Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Sample Scale-Free Gene Regulatory Network Using Gene Ontology

Guanrao Chen, Peter Larsen, Eyad Almasri, Yang Dai*

Abstract— **Currently there are various approaches to the reconstruction of gene regulatory networks from different sources of data. However, none of these methods incorporates explicitly scale-free property, one of the most important features of the targeted network, into their algorithms. In this paper, several network sampling strategies are explored on a set assembled from previous published gene interactions in yeast, expecting to reconstruct regulatory networks that are scale-free.**

I. Introduction

The reconstruction of gene regulatory networks is one of the most interesting topics in bioinformatics. Many approaches or models were proposed [1] [2] and different sets of data [3] [4] were tested to prove the validity and/or efficiency of these ideas. However, no method so far can be claimed to outperform others mainly because of the intrinsic noisy property of the data and "the curse of dimensionality".

As a network searching problem, if any property of the target network is known, it is beneficial to the reconstruction process. Recently, scale-free is found to be an underlying feature for genetic regulatory networks [5] [6], i.e., the probability of number of nodes having x edges ($p(x)$) follows the power-law distribution or the probability $p(x)=\alpha x^{-\gamma}$. But of all the computational models suggested so far, none of them has used this "universal mathematical law of life [7]" explicitly when algorithms were designed. Therefore, how to incorporate the scale-free property into the network sampling becomes an immediate challenge.

On the other hand, as a data mining problem, the nature and quality of data can in large part determine the quality of the mined network. Most of the data used to derive gene regulatory networks are time series gene expression data. Other information such as protein-protein interaction and/or sequence information have also been incorporated to assist the process [8] [9]. Recently, Gene Ontology (GO) annotation has been extensively used to provide such supporting information in gene clustering [10] [11]. A natural question then is, "How can GO knowledge be used in gene network reconstruction?"

Motivated by the above two questions and the emerging research on network searching or growing methods in reconstruction of gene regulatory networks, this paper (1) explores several network sampling strategies on a set of Gene Ontology annotations derived from previously published gene interactions, and (2) compares and analyzes the computational results. The network structures derived from scale-free searching schemes are more biologically relevant than those derived from threshold-based (restricted) random sampling approaches. Among those scale-free methods, a procedure based on simple K nearest neighbor rule outperforms the "classic" Preferential Attachment strategy.

II. Methods

A. Table of Likelihood of Interaction (LOI) scores

This study utilizes the table of Likelihood of Interaction (LOI) scores for GO annotation pairs, proposed by Larsen *et al.* [12]. The LOI score is a measure of the likelihood that a gene or a gene product with a particular molecular function influences the expression of another gene or a gene product. This likelihood is derived from the analysis of published gene interactions and their molecular functions. More specifically, if two genes closely resemble by their molecular functions from previously observed interaction pairs, then they will be considered likely to interact.

For the derivation of LOI scores, a set of 2457 yeast genes was selected from the *Saccharomyces cerevisiae* database of PathwayAssist® 3.0 and used to identify 4192 directed gene pairs of interaction types "Expression", "Regulation", and "Protein Modification" as defined in that software package [12]. These gene interactions are suggested by 4446 observed interactions from the automated PubMed literature search. The GO annotations (molecular function) of the regulator and the target genes were considered and 5014 pairs of GO annotations were observed. Due to the fact that some gene products have multiple GO annotations, this number exceeds the 4192 interaction pairs in the dataset. The distribution of GO interacting annotations is heterogeneous, dependent on both the number of observations of gene interactions and the frequency of GO annotations in the gene interactions. From this observation, it is necessary to determine if pairs of GO annotations found in gene interactions are at a frequency greater than random, given the distribution of those GO terms in the observed data. To address this question, the 5014 annotation pairs were randomly permutated 10,000

Manuscript received April 3, 2006.

G. Chen is with the Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607 USA. (e-mail: gchen4@uic.edu).

P. Larsen is with the Core Genomics Laboratory at University of Illinois at Chicago, Chicago, University of Illinois at Chicago, 845 W Taylor St Chicago, IL 60607-7058, USA (e-mail: plarsen@uic.edu).

E. Almasri is with the Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607 USA. (e-mail: ealmas1@uic.edu).

*Y. Dai is with the Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607 USA. (phone: 312-413-1487; fax: 312-413-2018; e-mail: yangdai@ uic.edu). Corresponding Author.

1-4244-0033-3/06/$20.00 ©2006 IEEE.

times. For each permutation, 4192 interacting gene pairs were drawn from the set of published gene interactions with a randomly selected gene for the regulator and randomly selected gene for the target. For each permutation, the calculated number of times a specific GO annotation regulated the expression of another specific GO annotation was recorded. An LOI score for each GO annotation was generated as a Z-score:

$$LOI_{ij} = (O_{ij} - X_{ij}) / S_{ij}, \qquad (1)$$

where LOI_{ij} is the Z-score for interacting GO molecular function annotation pair GO_i and GO_j, O_{ij} is the number of times that genes of annotations GO_i and GO_j were observed in regulator to target relationships in the literature-derived dataset. X_{ij} and S_{ij} are the average number of times and standard deviation respectively from resampling procedure for interacting pair of annotations GO_i and $GO_{j.}$ A negative LOI_{ij}-score indicates that a particular GO-annotation pair occurs less frequently than expected by random chance. A positive LOI score indicates an interaction between GO annotations occurs more frequently than expected at random. A score near zero indicates that the frequency occurs at a level near that expected by random.

A subset 798 cell-cycle dependent genes in the budding yeast *S. cerevisiae* microarray experiments was used for evaluation in this study [4]. In order to test the accuracy of the method proposed here, it is necessary to have a "true" expected gene interaction network. However, the "true" network is not known and depends upon the chosen definition of gene interaction. For this study, the "true" interactions were derived from the database of PathwayAssist® by submitting the list of genes and querying for instances published interactions between these genes limited to interaction types "expression" and "regulation". For example, 358 published gene interactions were found for 102 genes related to yeast cell cycle. A total of 1312 gene interactions for the 798-gene set were found. It should be noted that this is not a so-called "golden standard" set for a true evaluation of the learning outcome. Nevertheless, this list of previously published gene interaction pairs can be reasonably considered to be "true" gene interactions in this dataset.

To improve the accuracy and computational efficiency, a modularization approach is taken: the 798 genes are divided into 6 categories based on their biological process annotation as cellcycle, cytoskel, dnametabolism, metabolism, reproduction and transcription. There are overlaps among these 6 groups. More specifically, for every possible interaction-pair from the genes, their annotations were used to assign a LOI score for the likelihood of that interaction from the previously calculated table of GO LOI scores. If a gene possessed multiple annotations, then a LOI score was averaged between all possible pairs of annotations for a given potential interaction pair.

For the "true" interactions (called "truth" tables) of the 6 subnetworks, their histograms of node degree distribution are plotted in figure 1. As can be seen, for both incoming and outgoing connections, the distributions of node degree can be fitted by power-law distribution curves at a level of 95% confidence level. This evidence simply reasserts the underlying feature of scale-free for genetic networks and ultimately motivates the explicit use of this feature for network sampling in this paper.

```
Simple K (SK)
Specify K as the number of nearest neighbors;
For each gene i, make a list of its K nearest neighbors, L(i);
For each pair of gene i and j
    If i is within L(j) and j is within L(i), then set TGRN(i,j) =1;
Compare the resulting TGRN with the "truth" table.
```

B. Algorithms: Preferential Attachment and Simple K

The aim here is to produce some scale-free network explicitly based on a metric defined by the LOI scores. In the seminal paper, Barabasi *et al.* first introduced the "Preferential Attachment" strategy to grow a scale-free network [6]. It is a "rich get richer" procedure in which the "older" or early joined nodes will have more edges than the "younger" or late joined nodes will have. However, several recent papers [13] [14] suggest that for specific type of networks, although they might all show scale-free topology, the exact mechanisms that derive the networks can be quite different. As an alternative, another variation of the Preferential Attachment method was proposed [15]. A simple K nearest neighbor method that is popular in clustering was used in [16].

The methods of Preferential Attachment (PA) and simple K (SK) were tested using the LOI tables. The algorithms involved are listed below. For comparison, the Pure Threshold (PA) in [12] and Restricted Random Sampling (RRS) are also tested. Note that in algorithm RRS no metric

```
Pure Threshold (PT)
Initialize all elements in the Table of Gene Regulatory Network
(TGRN) to be 0;
Select a threshold value TV;
For every two genes i and j
  If the value of LOI (i, j) > TV then set TGRN (i, j) = 1;
Compare the resulting TGRN with the "truth" table.
```

```
Restricted Random Sampling (RRS)
Specify the maximum degree K (incoming and outgoing) for nodes;
Randomly connect pairs of nodes;
For the above network, check each node:
  If its degree > K, then randomly flip off a certain number of its
connected edges (e.g. 1);
  If its degree ==0, then randomly flip on a certain number of its
unconnected edges.
Repeat until the degrees of all nodes are in the range of [1, K];
Compare the resulting TGRN with the "truth" table.
```

```
Preferential Attachment (PA)
Define n as the number of nodes in the graph, m as the number of
edges, and r as the exponent of the power law distribution. Given
(n,m, r) do as follows:
Normalize LOI scores to be in the range of (0, 1];
For each pair of nodes (u, v)
   set $p(u,v) = LOI(u, v)^{-r}$
Repeat: m times
  Sample (u,v) with probability p(u,v)
  Set TGRN(i,j) =1;
Compare the resulting TGRN with the "truth" table.
```

is involved; only node restriction has the impact on the final structure of the network. The results are shown in table 1.

Special care must be taken in dealing with genetic regulatory network in which directionality of edges is intrinsic. Nodes in such network can be categorized into regulators and targets. Correspondingly, edges for a specific node or gene can be grouped as incoming and outgoing. Transcription factors, for example, which initiate the processes of transcription regulation, generally have more outgoing edges than incoming edges while those genes at the end of the regulation chain usually have less outgoing edges than incoming edges.

TABLE 1
SAMPLE RESULTS OF DIFFERENT ALGORITHMS

		Cellcycle	**Cytoskel**	**Dnametab**	**Metabolism**	**Reproduction**	**Transcription**
PT	Threshold	12	-2	8	0	0	-6
	True	25	109	51	19	24	13
	Total	510	3674	1430	1514	1047	957
	Precision	**0.049**	**0.030**	**0.036**	**0.013**	**0.023**	**0.014**
	Accuracy	**0.070**	**0.677**	**0.100**	**0.144**	**0.212**	**0.351**
RRS	Threshold	5	6	18	0	2	16
	True	45	38	50	26	30	13
	Total	1000	1477	1097	2071	1225	730
	Precision	**0.045**	**0.026**	**0.046**	**0.013**	**0.024**	**0.018**
	Accuracy	**0.126**	**0.236**	**0.098**	**0.197**	**0.265**	**0.351**
PA	PI	0.3	0.2	0.1	0.3	0.5	0.7
	PC	0.7	0.3	0.3	0.1	0.3	0.5
	PP	0.5	0.3	0.7	0.7	0.7	0.5
	True	47	28	43	24	56	29
	Total	1454	942	412	2272	2144	1270
	Precision	**0.032**	**0.030**	**0.104**	**0.011**	**0.026**	**0.023**
	Accuracy	**0.131**	**0.174**	**0.084**	**0.182**	**0.496**	**0.784**
SK	K	8	24	12	21	13	30
	True	30	49	52	26	30	18
	Total	216	772	255	886	368	900
	Precision	**0.139**	**0.063**	**0.204**	**0.029**	**0.082**	**0.020**
	Accuracy	**0.084**	**0.304**	**0.102**	**0.197**	**0.265**	**0.486**

True: number of "true" edges, Total: total number of edges in the graph, PI: Percentage of initial nodes, PC: Probability of connection inside the initial network, PP: Probability of connection outside the initial network. PA algorithm showed here is a variant of the listed one [see 15].

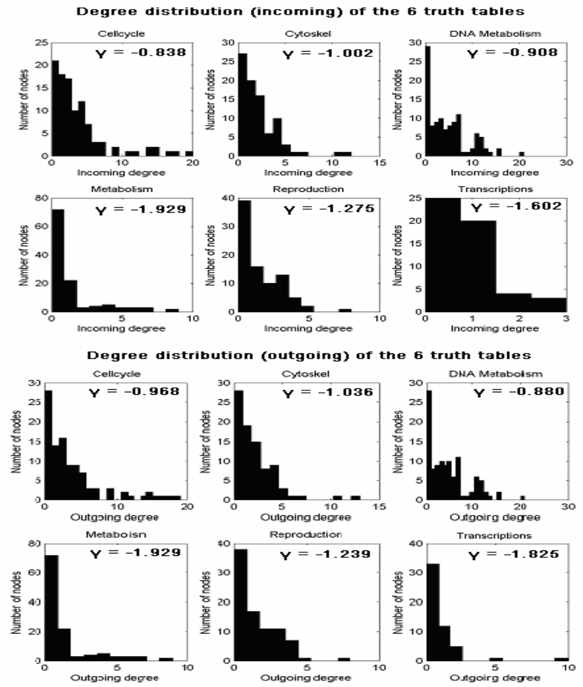

Fig. 1. Degree distribution of the "truth" tables. The values of γ are within 95% confidence bounds.

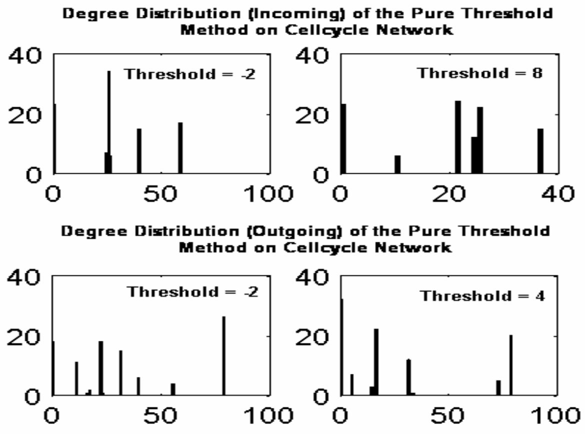

Fig. 2. Example of Degree Distribution of Pure Threshold method.

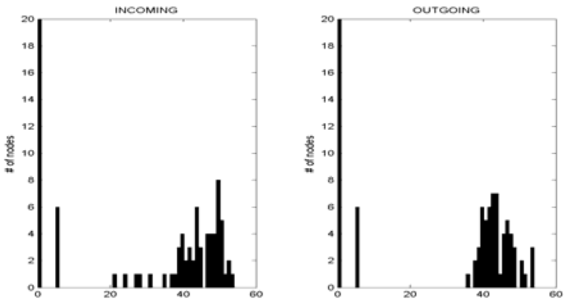

Fig. 3. Example of Degree Distribution by Restricted Random Sampling method on Cytoskel subnetwork.

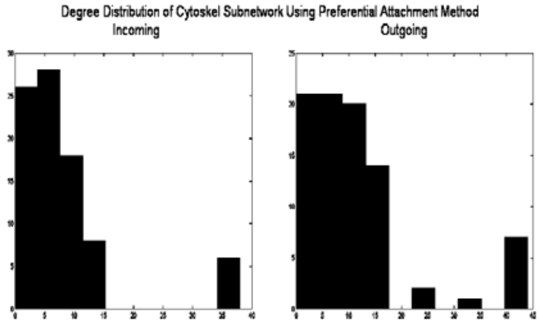

Fig. 4. Example of degree distribution of Cytoskel network using the Preferential Attachement algorithm.

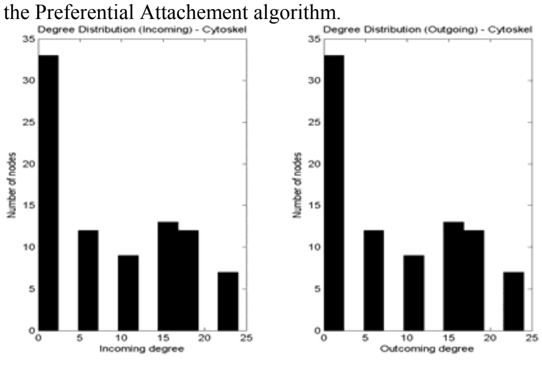

Fig. 5. Degree distribution of Cytoskel subnetwork (Simple K with K = 24)

III. RESULTS AND DISCUSSION

Table 1 provides the computation results: precision and accuracy were summarized. The precision was the primary interest in the study and accuracy values were calculated when the best precision values were obtained. It shows that the SK method outperforms the other 3 methods in precision while improved or maintained a similar accuracy level. For example, the precision increases from 0.049/0.045/0.032 to 0.139 in the Cellcycle subnetwork and from 0.036/0.046/0.104 to 0.204 in the Dnametabolism subnetwork. Particularly, the SK method exceeds the PA method greatly in performance. At first glance, this might be surprising given the simplicity of the former. However, it is very reasonable since in the PA, the selection of edges to be connected is like "half-random" in the sense that edges with high values are more likely to be selected, there is still some randomness involving a specific edge. While in the SK, it is always choose the "legal" highest ranked K edges for each node (those with high LOI scores). Therefore more biological knowledge was utilized

Figure 2 illustrates the edge connection distributions of the Pure Threshold method and figure 3 shows an example of the RRS method. As expected, the distributions do not follow any power-law pattern.

Figure 4 shows the incoming and outgoing degree distributions of Cytoskel network using the PA algorithm. The distribution reveals a scale-free pattern. As indicated by the results in Table 1, the PA algorithm is outperformed by the SK algorithm.

Figure 5 shows the incoming and outgoing degree distributions of Cytoskel network using the SK algorithm when $K = 24$. The distributions follow approximately the scale-free pattern. It can be seen that the degree distributions for both incoming and outgoing edges from the SK algorithm are the same. This is because in the SK, the directions of edges are not considered and the resulting network is symmetric.

IV. CONCLUSION

In this paper, several network sampling algorithms have been explored on a dataset assembled from previous published gene interactions in yeast. Some of them generate networks that resemble little in structure with those underlying "true" networks, that is, the constructed networks show no scale-free pattern at all. While the networks derived from methods using scale-free property do follow the power-law distribution, different algorithms have different performances. Careful analysis explains that the mechanism of some algorithm is more biologically relevant than that of other algorithms. Compared with randomly generated networks, the scale-free approaches improve the performance greatly. It should be noted that this study explores the scale-free network sampling methods using only previously published interactions and no gene expression data were used. Future research include (1) how to incorporate gene expression data into the network sampling procedure and (2) how to incorporate the sampling methods based on power-law distributions in network reconstruction models such as Bayesian networks approaches.

REFERENCES

[1] H. de Jong, "Modeling and Simulation of Genetic Regulatory Systems: A Literature Review", *Journal of Computational Biology*, *Vol.* 9, *No.* 1, 2002: p. 67–103.

[2] Bar-Joseph Z., "Analyzing time series gene expression data". *Bioinformatics*. 2004, Nov. 1; 20(16): p. 2493-503.

[3] R. Cho, *et al.,* "A genome-wide transcriptional analysis of the mitotic cell cycle", *Molecular Cell*, 2, July 1998: p. 65-73.

[4] P. T. Spellman, *et al.*, "Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization", *Molecular Biology of the Cell*, 9, 1998: p. 3273-3297.

[5] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabasi, "The large-scale organization of metabolic networks", *Nature*, (407), 2000: p. 651—654.

[6] A.-L. Barabasi and R. Albert, "Emergence of scaling in random networks", *Science*, *Vol.* 286, 1999: p. 509-512.

[7] Featherstone, D. E. and Broadie, K. "Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network", *Bioessays*, 24, 2002: p. 267-274.

[8] C.H. Yeang, T. Ideker and T. Jaakkola, "Physical network models", *J Comput Biol*, 11(2-3), 2004: p. 243-262.

[9] A.J. Hartemink, D. K. Gifford, T. S. Jaakkola and R. A. Young, "Combining location and expression data for principled discovery of genetic regulatory network models", *PSB*, 2002: p. 437 – 449.

[10] N. Speer, H. Fröhlich, C. Spieth and A. Zell, "Functional Distances for Genes Based on GO Feature Maps and their Application to Clustering", *Proc. Symp. on Comp. Int. in Bioinformatics and Comp. Biology (CIBCB)*, 2005: p. 142-149.

[11] N. Speer, H. Fröhlich and A. Zell, "Functional Grouping of Genes Using Spectral Clustering and Gene Ontology", *Proc. Int. Joint Conf. on Neural Networks (IJCNN)*, 2005: p. 298 – 303.

[12] P. Larsen, G. Chen1, Peter Larsen, E. Almasri and Y. Dai, "Gene Ontology Improves Prediction of Genetic Interactions Based on Bayesian Network Approach", preprint, 2006.

[13] A. Bhan, D. J. Galas and T. G. Dewey, "A duplication growth model of gene expression networks", *Bioinformatics, Vol*. 18, no. 11, 2002: p. 1486-1493.

[14] Yoram L., *et al.*, "Copying nodes vs. Editing links: the source of the difference between genetic regulatory networks and the WWW", *Bioinformatics*, 2006 22(5): p. 581-588.

[15] E.M. Airoldi and K.M. Carley, "Sampling Algorithms for Pure Network Topologies: A Study on the Stability and the Separability of Metric Embeddings", *SIGKDD Explorations: Special Issue on Link Mining*, 7(2), 2005: p. 13-22.

[16] H. Agrawal, "Extreme self-organization in networks constructed from gene expression data", *Phys. Rev. Lett.* 89, 268702 (2002).