Advanced Engineering Forum *Online: 2012-09-26*
ISSN: 2234-991X, Vols. 6-7, pp 824-830
doi:10.4028/www.scientific.net/AEF.6-7.824

Building Ensemble Classifier Based on Complex Network for Predicting Protein Structural Class

Peng Wu[a], Tao Xu[b], Likai Dong[c], Zhen Liu[d] and Yuehui Chen[e]

Shandong Provincial Key Laboratory of Network based Intelligent Computing, Jinan, 250022, China

[a]ise_wup@ujn.edu.cn, [b]ise_xut@ujn.edu.cn, [c]ise_donglk@ujn.edu.cn, [d]ise_liuz@ujn.edu.cn, [e]yhchen@ujn.edu.cn

Keywords: complex network, protein structural class, ensemble classifier.

Abstract. In recent years, complex network models were developed to solve classification and time series prediction problems. In this paper, ensemble classifier based on complex network (mainly scale-free network) is firstly used to predict protein structural class. For the classifier design, genetic programming and particle swarm optimization algorithm are used alternately to evolve the structure and encoding parameters. The experimental results validate the good performance of the proposed method.

Introduction

To predict protein structural class is a very useful, while full of challenging task in protein community. With the development of protein science, researchers have known that proteins always manifest their functions through the three-dimensional structures, and the amino acid sequence of proteins determined their three-dimensional structures. However, the three-dimensional protein structure information acquired by people is quite limited due to high computational cost and expensive experimental methods such as x-ray crystallography and nuclear magnetic resonance to get them. To overcome this, machine learning methods or so-called computer simulating methods are therefore emerging as alternative or complementary approaches. The core issue of such approaches is how to predict the specific protein structural class type of proteins using their information of amino acid sequence. According to the definition by Levitt and Chothia [1], protein structural class can be classified into the following four types: (i) all-α that is formed essentially by α-helices, (ii) all-β essentially by β-strands, (iii) α/β containing both α-helices and β-strands that are largely interspersed in forming mainly parallel β-sheets, and (iv) α+β containing also both of the two secondary structure elements that are largely segregated in forming mainly anti-parallel β-sheets. The prediction problem can be formulated as a regular multiple-classification problem. Actually many classifiers such as Artificial Neural Network (ANN), LogitBoost and Support Vector Machines (SVMs) have been applied in prediction of the structural classes of proteins from their amino acid sequence [2, 3, 4, 5]. For achieving good performance through these classifiers, a sample protein with good discriminative information representing is needed. To deal with this problem, the following five features from protein sequences are usually extracted: (i) chemical composition, which describes the global percent composition of each group of amino acids in a protein and contains 21 components, (ii) dipeptide composition, which is the normalized pair-wise occurrence frequency of the 20 native amino acids and consists of a total of 400 descriptor value, (iii) chemical distribution, which describes the distribution pattern of the attribute along the sequence and consists of a total of 105 descriptor value, (iv) quasi-sequence-order, which consists of a total of 100 descriptor value, and (v) conjoint triad, which contains 343 components [6].

In previous studies, complex network models were evolved to solve classification and time series prediction problems [13, 14, 15]. In this paper, a new ensemble classifier based on Complex Network (CN) is exploited to predict protein structural class, which has the same structure properties as CN (mainly scale-free network), and can be evolved by evolutionary algorithms such as Genetic Programming (GP) and Particle Swarm Optimization (PSO). The contributions of this paper includes:

(1) the classifier based on complex network is firstly used to predict protein structural class, which expands the range of applications of complex network model, and (2) instead of using a single feature type for all classes, combining a set of divers and complementary features is adopted.

The rest of this paper is organized as follow. In section 2, we have discussed some related work; the classifier based on complex network and details of combining of features are described in section 3; in section 4, a case study is reported, and finally the conclusion is given in section 5.

Related Work

The background knowledge related to proposed method, i.e. CN, GP and PSO, are briefly described in this section respectively.

CN is well studied in many fields of sciences [7, 8]. Undeniably, one can find that many systems in the world can be represented by models of CN such as the Internet, World Wide Web, brain, food webs and the network of relationships among people. All of them can be well characterized by structures with non-trivial topological features, which consist of nodes or vertices connected by links or edges and can not be found in simple networks such as lattices or random graphs. CN has been successfully applied to lot of fields such as, the analysis of metabolic and genetic regulatory networks, design of robust and scalable communication networks both (wired and wireless), development of vaccination strategies for the control of disease, and a broad range of other practical applications. Small-world network and scale-free network are the most intensively studied for complex networks. For the small-world network, its average number of edges between any two vertices is very small, while the clustering coefficient is large. For the scale-free network, its degree distribution follows a particular mathematical function called a power law.

Genetic programming [9] is a branch of genetic algorithm. Rather evolving a linear string as genetic algorithm does, genetic programming evolves computer programs which are usually tree structure. To evolve ensemble classifier based on complex network, the genetic operators, such as crossover and mutation are made some modification.

PSO is a powerful intelligence technique, motivated by social behavior of bird flocking or fish schooling, which was proposed by James Kennedy and Russell Eberhart [10]. A PSO algorithm shares many similarities with evolutionary algorithm such as genetic algorithm, in which a number of particles representing potential solutions and moving through the problem space are called swarm similar to population. During every searching, two important values are record, which are p_{bi}, that is referred as the particle's personal best position, and g_b, which is referred to as the swarm's global best position. Iteratively, the positions of all the particles in the swarm are updated and the fitness of the new positions of the particles evaluated. The positional update is traditionally done according to Eq. 1 and Eq. 2. Let $x_i(t)$ denote the position of particle i in the search space at time step t and $v_i(t)$ denote the velocity of particle i. The new velocity is calculated using Eq. 2, as follow,

$$v_i(t+1) = v_i(t) + c_1\varphi_1(p_{bi}(t) - x_i(t)) + c_2\varphi_2(g_b(t) - x_i(t)). \quad (1)$$

Here c_1 and c_2 are positive constant and φ_1 and φ_2 are uniformly distributed random number in range of [0, 1].

The position of the particle is changed by adding a velocity, $v_i(t)$, to the current position, i.e.

$$x_i(t+1) = x_i(t) + v_i(t+1). \quad (2)$$

In this study, PSO is used to tune the corresponding parameters of a complex network based classifier.

Ensemble classifier of complex network

This section will describe how to construct a classifier based on complex network and how to ensemble the basic classifiers for achieving improved performance.

The Description of Complex Network based Classifier. A directed and weighted graph is selected for representing a CN based classifier. We use the algorithm described in [11] to initialize a classifier (see Fig. 1). While learning the classifier, for any non-input node, i.e. represented by $Node_i$, where i real numbers are randomly generated and used to represent the connection strength between $Node_i$ and its adjacent nodes. In addition, two adjustable parameters a_i and b_i are randomly selected as flexible activation function parameters. In the learning process of classifier, the flexible activation function is used, which is shown as follow, i.e. Eq. 3,

$$f(a_i, b_i, x) = e^{-(\frac{x-a_i}{b_i})^2}. \tag{3}$$

The output of a non-input node, calculated as a flexible neuron. The total excitation of $Node_i$ is calculated as Eq. 4,

$$Node_i = \sum_{j=1}^{n} w_j * x_j . \tag{4}$$

Where x_j ($j = 1, 2, \ldots, n$) is the input to $Node_i$. The overall output of a classifier can be computed by depth-first method, recursively.

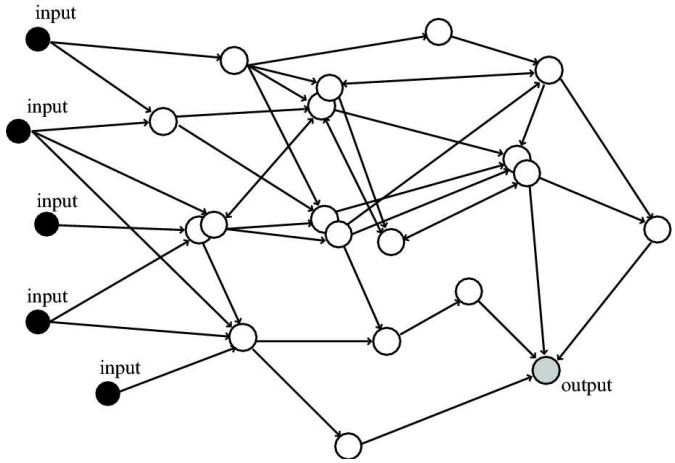

Fig. 1. An example of classifier based on CN (input nodes are black and the output node is grey).

The Optimization of Complex Network based Classifier. Finding an optimal or near-optimal complex network is formulated as a product of evolution. In this paper, GP with some modification is used to evolve the structure of a CN based classifier, which can be summarized as follows:

1) Initialization. A population of CN (including the embedded parameters) is randomly generated.
2) Evaluation. Each complex network genotype P_{gi} of current population is evaluated on the given task and assigned a fitness value Fit (P_{gi}) according to the defined fitness function, i.e. root mean square error.
3) Crossover. With a certain probability, P_c, using the tournament selection mechanism, i.e. m individuals are randomly selected from the population to go through a competition and the winner (in terms of fitness) is selected, to select two parents and create two children by swapping some vertexes with their adjacent edges between the two parent complex network genotypes. Fig. 2 depicts how crossover works.

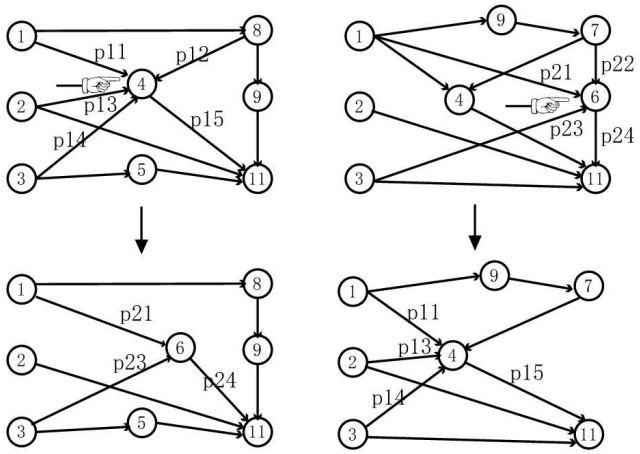

Fig. 2. Crossover (input nodes are $node_1$, $node_2$, $node_3$ and output node is $node_{11}$)

4) Mutation. With a probability, P_m, a complex network genotype is selected randomly for mutation. Four mutation operators are randomly chosen with the same probability, which are developed as follows:
   - Adding one edge: randomly select one node in the complex network and connect it to another randomly selected node.
   - Adding one node: add one node and connect it to a selected destination node with a preferential attachment function: this function defines the probability that a node in network receive a link from a newly inserted node [11]. The analytic form of this function, i.e. Eq. 5, is defined as follow,

$$\prod(k_i) = \frac{k_i}{\sum_{j=1}^{N} k_j} \quad . \tag{5}$$

   Where k_i is the degree of node i.
   - Deleting one edge: select a random node and delete one of its conjoint edges.
   - Deleting one node: randomly select a non-output node in the complex network and delete all of its conjoint edges.
   - Selection. Pair wise comparison is conducted for the union of μ parents and offspring. For each individual, q opponents are chosen uniformly at random from all the parents and offspring. For each comparison, if the individual's fitness is no smaller than the opponent's, it receives a selection. Select some individuals out of parents and offspring that has most wins to form the next generation. Steps 3-5 are repeated until a new population is constructed. If satisfied solution is found, then stop; otherwise go to step 2.

Parameters Optimization. For the parameters optimization of a CN based classifier, a number of global and local search algorithms that are GA, EP and gradient based learning method can be employed. The basic PSO algorithm is selected for parameter optimization due to its fast convergence and easy to implementation.

The General Hybrid Method. The general procedure of constructing a CN based classifier can be described as follow:

1) Create the initial population (complex network genotype and the corresponding parameters randomly);
2) Structure optimization by GP;
3) If the better structure is found, then go to step 4, otherwise go to step 2;
4) Parameter optimization by PSO algorithm. In this stage, the complex network structure is fixed, and it is the best genotype taken from the end of run of the structure search. All of the parameters encoded in the best genotype formulated a parameter vector to be optimized by PSO;

5) If the maximum number of PSO search is reached, or no better parameter vector is found for a significantly long time then go to step 6; otherwise go to step 4;
6) If satisfied solution is found, then stop; otherwise go to step 2.

Classifier Ensemble Scheme. Five different types of feature are extracted for each protein, i.e. chemical composition, dipeptide composition, chemical distribution pattern of the attribute, quasi-sequence-order, and conjoint triad. There are usually two ways of using these features, one of which is simply concatenating different feature vectors to a super vector, then constructing and integrating basic classifiers in the new feature vector space; the second is constructing basic classifiers in the different feature vector space separately, then integrating the basic classifiers. Considering the difference discriminative power of the five types of feature, simply concatenating is inappropriate, so the second method is adopted, in which the voting strategy is chosen when it comes to the last integrating stage. We can see it from Fig.3.

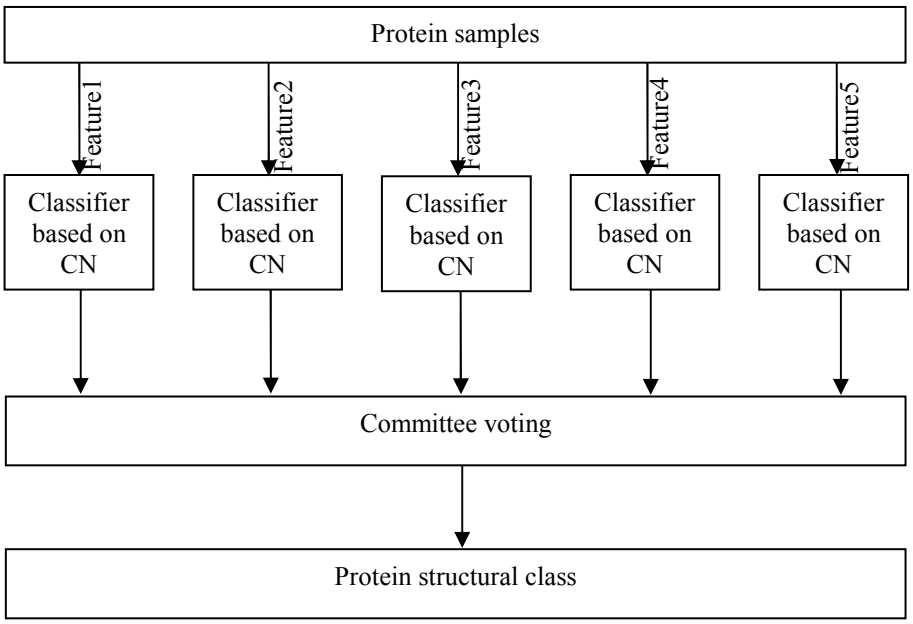

Fig. 3. Flowchart of the ensemble classifier based on CN

Case study

This section will describe how to construct a classifier based on complex network and how to ensemble the basic classifiers for achieving improved performance.

For comparison, we have used the same dataset with Chu [12], which consists 204 domains: 52 all-α domains, 61 all-β domains, 45 α/β domains, and 46 $\alpha+\beta$ domains. The predicting performance of classifier is evaluated by the success rate of cross-validation by jackknifing, which is thought the most objective and rigorous way in comparison with the other two, which are sub-sampling test and independent dataset test. During the process of jackknife test, each protein in the dataset is taken as independent test sample and all the corresponding parameters are evolved without using this protein.

The results of prediction by independently using the five types of features and integrating all of them are shown in Table 1. We can see from Table 1, using quasi-sequence-order yields the highest success overall rate, 91.7%, due to its combination of chemical composition and the normalized occurrence of them. On the contrary, the chemical composition alone gives the worst result which is only 73.5%, due to the lack of sufficient information. The eventual result of ensemble classifier based CN is 94.1%, which is higher than the other five results due to the efficiently exploiting all the five types of feature.

Table 1. Summary of success rate using different feature under the jackknife test

Feature	Success rate (%)				
	all-α	all-β	α/β	$\alpha+\beta$	overall
Chemical composition	80.8	88.5	68.9	50	73.5
Dipeptide composition	90.4	91.8	86.7	73.9	86.3
Chemical distribution	76.9	83.6	93.3	45.6	75.5
Quasi-sequence-order	92.3	95.1	97.8	80.4	91.7
Conjoint triad	86.5	85.2	91.1	69.6	83.3
Ensemble of all(our method)	94.2	96.7	97.8	86.9	94.1

Table 2. Comparison of different methods by the jackknife test

Method	Success rate (%)				
	all-α	all-β	α/β	$\alpha+\beta$	overall
LogitBoost [6]	90.4	88.5	80.0	73.9	83.8
Complexity measure factor [6]	82.7	90.2	100	87.0	89.7
Autocorrelation factor[6]	88.5	96.7	77.8	73.9	85.3
AAPCA[6]	82.0	97.0	82.0	78.0	85.0
BTSVM[6]	90.4	100.0	97.8	73.9	91.2
IDQD[6]	90.4	93.4	100.0	89.1	93.1
ApEn, hydrophobicity pattern[6]	96.2	98.4	100.0	93.5	97.0
Best-first searching[6]	92.3	93.4	95.6	78.3	90.2
Our method	94.2	96.7	97.8	86.9	94.1

In order to compare with other methods, other experimental results on same dataset are listed in Table 2, in which the highest Jackknife-tested overall success rate is 97.0%. The result of ensemble classifier based on CN is very close to the state-of-the-art one due to the more flexible non-linear mapping power achieved by representing with a kind of sophisticated structure.

Conclusion

In this research, the ensemble classifier based on CN was used to predict protein structural class, and the promising experimental results validate the efficiency of our method. In the evolutionary algorithms which are genetic programming and particle swarm optimization are used alternately to evolve the structure and parameters of a classifier. We believe that the current research might help to open a new avenue to further study the complex networks in theory and application.

Acknowledgments

The authors acknowledge the financial support from the National Natural Science Foundation of University of Jinan (Nos. XKY1025, XKY0927), the Natural Science Foundation of Shan dong Province (No. ZR2009FL002).

References

[1] M. Levitt, C. Chothia, Structural patterns in globular proteins, Nature. 261 (1976) 552-558.

[2] Y.C. Dong, G.P. Zhou, Prediction of protein structural classes by neural network, Biochimie. 82 (2000) 783-785.

[3] Y.C. Dong, G.P. Zhou, Prediction of protein structural classes by support vector machines, Comp. Chem. 26 (2002) 293-296.

[4] G.P. Zhou, N. Assa-Munt, Some insights into protein structural class prediction, Pro. Stru. Fun. Gen. 44 (2001) 57-59.

[5] G.P. Zhou, Y.D. Cai, Predicting protease types by hybridizing gene ontology and pseudo amino acid composition, Pro. Stru. Fun. Bio. 63 (2006) 681-684.

[6] C. Chen, L.X. Chen, X.Y. Zou and P.X. Cai, Predicting protein structural class based on multi-features. J. Theo. Bio. 253 (2009) 388-392.

[7] R. Albert, A.L. Barabasi, Statistical Mechanics of Complex Networks, Rev. Mod. Phy. 74 (2002) 47-97.

[8] Y.C. Lai, A.E. Motter and T. Nishikawa, Attacks and Cascades in Complex Networks, Lect. Not. Phy. 650 (2004) 299-310.

[9] J.R. Koza, Genetic Programming: On the Programming of Computers by Means of Natural Selection, MIT Press, 1992.

[10] J. Kennedy, R.C. Eberhart, Particles swarm optimization, In: IEEE international conference on neural networks, IEEE Press, New York, 1995, pp. 1942-1948.

[11] A. Mauro, B. Ilaria, D.F. Matteo and P. Stefano, Evolving Complex Neural Networks, LNAI. 4733 (2007) 194-205.

[12] K.C. Chou, A key driving force in determination of protein structural classes, Bio. Bio. Res. Com. 264 (1999) 216-224.

[13] P. Wu, Z. Liu, Classification of DNA Microarray for Cancer Diagnosis Using Complex Network, In: 2nd International Conference on Intellectual Technique in Industrial Practice, IEEE Press, New York, 2010, pp. 266-269.

[14] P. Wu, Y.H. Chen, Q.F. Meng and Z. Liu, Small-Time Scale Network Traffic Prediction Using Complex Network Models, In: 5th International Conference on Natural Computation, IEEE Press, New York, 2009, pp. 303-307.

[15] P. Wu, Y.H. Chen, T. Xu and H.K. Tang, Evolving Complex Network for Classification Problems, In: International Conference on Computational Intelligence and Natural Computing, IEEE Press, New York, 2009, pp. 287-290.