International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:4, No:12, 2010

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

Xiaoli Shen, and Yuehui Chen,

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 neutral 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 decision fusion (DF) and Genetic Algorithm based Selected Ensemble (GASEN). A better accuracy was got by DF and GASEN. So, the integrated methods were proved to be effective.

Keywords—protein interaction sites, features, sliding windows, radial basis functional neutral networks, genetic algorithm based selected ensemble.

I. Introduction

PROTEINS are polymers that are made up of amino acids, and they are the critical macromolecules in vivo. There are tens of thousands of different proteins in each organism. They all have their unique three-dimensional structures and implement specific functions respectively [1]. But they don't function alone. They complete a particular function via their interactions. Protein interactions control the various processes of life. For example, metabolism and signal transduction, DNA synthesis, gene transcription, protein translation, modification and positioning, cell cycle regulation and so important biological processes all need the interactions of proteins [2]. We can say that structure of cells, tissues and bodies are all related to proteins, and they participate in each activity in vivo.

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 an non-interaction site [3].

Here, we focused on predicting whether an amino acid residue is an interaction site. With the development of bioinformatics, there are many effective methods were introduced to this research. It is a two-type classification problem to predict protein interaction sites based on bioinformatics. If an amino acid residue was an interaction site, it was labeled as '1'. Otherwise it was labeled as '0' [3]. The main steps of this work are as follows:

Xiaoli Shen is with the School of Information Science and Engineering University of Jinan Jinan 250022, P.R.China e-mail: (xiaoli_shen@sina.com).
Yuehui Chen is with the School of Information Science and Engineering University of Jinan Jinan 250022, P.R.China e-mail: (yhchen@ujn.edu.cn).
Manuscript received April 19, 2005; revised January 11, 2007.

1) selected an appropriate data set;
2) extracted important features;
3) defined interaction sites;
4) generated sample sets;
5) used predictors to predict interaction sites;
6) determined the evaluation methods;
7) evaluated the results via evaluation methods;

In the above steps, 2 and 5 are critical to this work. Many authors focused on the two points: Sébastien Fiorucci etal [4] used electrostatic desolvation profiles as feature. Consuelo Latorre Fortes-Dias etal [5] used peptide arrays. Mile Sikic etal [6] listed 17 different features and used random forest to predict. Man Lan etal [7] 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) [8] [9] neutral networks as the classifiers. We also adopted Bagging and Adaboost [10] [11] to improve the effect. Finally, decision fusion (DF) [12] and Genetic Algorithm based Selected Ensemble (GASEN) [13] were used to integrate the results generated by the single classifier. A better accuracy was got.

II. Materials and Methods

A. Data Set

There are many different data sets can be used to this study. In our work, we selected a non-redundant control data set that contains 149 protein molecules (S149), because it is available and appropriate. It includes 92 hetero-complexes and 57 homo-complexes. The data set can be available on the SPPIDER web site (http://spider.cchmc.org) [14].

B. Features

Feature is the first key to predict successfully, because some important features can improve the accuracy. In this paper, we extracted 10 different features:

1) sequence profiles (SP) [15] : it represents the relative frequency of an amino acid type at each position. It can be generated by multiple sequence alignment.

2) entropy (E) [16] : it is the measurement of sequence variability at one position. Here, it expressed the order among elements (amino acid residues).

3) relative entropy (RE) : it is the normalization of entropy. It is changed between 0 and 100.

4) conservation weight (CW) : it is the measurement of sequence conservation at one position. It is changed between 0 and 1.

5) complex accessible surface area (CASA) [3] : it expresses the total solvent exposure in a bound complex. It can be

calculated by SURFACE, AREAIMOL [17] or PSAIA [6] [18].

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

7) back-bone ASA (b-ASA) : it was calculated by PSAIA.

8) side-chain ASA (s-ASA) : it was calculated by PSAIA too.

9) polar ASA (p-ASA) : it was calculated by PSAIA too.

10) non-polar ASA (n-ASA) : it was calculated by PSAIA too.

The preceding 6 features can be downloaded from HSSP (Homology derived Secondary Structure of Proteins) (ftp://ftp.ebi.ac.uk/pub/databases/hssp/).

C. Definition of Protein Interaction Sites

The reason that we defined the protein interaction sites was to create sample sets. Usually, there are two methods can be used to define an interaction site [19] [20] :

1) based on reduction of residue solvent ASA before protein complex forming and after formed.

2) based on the distance between α carbon atoms of residues.

Here, we chose the first one. Before protein complex forming, a residue exists in a monomer (a chain). So we call ASA of it as MASA (monomer ASA). After complex (contain one or several chains) formed, we call ASA of it as CASA (complex ASA). The software of PSAIA can be used to calculate MASA and CASA. Then a residue in the complex was defined as a surface residue, if MASA / total ASA of a free amino acid $\geq$ 20% [21]. The value of cut-off can also be others (5%, 10%, 16%) [14]. Total ASA of 20 amino acids were calculated by Huanxiang Zhou [22]. Finally, a residue was defined as an interaction site in the surface residues, if MASA - CASA $\geq$ 1($\mathring{A}^2$) [3]. The others were non-interaction sites.

D. Creation of Sample Sets

The sample sets were made up of features. They were the input of classifiers. 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;

Every position of SP contains 20 values. The other features all include one value. Mile Sikic etal demonstrated that accuracy was the best when joined a residue and its 8 neighbor residues (9 sliding windows). In our previous experiments, we also demonstrated this. So, 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.

E. RBF Neural Network and Integration

Classifier is the second key of this subject, because a good classifier can improve the accuracy too. In this paper, we adopted RBF neural network as the classifier. It has many advantages. First, its structure is simple and it only contains 3 layers. Second, it is a forward network and can calculate an arbitrary nonlinear mapping. There are many functions can be used as RBF. Here, we adopted the usual one:

$$H_i(x) = \exp(-\|x - c_i\|^2/2\sigma_i^2), \quad i = 1, 2, ..., I; \quad (1)$$

$H_i(x)$ are the results of the second layer (hidden layer). x are the input vectors. c_i are the centers of the function. σ_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:

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

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

There are some parameters need to be adjusted in a RBF neural network. Here, we used Particle Swarm Optimization (PSO) [23] [24] to optimize these parameters. PSO has been proved to be effective and the implement of it was easy. In PSO, the most important parts are the update formulas of velocity and position of each particle. We adopted the original two formulas:

$$v_{in} = \omega * v_{in} + c_1 * r_1 * (Pbest_{in} - x_{in}) + c_2 * r_2 * (Gbest_n - x_{in}); \quad (3)$$

$$x_{in} = x_{in} + v_{in}; \quad (4)$$

v_{in} and x_{in} are the velocity and position of particle i respectively. They are changed in a n-dimensional space. ω is an inertia factor and it is non-negative. It is used to adjust the scope of solution space. c_1 and c_2 are two learning factors and they are set as 2 usually. They are used to adjust the maximum learning step. r_1 and r_2 are two random decimal fractions and they are changed between 0 and 1. They can increase the randomness of search. $Pbest_{in}$ is the best position of particle i so far and $Gbest_n$ is the best position of all particles so far.

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. Above, we had made 4 sample sets, thus we need 4 RBF neutral networks to train them. Every neutral network generated a result and the 4 groups of results were compared.

Next, we used the methods of constructing combined classifier to train the 4 sample sets. Here, we adopted bootstrap aggregation (Bagging) and AdaBoost. They all focus on dealing with the training data set. They have been proved they can improve the classification accuracy by a lot of experiments.

Finally, we used two integrated methods to integrate the results of 4 RBF neural networks. Usually, the generalization ability of an integrated method is better than the corresponding single classifier and it can also improve performance.

The first integrated method we used is called decision fusion (DF). It is represented as: R = W * F. They express 3 matrices respectively. R is the final result of this method. W is weight. F is the result of 4 RBF neural networks. In the above 4 sample sets, we added several features in turn. In our opinion, the more

features the sample set contained, the more it contributed to the final result. So, the weights were set as 0.1, 0.2, 0.3 and 0.4 in turn and the sum of them was 1. There was only one node in the third layer of neural network, because our work is a two-type classification problem. Thus, W was showed as 4*1 and F was showed as 1*4. So, R only contained one value (final result of each sample).

The second integrated method we used is called Genetic Algorithm based Selected Ensemble (GASEN). It is the extension of generalized ensemble method (GEM). GEM is calculated by the following formula:

$$f_{GEM} = \sum_{i=1}^{n} \omega_i f_i(x); \tag{5}$$

ω_i is the weight and it is changed between 0 and 1. The sum of n ω_i is 1. In GASEN, the weights were optimized by Genetic Algorithm. It simultaneously optimized 4 weights each cycle.

III. Experiments and Results

A. Evaluation of The Results

First, we defined the followings:

TP (true positives): it represents the number of interaction sites that were predicted correctly.

TN (true negatives): it represents the number of non-interaction sites that were predicted correctly.

FP (false positives): it represents the number of non-interaction sites that were predicted as interaction sites.

FN (false negatives): it represents the number of interaction sites that were predicted as non-interaction sites.

N: it represents the number of all sites in a protein molecule.

The following metrics were used to evaluate the prediction results [3] [14]:

sensitivity of the positive data:

$$sensitivity^{+} = TP/(TP + FN); \tag{6}$$

specificity of the positive data:

$$specificity^{+} = TP/(TP + FP); \tag{7}$$

specificity of the negative data:

$$specificity^{-} = TN/(TN + FN); \tag{8}$$

false alarm rate:

$$FA - Rate = FP/(FP + TN); \tag{9}$$

accuracy of prediction:

$$accuracy = (TP + TN)/N; \tag{10}$$

Matthews Correlation Coefficient:

$$MCC = \frac{TP*TN - FN*FP}{\sqrt{(TP+FN)(TN+FP)(TP+FP)(TN+FN)}}; \tag{11}$$

MCC often provides a better-balanced evaluation of prediction.

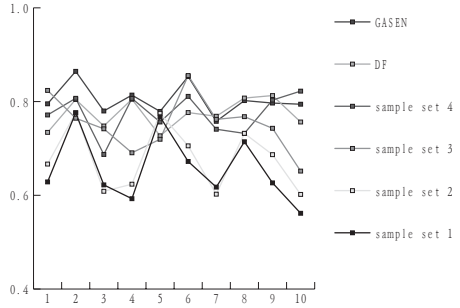

Fig. 1. The accuracy of RBF neural network and their integration.

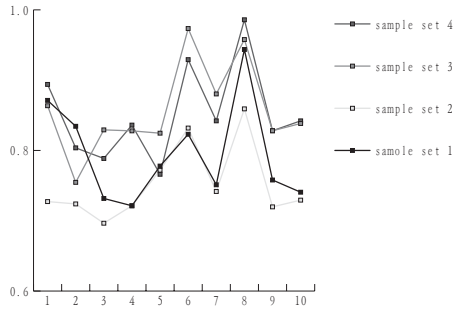

Fig. 2. The accuracy of Bagging.

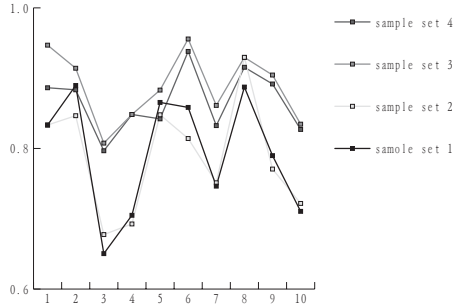

Fig. 3. The accuracy of AdaBoost.

B. Experimental Results

In our experiments, we adopted 10 cross-validation. The 149 protein molecules were divided into 10 groups and every group contained 15 molecules (the last group contained 14 molecules). Each time, one group was selected as test set and the remaining groups were train set. Thus every method was carried out 10 times.(see Fig.1-3)

Every time, PSO and GASEN iterated 1000 times to generate the final result. We chose 10 particles in PSO. In Bagging, the number of self-sample set was set as 10 and each set contained 63% data of the original sample set. In AdaBoost, the number of boosting was set as 10 and we also chose 63% data of the original sample set in each boosting.

In GASEN, we simultaneously optimized 4 weights every time and they were changed between 0 and 1. So, one weight corresponded 14 seats according to the principle of Genetic Algorithm. Thus, the chromosome in each individual included 14*4 seats. The size of population was set as 40.

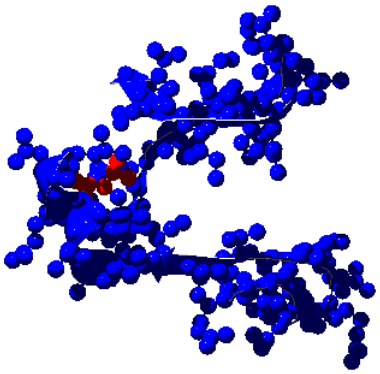

Fig. 4. Visualization of interaction sites (blue) and non-interaction sites (red: only one) of 1mvf that we defined in "Definition of Protein Interaction Sites".

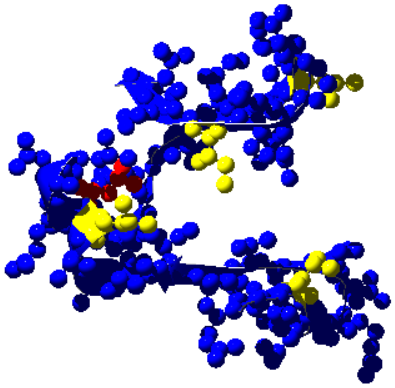

Fig. 5. Visualization of the result of 1mvf in GASEN. Blue showed the interaction sites that were predicted correctly. Red showed the non-interaction sites that were predicted correctly. Yellow (4 ones) showed the residues that were not predicted correctly.

The probability of cross was set as 0.8 and the probability of mutation was set as 0.05.

The final experimental results were the average of 10 times. (see Table 1-3)

In Table 1, accuracy of the 4 sample sets increased in turn and the other measurements were also generally better than the former one in turn. This proved that the added features were useful for predicting. The two integrated methods generated a better result respectively and GASEN was better than DF. On the SPPIDER web site, their final results were 72.48% (version 1) and 74.18% (version 2). Our posterior 4 results were all better than their's.

In Table 2 and 3, the results were generally better than the results in Table 1. As mentioned above, Bagging and AdaBoost focus on dealing with the training data set and constructing combined classifiers. So they generated better results than a single classifier and AdaBoost was a little better than Bagging.

In the end, we used two protein molecules to validate our methods. The PDB ID of them are 1mvf and 1qz8. 1mvf is related to immune system. It contains 4 chains (A, B, D, E) and we used the third one. There are 44 amino acid residues in this chain and we predicted 40 ones by GASEN correctly. 1qz8 is a fragment of SARS corona virus NSP9. It is a new molecule, so we do not know its type and function currently. There are two chains (A, B) in this molecule and we used the first one. The chain contains 111 amino acid residues and we predicted 96 ones by GASEN correctly. (see Fig.4-7)

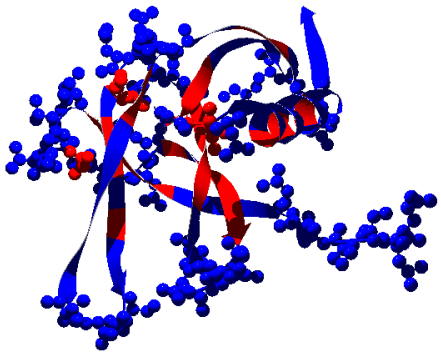

Fig. 6. Visualization of interaction sites (blue) and non-interaction sites (red) of 1qz8 that we defined in "Definition of Protein Interaction Sites".

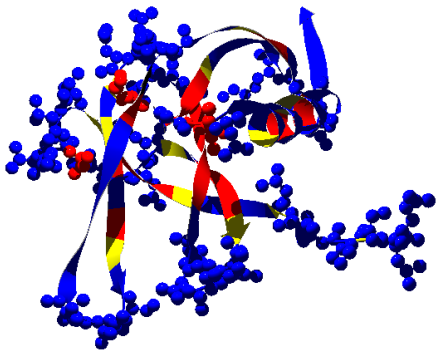

Fig. 7. Visualization of the result of 1qz8 in GASEN. Blue showed the interaction sites that were predicted correctly. Red showed the non-interaction sites that were predicted correctly. Yellow (15 ones) showed the residues that were not predicted correctly.

IV. 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 via many different methods. The next work, we want to use more unique and critical features to predict protein interaction sites. We also hope that more and more different and new methods about computational intelligence and biology can be applied to this subject.

Acknowledgment

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

TABLE I
THE FINAL EXPERIMENTAL RESULTS OF RBF NEURAL NETWORKS AND THEIR INTEGRATION

methods	$sensitivity^+$	$specificity^+$	$specificity^-$	FA-Rate	accuracy	MCC
Sample Set 1	0.933791	0.801388	0.902411	0.539768	0.657926	0.475746
Sample Set 2	0.941524	0.841441	0.890134	0.425145	0.677865	0.587108
Sample Set 3	0.943806	0.962823	0.919847	0.148154	0.752054	0.824262
Sample Set 4	0.952763	0.951528	0.926355	0.185355	0.773842	0.791662
DF	0.979063	0.965711	0.968608	0.231678	0.774042	0.752559
GASEN	0.963019	0.967333	0.942339	0.18466	0.803705	0.807351

TABLE II
THE FINAL EXPERIMENTAL RESULTS OF BAGGING

methods	$sensitivity^+$	$specificity^+$	$specificity^-$	FA-Rate	accuracy	MCC
Sample Set 1	0.965628	0.885068	0.962815	0.2936	0.79528	0.716172
Sample Set 2	0.946969	0.858953	0.950921	0.278801	0.752263	0.694826
Sample Set 3	0.967892	0.951217	0.963004	0.143537	0.857789	0.828646
Sample Set 4	0.975817	0.940382	0.969728	0.173779	0.851574	0.814278

TABLE III
THE FINAL EXPERIMENTAL RESULTS OF ADABOOST

methods	$sensitivity^+$	$specificity^+$	$specificity^-$	FA-Rate	accuracy	MCC
Sample Set 1	0.963229	0.870216	0.936981	0.330586	0.79362	0.709378
Sample Set 2	0.968067	0.893375	0.956177	0.345422	0.788548	0.718661
Sample Set 3	0.974602	0.965497	0.970935	0.116172	0.888569	0.894188
Sample Set 4	0.965665	0.975666	0.953477	0.096172	0.866188	0.892111

REFERENCES

[1] W. Xiangyu, C. Shouliang and G. Mingde, *General Biology*,2rd ed. Beijing: Higher Education Press, 2005.
[2] G. Drewes and T. Bouwmeester, *Global approaches to protein-protein interactions*. Curr.Opin.Cell.Biol. 15, 1-7, 2003.
[3] Y. Changhui, H. Vasant and D. Drena, *Predicting Protein-Protein Interaction Sites From Amino Acid Sequence*. Technical report, Iowa State University, 2002.
[4] F. Sébastien and Z. Martin, *Prediction of Protein-Protein Interaction Sites Using Electrostatic Desolvation Profiles*. Biophys. J. 98, 1921-1930, 2010.
[5] L. F. Consuelo, M. M. dos S.Roberta, J. M. Angelo, R. de M.F. Marcos, C. Carlos and G. Claude, *Identification of continuous interaction sites in PLA2-based protein complexes by peptide arrays*. Biochimie. 91, 1482-1492, 2009.
[6] S. Mile, T. Sanja and V. Kristian, *Prediction of Protein-Protein Interaction Sites in Sequences and 3D Structures by Random Forests*. PloS. Comput. Biol. 5, 1-9, 2009.
[7] L. Man, L. T. Chew and S. Jian, *Feature generation and representations for protein-protein interaction classification*. J. Biomed. Inform. 42, 866-872, 2009.
[8] 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. 60, 659-672, 2010.
[9] F. A.H. Mohamed, S. Friedhelm, P. Günther, *Semi-supervised learning for tree-structured ensembles of RBF networks with Co-Training*. Neur. Netw. 23, 497-509, 2010.
[10] M. M. Gonzalo and S. Alberto, *Out-of-bag estimation of the optimal sample size in bagging*. Pattern. Recognit. 43, 143-152, 2010.
[11] S. Noritaka, M. Hiromi, M. Michiharu and M. Lixin, *Bagging and AdaBoost algorithms for vector quantization*. Neurocomputing. 73, 106-114, 2009.
[12] C. Guo and Z. Hongfu, *Fusion Diagnosis for EngineWear Fault Based on Integrated Neural Network*. J. Nanjing Univ. Aero & Astr. 36, 278-283, 2004.
[13] Z. Zhihua, W. Jianxin and T. Wei, *Ensembling neural networks: Many could be better than all*. Artif. Intell. 137, 239-263, 2002.
[14] 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. 10, 233-246, 2009.
[15] L. Yang, T. Zhengquan and W. Yifei, *SVM-based protein interaction sites prediction*. Technical report, Shanghai University, 2006.
[16] L. Chun and Q. Weiyi, *Mathematical description of biological macromolecules and its application*. Liaoning: Dalian University of Technology Press, 2009.
[17] *Comparison of SURFACE and AREAIMOL for accessible surface area calculations*, http://www.ccp4.ac.uk/Newsletters/newsletter38/03_surfare-a.html
[18] J. Mihel, M. Sikic, S. Tomic, B. Jeren and K. Vlahovicek, *PSAIA-Protein Structure and Interaction Analyzer*. BMC. Struct. Biol. 8, 21, 2008.
[19] W. Kabsch and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features*. Biopolymers. 22, 2577-2637, 1983.
[20] P. Fariselli, A. Zauli, I. Rossi, M. Finelli, P. L. Martelli and R. Casadio, *A neural network method to improve prediction of protein-protein interaction sites in heterocomplexes*. In: 13th IEEE Workshop on Neural Networks for Signal Processing, pp. 33-41. IEEE Press, 2003.
[21] M. Wei, W. Feifei and P. Xinjun, *Prediction of Protein-Protein Interaction Sites Using Support Vector Machine*. J. App. Scie. 26, 403-408, 2008.
[22] Z. Huanxiang and S. Yibing, *Prediction of Protein Interaction Sites From Sequence Profile and Residue Neighbor List*. Proteins. 44, 336-343, 2001.
[23] E. Mohammed and S. K. Mohamed, *PSO_Bounds: A New Hybridization Technique of PSO and EDAs*. Stud. Comp. Intell. 203, 509-526, 2009.
[24] A. Marco, de O. Montes, S. Thomas, B. Mauro and D. Marco, *Frankenstein's PSO: A Composite Particle Swarm Optimization Algorithm*. IEEE Trans. Evol. Comput. 13, 1120-1132, 2009.

Xiaoli Shen was born in 1984. He received the B.S. degree in computer science from University of Jinan, Jinan, China, in 2008. Since 2008, he has been a master student in the School of Information Science and Engineering, University of Jinan. His research interests include neural network and bioinformatics.

Yuehui Chen was born in 1964. He received his B.Sc. degree in the Department of Mathematics (major in control theory) from the Shandong University of China in 1985, and Ph.D. degree in Department of Electrical Engineering and Computer Science from the Kumamoto University, Japan in 2001. During 2001-2003, he had worked as a Senior Researcher of the Memory-Tech Corporation at Tokyo. Since 2003 he has been a member at the Faculty of School of Information Science and Engineering in University of Jinan, where he is currently heads the Laboratory of Computational Intelligence. His research interests include evolutionary computation, neural networks, fuzzy systems, hybrid computational intelligence and their applications in time-series prediction, system identification and intelligent control. He is the author and co-author of more than 70 papers. Professor Yuehui Chen is a member of IEEE, the IEEE Systems, Man and Cybernetics Society and the Computational Intelligence Society. He is also a member of the editorial boards of several technical journals and a member of the program committee of several international conferences.