A Problog Model For Analyzing Gene Regulatory Networks

António Gonçalves[1], Irene M. Ong[2], Jeffrey A. Lewis[3] and Vítor Santos Costa[1]

[1] Faculty of Sciences, Universidade do Porto
CRACS INESC-TEC and Department of Computer Science
Porto, Portugal 4169-007
Email: up100378013@alunos.dcc.fc.up.pt, vsc@dcc.fc.up.pt
[2] Great Lakes Bioenergy Research Center, University of Wisconsin
Madison, WI 53706
Email: ong@cs.wisc.edu
[3] Department of Genetics, University of Wisconsin
Madison, WI 53706
Email: jalewis4@wisc.edu

Abstract. Transcriptional regulation play an important role in every cellular decision. Gaining an understanding of the dynamics that govern how a cell will respond to diverse environmental cues is difficult using intuition alone. We introduce logic-based regulation models based on state-of-the-art work on statistical relational learning, to show that network hypotheses can be generated from existing gene expression data for use by experimental biologists.

1 Introduction

Transcriptional regulation refers to how proteins control gene expression in the cell. Many major cellular decisions involve changes in transcriptional regulation. Thus, gaining insight into transcriptional regulation is important not just for understanding the fundamental biological processes, but also will have deep practical consequences in fields such as the medical sciences. With the advent of high-throughput technologies and advanced measurement techniques molecular biologists and biochemists are rapidly identifying components of transcriptional networks and determining their biochemical activities. Unfortunately, understanding these complex multicomponent networks that govern how a cell will respond to diverse environmental cues is difficult using intuition alone.

In this work, we aim at building probabilistic logical models thatwould uncover the structure and dynamics of such networks and how they regulate their targets.

Despite the challenge of inferring genetic regulatory networks from gene expression data, various computational models have been developed for regulatory network analysis. Examples include approaches based on logical gates [1, 2], and probabilistic approaches, often based on Bayesian networks [3]. On one hand, logic gates provide a natural, intuitive way to describe interactions between

proteins and genes. On the other hand, probabilistic approaches can handle incomplete and imprecise data in a very robust way.

Our main contribution is in introducing a model that combines the two approaches. Our approach is based on the probabilistic logic programming language ProbLog [4, 5]. In this language, we can express true logical statements (expressed as *true rules*) about a world where there is uncertainty over data, expressed as *probabilistic facts*. In the setting of gene expression, this corresponds to establishing:

(1) a set of *true rules* describing the possible interactions existing in a cell;
(2) a set of *uncertain facts* describing which possible rules are applicable to a certain gene or set of genes.

Given time-series gene expression data, we want to choose the probability parameters that best describe the data. Our approach is to reduce this problem to an optimization problem, and use a gradient ascent algorithm to estimate a local solution [6] in the style of logistic regression. We further contribute an efficient implementation to this algorithm that computes both probabilities and gradients through binary decision diagrams (BDD). We validate our approach by using it to study expression data on an important gene-expression pathway, the Hog1 pathway [7].

Related Work Logic-based modeling is seen as an approach lying midway between the complexity and precision of differential equations on one hand and data-driven regression approaches on the other[8].

Despite the difficulty of deciphering genetic regulatory networks from microarray data, numerous approaches to the task have been quite successful. Friedman *et al.* [3] were the first to address the task of determining properties of the transcriptional program of *S. cerevisiae* (yeast) by using Bayesian networks (BNs) to analyze gene expression data. Pe'er *et al.* [9] followed up that work by using BNs to learn master regulator sets. Other approaches include Boolean networks (Akutsu *et al.* [10], Ideker *et al.* [11]) and other graphical approaches (Tanay and Shamir [12], Chrisman *et al.* [13]).

The methods above can represent the dependence between interacting genes, but they cannot capture causal relationships. In our previous work [14], we proposed that the analysis of time series gene expression microarray data using Dynamic Bayesian networks (DBNs) could allow us to learn potential causal relationships.

2 Methods

Recently, there has been interest in combining logical and probabilistic representations within the framework of Statistical Relational Learning [15]. This framework allows the compact representation of complex networks, and has been implemented over a large variety of languages and systems. Arguably, one of the most popular SRL languages is the programming language ProbLog [4, 5]. This

language was initially motivated by the problem of representing a graph where there is uncertainty over whether edges exist or not. As a straightforward example consider the directed graph in Figure 1.

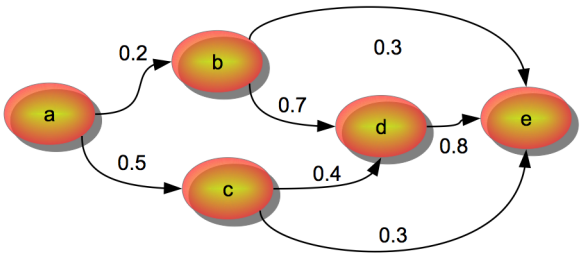


Fig. 1. A simple directed graph, where each edge has a probability of being true.

Notice that each edge has a probability of being true. As an example, starting from `a` we can reach `b` with probability 0.2 and `c` with probability 0.5. We assume that all the different probabilities are *independent*.

Given the example in Fig 1, ProbLog allows one to answer several queries, such as *what is the most likely path between two nodes*, and *what is the total probability that there is a path between two nodes*. The algorithm takes advantage of independence between probabilistic facts.

Note that computing the probability is not simply the sum if different paths have a common edge. As an example, consider $Pr(\texttt{ae})$. The path `abde` shares the edge `de` with `acde`, and the edge `ab` with `abe`. Summing these three paths would count two edges twice.

Kimmig and de Raedt proposed an effective solution to this problem. The idea is that probability can be computed as a sum if the paths do not share edges. This can be obtained by selecting an edge (or fact), and splitting into the case where the edge is true and the case where the edge is false. The process can be repeated recursively until we run out of facts to split. Kimmig and de Raedt's key observation is that this idea is indeed the same one that is used to construct binary decision diagrams (BDDs): the total probability can be obtained by generating a BDD from the proofs.

Binary decision diagrams provide a very efficient implementation for probability computation over small and medium graphs. Unfortunately, they do not scale to larger graphs with thousands of nodes. In this case, ProbLog implementations rely on approximated solutions, either Monte Carlo methods or often by approximating the total probability by the probability of the best k queries [5].

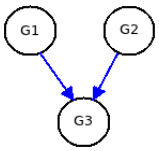
3 Experimental Methodology

We obtained time-series gene expression data from Lee et al. [16] for our experiments. The experiments followed the response of actively growing Saccharomyces cerevisiae to an osmotic shock of 0.7 M NaCl. The dose of salt was selected by the experimentalists to provide a robust physiological response but allow high viability and eventual resumption of cell growth. The samples were collected

before and after NaCl treatment at 30, 60, 90, 120, and 240 min (measuring the peak transcript changes that occurs at or after 30 min) [17]. We focused our attention on the 270 genes of the Hog1 Msn2/4 pathway from Capaldi [7] for which we have expression data and utilized the temporal data to better estimate the relationships from the data.

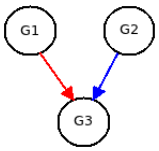
Our experiments aim for a more detailed picture of the learned network by using the temporal nature of the data. The output generated is a weighted, directed gene network, but nodes are connected as a gated network:

- **AND:** two promoter genes need to be active in order to activate a gene, as shown in the graph. We also show the ProbLog code for the temporal model:



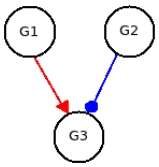

```
active(G3,T1,Z) :-
      next_step(T0,T1),
      and(G1,G2,G3),
      active(E,T0,G1),
      active(E,T0,G2).
```

- **OR:** either promoter gene needs to be active in order to activate a gene, as shown in the graph. We also show the corresponding ProbLog code for the temporal model.



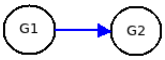

```
active(G3,T1,Z) :-
      next_step(T0,T1),
      or(G1,G2,G3),
      active(E,T0,G1).
active(G3,T1,Z) :-
      next_step(T0,T1),
      or(G1,G2,G3),
      active(E,T0,G2).
```

- **XOR:** one promoter gene needs to be active and one repressor gene needs to be inactive in order to activate a gene, as shown in the graph.




```
active(G3,T1,Z) :-
      next_step(T0,T1),
      xor(G1,G2,G3),
      active(E,T0,G1),
      not_active(E,T0,G2).
```

  This is the only case where we allow the possibility of negative regulation.
- **SINGLE:** a unique promoter gene regulates the target gene.




```
active(G2,T1,Z) :-
      next_step(T0,T1),
      single(G1,G2),
      active(E,T0,G1).
```

We use two different forms of temporal data: expression level (E), and variation (Δ). We experimented with three different approaches:

(1) Level influences variation (LV).
(2) Variation influences variation (VV).
(3) Level influences level (LL).

One important advantage of the approach is that it allows us to implement *soft constraints* on the probability distribution. These constraints are implemented by saying that satisfying some rule must have probability 1 or 0. In our experiments, we implement constraints saying that a *gene must be explained by a single rule*. Two example constraints for **OR** are of the form: The next constraint says that there must be a single set of parents for a gene defined with the LV $\vee$ rule:

$$\begin{array}{c} E_t(G_1) \vee E_t(G_2) \Rightarrow \Delta_{t+1}(G) \\ \wedge \\ E_t(G_3) \vee E_t(G_4) \Rightarrow \Delta_{t+1}(G) \\ \rightarrow \\ G_1 = G_3 \wedge G_2 = G_4 \end{array}$$

The second constraint ensures that we cannot use two rules of different types at the same time:

$$\begin{array}{l} \neg(\ E_t(G_1) \vee E_t(G_2) \Rightarrow \Delta_{t+1}(G) \\ \qquad\qquad \wedge \\ \quad E_t(G_3) \oplus E_t(G_4) \Rightarrow \Delta_{t+1}(G) \\ \ ) \end{array}$$

In practice, we must be careful not to flood the system with soft constraints. In our experiment we implemented one joint soft constraint per gene.

4 Conclusion

Learning regulatory networks from gene expression is a hard problem. Data is noisy and relationships between genes highly complex. We present a statistical relational approach to modeling pathways. Our approach allows us to design a coarser and a more fine grained model, based on probabilistic gates.

We plan to continue improving the model quality and experiment with new data. Specifically, we would like to experiment with implementing a regression based approach, as it fits our framework naturally. Last, but not least, we would like to investigate how to reduce the number of parameters in the model by exploiting strong correlations between gene expression.

Acknowledgments

This work is financed by the ERDF European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) FCOMP-01-0124-FEDER-010074 and by National Funds through the FCT Fundacão para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project HORUS (PTDC/EIA-EIA/100897/2008) and by the US 760 Department of Energy (DOE) Great Lakes Bioenergy Research Center (DOE BER 761 Office of Science DE-FC02-07ER64494).

References

1. Glass, L., Kauffman, S.: A logical analysis of continuous, non-linear biochemical control networks. Journal of Theoretical Biology **39** (1973) 103–129
2. Thomas, R.: Boolean formalization of genetic control circuits. Journal of Theoretical Biology **42** (1973) 563–585
3. Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using Bayesian networks to analyze expression data. Journal of Computational Biology **7**(3/4) (2000) 601–620
4. Raedt, L.D., Kimmig, A., Toivonen, H.: Problog: A probabilistic prolog and its application in link discovery. In Veloso, M.M., ed.: IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007. (2007) 2462–2467
5. Kimmig, A., Santos Costa, V., Rocha, R., Demoen, B., Raedt, L.D.: On the Implementation of the Probabilistic Logic Programming Language ProbLog. Theory and Practice of Logic Programming Systems **11** (2011) 235–262
6. Gutmann, B., Kimmig, A., Kersting, K., Raedt, L.D.: Parameter learning in probabilistic databases: A least squares approach. In: ECML/PKDD–08. Volume LNCS 5211., Antwerp, Belgium, Springer (September 15–19 2008) 473–488
7. Capaldi, A., Kaplan, T., Liu, Y., Habib, N., Regev, A., Friedman, N., O'Shea, E.: Structure and function of a transcriptional network activated by the mapk hog1. Nature Genetics **40** (2008) 1300–1306
8. Morris, M., Saez-Rodriguez, J., Sorger, P., Lauffenburger, D.: Logic-based models for the analysis of cell signaling networks. Biochemistry **49** (2010) 3216–3224
9. Pe'er, D., Regev, A., Tanay, A.: Minreg: Inferring an active regulator set. In: Proceedings of the 10th International Conference on Intelligent Systems for Molecular Biology, Oxford University Press (2002) S258–S267
10. Akutsu, T., Kuhara, S., Maruyama, O., Miyano, S.: Identification of gene regulatory networks by strategic gene disruptions and gene overexpressions. In: Proc. the 9th Annual ACM-SIAM Symposium on Discrete Algorithms. (1998) 695–702
11. Ideker, T., Thorsson, V., Karp, R.: Discovery of regulatory interactions through perturbation: Inference and experimental design. In: Pacific Symposium on Biocomputing. (2000) 302–313
12. Tanay, A., Shamir, R.: Computational expansion of genetic networks. In: Bioinformatics. Volume 17. (2001)
13. Chrisman, L., Langley, P., Bay, S., Pohorille, A.: Incorporating biological knowledge into evaluation of causal regulatory hypotheses. In: Pacific Symposium on Biocomputing (PSB). (January 2003)
14. Ong, I., Glasner, J., Page, D.: Modelling regulatory pathways in *Escherichia coli* from time series expression profiles. Bioinformatics **18** (2002) S241–S248
15. Taskar, B., Getoor, L.: Introduction to Statistical Relational Learning. MIT Press (2007)
16. Lee, V., Topper, S., Hubler, S., Hose, J., Wenger, C., Coon, J., Gasch, A.: A dynamic model of proteome changes reveals new roles for transcript alteration in yeast. Molecular Systems Biology **7**(514) (2011)
17. Berry, D.B., Gasch, A.P.: Stress-activated genomic expression changes serve a preparative role for impending stress in yeast. Molecular Biology of the Cell **19**(11) (2008) 4580–4587