

Inferring gene interactions with Bayesian networks

Dirk Husmeier

Biomathematics and Statistics Scotland, Edinburgh EH9 3JZ, UK ¹

Keywords: Microarray experiments, genetic networks, Bayesian networks

1 Introduction

A central goal of molecular biology is to understand the regulatory interactions of gene transcription and protein synthesis. [1] proposed the application of Bayesian networks (BNs) to infer genetic interactions from microarray experiments. The main challenge for the inference procedure is that interactions between hundreds of genes have to be learned from sparse data (typically containing only about a dozen measurements). The inevitable consequence is that the posterior probability of network structures becomes vague, and it is the objective of the present work to experimentally quantify how much can be learned from the data.

2 Method and Results

We use the model regulatory network proposed by [2], which is shown in Figure 1, left, and which contains several structures similar to those in the literature, like a hysteretic oscillator, a genetic switch, as well as a ligand binding mechanism that influences transcription. Most transcription factors dimerize before they are active, each gene has more than one rate of transcription, depending on whether promoters are bound or unbound, and the presence of different time scales makes this model representative of a real biological system and a suitable challenge for the Bayesian network inference algorithm. The system of differential equations describing Figure 1, left, is integrated numerically. Then, the signals are sampled (taking measurements at 12 time points) and discretized, as shown in Figure 1, right. Starting from different priors $P(\mathcal{M})$ on the BN structure \mathcal{M} (imposing different bounds on the maximum fan-in to a gene), BNs are sampled with Markov chain Monte Carlo (MCMC) from the conditional probability $P(\mathcal{M}|\mathcal{D})$, given the data \mathcal{D} thus obtained. Denote by $P(e_{ik}|\mathcal{D})$ the posterior probability for an interaction (edge) between genes i and k , which is given by the proportion of networks in the MCMC sample that contain this edge. Let $\mathcal{E}(\theta) = \{e_{ik} | P(e_{ik}|\mathcal{D}) > \theta\}$ denote the set of all edges whose posterior probability exceeds a given threshold $\theta \in [0, 1]$. From this set we compute (1) the sensitivity (the proportion of recovered true edges), and (2) the (complementary) specificity (the proportion of erroneously detected spurious edges), from which the *receiver operator characteristics* (ROC) curves of Figure 2 can be obtained. This allows estimating the proportion of spurious gene interactions typically incurred for a specified target proportion of recovered true gene interactions, and thus offers practical guidelines and support in selecting the respective decision thresholds.

References

- [1] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using Bayesian Networks to Analyze Expression Data. *Journal of Computational Biology*, 7:601–620, 2000.
- [2] Daniel E. Zak, Francis J. Doyle, Gregory E. Gonye, and James S. Schwaber. Simulation Studies for the Identification of Genetic Networks from cDNA Array and Regulatory Activity Data. *Proceedings of the Second International Conference on Systems Biology*, pages 231–238, 2001.

¹E-mail: dirk@bioass.ac.uk

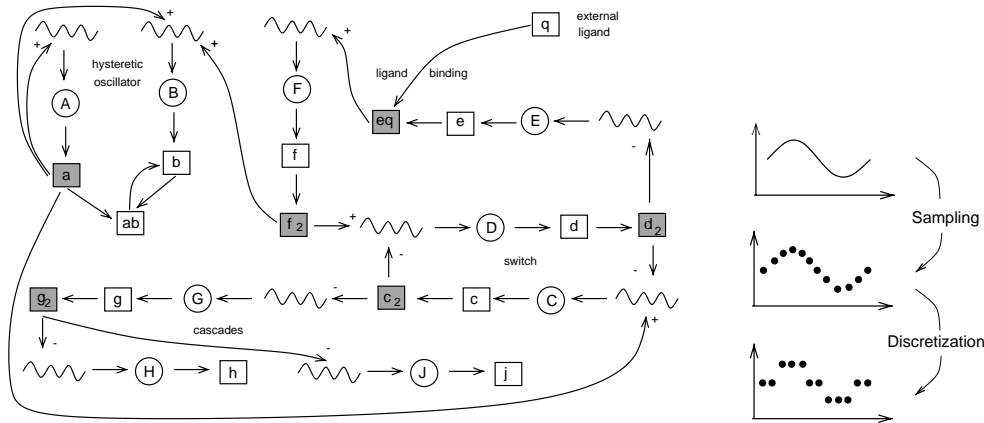


Figure 1: **Realistic simulation.** *Left:* Transcription factor dimers (shaded squares) bind to the *cis* regulatory site in the promoter region upstream of a gene (oscillating lines), influencing its rate of transcription. The transcribed mRNAs (circles) are translated into proteins (squares), which dimerize to form new active transcription factors that can bind to other *cis* regulatory sites. *Right:* Information contained in the true time dependent mRNA abundance levels is partially lost due to sampling and discretization.

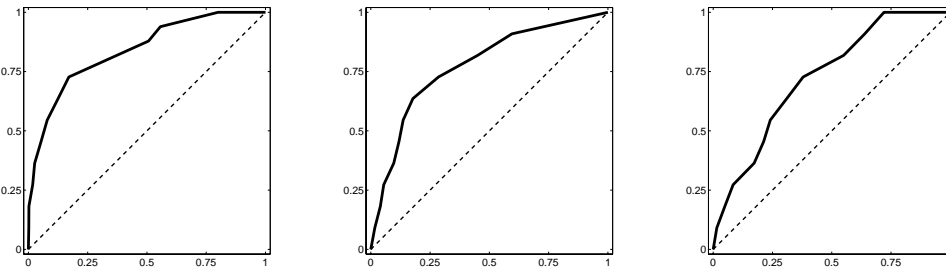


Figure 2: **ROC curves.** In each subfigure the sensitivity (proportion of recovered true edges) is plotted against the complementary specificity (proportion of false edges). This shows which price in terms of erroneously predicted spurious edges has to be paid for a desired recovery rate of true edges. The thin, diagonal dashed line is the ROC curve of a random predictor. The solid line shows the ROC curve for a Bayesian network trained with MCMC on 12 time points of gene expression data simulated from the system in Figure 1. *Left column:* max fan-in= 2; *middle column:* max fan-in= 3; *right column:* max fan-in= 4. Note that a larger area under the ROC curve indicates a better performance. The most restrictive prior (maximum fan-in=2) gives the best results, which is in agreement with the true network structure of Figure 1. This underlines the obvious fact that, for small data sets, the inclusion of available prior knowledge improves the performance of the inference scheme, and the amount of improvement is quantifiable from the ROC curves. The ROC curves suggest that local gene interactions can, to a certain extent, be inferred from microarray experiments, but that they are inevitably obscured by a considerable amount of spurious interactions. This should be taken as a cautioning note for those trying to back up detected interactions with circumstantial evidence from the biological literature.