Clustering PPI Data Based on Bacteria Foraging Optimization Algorithm

Xiujuan Lei[*1] Shuang Wu[2]

[1,2]*College of Computer Science*
Shaanxi Normal University
Xi'an, Shaanxi Province, China, 710062
[1]xjlei@snnu.edu.cn
[2]persistencewu@126.com

Liang Ge[3] Aidong Zhang[4]

[3,4]*Department of Computer Science and Engineering*
State University of New York at Buffalo
NY 14260-2000, USA
[3]liangge@buffalo.edu
[4]azhang@buffalo.edu

* *Corresponding author*

Abstract—**This paper proposed a novel method using Bacteria Foraging Optimization(BFO) algorithm to avoid the influence of cluster number on experimental result of clustering PPI networks. The initial position that the bacterium located in was considered to be the cluster center and the positions that the bacterium moved were regarded as the adjacent nodes of cluster center. The algorithm classified the nodes selected in the chemotactic operation into cluster when executing the reproduction and elimination-dispersal operations. The procedure kept on creating new clusters until all the nodes were grouped into the clusters. The simulation result showed that the algorithm not only effectively improved the accuracy of cluster result, but also automatically determined the cluster number.**

Keywords-bacteria foraging optimization algorithm; PPI networks; accumulation coefficient of edge

I. INTRODUCTION

As is known, protein normally interacts with other proteins to perform some function in certain time and space. Recently the rapid development of proteomics has attracted more and more researchers to analyze PPI networks. Several researchers found that PPI networks had the characters of small-world[1] and scale-free[2] which suggested an important topological structure known as modularity. Therefore it is essential to adopt clustering methods to predict the functional modules. However, the traditional clustering methods[3] do not perform well. Elena Nabieva *et al*[4] firstly proposed a functional flow model which was time consuming. Young-Rae Cho *et al*[5] adopted another flow based algorithm. Even if the algorithm had higher accuracy compared with other competing approaches, the *f-measure* value[6] of clustering result was low. Apurv Goel *et al*[7] developed a software for the generation and analysis of dynamic four-dimensional PPI networks. Our research team[8] had adopted artificial bee colony and particle swarm optimization algorithms to predict functional modules.

Drove by the rapid development of intelligence algorithms, Passino[9] proposed Bacteria Foraging Optimization(BFO) algorithm. Several researchers combined it with intelligent methods[10] and it turned out that the performance was superior to the improved genetic and particle swarm optimization algorithms.

In this paper we propose a novel method taking the principle of BFO algorithm into account to predict the functional modules of PPI networks. In Section Ⅱ, the principle of BFO algorithm and several concepts related to PPI networks are briefly introduced. In Section Ⅲ, the improved algorithm which integrates the principle of BFO algorithm is described in details. We make comparison with flow algorithm in Section Ⅳ. The experimental results show that the algorithm is superior to flow algorithm.

II. BASIC CONCEPTS

A. Principle of BFO algorithm

BFO algorithm is an evolutionary algorithm which consists of chemotactic, reproduction, and elimination-dispersal operations[11]. The bacterium moves in two different ways to avoid noxious environment which is regarded as chemotactic behavior. In general, several bacteria which are becoming more and more incapable of searching food are obliged to be eliminated. In order to maintain the scale of population, the remained bacteria will reduplicate and generate new individual which is considered to be reproduction behavior. Because of the sudden change in the local environment, the bacteria population may be gradually inadaptable to the environment which leads to a fact that a group of bacteria are either killed or dispersed into a new location. This phenomenon is the elimination-dispersal behavior which can prevent the algorithm from trapping into the local optimal solution and search for a new individual which is much closer to the global optimal solution.

978-0-7695-4574-5/11 $26.00 © 2011 IEEE
DOI 10.1109/BIBM.2011.18

B. Relevant concepts of PPI networks

The weighted degree of node is defined as the summation of the weight value of edges between nodes *i* and its neighbors. The clustering coefficient of node which is used to assess the quality of cluster result is calculated by the following equation[12].

$$C_i = 2n_i / k_i(k_i - 1), \tag{1}$$

where k_i represents the degree of node *i*, n_i refers to the number of edges connecting all the neighbor nodes of *i* with each other. Recently the concept of clustering coefficient of node is extended to edge and the accumulation coefficient of edge[13] is defined as follows:

$$WC_{u,v} = \frac{\sum_{k\in I_{u,v}} w(u,k)\cdot\sum_{k\in I_{u,v}} w(v,k)}{\sum_{s\in N_u} w(u,s)\cdot\sum_{t\in N_v} w(v,t)}. \tag{2}$$

where the sets N_u and N_v represent the sets of directly adjacent nodes of node *u* and node *v* respectively. The symbol $w(u,s)$ refers to the weight value of edge linking nodes *u* with *s*. The set $I_{u,v}$ stands for the set of common nodes between the adjacent nodes of nodes *u* and *v*.

The Comprehensive Network Feature Value(*CNFV*) of node[14] can reveal the joint strength among this node and other nodes. The *CNFV* of node *i* is defined as follows:

$$CNFV_i = \beta * C_i + (1-\beta) * w(i) / n. \tag{3}$$

The parameter β is a random number within 0 and 1, $w(i)$ refers to the weighted degree of node *i*, and *n* stands for the number of protein nodes in PPI network.

C. Object function

The clustering coefficient of a cluster module is defined as the average clustering coefficient of all the protein nodes belonging to this cluster module[14]. The equation is as follows:

$$C_B = \frac{\sum_{j=1}^{h} C_{Bj}}{h}. \tag{4}$$

In Eq. (4), the parameter C_{Bj} represents the cluster coefficient of node *j* and *h* stands for the number of nodes which belong to cluster *B*.

D. Evaluation criteria of cluster result

In general, a large number of studies on clustering analysis adopt *precision* and *recall* values to evaluate cluster result[6]. Suppose that *X* represents one cluster module in the cluster results, F_i stands for the matched cluster module in the standard PPI dataset.

$$precision(X, F_i) = \frac{|X \cap F_i|}{|X|}, \tag{5}$$

$$recall(X, F_i) = \frac{|X \cap F_i|}{|F_i|}, \tag{6}$$

where the expression $|X\cap F_i|$ stands for the number of common proteins between cluster modules *X* and F_i. However, these two criteria have drawbacks in facing with the unexpected circumstances of larger and smaller cluster modules. Therefore, we assess the accuracy of modules with the *f-measure* value which balances *precision*, *recall*, and running time:

$$f - measure = \frac{3}{\frac{1}{precision} + \frac{1}{recall} + time}. \tag{7}$$

III. Algorithm

A. Data preprocessing

With regard to the data attribute of PPI networks, protein name can be transformed into the positive integer in turn and the data is converted into an adjacent matrix *p*. Assume that the number of protein nodes is *n*, X_i represents the *i*-th protein which is denoted as X_i=(p_{i1}, p_{i2},…, p_{in}) and X_{ij} stands for the inner product of two protein nodes. The similarity between nodes *i* and *j* is defined as:

$$S_{ij} = \frac{\sum_{k=1}^{m} \min(X_{ik}, X_{jk})}{\sum_{k=1}^{m} \max(X_{ik}, X_{jk})}, \tag{8}$$

The protein also has interactions with other proteins via some protein or some several proteins. Therefore, this paper utilizes the weighted similarity coefficient of node which takes all the interactions of each protein node into consideration[15]. The equation is as follows:

$$Y_i = w_i \sum_{j=1}^{n} S_{ij} + \frac{1}{2} B_i \sum_{j=1}^{n} S_{ij}, \tag{9}$$

In Eq. (9), the parameter w_i represents the weight value of the *i*-th protein and B_i stands for number of protein nodes which have interactions with the *i*-th protein.

$$\alpha_i = \frac{Y_{\max} - Y_i}{Y_{\max}} \times 100\%. \tag{10}$$

The symbol α_i represents the deviation degree of *i*-th protein node[15]. If the value α_i is higher than the threshold α, then the node is considered to be the sparse node.

B. The principle of algorithm

This paper takes advantage of the mechanism of BFO algorithm to deal with clustering problem of PPI networks. The initial position that the bacterium locates in stands for the initial selected cluster center. Then the bacterium will move in the adjacent area of the initial position in the phase of chemotactic operation. This phenomenon can be regarded as the procedure of selecting the directly connected nodes of initial cluster center. Assume that the selected protein nodes are saved in a set *neighbor*, and then the bacterium reduplicates the superior individuals to maintain the stability of population which is regarded as the clustering procedure of merging the protein nodes in the set *neighbor* into the cluster that initial cluster center belongs to. This paper takes advantage of the clustering coefficient of node and accumulation coefficient of edge to classify protein nodes into cluster module. Finally, the remained protein nodes in the set *neighbor* will be accepted at a random probability in the elimination-dispersal operation. Unfortunately, the added protein node may destroy the original optimal solution due to the randomness of acceptance probability. This paper adopts an object function which has been discussed in the former section to judge whether accept the newly added protein node to take part in this cluster.

C. The implementation step of algorithm

Step 1: The algorithm firstly utilizes the weighted similarity coefficient of node to eliminate sparse nodes. And assign values to several parameters: the thresholds of node degree and *CNFV*, the maximal times of chemotactic operation *NC*, the index of the iterations of the chemotactic operation *N*=1, and the cluster number is denoted as *clu_num*=0.

Step 2: Select a cluster center according to degree and *CNFV* of node.

Step 3: Protein nodes which are directly connected with cluster center are preserved in a set *neighbor*.

Step 4: If the clustering coefficient of a node is high and the accumulation coefficient of edge between the node and its cluster center is also high, then it can be classified into this cluster. This procedure terminates until all the protein nodes in the set *neighbor* have been judged whether they have access to be grouped into this cluster.

Step 5: With regard to remained protein nodes in the set *neighbor*, this algorithm adopts the acceptance strategy of a random probability and evaluates the obtained cluster module according to Eq. (4) to judge whether accept this node to participate in this cluster.

Step 6: Set *N*=*N*+1. If the value *N* arrives at the maximal iterations *NC*, go to **Step 7**, else go back to **Step 3**.

Step 7: Obtain a cluster module, meanwhile set the cluster number as *clu_num*=*clu_num*+1.

Step 8: If all the protein nodes which have the higher degrees and comprehensive network feature values are merged into the cluster modules. Then assess the performance of algorithm in terms of *precision*, *recall*, *f-measure* and so on, meanwhile output the cluster modules. Else go back to **Step 2**.

D. The time complexity of algorithm

Suppose that the number of protein nodes in PPI dataset is *n*, and the number of protein nodes which satisfy the selection requirements of cluster center is *m*, the maximal times of chemotactic operation is *NC*. The time complexity of merging protein nodes into cluster modules in the chemotactic and reproduction operations is $O(n^2)$; Afterwards adopt an object function in the elimination-dispersal operation, the time complexity is $O(n)$. Therefore, the time complexity of executing the three main operations for one time is $O(n^2)$; One cluster module is obtained after executing three operations for *NC* times, the time complexity is $O(NC \times n^2)$; The number of protein nodes which have the potential to be selected as cluster centers is *m*, so it is likely that there are *m* cluster modules in the end, the time complexity is $O(m \times NC \times n^2)$.

IV. EXPERIMENTAL RESULTS

A. Parameter analysis

To assess the performance of algorithm, we use MIPS PPI data sets to evaluate the cluster modules predicted by our method[16]. There are several parameters which may have effect on the cluster results, such as the thresholds of node degree, accumulation coefficient of edge *lambda* and clustering coefficient of node *mu*.

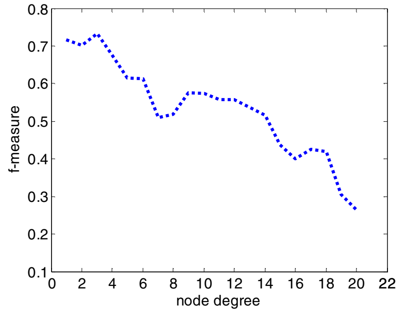

Figure 1. The influence of node degree

As the node degree increases, the corresponding *f-measure* value of the cluster result roughly descends which results from a fact that higher node degree may discard more protein nodes. When the node degree is set as 3, we can obtain the optimum value.

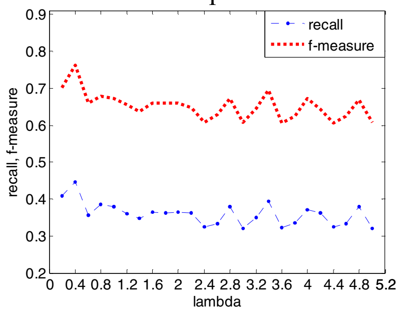

Figure 2. The influence of accumulation coefficient of edge

Fig. 2 shows that the *recall* and *f-measure* values of cluster result are inclined to gently fluctuate as the parameter lambda varies from 0.1 to 5. It is obvious that when the parameter lambda is set as 0.4, all the values of the evaluating criteria arrive at the highest point.

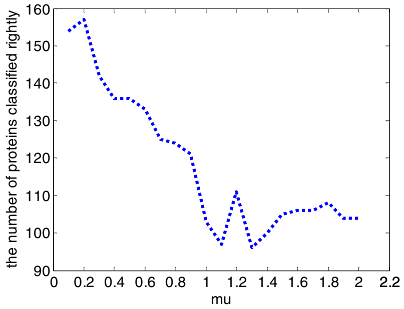

Figure 3. The influence of clustering coefficient of node

Fig. 3 shows that as the threshold of clustering coefficient of node mu increases, the algorithm can find the less correct proteins. The experimental results reveal that when the parameter mu is set as 0.2, we can get the optimal cluster result.

B. *Performance comparison*

The functional flow algorithm which is referred to in the section of introduction is an effective method in solving clustering problem of PPI networks. Therefore this paper makes comparisons between BFO and Flow algorithms as follows:

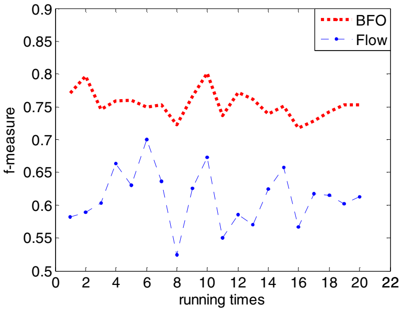

Figure 4. Comparison between BFO and Flow algorithms

The flow method which is based on the concept that the functional information of a protein flows through every possible path in PPI networks requires cluster number. However, the approach proposed in this paper overcomes this drawback. Fig.4 describes the cluster result of BFO algorithm is superior to the Flow algorithm all the time in terms of *f-measure* value. The top 5 cluster modules of cluster result which is obtained by means of BFO algorithm are showed as follows.

TABLE I. THE TOP 5 MODULES IN THE CLUSTER RESULTS

Cluster module	The proteins classified rightly	The proteins classified wrongly
1	YGL240w, YFR036w, YHR166c, YDR118w, YBL084c, YKL022c, YOR249c, YNL172w, YDL008w, YLR102c, YLR127c	YIR025w
2	YML077w, YGR166w, YBR254c, YOR115c, YMR218c, YDR472w, YDR246w, YKR068c, YDR108w, YDR407c	——
3	YNL121c, YMR203W, YOR045w, YGR082w, YNL070w, YNL131w, YPR133w-a	——
4	YHR107c, YCR002c, YJR076c, YDL225w, YLR314c	YDR507c
5	YJR093c, YLR277c, YAL043c, YLR115w, YDR301w	YPR107c,YKR002w, YNL317w

ACKNOWLEDGMENT

This paper was supported by the National Natural Science Foundation of China(No.61100164), the Natural Science Foundation of Shaanxi Province of China in 2010(No. 2010JQ8034), and the Fundamental Research Funds for the Central Universities(No. GK200902016), and the Graduated Student Innovation Foundation of Shaanxi Normal University(No. 2011CXS030).

REFERENCES

[1] Watts, D.J. and Strogatz, S.H. Collective dynamics of 'small-world' networks. Nature,1998,393: 440-442

[2] Barabási, A.L. and Oltvai, Z.N. Network biology: understanding the cell's functional organization. Nature Reviews: Genetics, 2004,5: 101-113

[3] Penggang Sun, Lin Gao, Shanshan Han. Identification of overlapping and non-overlapping community structure by fuzzy clustering in complex networks. Information Sciences, Mar.2011, 181(6): 1060-1071

[4] Elena Nabieva, Kam Jim, Amit Agarwal, Bernard Chazelle, and Mona Singh. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics, 2005, 21(1): i302-i310

[5] Young-Rae Cho, Woochang Hwang, Murali Ramanathan, and Aidong Zhang. Semantic integration to identify overlapping functional modules in protein interaction networks. BMC Bioinformatics, 2007, 8(265)

[6] Aidong Zhang. Protein Interaction Networks, New York, USA: Cambridge University Press. 2009

[7] Apurv Goel, Simone S. Li, Marc R. Wilkins. Four-dimensional visualisation and analysis of protein–protein interaction networks. Proteomics, July 2011, 11(13):2672–2682

[8] Xiujuan Lei, Xu Huang, Aidong Zhang. Improved Artificial Bee Colony Algorithm and Its Application in Gene and PPI Data Clustering. The IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA2010), Changsha, China, Sep.23-26, 2010, 514-521

[9] Passino K M. Biomimicry of bacterial foraging for distributed optimization and control[J]. IEEE Control Systems Magazine , 2002, 22: 52-67

[10] Dong Hwa Kim, Ajith Abraham, Jae Hoon Cho. A hybrid genetic algorithm and bacterial foraging approach for global optimization. Information Science, 2007: 3918-3937

[11] Swagatam Das, Arijit Biswas, Sambarta Dasgupta and Ajith Abraham. Bacterial Foraging Optimization Algorithm: Theoretical Foundations, Analysis, and Application. Studies in Computational Intelligence. 2009(203):23-55

[12] Radicchi F, Castellano C, Cecconi F. Defining and identifying communities in networks. Proc. Natl. Acad. Sci. USA, 2004, 101(9): 2658-2663

[13] D. S. Modha and W. S. Spangler. Feature Weighting in K-means Clustering. Machine Learning, 2003: 217-237

[14] Xiaoli Li. Biological data mining in protein interaction networks. IGI publishing. 2009

[15] Letovsky S., Kasif S.. Predicting protein function from protein-protein interaction data: a probabilistic approach. BMC Bioinformatics, 2003, 19(6): 197-204

[16] U.Güldener, M. Münsterkōtter, G. Kastenmüller, N. Strack, J. van Helden, C. Lemer, *et al*. CYGD: the comprehensive yeast genome database. Nucleic Acids Research, 2005, 33:D364–D368