Scie. Forum Vol. 26 (2008), p. 403-408
[13] Z. Huanxiang and S. Yibing: Prediction of Protein Interaction Sites From Sequence Profile and Residue Neighbor List. Proteins. Forum Vol. 44 (2001), p. 336-343
[14] E. Mohammed and S. K. Mohamed: PSO Bounds: A New Hybridization Technique of PSO and EDAs. Stud. Comp. Intell. Forum Vol. 203 (2009), p. 509-526