Analyzing the Effect of Prior Knowledge in Genetic Regulatory Network Inference

Gustavo Bastos and Katia S. Guimarães

Center of Informatics, Federal University of Pernambuco, Brazil
gbs@cin.ufpe.br, katia@cin.ufpe.br

Abstract. Inferring the metabolic pathways that control the cell cycles is a challenging and difficult task. Its importance in the process of understanding living organisms has motivated the development of several models to infer gene regulatory networks from DNA microarray data. In the last years, many works have been adding biological information to those models to improve the obtained results. In this work, we add prior biological knowledge into a Bayesian Network model with non parametric regression and analyze the effects of such information in the results.

1 Introduction

Gene regulation in eukaryotes is the result of interactions between proteins, genes, metabolites, enhancers, promoters, transcription factors and other biological elements. These interactions control when and with what intensity each gene is expressed in the genome and transcribed into RNA. In order to make it easier to discover the metabolical pathways which schematically describe such interactions, we can focus on just one kind of the elements cited above. For example, we can look only at genes, obtaining a network with gene-gene interactions, which is called a gene regulatory network. To infer the architecture of this type of network from gene expression microarray data is still a challenge in Bioinformatics.

Several mathematical models have been proposed to infer gene regulatory networks from microarray data: Boolean networks [1], differential equations [2], Bayesian networks [3,4], and others. These methods have achieved good results, but they still face hard problems, such as large computational demands and relatively poor quality in the results obtained, in the form of wrong edge directions and gene bypassing. Those problems may be due to the volume of data available to train the networks, usually far less than the proven number of samples needed, which for networks with binary nodes is $O(n^2 \log n^2 \log n^{k+1})$ [5], where n is the total number of genes and k is the maximum input degree of a node.

In order to try to overcome such problems, prior biological information is being added to some models. Hartemink *et al.* [6] have used Bayesian network with simulated annealing and Bayesian Scoring Metric (BSM) to choose the best network, and genomic location information to add prior knowledge to the model. Imoto *et al.* [7] designed a general framework for combining microarray

S.K. Pal et al. (Eds.): PReMI 2005, LNCS 3776, pp. 611–616, 2005.

data with biological information, and Tamada *et al.* [8] expanded this framework using motif detection to improve the model results. Kightley *et al.* [9] used an algorithm based on an epistemic approach and added prior knowledge to the input data.

In this work we perform a careful analysis of the effect that prior knowledge may have in the quality of the results obtained. We chose to use as reference a Bayesian network model and a combination of nonparametric regression with Bayesian Information Criterion (BIC), which was developed to infer the tricarboxylic (TCA) cycle of the *Saccharomices cerevisiae* cell cycle, with relatively good results [10]. Our analysis shows that, while adding prior knowledge actually yields better results to a certain extent, the degree of improvement is more attached to the quality of the information added than to its volume.

2 Network Inference Model

Under the Bayesian network model, nodes represent genes and edges represent relations between genes. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ be a random n-dimensional vector containing the genes to be analyzed, and assume that G is a directed graph. We use a joint density distribution where each gene follows a normal distribution with a β-spline non-parametric model with Gaussian noise. We define the density distribution as:

$$f(x_i|\theta_G) = \prod_{i=1}^{n} \prod_{j=1}^{s} f_i(x_{ij}|\mathbf{p}_{ij}; \theta_i),$$

where x_i is the vector of observations of the ith gene, s is the number of observations of a gene, x_{ij} is the jth observation of the ith gene, $\mathbf{p}_{ij}$ is the observation vector of parent genes, $\theta_G = (\theta_1, \ldots, \theta_n)^T$ is a parameter vector in graph G, and θ_i is a parameter vector in the model f_i, *i.e.*, the model of the ith gene.

In order to choose the network that best reproduces the relations between the genes, given the observations, we use BIC as follows:

$$\begin{aligned} \log p(G|\mathbf{X}) = &- log\, p(L_i) \\ &+ \sum_{i=1}^{n} \sum_{j=1}^{s} \left\{ -log\, f_i(x_{ij}|\mathbf{p}_{ji}, \theta_i) + \frac{d_i}{2} \log n \right\}, \end{aligned} \tag{1}$$

where $p(G|\mathbf{X})$ is the probability of the graph, given the observations $\mathbf{X}$, and d_i is the dimension of θ_i. The chosen final graph is the one which minimizes (1), minimizing each node individually.

We used a 'voting' criterion to choose the edges of the final graph. The program was run a certain number of times and, after that, the edges which had a score above a threshold were selected. This score corresponded to the sum of the two possible relations between the two genes considered. The direction of the edge was chosen as the one which occurred the most. For example, suppose that we ran the program 10 times, and the results were such that in 3 of the

executions, we obtained the edge (Gene_1, Gene_2); in 4 of the executions, we had (Gene_2, Gene_1); (Gene_3, Gene_4) was obtained only once, and (Gene_4, Gene_3) never appeared. According to our criterion, if the threshold is set to 5, (Gene_2, Gene_1) is the only chosen edge, as the relations between those two genes added up to 7, and (Gene_2, Gene_1) appeared more than (Gene_1, Gene_2).

3 Experiments and Results

We used the model described in the previous section to infer the aerobic respiration cycle (TCA cycle) of *Saccharomyces cerevisiae*. According to Hoffgen [5], it would be necessary to have more than 700 samples in order to guarantee estimation of the simplified network architecture. Such number of samples is too large and there are still no databases to provide them. Nevertheless, fairly good results can been obtained from the available microarray data, as shown in this work. In the experiments presented in this section, we used the *alpha* time series consisting of 18 time observations, which is one of the series of Spellman [11], using 600 iterations.

We build our analysis around a Bayesian network model that uses a combination of nonparametric regression with Bayesian Information Criterion (BIC), with no previous knowledge, which was developed to infer the tricarboxylic (TCA) cycle of the *Saccharomices cerevisiae* cell cycle [10]. We call that model Experiment 1, and its result is shown in Figure 1, where the reference network is shown on the left, while the network on the right was inferred by the model with threshold set to 50% of the total number of iterations.

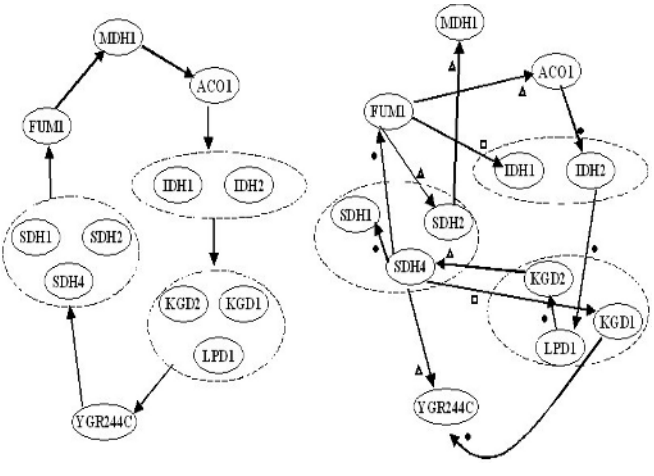

Fig. 1. Partial representation of respiration metabolic pathway of *S. cerevisiae*. On the left, the reference network; on the right, the network generated by the model without prior knowledge.

In all figures in this text, a circle represents a correct edge, while a triangle indicates inverted direction or gene bypassing, and a square represents an extra edge (not in the reference network).

In order to evaluate to which extent prior knowledge information could improve the results of the model, we ran several series of experiments fixing for each one the presence of a certain number of edges. The results in general clearly point to directions that can help one to choose which information to add when building a model.

An important general observation is that while the results almost always present some improvement, that is not highly significant, meaning that the number of samples is still a big factor. Another important result is that the number of edges informed is not as important as the correctness of those. That is, if one incorrect edge is informed, the model may have a performance that is actually worse than that of the model without prior knowledge.

To illustrate what may happen, we present the results of two experiments.

Experiment 2 consisted on adding biological information to the model of Experiment 1 through the preselection of five edges which had to appear in the final graph. We chose three edges corresponding to relations between co-regulated genes: (IDH2, IDH1), (LPD1, KGD1) and (SDH4, SDH2), and two edges corresponding to the beginning and end of the TCA cycle: (MDH1, ACO1) and (FUM1, MDH1).

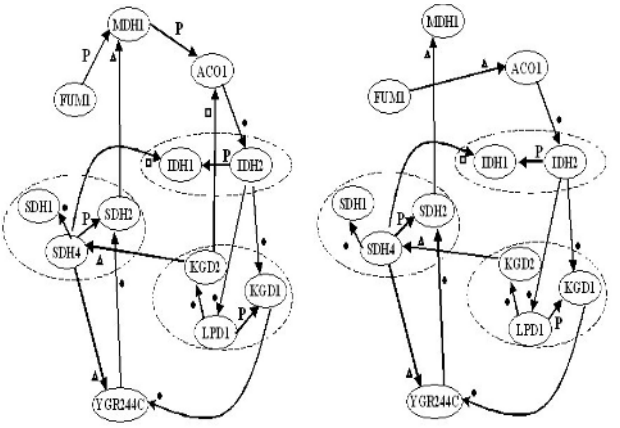

Fig. 2. Respiration metabolic pathway generated by our model. On the left, network inferred in Experiment 2. On the right, network inferred in Experiment 3.

The final network of Experiment 2 is shown in Figure 2, on the left. Circles, triangles and squares in this figure have the same meaning as in Figure 1, and a **P** indicates a preselected edge. The standard name of gene YFR244C, shown in Figure 2, is LSC2, but the systematic name was kept for having been used in Kim's work [12] and others [10]. We can notice that, by adding prior

knowledge, we lost one of the correct edges [(SDH4, FUM1)] found in Experiment 1. However, we detected two other correct edges [(IDH2, KGD1) and (YGR244C, SDH2)] and eliminated two cases of bypassing or inverted edges [(FUM1, ACO1) and (FUM1, SDH2)]. The number of incorrect edges remained the same, although those edges differed: we had (FUM1, IDH1) and (SDH4, KGD1) in Experiment 1, and (SDH4, IDH1) and (KGD2, ACO1) in Experiment 2. This second model had 7 correct edges, 2 incorrect ones, and 3 occurrences of bypassing or inverted direction. One can say that some edges really improved the model's result, specially (LPD1, KGD1) and (SDH4, SDH2).

Experiment 3 consisted on choosing a subset of edges used in Experiment 2, namely: (LPD1, KGD1), (SDH4, SDH2) and (IDH2, IDH1). The result of Experiment 3 is shown in Figure 2, on the right, and is comparable to that of Experiment 2. An incorrect edge [(KGD2, ACO1)] was eliminated, while an occurrence of bypassing was introduced [(FUM1, ACO1)]. The number of correct edges of Experiment 3 was the same as in Experiment 2, in spite of the former's using less prior knowledge.

The experiments in general showed that relations of co-regulation are more effective information than causal relations. One difficulty common to all experiments was the fact that the part of the network next to the external "interference" signal was never correctly rebuilt. That can probably be explained by the large number of nodes (genes) in that region that were ignored in the reference model. That point will require more careful analysis.

More information on the experiments made can be found at [13].

4 Conclusion

Many mathematical models are being used to infer gene networks nowadays but all of them struggle to overcome two main problems, namely: determining the direction of edges and deciding whether a relation between genes is direct or not. These problems mainly arise due to the difficulty in obtaining reliable data in large quantities. The addition of prior knowledge to the models is a promising way of trying to overcome microarray data flaws.

In this work we presented the analysis of the behavior of a Bayesian Network model with non parametric regression and Bayesian Information Criterion in the presence of prior knowledge. Despite the small number of available samples to train the model, the results were encouraging. The experiments showed that relations of co-regulation is more effective information than causal relations. Furthermore, the correctness of the informed edges has a major impact on the results: informing incorrect edges produced worse results than those obtained with no prior knowledge at all.

As future work, the authors will add to the model more prior information about the beginning and end of the TCA cycle (upper part of the network shown in Figure 2), as this region presents the smallest number of correct guesses of the model. Experiments with other metabolic pathways are also being done.

References

1. Akutsu, T., Miyano, S., Kuhara, S.: Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. In: Proc. of the Pacific Symp. on Biocomputing. Number 4 (1999) 17–28
2. Chen, T., He, H.L., Kuhara, S.: Modeling gene expression with differential equations. In: Proc. of the Pacific Symp. on Biocomputing. Number 4 (1999) 29–40
3. Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using Bayesian Networks to Analyze Expression Data. In: Proc. of the 4th Annual Intern. Confer. on Computational Molecular Biology, Tokyo, Japan, ACM, ACM Press (2000) 127–135
4. Imoto, S., Goto, T., Miyano, S.: Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. In: Proc. of the Pacific Symp. on Biocomputing. Number 7 (2002) 175–186
5. Hoffgen, K.: Learning and robust learning of product distributions. In Pitt, L., ed.: Sixth Annual Workshop on Computational Learning Theory, New York, NY, ACM Press (1993) 77–83
6. Hartemink, A.J., Gifford, D.K., Jaakkola, T.S., Young, R.A.: Combining Location and Expression Data for Principled Discovery of Genetic Regulatory Network Models. In: Proc. of the Pacific Symp. on Biocomputing. Number 7 (2002) 437–449
7. Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S., Miyano, S.: Combining Microarrays and Biological Knowledge for Estimating Gene Networks via Bayesian Networks. In: Proc. 2nd Computational Systems Bioinformatics. (2003) 104–113
8. Tamada, Y., Kim, S., Bannai, H., Imoto, S., Tashiro, K., Kuhara, S., Miyano, S.: Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. BioInformatics **19**(Suppl. 2) (2003) ii227–ii236
9. Kightley, D., Chandra, N., Elliston, K.: Inferring Gene Regulatory Networks from Raw Data: A Molecular Epistemics Approach. In: Proc. of the Pacific Symp. on Biocomputing. Number 9 (2004) 510–520 Website: http://helix-web.stanford.edu/psb04/.
10. Bastos, G., Guimarães, K.S.: A Simpler Bayesian Network Model for Genetic Regulatory Network Inference. In: Proc. of the Intern. Joint Conference on Neural Networks 2005 (IJCNN'05), Montréal, Québec, Canada (2005) To be published.
11. Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., Futcher, B.: Comprehensive Identification of Cell Cycle-regulated Genes of Yeast *Saccharomyces cerevisiae* by Microarray Hybridization. Molecular Biology of the Cell **9** (1998) 3273–3297
12. Kim, S., Imoto, S., Miyano, S.: Dynamic Bayesian Network and Nonparametric Regression for Nonlinear Modeling of Gene Networks from Time Series Gene Expression Data. In Priami, C., ed.: Proc. of First Computational Methods in Systems Biology (CMSB). (Number 2602)
13. BioLab: Bioinformatics Laboratory at CIn/UFPE. Website: http://biolab.cin.ufpe.br/tools/ (2000)