Accuracy of the Models for Gene Regulation — a Comparison of Two Modeling Methods

Kimmo Palin[*]

Keywords: Gene regulatory network, gene disruption network, minimax graph, REVEAL algorithm, Floyd–Warshall algorithm.

Introduction

Many interacting genes control the gene expression resulting in a complex regulation structure called *the gene regulatory network* (GRN). Not much is known about the GRN as a whole but we can observe its effects by microarrays. The microarray experiments measure the expression of all the genes of a genome under various experimental conditions.

This study compares two different GRN modeling methods to see whether one of them gives better results in the perspective of studying biological gene regulation. We infer the GRN models by microarray measurements in time series and in gene disruption mutants [3, 2]. The two methods approach the "real" GRN from two different directions: on one hand by removing edges from a highly connected network and on the other hand by adding edges into an empty network. These two methods result in different networks thus giving rise to a problem of comparing them.

Concepts

We see the GRN as a graph that qualitatively represents the regulatory interactions within the genome. The GRN has a node for each gene and a directed edge between two genes if *the product of the source gene physically interacts with the expression machinery of the target gene.* Thus there are outgoing edges for transcription initiation and elongation factors, translation factors, interfering RNAs and other factors directly effecting the RNA expression; but not for mediated effects involving a third gene.

Our first method for inferring the GRN uses *gene disruption networks* (GDN) [8, 6] as a starting point for the inference. A GDN has a node for each gene as does the GRN but it has an edge if *the disruption of the source gene significantly alters the expression of the target gene.*

We use the GDN because of the difficulty of directly finding the edges of the GRN. In contrast, we can find the outgoing edges for a gene in the GDN by standard laboratory procedures involving a gene disruption and a microarray measurement.

Minimax Method

If one takes a transitive reduction of the GDN [8] one most certainly stands with more edges than in the GRN. This is because some of the mediating genes probably lack an edge due to random noise. Here we present a method to avoid the problem when we have a measure of "directness" of the edges in the GDN.

We measure the directness of an edge by a cost function c. The function c should give a low cost for edges in the GRN and a high cost for edges whose effect some third gene mediates. Since the function c contains much of the information used in the pruning of the GDN, the choice of c is crucial for the method's success. Therefore the practitioner should take great care in developing a sound and robust cost function c.

[*]Department of Computer Science, P.O. Box 26 (Teollisuuskatu 23) FIN-00014 UNIVERSITY OF HELSINKI

Given an edge $i \rightarrow j$ in the GDN, there must exist a regulatory path $i \rightsquigarrow j = i \rightarrow x \rightarrow \cdots \rightarrow j$ in the GRN. This is clear because we must have a regulatory path such that the disruption of i effects the expression of j. Otherwise there would not be an edge $i \rightarrow j$ in the GDN.

Now we argue that a mediated edge in the GDN results from a regulatory path that contains the most likely direct edges. This means that, for each edge $i \rightarrow j$ in the GDN, we need a path $i \rightsquigarrow j$ in the GRN, such that the most expensive edge on the path is as cheap as possible according to c.

The paths that minimize the maximum edge cost between two nodes of the GDN form a minimax graph. We compute the pairwise minimax paths in time $O(n^3)$ with the Floyd–Warshall algorithm applied on the closed semiring $(\min, \max)$ [1].

We can use one of the standard expression distance metrics as a simple way of computing $c(i \rightarrow j)$. To this end, we use the cosine correlation metric as our first cost function.

We obtain a bit more intricate method by assuming that the genes i and j are closely related if their expression codes for the same information. Thus for the second cost function we use their mutual information. As the probability model for the expression we assume the Bernoulli distribution with the parameter approximated from the data.

Our third cost function uses also the information about the genes' regulatory regions and the transcription factors' binding sites. We compute the cost $c(i \rightarrow j)$ by adding the log odds of the TF binding motif i in the upstream j to the log odds of the differential expression of j given i versus not given i.

Adapted REVEAL

Our second notably different model for the GRN adds new genes into an initially empty network. The algorithm REVEAL [4] uses this approach by selecting the optimal set of regulators for each individual gene.

The REVEAL algorithm has a major problem with the assumption about noiseless measurements. Under this assumption REVEAL finds the smallest set of regulators whose mutual information equals the regulated gene's entropy. This is a problem in two ways: first, if there is noise in the expression measurements, adding more regulators always increases the mutual information; second, if there is only modest amount of data available, there can be several regulator sets explaining the expression behavior.

We adapt the REVEAL for more realistic data with two alterations. First, we limit the possible regulators to the ones whose disruption effects the expression of the gene in question, and second, we select the set of regulators minimizing the Akaike Information Criterion (AIC) [7] instead of maximizing the mutual information.

Results

The three different edge cost functions result in different outputs of the algorithm [5]. Both the cosine correlation and the mutual information costs give roughly the same number of edges but just over half of them are present in both networks. The network generated with sequence based pricing is significantly smaller than the others due to the lack of binding site information for some of the genes present in other networks. Note that only less than one fourth of the edges after pruning are present in all three networks. These numbers suggest that these metrics are unable to obtain strong and coherent signal to discriminate between direct and transitive edges.

The modified REVEAL algorithm results in yet another candidate as GRN. As seen on table 1 it has a quite narrow in-degree distribution. This is biologically unexpected as the GRNs are generally believed to be scale free [6] thus the nodes' degree distribution should follow the power law. The power law should give a heavy tail toward large numbers. One observes this phenomenon for example in the purged networks' degree distributions and also in the REVEAL network's out–degree distribution [5].

The explanation for the REVEAL network's unexpected topology lies in the AIC–criterion. The criterion depends on the size of the data available and the number of parameters in the model. The number of parameters in the REVEAL network grows exponentially so any realistically sized dataset looks tiny even with a fairly simple network.

Table 1: Distribution of the number of regulator and regulated genes in graph by REVEAL

Nodes with Edges	10%	50%	Mean	90%	100%
In-degree	1	2	2.4	4	4
Out-degree	1	2	4.3	10	34

The four generated GRN candidates contain only two common edges (ECM18 → RAD6 and ERG4 → UBR1, neither look biologically sound). This results might be expected but nevertheless disappointing.

Discussion

It is difficult to infer GRN from large scale, non–targeted, measurements. The inherent difficulty rises from the complex model of networked interactions. Inferring or even verifying the complex interactions requires a lot of experimental evidence. The amount of data needed is not currently available. The future of GRN modeling will most likely be based on small advances in well planned and well targeted experiments, not on overly large scale explorative analysis.

It is also difficult to evaluate the goodness of proposed GRNs. We lack theoretically or practically satisfactory methods to compare the goodness of the GRN candidates to one another or to somehow biologically correct network.

One of the main reasons for the difficulty of dealing with GRNs is that the assumption of simplicity does not hold. This feature renders the traditional model selection methods useless. As result we need either to develop new methods for evaluating the biological significance of our results or to restrict our models to small and relatively well known subsystems such as the cell cycle.

Acknowledgments

I would like to thank Esko Ukkonen and Juho Rousu for many helpful comments.

References

[1] T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to Algorithms.* The MIT Press and McGraw-Hill Book Company, 1989.

[2] J. Gollub, C. A. Ball, G. Binkley, J. Demeter, D. B. Finkelstein, J. M. Hebert, T. Hernandez-Boussard, H. Jin, M. Kaloper, J. C. Matese, M. Schroeder, P. O. Brown, D. Botstein, and G. Sherlock. The Stanford Microarray Database: data access and quality assessment tools. *Nucl. Acids. Res.*, 31(1):94–96, 2003.

[3] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–126, Jul 2000.

[4] S. Liang, S. Fuhrman, and R. Somogyi. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. *Pac Symp Biocomput*, pages 18–29, 1998.

[5] K. Palin. Geenisäätelyverkkojen mallintaminen keskeytysmutanttien ja aikasarjojen ekspressiomittauksista. Technical Report C–2003–5, University Of Helsinki, Department of Computer Science, February 2003. In Finnish. Title: Modelling Gene Regulatory Networks by Expression Measurements in Disruption Mutants and in Timeseries.

[6] J. Rung, T. Schlitt, A. Brazma, K. Freivalds, and J. Vilo. Building and analysing genome-wide gene disruption networks. *Bioinformatics*, 18(90002):202S–210, 2002.

[7] T. Van Allen and R. Greiner. Model selection criteria for learning belief nets: An empirical comparison. In *Proc. 17th International Conf. on Machine Learning*, pages 1047–1054. Morgan Kaufmann, San Francisco, CA, 2000.

[8] A. Wagner. How to reconstruct a large genetic network from n gene perturbations in fewer than n^2 easy steps. *Bioinformatics*, 17(12):1183–1197, Dec 2001.