INFERRING GENE REGULATORY NETWORKS FROM RAW DATA – A MOLECULAR EPISTEMICS APPROACH

D. A. KIGHTLEY, N. CHANDRA AND K. ELLISTON

Genstruct Inc., 125 Cambridgepark Drive,

Cambridge, MA 01702, USA

Biopathways play an important role in the functional understanding and interpretation of gene function. In this paper we present the results of an iterative algorithm for automatically generating gene regulatory networks from raw data. The algorithm is based on an epistemics approach of conjecture (hypothesis formation) and refutation (hypothesis testing). These operations are performed on a matrix representation of the gene network. Our approach also provides a way of incorporating external biological knowledge into the model. This is done by pre-assigning portions of the matrix - which represent previously known background knowledge. This background knowledge helps make the results closer to a human's rendition of such networks. We illustrate our approach by having the computer replicate a gene regulatory network generated by human scientists at an academic lab.

1 Introduction

Gene regulation in eukaryotes is the result of a complex interaction of numerous elements that combine to determine the expression of genes. The bindings of multiple transcription factors at cis-regulatory sites act in combination to determine the level of gene transcription. Discovering the nature of these interactions remains a challenging problem. Elucidation of the regulatory network architecture from a set of experimental data is a complex problem and development of an automated process can help in generating networks that are too large and too complex for humans to handle.

Algorithms for automatically generating a genetic regulatory network have been used on a number of different data types. Microarrays [5] give a measure of levels of gene expression in a cell and these data have been used to generate the underlying genetic network [17]. However, the cost of analysis of each interaction in the network is high. The complete set of data is rarely produced and data are frequently sparse. As a result, network inference algorithms are typically applied for recreating complex functional network structures from limited datasets [11, 13, 15].

A different technique measures changes in mRNA transcription of various target genes, measured by PCR, when another gene is perturbed. These perturbation studies [8, 10] can yield information as to which genes are regulated, either directly or indirectly, by another. Thus by combining the interactions it is possible to build up a regulatory network. However, an effect can be the result of a direct interaction or an indirect action through intermediate genes. Therefore, it is necessary to incorporate prior knowledge of the system to infer the network structure; a Bayesian network has been used for this purpose [14, 16].

An alternative approach for generating gene regulatory networks has been to use reverse engineering of data using generative algorithms [6, 7, 12]. This approach starts with a set of observations and generates networks that approximate the solution. Through modification and refinement the network that best explains the data is arrived upon (see Section 3.1).

2 Gene Perturbation Data

2.1 Source of the Data

The data relating to gene regulation of purple sea urchin (*Strongylocentrotus purpuratus*) embryo development has been made available on the Internet [2], from where the data was transcribed. Figure 1 is a sample of the data giving the effects on two transcription factors out of a total of 60 genes. The dataset relates to experiments performed at the Davidson Laboratory at the California Institute of Technology that involved quantitative PCR studies on embryos during the early stages of development (< 72 hr). Details of the findings from the studies have been published [3].

2.2 Gene Perturbation

The experiments performed on the Sea Urchin embryos involved perturbation of genes and measurement of changes in expression of a second, target gene. In the absence of other influences, perturbation of a gene that is an activator of another will cause the expression of the second gene to be decreased. Alternatively, if the perturbed gene is inhibitory, the expression level of the latter will be increased.

The numerical values refer to the cycle number in the PCR experiment and this relates back to the starting level of mRNA, which is amplified exponentially during PCR. A value of 1 represents an approximate doubling of initial mRNA level. Thus, if a value of 3 is reported for an interaction, perturbation of the gene resulted in an 8 fold increase in the gene product compared with the unchanged cell. The convention used in the data is that negative values mean less starting mRNA. Thus, if perturbation of a gene results in lower quantities of mRNA transcribed from target genes, the relationship must have been activation. Similarly positive values indicate inhibition.

Transcription regulation involves a complex network of genes that encode transcription factors which, in turn, regulate other genes. A specific transcription factor can regulate multiple genes and there are chains of interactions which form a cascade. Thus perturbation of a single gene can affect the expression of many other genes both directly and indirectly. Consequently, an observed change in gene expression is the result of the combined effects on all of the regulatory genes that influence its transcription. Being able to determine whether an interaction is direct or indirect is a hurdle in deciphering causality in gene regulatory networks.

2.3 A look at the data from the Davidson lab

The experimenters presented data relating to three types of perturbations:

- Morpholino-subsituted antisense oligonucleotide (MASO) - the mRNA transcribed from a gene binds to the complimentary RNA strand, thereby preventing translation of the gene product.
- Messenger RNA overexpression (MOE) - involves amplification of gene products from the perturbed gene.
- Engrailed repressor domain fusion (En) - the transcription factor is converted into a form in which it becomes the dominant repressor of all target genes.

The three techniques represent distinctly different methods for gene perturbation. However we do not have enough details on them to determine whether there are any useful differences in the results. Therefore, no distinction between techniques was made, results having been taken as being equivalent, and data for the same perturbation, but from different experimental techniques, were combined.

The results for each perturbation experiment were reported as up to 7 individual values that relate to both replicate measurements of the same cDNA batch and separate experiments. These values were averaged to provide a single value for equivalent samples. Results recorded as Not Significant (NS) were treated as zero.

Gene	Perturbation	12-16 h	18-21 h	24-27 h	30-36 h	41-48 h	60-72 h	Data of:
gatac	Cad MOE[31]		-2.4/NS	-6.4/-3.6/-4.9				A. Ransick, T. Minokawa & C. Livi
	Elk-En			-6.0, -5.2, -5.7				M. Arnone
	Otx-En[21]			-2.1/-3.3				A. Ransick & T. Minokawa
	GataE MASO[14]		NS/NS/NS/NS/ NS	-2.2, -2.5/NS, NS/-2.4, -2.5				P. Y. Lee
	Dri MASO[16]		-2.1/-2.3		NS/NS			G. Amore
	N MASO[30]		-2.6		NS (29 h)	NS		C. Calestani
	GataC MASO			+3.1, +3.7/NS				P. Oliveri
	Hnf6 MASO		-2.5/-3.0					G. Amore & O. Otim
	Gcm MASO[32]		-2.2	-2.5				A. Ransick
gatae	Cad MOE	NS/NS	-5.1/-3.1	-4.7/-3.6/-3.1				A. Ransick, T. Minokawa & C. Livi
	DnN MOE		-2.4/-1.4					C. Calestani
	Otx-En		-3.3, -2.2/-2.8, -3.0/-3.2	NS/-1.5, -2.4/ -2.4/-4.1/-2.4				A. Ransick & T. Minokawa
	Hox11/13b MASO	+1.6		+2.4/+1.7				C. Arenas-Mena
	Soxb1 MOE	-1.7, -3.2/ -2.0, -2.1[25]						J. Rast & K. Young
	Krl MASO		-2.1/NS, -2.0/ -2.6, -3.7[23]	NS, NS/+4.4, +2.2				J. Rast & K. Young
	FoxA MASO					-2.8, -1.6		K. Young & P. Oliveri
	Hox11/13b MASO			+2.4/+1.7		NS	NS	C. Arenas-Mena
	Hnf6 MASO				NS/NS	-2.1/-2.6		G. Amore & O. Otim

Figure 1. A sample of the data presented on the Davidson Lab website. This portion of the data relates to perturbation of multiple genes and the effect on the transcription factors, *GataC* and *GataE*.

The original data used ±1.6 as the significance threshold. However by treating non-significant samples as zero, time-averaged samples were reduced in value so a

lower threshold was needed. After analysis of the data, values that fell below ±0.75 were taken to indicate no significant interaction.

Data are presented as a set of time slices that cover intervals in embryo development between 12 and 72 hours after fertilization. However, most data are for three time slices between 12 and 28 hr and the remaining information is very sparse. For the majority of the work, mean values for the first 4 ranges were combined to yield an average across these times.

In addition to gene perturbation results, there is a table of genes that are not affected by perturbation during the first 24 hrs and also footnotes that provide information about gene interactions, many highlighting possible indirect effects. This additional information was incorporated into the experimental data to yield a single value for the effect of one gene on another. Data were available for only around 12.8% (460 out of 3600) of the possible interactions. Some of the remainder may be filled in by future experimentation but, for the purpose of this analysis, these 'unknowns' were taken to indicate no interaction unless there were indications to the contrary.

2.4 Gene Selection

The overall dataset contained 60 genes identified to regulate gene expression in Sea Urchin embryos. To simplify the system, a decision to concentrate on the Endomesoderm was made since there was the greatest quantity of data relating to these cells. The remainder of the embryonic regions had considerably less experimental coverage. Twenty-one regulatory genes are active in the Sea Urchin endomesoderm during the chosen developmental stages and, of the 441 possible interactions, there are 162 data points or 36.7% coverage.

In addition to the 21 genes, the published endomesoderm regulatory network also includes complexes (e.g. Su(H)-N^{IC}, n-TCF) involving endomesoderm gene products. However, no data were presented that supported the formation of these complexes, nor was there any data for their action within the cell. Therefore, complexes were omitted from the analysis.

3 Algorithm for Network analysis

3.1 The flowchart

The algorithm used is based on exploring the state space of all possible gene networks (models) in a systematic, iterative fashion. The first step involves generating a model from a given set of components. The components for the gene network are:

- An activation
- An inhibition
- No effect

These three relations between genes are represented as +1, -1 and 0 in a matrix of gene-to-gene interactions. The initial model generated represents a hypothesis that has to be tested and scored. The next step involves simulation. The model, which represents a set of regulatory connections between genes, can be simulated qualitatively. For example, the network contains the relation: A activates B which activates C. The experimental data are checked to see what experiments have been done. Assume that one of the experiments involved overexpressing A then, according to our hypothesized model, an overactivation of A will result in an increase in B and C. The results of the simulation are tested against the actual data. As indicated below, the actual data will show that B increases and C decreases.

This comparison is then used to score the model. The model is then modified using a state space search algorithm to create a new model. The process is followed iteratively till the score does not improve any more. To avoid local minima, the modified models are randomly perturbed using an annealing method.

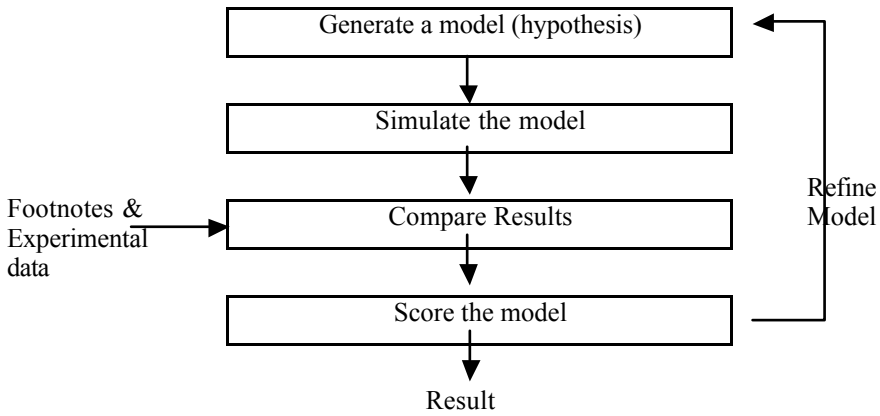

Figure 2. Molecular Epistemics Algorithm of Conjecture and Refutation.

3.2 Handling non-numerical biological knowledge outside the raw data

The process of scientific discovery involves experimentation, but interpretation of the results involves bringing to bear ones prior knowledge of the underlying biology. Our approach allows for outside literature, footnotes and personal knowledge to be added to the model before it runs. This is achieved in two ways. The first approach is to incorporate externally known regulatory knowledge into the input data prior to running the algorithm. Another approach involves incorporating the known prior knowledge into the initial model. The idea here is to make some of the gene-to-gene connections 'fixed' or pre-set before the model generation process is started. If this cannot be done for all the knowledge, it can be incorporated into the scoring algorithm [1].

4 Endomesoderm Gene Regulatory Networks

4.1 Representation of the Regulatory Networks

Networks generated by the algorithm were displayed graphically using Netbuilder, a tool for construction of computation models developed by Science and Technology Research Centre, University of Hertfordshire, UK. This tool was also used by the Davidson Lab team to display their network results. The colors and overall network layout presented here were chosen to closely resemble those used in the Davidson paper and so make for easier comparison.

4.2 The Complete Regulatory Network

By using a straight substitution of the data with values greater than or equal to the threshold taken to mean activation or inhibition depending on the sign, and all other values to signify no connection, a simple representation of the entire network of connections was obtained (Figure 3). This interpretation takes into account the additional information provided in the footnotes to the data (incorporated into the values), but is doing no interpretation or analysis of the data. The generated network comprises 56 links between the genes of which 45 were activations and 11 inhibitions.

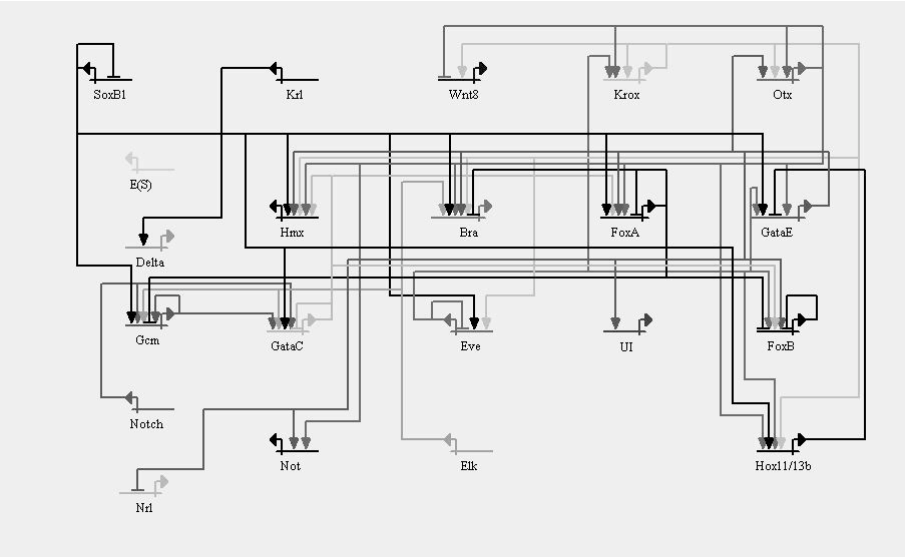

Figure 3: Automatically generated Endomesoderm gene regulatory network that directly reflects the raw data.

The complete network generated directly from the data is similar to the Endomesoderm network published by Davidson, however there are some notable differences which may not be related directly to interpretation of the information. Firstly, the data available on the website is constantly under review and is augmented as new results become available. The dataset used in this study was dated October 28th, 2002 and so was considerably newer than that used to construct the network for the article that appeared in the March 1 issue of Science [3]. Although the network displayed on the website is also being updated, it is changed less frequently than the data and may not reflect all the updates. Secondly, the Davidson Lab's network represents the regulatory network for the organism and includes many genes that are not active in the endomesoderm. These genes will have interactions with the 21 genes under study which may have effects that are not apparent when the endomesoderm is viewed in isolation.

Nevertheless, there are still discrepancies. Some links are present on the published network even though the dataset indicates they should not be there. For instance, there are data to suggest an activation link between *bra* and *nrl*, however a footnote states that this must be an indirect link since *bra* is not active in the cell at this time. The data used for this work took all of the footnotes into account and so does not show this link, whereas the published network included it. On the other hand, there is data to support an activation link between *eve* and four other genes, yet the published networks only show a single effect. Thus, while these networks and the Davidson Lab published networks show similar information, they show some differences which are, at least partly, due to differences in the source data.

4.3 Network reduction

The scoring mechanism in the underlying algorithm was modified to give a low score to links that can be explained by intermediate genes. This was done to remove indirect links – thereby generating a minimal network that explained the raw data faithfully. For instance, *elk, Sox_1* and *Notch* all activate both *GataC* and *gcm*, and *gcm* activates *GataC* (Figure 3). Therefore, it is possible that the observed effects on *GataC* were really a result of an indirect effect through *gcm*. This suggests that the three links from *elk, Sox_1* and *Notch* to *GataC* could be removed without contradicting information contained in the data.

By eliminating the maximum number of links without breaking any of the connections between genes or making a link with too many intermediates, it was possible to remove 13 links from the network (all activations) and reduce the total number of links from 56 to 43 (Figure 4). In separate runs of the algorithm it was possible to get slightly different sets of links removed, but the minimum number of links necessary to explain all of the data was still 43.

The algorithm was also run in a configuration that permitted the removal of links that can be explained through pathways of up to 2 intermediate genes. In this way 3 extra edges could be removed, however the more intermediates there are the harder it is to justify that the link has been retained and the observed effect is still valid.

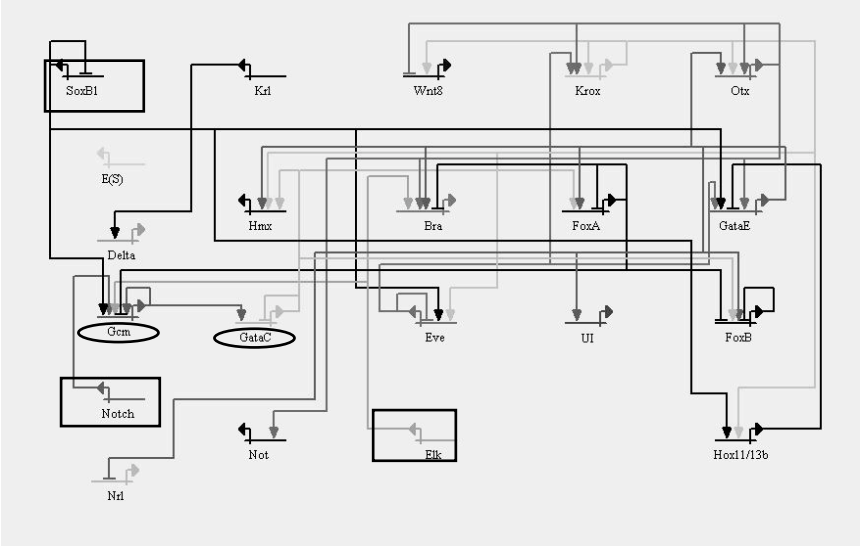

Figure 4: Automatically generated minimal Endomesoderm network with links removed where a connection is already present through a single intermediate node. On the complete network, genes highlighted in rectangular boxes have links to both *GataC* and *gcm* (ellipses). In the minimal network, their actions on *GataC* are all through *gcm*.

4.4 Networks from separate stages of embryo development

Data for the 21 endomesoderm genes at each time period was rendered into a separate network to compare expression profiles at each time. This yielded a set of networks that contained 15 (12-16 hr), 30 (18-21 hr), 45 (24-28 hr), 6 (32-36 hr), 2 (40-48 hr) and 0 (60-72 hr) links. Although gene expression does change through the development stages, it is unlikely that these results represent an accurate picture of the regulatory system, rather an indication that the dataset is incomplete. Thus, without additional data to indicate that genes operational at one period are turned off in another (there are some data), it will be very difficult to draw any conclusions from these observations.

5 Next steps

5.1 Probabilistic assignment of effects

The approach taken for this study relied on definitive assignment of a link (or no link) between two genes based on the data. The output from the algorithm is trinary and, therefore, relies heavily on the thresholding function to define whether a gene is activated or inhibited. There is no indication as to the certainty of these predictions and this all-or-nothing approach leads to the possibility that a small change in the threshold level can create or eliminate links.

The idea here is to generate networks with links with varying levels of confidence. This may be done in our platform by placing link values on a continuous scale, for example from -10 to +10. The output value is a measure of the certainty that the algorithm can predict the presence of a link. For instance, a value of -10 would mean an activation relationship with absolute certainty, likewise +10 for a certain inhibition. A value closer to zero is less certain. A threshold function will still be required to apply the cut-off that defines an interaction with no link. Nevertheless, a value just exceeding the threshold will be labeled as uncertain, rather than all links having equal validity.

5.2 Incorporation of auxiliary information

A mechanism for incorporating external auxiliary knowledge of biology is needed. An example of where auxiliary information could be used is in the action of *Otx* on *wnt8*. The data indicates that this should be a straight forward inhibition. However, the published network indicates that *Otx* activates an intermediate gene labeled 'Rep. of wnt8' [Repressor] and that this gene inhibits *wnt8*. There is no footnote with the data that could indicate why the link was drawn like this, yet evidence can be found in another publication by the group at the Davidson

Laboratory [4]. This paper reported that introduction of an obligate repressor of *Otx* target genes resulted in a many fold increase in the transcripts of *wnt8*. Thus, this information is showing that the action of *otx* on *wnt8* is a two (or more?) step process. This knowledge could have been incorporated into the algorithm to improve accuracy of the output.

A future development of the module would, therefore, utilize the auxiliary information known about interactions and incorporate this into the decisions to include a link or not. Thus, additional knowledge could be used to strengthen the case for a particular configuration of the network over another.

6 Discussion

Automated generation of biopathways can help generate large complex gene regulatory networks that can be minimized to best explain the raw data. These methods can incorporate knowledge gleaned from the literature, footnotes and other sources. This makes the approach closer to how a human would work: bringing to bear knowledge and prior experiences when interpreting results from experiments.

7 Acknowledgements

We would like to thank our scientific advisors Atul Butte and Trey Ideker for their inputs and direction in selecting the data set and developing the approach.

References

1. Chandra et. al. "Epistemics Engine", U.S. Patent application, (Nov 2002)
2. Davidson Laboratory Website. http://its.caltech.edu/~mirsky/qpcr.html
3. Davidson *et al*. A genomic regulatory network for development. Science **295**, 1669-1678 (2002)
4. Davidson et al. A provisional regulatory gene network for specification of endomesoderm in the sea urchin embryo. Developmental Biology **246**, 162-190 (2002)
5. Kohane IS, Kho A, Butte AJ. Microarrays for an Integrative Genomics, MIT Press (2002)
6. Kosa, *et al*. Reverse engineering of metabolic pathways from observed data using genetic programming. Pacific Symposium on Biocomputing **6**, 434-445 (2000)

7. Kosa, *et al*. Reverse engineering and automatic synthesis of metabolic pathways from observed data using genetic programming. Stamford University Technical report SMI-2000-0851 (2000)

8. Ideker TE, Thorsson V, Karp RM. Discovery of regulatoryinteractions through perturbation: inference and experimental design. Pacific Symposium on Biocomputing **5**, 302-313 (2000)

9. Wessels L.F.A., Van Someren, E.P. and Reinders, M.J.T. A comparison of genetic network models. Pacific Symposium on Biocomputing **6**, 508-519 (2001)

10. Maki, Y. *et al*. Development of a system for the inference of large scale genetic networks. Pacific Symposium on Biocomputing **6**, 446-458 (2000)

11. Smith VA, Jarvis ED, Hartemink AJ. Evaluating functional network inference using simulations of complex biological systems. Bioinformatics **18(Suppl. 1)**, S216-24 (2002)

12. Liang S, Fuhrman S, Somogyi R. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. Pacific Symposium on Biocomputing **3**, 18-29 (1998)

13. Imoto S, Goto T, Miyano S. Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. Pacific Symposium on Biocomputing **7**, 175-186 (2002)

14. Chrisman *et al*. Incorporating biological knowledge into evaluation of causal regulatory hypothesis. Pacific Symposium on Biocomputing **8**, *In press (2003)*

15. Akutsu T, Miyano S, Kuhara S. Algorithms for inferring qualitative models of biological networks. Pacific Symposium on Biocomputing **5**, 290-301 (2000)

16. Hartemink AJ *et al*. Combining location and expression data for principled discovery of genetic regulatory network models. Pacific Symposium on Biocomputing **7**, 437-449 (2002)

17. Wimberly FC, Glymour C, Ramsey J. Experiments on the accuracy of algorithms for inferring the structure of genetic regulatory networks from associations of gene expressions, I: algorithms using binary variables. Submitted to the Journal of Machine Learning Research. (2002)