

A Bayesian Method for Biological Pathway Discovery from High-Throughput Experimental Data

Wei Wang

Center for Biomedical Informatics and the
Intelligent Systems Program
University of Pittsburgh
wew@cbmi.pitt.edu

Gregory F. Cooper

Center for Biomedical Informatics and the
Intelligent Systems Program
University of Pittsburgh
gfc@cbmi.pitt.edu

Abstract

This poster describes a novel Bayesian method for discovering intra-cellular pathways from high throughput data. This Bayesian method is generalized from a deterministic algorithm [1], and it combines experimental data with prior belief to produce as output a probability distribution over the possible causal relationships between each pair of variables.

We applied this algorithm to gene expression data [2] on galactose metabolism in yeast. The Area Under ROC curve (AUROC) for the Wagner algorithm is 0.64. For the Bayesian algorithm the AUROC is 0.87, with a 95% confidence interval of (0.77 0.94). Thus, the Bayesian algorithm performs statistically significantly better than Wagner algorithm.

1. Introduction

Researchers have applied a variety of approaches to model the biological pathways, such as differential equations, Boolean networks, and Bayesian networks. The current poster presents a novel model of pathways that is designed for finding direct causation and feedback cycle relationships.

We assume that there are biological processes and states that are represented by variables. Let V be a biological pathway system that contains n biological variables that are designated as V_i , for i from 1 to n . For any two variables V_i and V_j ($i \neq j$) in V , we assume that experimental data allow us to distinguish in principle the following pairwise relationships:

- *Causal influence (C)*: V_i causally influences V_j if some change in the state of V_i leads to a change in the state of V_j under given

experimental conditions. Notation: $C(V_i, V_j) = true$.

- *No Causal influence (NC)*: V_i does not causally influence V_j if $C(V_i, V_j) = false$.

The Wagner Model allows us to distinguish the following pairwise relationships:

- *Direct Causation (DC)*: V_i directly causes V_j outside a feedback loop if (1) V_i influences V_j not through intermediate variables, and (2) V_i and V_j are not in a feedback cycle (see below).
- *Feedback cycle (F)*: V_i and V_j are in a feedback cycle if $C(V_i, V_j) = C(V_j, V_i) = true$.

The Wagner algorithm takes as input C/NC , and outputs DC/F . For the detailed algorithm see Wagner [1].

2. Bayesian algorithm

The Wagner algorithm assumes that the causal effects at the C/NC level are known with certainty. However, high throughput biological experiments can leave us with some uncertainty. We developed a Bayesian algorithm that takes as input C/NC under uncertainty, and outputs DC/F under uncertainty.

As a first step, the algorithm takes as input DNA microarray data D from which it derives the probabilities $P(C(V_i, V_j) = true | D)$ for each pair (i, j) . It then uses these probabilities as described in the remainder of this section.

Let M be an n by n *causality matrix* that represents all the binary cause-effect relationships (C/NC) in V . If V_i causally influences V_j ($i \neq j$), $M^{i,j} = 1$ otherwise $M^{i,j} = 0$. Because there are two choices 0 and 1 for each i,j ($i \neq j$) position in M , the total number of elements in the M is $2^{n(n-1)}$. Let M_k be an arbitrary instance of M , and let S denote the set of $2^{n(n-1)}$ unique

instances given by $M_1, M_2, \dots, M_{2^{n(n-1)}}$.

$$M_k = \begin{pmatrix} - & M_k^{1,2} & \dots & M_k^{1,n-1} & M_k^{1,n} \\ M_k^{2,1} & - & \dots & M_k^{2,n-1} & M_k^{2,n} \\ & & \dots & & \\ M_k^{n-1,1} & & \dots & - & M_k^{n-1,n} \\ M_k^{n,1} & & \dots & M_k^{n,n-1} & - \end{pmatrix},$$

$$M_k^{i,j} = \begin{cases} 0, & \text{if } C(V_i, V_j) = \text{false} \\ 1, & \text{if } C(V_i, V_j) = \text{true} \end{cases} \text{ for } i \neq j.$$

Let

$$P(C(V_i, V_j) | k, D) \equiv \begin{cases} P(C(V_i, V_j) = \text{true} | D), & \text{if } M_k^{i,j} = 1 \\ 1 - P(C(V_i, V_j) = \text{true} | D), & \text{if } M_k^{i,j} = 0 \end{cases}$$

for $i \neq j$.

Assuming independence of the causal relationship between each pair of variables,

$$P(M_k | D) = \prod_{i \neq j} P(C(V_i, V_j) | k, D). \quad (1)$$

Let $W(V_i, V_j | M_k)$ be a function that returns the relationship R that is output by the Wagner algorithm for variables V_i and V_j , when given M_k . Recall that R is a value in $\{DC, F\}$. Let $I_R(W(V_i, V_j | M_k)) = 1$ if $W(V_i, V_j | M_k)$ returns R , otherwise $I_R(W(V_i, V_j | M_k)) = 0$. The Probabilistic Wagner (PW) algorithm returns for each value of V_i, V_j , and R the expectation given by Equation 2:

$$PW(V_i, V_j, R, D) = \sum_k I_R(W(V_i, V_j | M_k)) \cdot P(M_k | D) \quad (2)$$

The PW algorithm involves summing over the $2^{n(n-1)}$ terms in Equation 2. In practice, we can approximate PW (APW) by drawing stochastic samples from the distribution of M to approximate Equation 2. Let $M_{r(t)}$ be a randomly sampled value of M according to Equation 1. Then for T samples, APW is defined as follows:

$$APW^T(V_i, V_j, R, D) = \frac{1}{T} \sum_{t=1}^T I_R(W(V_i, V_j | M_{r(t)})),$$

where $M_{r(t)}$ is a randomly drawn member from S . In the large sample limit we have the following result:

$$\lim_{T \rightarrow \infty} APW^T(V_i, V_j, R, D) = PW(V_i, V_j, R, D).$$

3. Results and Discussion

We applied APW to the cDNA microarray data obtained from experiments that focused on the

galactose utilization pathway in the yeast *Saccharomyces cerevisiae* as reported by Ideker, et al. [2]. The experiments involved single gene deletions of nine of the key genes that participate in yeast galactose metabolism. This pathway has been studied extensively and the causal regulatory relationships shown in Figure 1 form a reasonable gold standard.

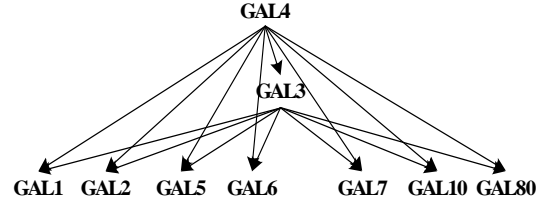


Figure 1. The relationships among genes in the yeast galactose utilization pathway, when extracellular galactose is present.

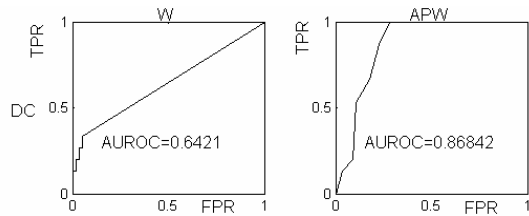


Figure 2. The comparison of the performance of W and APW .

We use AUROC (Area Under ROC curve) to evaluate APW and W . For APW , the AUROC is 0.87, with 95% confidence interval (0.77 0.94). W has an AUROC of 0.64. The performance of APW therefore is statistically significantly better.

4. Acknowledge

We thank Yao Zhang and Shyam Visweswaran for helpful discussions. This research was supported by NASA grant NRA2-37143.

5. References

- [1] A. Wagner, A. "How to reconstruct a large genetic network from n gene perturbations in fewer than n^2 easy steps.", *Bioinformatics* 17, 2001, pp1183-1197.
- [2] T. Ideker, et al. "Integrated genomic and proteomic analyses of a systematically perturbed metabolic network.", *Science* 292, 2001, pp929-93.