PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2012, 39(6): 581~590
www.pibb.ac.cn

A Novel Scale-free Network Construction Method and Its Application in Gene Expression Profiles Simulation*

ZHENG Ming[1)], HUANG Yan-Xin[2)], SHEN Wei[1,3)], ZHONG Yi[1)], WU Jia-Nan[1,4)], LIU Gui-Xia[1)], ZHOU You[1)**]

([1)] *Key Laboratory of Symbol Computation and Knowledge Engieering of The Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun* 130012, China;
[2)] *National Engineering Laboratory for Druggable Gene and Protein Screening, Northeast Normal University, Changchun* 130024, China;
[3)] *College of Computer Science and Technology, Beihua University, Jilin* 132021, China;
[4)] *College of Computer Science and Technology, Changchun University, Changchun* 130012, China)

Abstract In this paper, a novel scale-free network construction algorithm based on reconnection method was proposed. The regulatory node of the new node will be reselected according to the reconnection method. The probability of reconnection depends on the gamma in the power-law distribution model parameters. The constructed network with our algorithm was used for simulating gene expression profiles using differential equation model with two heuristic search algorithms, GA and PSO, and new algorithm GFA to optimize the criterion. The candidate old node can be selected as regulatory node based on the number of links the old node already has. The network in the experiment was testified using log-log graph. And the simulated gene expression profiles were also tested with three different well developed algorithms' software available free from internet by reconstructing the network. PPV and Se of the links were calculated and visualized. A part of the results and the full version program written by java could be downloaded from our website: http://ccst.jlu.edu.cn/CSBG/ourown/.

Key words gene expression profiles, scale-free network, reconnection, differential equation, heuristic search algorithm
DOI: 10.3724/SP.J.1206.2011.00311

1 Introduction

Gene expression profiles yield quantitative and semi-quantitative data on the concentration of protein expressed by the corresponding genes in a specific condition and time, especially ncRNA. One of the high-throughput data applications is to infer or reverse-engineer gene regulatory networks (GRNs) among genes using various mathematical approaches[1]. But various expression profiles could not be enough for testifying the increasing number of inferring algorithms. Simulation profiles could be one of many measures used for the given inference algorithm.

The gene regulatory network is a kind of complex network according to the researches[2]. The reality of the network in biology gene regulatory level could be found according to various experiments[3]. The scale-free property is very common in the gene and

*This work was supported by grants from The National Natural Science Foundation of China(60873146, 60973092, 60903097, 61172183), Hi-Tech Research and Development Program of China (2009AA02Z307), Project of Science and Technology Innovation Platform of Computing and Software Science (985 Engineering), The Key Laboratory for Symbol Computation and Knowledge Engineering of the National Education Ministry of China, Graduate Innovation Fund of Jilin University (20111062) and Natural Science Foundation of Jilin Province (20101503).

**Corresponding author. Tel: 86-18686660880,
E-mail: zyou@jlu.edu.cn

Received: September 19, 2011 Accepted: November 4, 2011

pathway network. The properties of scale-free are contained by the network. Hence, scale-free network can be constructed for generating and simulating gene expression profiles. On the other hand, Internet, social network and other common networks also follow a scale-free power-law distribution[4]. For these reasons, modeling the scale-free network with some suitable methods is one of the most challenging tasks in the complex network research. The existing algorithm includes BA model[5], fitness model[6], other important models. All these scale-free network models have two basic mechanisms. The first one is growth, which means network should be expanded continuously by the addition of new nodes to the existing network. The other is preferential attachment, which means new nodes should attach preferentially to the old nodes that already have many links.

The proposed scale-free construction model defined by our research focuses on the growth part, proposed reconnection method for emphasizing the core role of the hub regulatory genes.

The more important part of our research is the generation of the gene expression profiles. To simulate the suitable values in the profiles, a well fitted function should be used and testified. And the method used for generating profiles should be dynamic. The differential equation (DE)[7], model was used to be the simulation functions. Simulated profiles fit developed models to the network. Fitting adjusts the unknown parameters so that an optimal value for a fitness criterion is ensured. The least square is used as criterion value. And some heuristic search algorithms described later were used to optimize the criterion. The formula of the DE model was perfect for time series profiles data simulation.

As the number of the network and simulation profiles parameters for even medium sized networks may be very large to a computer, a suitable fitting algorithm for underdetermined problems have to be applied. Among different fitting strategies the forward selection fitting algorithm [8] has shown reasonable performance, in particular for sparse networks, which is also the form of our network. Therefore, it has been adopted in our research.

We tested the proposed network algorithm with power-law property and visualized the distribution graph. And we reconstructed the network with gene expression profiles simulated by our method with three kinds of GRN inference algorithms' software.

2 Scale-free network construction

The computer simulation is performed in two major parts. The first part is the construction of the regulatory network, which corresponds to the static part of the simulation since the network is unchangeable during the dynamics. Actually, sparseness of GRNs can be found as a general property[9]. So the most common and important design rule for simulating GRNs is that their topology should be sparse. Sparseness reflects the fact that genes are regulated only by a limited number of genes [10]. Sparseness means that regulatory inputs per gene should be limited, thus a low in-degree is desired. However, some regulatory genes may control a large part of the entire network, thus the out-degree per gene is unrestricted. In the algorithm aspect of the mathematics, enforcing the sparseness property during network construction has the benefit that it significantly reduces the number of model parameters to be estimated and consequently improves the efficiency of network construction and gene expression matrix simulation. Because of the sparseness, scale-free networks grow with connectivity $k=1$ by using the growing network with redirection algorithm [11]. The connectivity corresponds to the number of regulators that a new node can have when it connects to the network. The network is grown as follows.

Another difference in the novel scale-free network construction is that direction property can be found in the network. The initial network condition consists of only one node, or a seed network composed of a small number nodes. Because of $k=1$, a new node is linked to the only one old node. And the old node is the regulator of the new node and the source of the directed line between them. The regulatory node is selected by a special probability selected method. The details will be described later. As the regulatory node is found, the new directed link from the selected regulatory old node to the new node is established with probability r. Or the directed link from another old node calculated basis on the selected old node and the new of node is established with probability $1-r$. We call this approach is reconnection method. The work flow of reconnection method can be described as follows:

(1) If the selected node hasn't ancestor node, the

regulatory node is the selected node itself.

(2)If the selected node has and only has one ancestor (k =1), the regulatory node becomes the ancestor node at this epoch with probability 1−r. If the arrowed line is reconnected, go to (1). Else go to (3).

(3)The regulatory node of the new node according to (2) is the old node itself.

The links in the network are directed arrow line from regulatory node to the target node, which means the regulator node's behavior control the target node's behavior, for instance, the gene expression.

In one hand, the reconnection method is used for emphasizing the hub regulatory node in the network. The phenomenon of various mutations in the genetic world is very common, which means that statistically speaking the nodes with less connectivity have more chances to be the ones that fail. This allows the system to maintain itself very easily since it is still well-connected. The hub regulatory was found as a general model in the gene networks, p53 in the human genome for example [12]. In other hand, the novel method is used because the real networks such as social network, relationship network, especially the gene expression network, have the ability that the control behavior has lag time. It seems that one thing is controlled by the other, but in fact, the mutual information [13] between the two things is 0. Both of them are regulated by the same regulator.

The parameter gamma in the power-law distribution function could be defined by Equation (1) according to the research[14].

$$\gamma = 1 + r^{-1} \tag{1}$$

Where r is the probability that a new node connects to the selected old node. For instance, when γ=3, the network has r=0.5, corresponding to a growth with linear preferential attachment.

The selected method is based on the well connected nodes priority. The probability that one old node is selected as a candidate linked node depends on the number of links the node already had. But the nodes which have less number of links or even have no links also have the change to link to the new one. The selected weight value 1 is evaluated to each node even they have no links. What's more, the selected weight value 2 is evaluated to each link the old nodes have. For example, the old node that has 4 links will be evaluated selected weight value 2×4+1=9. The selected probability p of old node i in the network could be defined as Equation(2):

$$p_i = \frac{2\times n_i + 1}{\sum_i (2\times n_i + 1)} \tag{2}$$

Where n_i is the number of links node i already had. But if the node is the target of the link, the link between the source and this node doesn't count. The denominator of Equation(2) is the total selected weight value of the network before the new node is added.

The flow chart of scale-free network construction procedure is shown as Figure 1.

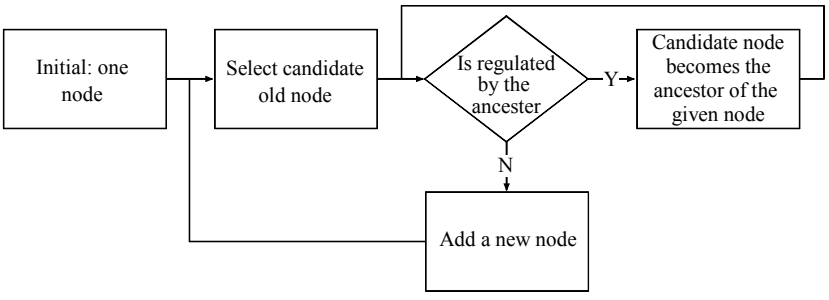

Fig. 1 The flow chart of the network construction part

According to the modeling algorithm details described above, the nodes in the network only know his child nodes. But the nodes don't know which node is their ancestor, this is more sense to the common network. And because of k=1, each node only has one ancestor to simply the network model. The out-degree per gene in the model is unrestricted.

The relation between topological property and the functional property of the control network can be seen from the network. The most important control node of

the network is the node which has the most number of the controlled nodes in the network. Observed directly from the simulated network, this node is the first produced node in the figure. What's more, the relation can be seen that the number of control nodes is very small and the control nodes and the controlled nodes are power-law distribution.

3 Simulation of gene expression profiles

The novel scale-free network construction algorithm proposed in this paper could be used for the simulation of the gene expression profiles. The second part of the computer implements is the simulation of the gene expression profiles.

When the network is composed of N nodes connected by directed links, each node i is assigned an expression variable $v_i(t)$ at time t. The state of the network at time t is represented by a set of expression variables ($v_0(t)$, $v_1(t)$,……, $v_N(t)$). And each node has T time steps, so do the network. At each time step t, the value of one node i is defined as follows: $v_i(t)=F(v_0(0), v_1(0),\cdots\cdots, v_N(0), v0(1), v_1(1),\cdots\cdots, v_N(1), \cdots\cdots, v_0(T-1), v_1(T-1),\cdots\cdots, v_N(T-1))$. The parameters include the node i value itself at time 0, 1,……, $T-1$. This is a classical mathematical Markov problem [15]. But too many parameters make the problem become $N-P$ hard problem. The first-order Markov chain[16] is used in our research to simply the problem. That is, the value of node i at time t, $v_i(t)$ just depends on the value of regulatory node, which may be the node itself, in the network at time $t-1$. And t is greater than or equal to 1 absolutely. When t is 0, random value is used for the node i value.

The quantity relationship between node i at time t and the ancestor of node i at time $t-1$ can be found with many methods. Many simulation functions are testified to simulate gene expression profiles. Finally, in our research, the differential equation (DE) approach was used as our basis simulation model. For simplicity, we considered only systems that are operating near a steady state, so that the dynamics can be approximated by a linear system of a set of ordinary differential equations as Equation(3):

$$\mathrm{d}x_i(t)/\mathrm{d}t = \sum_j^N w_{ij}x_j(t-1)+b_i \qquad (3)$$

Where the $x_i(t)$ is the concentration of the mRNAs at time t that reflect the expression levels of the gene at time t, N is the number of measured genes; the w_{ij} is the coefficient representing the influence of node j on the regulation of node i, with a positive sign indicating activation, a negative sign indicating repression, and a zero indicating no interaction; and the b_i is the constant output observed in the absence of regularity inputs, especially presents the algebraic sum of the external stimulus and noise for gene i.

When the network construction is completed, the construction of the sparseness control weight value matrix W is known. If the directed line from gene j to gene i exists, the value of w_{ij} is non-zero, positive or negative, and if the directed line from gene j to gene i doesn't exist, w_{ij} is zero.

For simplicity and discretization of differential equation, we converted the Equation(3) to Equation(4):

$$x_i(t)-x_j(t-1) = \sum_j^N w_{ij}x_j(t-1)+b_i \qquad (4)$$

In an experiment, we can apply a prescribed random stimulus b_i and use a set of random values to the non-zero w_{ij} for gene i, the value of x_i is what we want. It's the concentrations of all T different measured times. The matrix X can be presented as Equation(5):

$$X_{N\times T} = \begin{pmatrix} x_{11} & \cdots & x_{1T} \\ \vdots & \cdots & \vdots \\ x_{N1} & \cdots & x_{NT} \end{pmatrix} \qquad (5)$$

Where x_{ij} is the value of concentration of gene i at the time j.

The solution space is too big to fit the Equation(4) exactly. So two kinds of heuristic research algorithms, which are Genetic algorithm (GA) [17], and Particle Swarm Optimization (PSO) [18], were used to find the matrix X.

For each gene, least square[19], of Equation(4) can be calculated. And the algebraic sum of all the genes can be used for the fitness for GA and PSO. When the fittest Matrix X, according to the smallest fitness value, was found, or the max number of iteration was reached, the algorithm ends.

From the Equation(3) of DE model, we can see this model is specially used for presenting the dynamic system. The time series is the essence of the model. Though the time complexity is bigger than some other algorithm, DE model allows us to consider more accurate networks and gene expression matrix. Because of increasingly development of the hardware of the computers, larger networks even include hundreds of genes also can be constructed and the relevant gene expression matrix can be exactly

simulated.

4 Experiments

To show the efficiency of the proposed scale-free network construction method and simulation of gene expression profiles approach, we implemented the algorithm with java, and used a series of parameters to investigate the properties of the network and the profiles values and compare the efficiency of our algorithm with other existed traditional algorithms.

4.1 Scale-free Network construction

Artificial scale-free regulatory networks were generated with the parameters in Table 1. In Table 1: nodes indicate the number of vertices in the final generated networks; γ is the exponent value, which was 3 for every node in our experiment, in the degree distribution function for each node; times is the number of time spans; steps means the time span unit; overFlow is the threshold of the gene expression profiles values which are subject to $0 \leqslant x \leqslant$ overFlow; outTimes means the number of time spans in one merge group, which indicate that there are 20 time points in the gene expression matrix in our experiment; signal is the level of the noise added to each value; seed is used to generate the random noise value. If seed = −1, the procedure will use the number of millisecond at current time, otherwise the program will use the seed value directly.

Table 1 The parameters for generating network and simulating gene expression profiles

Parameters	Nodes	γ	Times	Steps	overFlow	outTimes	Signal	Seed
Value	100	3	1000	0.01	50	50	0.05	−1

One of the network construction algorithm results was shown as Figure 2. In Figure 2, the most right node is the node 0. The shape is round and the size is the biggest, the other nodes are clockwise from it. The lines, from the regulatory node to its target node, have rectangle resource part, which indicate the regulatory node. The hollow circle means that the regulator of the node is the node itself. That is, it regulates itself.

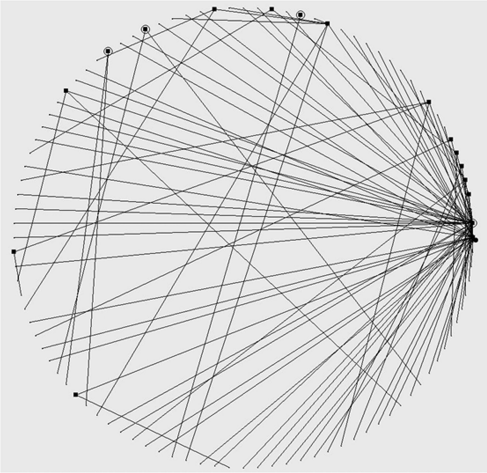

Fig. 2 The graph of the network constructed by the computer program according to the Table 1 parameters

The in-degree for each node was 1 because $k = 1$, but the out-degrees were different, and the values were unrestricted. In our algorithm, the out-degree distribution of the network is power-law. The scale-free network has many nodes with only a few links and a few nodes with many links. So when the number of nodes is big enough, the curve of degree distribution approximated to the axes of the 2-dimention graph. Log-log graph [20] was used to display our distribution curves. Log-log graph use numerical data that are logarithmic scales on both the horizontal and vertical axes. Because of the nonlinear scaling of the axes, a function of the form $y = x^{\gamma}$ will appear as a straight line on the graph, in which γ will be the slope of the line (gradient). We used Matlab to display log-log graph of the out-degree of network of Figure 2. The graph is described as Figure 3.

In Figure 3, Frequency is the number of out-degree, Degrees is the number of node which has Frequency links. The fit curve in Figure 3 is generated by Matlab according to the points in log-log graph, the parameter γ is −3.321, the absolute value of γ approximated to our parameters, which is 3 according to Table 1. Power-law distribution could be found and the efficiency of the reconnection algorithm performs.

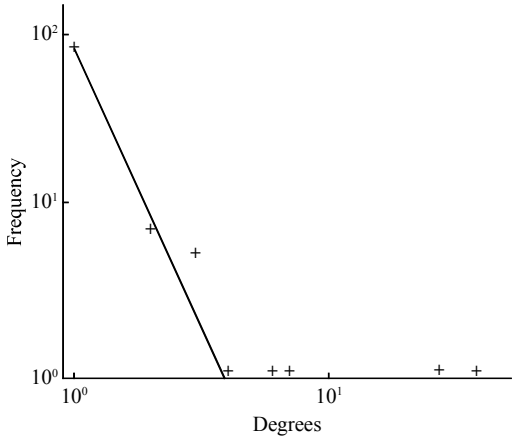

Fig. 3 The log-log graph of degree distribution function generated by computer program according to Figure 2

The out-degree of network and the result network matrix which is discrete to Boolean network [21], generated according to Table 1 can be downloaded from our website. And the matrix file is the one used to generate Figure 3.

The log-log graph of the BA network model with 100 nodes was shown as Figure 4. The corresponding network was shown as Figure 5. Power-law distribution also could be found in BA model. But the slope of the fitted linear function always approximates −1. So the slope can't be selected by the traditional algorithm. And the network is undirected which can't match many genetic networks. The corresponding BA Matlab codes can be downloaded from our website.

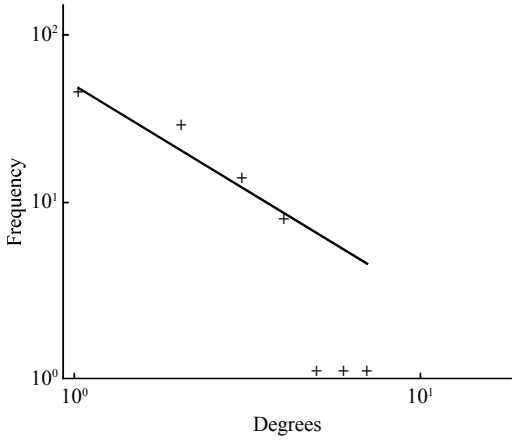

Fig. 4 The log-log graph of degree distribution function generated by BA algorithm

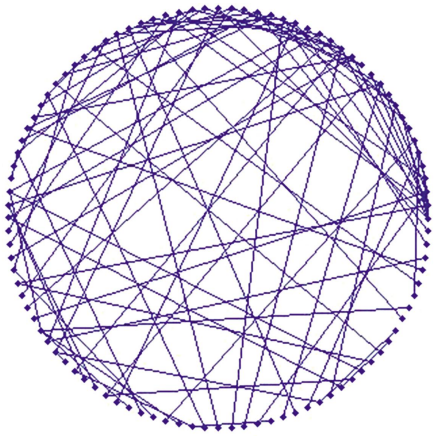

Fig. 5 The undirected network which corresponds to Figure 4 generated by Matlab

4.2 Profiles simulation

The efficiency of gene expression profiles simulation algorithm is the core part of our research. Some gene regulatory network construction methods with corresponding software available free on the internet were used to test the profiles. These algorithms' software includes Banjo [22], ARACNE [23] and NETI[24].

Banjo is gene network inference software that is based on Bayesian networks formalism. ARACNE belongs to the family of information-theoretic approaches. The NETI identifies the gene network as well as the direct targets. It is based on differential equation and is applied when gene expression data are dynamic (time-series).

Run these algorithms' software with our gene expression profiles data and some real data with 50 variables because of lack of biological experiments[25], we can know about the efficiency of our algorithm and also recognize the merits and drawbacks of the 3 kinds of network inference algorithm: Bayesian, mutual information and differential equation. The real data also can be downloaded from our website.

When we reconstructed the networks from the generated time series using the software described above. We will be in hypothesis that the lines generated by these three algorithms were true lines. Then the number of links for True Positives (*TP*), False Positives (*FP*) and False Negatives (*FN*) will be calculated which can be used to be combined to estimate Positive Predictive Value(*PPV*) and Sensitivity value (*Se*) defined as in[1]. And they can be described as Equation(6).

$$PPV = \frac{TP}{TP+FP};\ Se = \frac{TP}{TP+FN} \quad (6)$$

The main configuration settings of Banjo, ARACNE and NETI were shown as Table 2, Table 3 and Table 4.

Table 2 The configuration settings Banjo with our simulation profiles

discretizationPolicy	maxTime	searcherChoice	maxParentCount
q5	10 min	Greedy	5

Table 3 The configuration settings ARACNE with our simulation profiles

MI P value	DPI tolerance	Kernel width	MI threshold
0.05	0.1	0.25	0.06

Table 4 The configuration settings NETI with our simulation profiles

Background function	Kernel function	Times
$10^{-t^{0.9}} + t1.2$	$10^{-t^{0.3}} \times (1+t)0.4$	0.9

In Table 2, discretizationPolicy means the Banjo will discrete our gene expression profiles values to 5 different discrete values, this method will discards many important information about the corresponding genes. Our profiles values, which were from 0 to 50, can't discrete more discrete types because of the restriction of the program. Greedy was selected as heuristic search algorithm in the Banjo because of only two of them can be selected (the other is simulated annealing).

In Table 3, the MI *P* value means Significance level for a MI estimate to be considered statistically different from zero. DPI tolerance means the percentage of MI estimation considered as sampling error. The kernel width of the Gaussian Kernel Estimator (estimates the probability density function of the dataset) affects the smoothness of the density function. Default value is 0.15MI threshold is for a mutual info (MI) estimate to be considered statistical different from zero.

In Table 4, two functions are used as parameters. NETI generalize the weight value matrix and background stimulus matrix by converting differential equations into integral equations with adjustable kernel functions. This method simplified the number of parameters. These functions were selected according to the manual of the NETI.

To test the efficiency of our novel algorithm and assess the different GRN inference algorithms, 100 runs with different links which is from 1 to 100 were performed and *PPV* and *Se* values were calculated. The result curves were presented and shown as Figure 6.

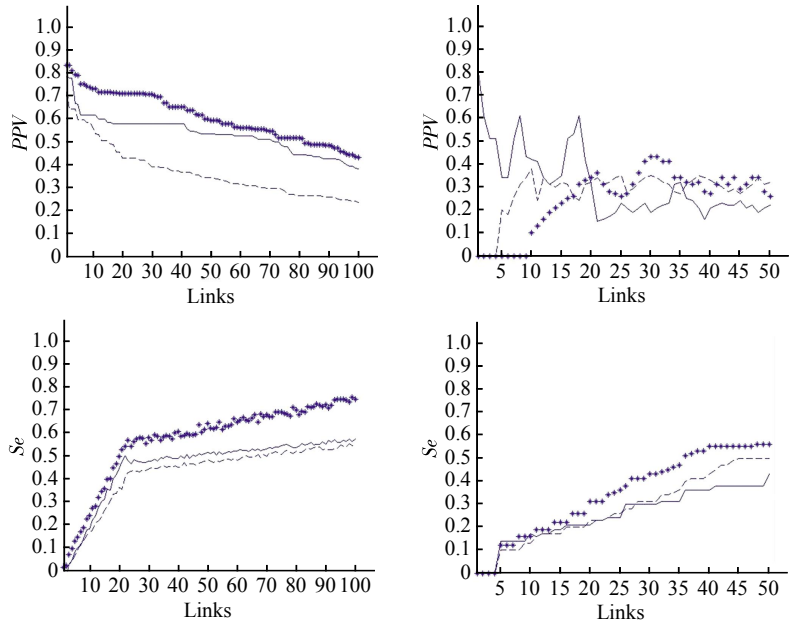

Fig. 6 The *PPV* and *Se* graph of the three algorithms on the total number of links for arbitrary networks and real data

The right graph is our generated data. The left graph is the real data from[25]. The three algorithms is Inference models NETI (* line), ARACNE (dash line) and Banjo (solid line).

We found that our novel simulation profiles algorithm performs well. NETI was superior for our algorithm model to both Banjo and ARACNE, demonstrating higher predictive power and sensitivity. The networks with our algorithm were reconstructed using NETI with greater accuracy (*i.e*. with smaller number of *FP* and *FN*). The fluctuation can be found from the real data. It may be the perturbation of the real gene world. If the number of links is greater than twenty, many of those links were false positives, decreasing the *PPV* values. And the values of *Se* for all algorithms were not increased any more.

If the configuration settings for the three kinds of algorithms were equivalent, the differential equation model is the best one for time series data.

The data for all the three kinds of algorithm in the case of links =100 according to Table 1 is also can be downloaded from our website.

4.3 Software

The developed algorithm software of the proposed method for generating gene expression profiles was implemented with MyEclipse 7.0. The program, named Simulator, which could be downloaded from our website, could run on your own computer with JRE 1.5.0.011 or higher version. JGAP [26], JSwarm-PSO[27] and GFA algorithm[28] were used as our heuristic search algorithms in the software. But JSwarm-PSO could only create the fixed number of genes, so this code was rewritten by our research group for any number of genes. The parameters settings of the GA and PSO were the same as their sample codes on their websites. If you want our algorithm's java codes, you can e-mail the authors.

5 Conclusion

In this paper we proposed a novel scale-free construction algorithm named reconnection method, and used the constructed network, which was large, hub regulatory and sparsely connected on a scale-free property, to simulate the gene expression profiles with differential equation model. The GA and PSO were used to optimize and find the fittest solution of the gene expression profiles. We tested the simulated profiles with Banjo, ARACNE, NETI, and *PPV* and *Se* were calculated with different number of links, and compare the real data.

Our algorithm performed in biochemical network which can be observed from Figure 3～6. Differential equations model is one of the well advanced formalizations in biochemical systems network modeling through the comparison of the algorithms.

Time-series data simulated by our software allow one to investigate the dynamics of activation (inhibition) of genes. These data can be useful to infer the direct molecular mediators (targets) of the regulatory gene in the cell. But network inference from time-series data does not yield acceptable results. Reverse-engineering algorithms need to be improved. One of the reasons for the poor performance of inference algorithms is the smaller amount of information contained in time-series data when compared with steady state data. One way to improve performance in the time-series case is to get more time series gene expression profiles. Our algorithm's software can be used for the data.

References

[1] Bansal M, Belcastro V, Ambesi-Impiombato A, *et al*. How to infer gene networks from expression profiles. Molecular Systems Biology, 2007, **3**(3): 78

[2] Gu X. An Evolutionary model for the origin of modularity in a complex gene network. J Experimental Zoology Part B-Molecular and Developmental Evolution, 2009, **312B**(2): 75−82

[3] 韦晓兰. 细胞黏附分子互作网络的构建与分析. 生物化学与生物物理进展, 2011, **38**(4): 347−352
Wei X L. Prog Biochem Biophys, 2011, **38**(4): 347−352

[4] Tanaka T, Aoyagi T. Weighted scale-free networks with variable power-law exponents. Physica D-Nonlinear Phenomena, 2008, **237**(7): 898−907

[5] Barabasi A L, Albert R. Emergence of scaling in random networks. Science, 1999, **286**(5439): 509−512

[6] Bianconi G, Barabasi A L. Bose-Einstein condensation in complex networks. Physical Review Letters, 2001, **86**(24): 5632−5635

[7] de Jong H. Modeling and simulation of genetic regulatory systems: a literature review. J Computational Biology: A J Computational Molecular Cell Biology, 2002, **9**(1): 67−103

[8] Zheng M, Liu G X, Wang H, *et al*. Gene regulatory network reconstruction of P38 MAPK pathway using ordinary differential equation with linear regression analysis. Advances in Computational Intelligence, 2009, **116**(61): 299−308

[9] Hecker M, Lambeck S, Toepfer S, *et al*. Gene regulatory network inference: Data integration in dynamic models-A review. Biosystems, 2009, **96**(1): 86−103

[10] Quignodon L, Grijota-Martinez C, Compe E, *et al*. A combined approach identifies a limited number of new thyroid hormone target genes in post-natal mouse cerebellum. J Mol Endocrinol, 2007, **39**(1): 17−28

[11] Krapivsky P L, Redner S. Finiteness and fluctuations in growing networks. J Physics a-Mathematical and General, 2002, **35** (45): 9517−9534

[12] Zhi-Gang X, Jian L, Biao Y, *et al*. p53 anti-tumor research in Bel-7402 by using human-derived vector. Prog Biochem Biophys, 2007, **34**(5): 465−470

[13] Biswas A, Guha A. Time series analysis of categorical data using auto-mutual information. J Statistical Planning and Inference, 2009, **139**(9): 3076−3087

[14] Silva A C E, da Silva J K L, Mendes J F F. Scale-free network with Boolean dynamics as a function of connectivity. Physical Review E, 2004, **70**(6): 066140

[15] Malikopoulos A A. Convergence properties of a computational learning model for unknown markov chains. J Dynamic Systems Measurement and Control-Transactions of the Asme, 2009, **131**(4): 041011

[16] Shamshad A, Bawadi M A, Hussin W M A W, *et al*. First and second order Markov chain models for synthetic generation of wind speed time series. Energy, 2005, **30**(5): 693−708

[17] Ribeiro A, Ranz J, Burgos-Artizzu X P, *et al*. An image segmentation based on a genetic algorithm for determining soil coverage by crop residues. Sensors, 2011, **11**(6): 6480−6492

[18] Zhu X J, Shen J, Wang Y L, *et al*. The reconstruction of particle size distributions from dynamic light scattering data using particle swarm optimization techniques with different objective functions. Optics and Laser Technology, 2011, **43**(7): 1128−1137

[19] Pan Z W, Xiao Q W. Least-square regularized regression with non-iid sampling. J Statistical Planning and Inference, 2009, **139**(10): 3579−3587

[20] Fraigniaud P, Lebhar E, Lotker Z. A doubling dimension threshold Theta (log log n) for augmented graph navigability. Algorithms-Esa 2006, Proceedings, 2006, **4168**: 376−386

[21] Tambayong L. Boolean network and dimmelian tie in the Co-author model: A study of dynamics and structure of a strategic alliance model. Advances in Complex Systems, 2011, **14**(1): 1−12

[22] Yu J, Smith V A, Wang P P, *et al*. Advances to Bayesian network inference for generating causal networks from observational biological data. Bioinformatics, 2004, **20**(18): 3594−3603

[23] Basso K, Margolin A A, Stolovitzky G, *et al*. Reverse engineering of regulatory networks in human B cells. Nature Genetics, 2005, **37**(4): 382−390

[24] Novikov E, Barillot E. Regulatory network reconstruction using an integral additive model with flexible kernel functions. Bmc Systems Biology, 2008, **2**(2): 5

[25] Spellman P T, Sherlock G, Zhang M Q, *et al*. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell, 1998, **9**(12): 3273−3297

[26] Chen D Y, Chuang T R, Tsai S C. JGAP: a Java-based graph algorithms platform. Software-Practice & Experience, 2001, **31**(7): 615−635

[27] Brits R, Engelbrecht A P, van den Bergh F. Locating multiple optima using particle swarm optimization. Applied Mathematics and Computation, 2007, **189**(2): 1859−1883

[28] Zheng M, Liu G X, Zhou C G, *et al*. Gravitation field algorithm and its application in gene cluster. Algorithms for Molecular Biology, 2010, **5**: 32

一种新的无标度网络构建方法及其在基因表达谱模拟上的应用 *

郑 明[1)] 黄艳新[2)] 沈 威[1, 3)] 钟 毅[1)] 吴佳楠[1, 4)] 刘桂霞[1)] 周 柚[1)**]

([1)] 吉林大学计算机科学与技术学院，符号计算与知识工程教育部重点实验室，长春 130012；
[2)] 东北师范大学药物基因和蛋白质筛选国家工程实验室, 长春 130024；
[3)] 北华大学计算机科学与技术学院, 吉林 132021；[4)] 长春大学计算机科学与技术学院, 长春 130012)

摘要 本文提出一种新的基于重连接方法的无标度网络构建算法．根据重连接方法新节点的调控节点会被重选，重连接概率取决于幂率分布模型参数 gamma．用本文算法构建的网络通过微分方程模型来模拟基因表达谱数据，所用的优化算法为 GA 与 PSO．候选节点的选择可以根据已有节点的连接数决定．实验的网络可以用 log-log 图，模拟的基因表达谱也用微分方程模型来验证效果．每个连接的正确性将会通过实验验证，完整的程序可以通过我们的官方网站获得：http://ccst.jlu.edu.cn/CSBG/ourown/.

关键词 基因表达谱，无标度网络，重连接，微分方程模型，启发式搜索
学科分类号 Q67，TP18 **DOI**: 10.3724/SP.J.1206.2011.00311

* 国家自然科学基金(60873146, 60973092, 60903097, 61172183), 国家高技术研究发展计划(2009AA02Z307)，计算与软件科技创新平台(985 工程)，教育部符号计算与知识工程国家重点实验室资助项目，吉林大学研究生创新项目(20111062)，吉林省科学基金(20101503)资助项目.
** 通讯联系人. Tel: 18686660880, E-mail: zyou@jlu.edu.cn
收稿日期：2011-09-19，接受日期：2011-11-04