Bayesian Consensus Pathway Construction and Expansion Using Microarray Gene Expression Data

Andrew P. Hodges, Zuoshuang Xiang, Peter Woolf, and Oliver He, University of Michigan Medical School.

Abstract

Signaling and regulatory pathways that guide gene expression have only been partially defined for most organisms. However, given the increasing number of microarray measurements, it may be possible to reconstruct such pathways and uncover missing connections directly from experimental data. Using a compendium of microarray gene expression data obtained from *E. coli*, we constructed a series of Bayesian network models for the reactive oxygen species (ROS) pathway as defined by EcoCyc. Three consensus Bayesian network models (large, medium, and small) were generated based on consensus stringency. While less stringent consensus models diverged from the literature model, more stringent consensus models with fewer genes better approximated the known ROS pathway. Networks at the three consensus levels were expanded to predict genes that enhance the Bayesian network model using an algorithm termed 'BN+1'. Expansion of each of the three ROS-based networks predicted many stress-related genes and their possible interactions with other ROS pathway genes. For example, BN+1 expansion of the large network predicted a potential important role for *uspE* in regulating the ROS pathway and biofilm stress responses. The medium network expansion identified several genes (e.g., *sra* and *yodD*) and their possible interactions with other genes in the ROS pathway. The majority of known acid fitness island genes were recovered within the top 10 predicted genes by expansion of the small network containing *gadE*, *gadW* and *gadX*. The presently reported consensus and BN+1 expansion method is a generalized approach applicable to the study of other biological pathways and living systems.

Introduction

In this study, we address two fundamental questions in systems biology: 1) Is a network derived from transcriptional microarray gene expression data similar to a literature derived network. The first question is important because an increasing number of systems biology tools such as GSEA presuppose that literature derived networks should be the same as an expression derived network. Surprisingly, this hypothesis has not been rigorously tested. 2) Can we reliably predict new genes that can be added to an existing pathway based on microarray data?

Methods

Data Preprocessing: A compilation dataset comprising 305 gene expression microarray observations and 4,217 genes from *Escherichia coli* MG1655 was obtained from the M3D database (1). A coefficient of variation threshold (c.v. ≥ 1.0) was used to select 4,205 genes for analysis. Twenty-seven genes were identified from the EcoCyc ROS detoxification pathway (downloaded on March 26, 2008) and matched to unique features found in the 305 available gene expression microarray chips. Expression profiles for each gene were discretized using a maximum entropy approach that uses three equally-sized bins. All analyses were completed in MARIMBA (http://marimba.hegroup.org).

Learning Bayesian Network Pathway Models: Given the set of 27 genes, Bayesian network analysis was used to learn the structure of the large model which served as our core starting topology. To maximize the network search space, 4000 independent simulations with random starts were used to search 2.5×10^7 networks per start for a total of 1×10^{11} networks. Five top networks were saved from each run, thereby generating a final list of 2×10^4 top-scoring networks. These networks were used to estimate the posterior distribution.

Consensus Network Selection: During the search, each network was scored using log of the BDe score (2) which is the natural log of posterior probability (P(D|M)), The calculation of this score was implemented using the public software BANJO (3). To reduce the large 27 gene network down to medium and small networks, we trimmed the networks using a parameter we termed *B*-value. A *B*-value can be thought of as a normalized posterior score and is defined as follows:

$$B-value = 1 - P = 1 - \sum_{k=1}^{j} \left(\frac{e^{S_k}}{\sum_{i=1}^{x} e^{S_i}} \right)$$

Here *j* is the number of top unique scores (natural log of posterior probability) chosen for inclusion in the consensus network calculation, while *x* is the number of all unique scores saved for network analysis. *Sk* is the natural log of posterior probability for a unique score *k* that appears for at least one saved network. *P* is the sum of posterior probabilities for the top *j* scores normalized across all unique posterior probabilities (scores); i.e. *P* is a cumulative density function (CDF) value that represents the coverage of the best networks relative to all possible networks. The *B*-value measures the strictness of a "top" network compared to the total networks stored. Three consensus networks (large, medium, and small) were selected based on different *B*-values from the ROS network study.

Network Expansion Using BN+1: The BN+1 algorithm is defined in Fig. 1.

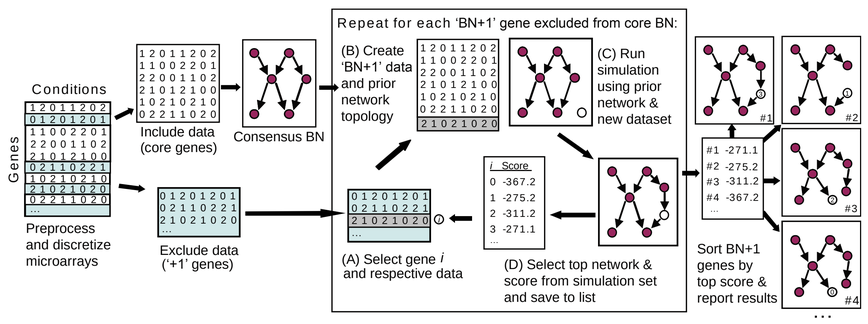

Fig. 1. Schema for the consensus network generation and BN+1 algorithm in MARIMBA web system. Basically, after a consensus network is selected, the network is iteratively expanded by adding one new gene to the core network followed by BN execution. The top BN+1 genes are defined as those that maximized the BN scores.

Results

Consensus Network Analysis

We created a novel *B*-value as a cutoff to select the number of networks for inclusion in consensus network generation. Stricter consensus networks (medium and small networks) defined by decreased *B*-values better match the known pathways in EcoCyc, RegulonDB, and literature than looser networks (Fig. 2).

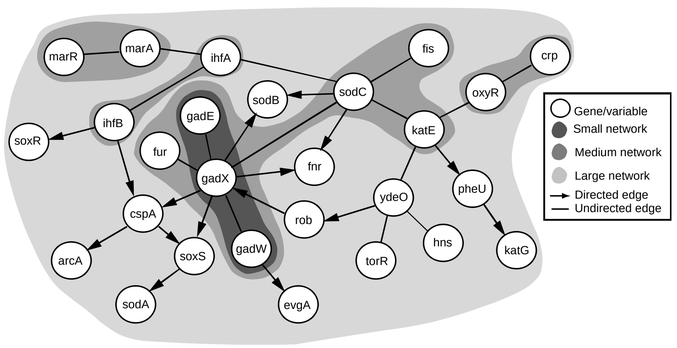

Fig. 2. Consensus networks for the ROS detoxification pathway based on gene expression data. The large consensus network (27 genes) is the most permissive (*B*–value=0.247) for edge inclusion and was derived from the consensus of the 33 top networks that shared the best identical posterior probability. A medium consensus network (13 genes) of intermediate stringency (*B*-value=10^{-3}) was derived from the top 3,644 simulated networks. The small network (3 genes) was derived by including all 20,000 networks (*B*-value = 0) and additional expert curation.

BN Expansion using BN+1

Our BN+1 expansion analysis resulted in identification of known and predicted genes important for ROS and stress responses.

Rank	Large Network (27 gene)	Medium Network (13 gene)	Small Network (3 gene)
1	**dusB** (tRNA-dihydrouridine synthase B); S=-8295.81	**dusB** (tRNA-dihydrouridine synthase B); S=-3821.20	**slp** (outer membrane lipoprotein); S=-949.65
2	**fdhE** (formate dehydrogenase formation protein); S=-8298.44	**sra** (30S ribosomal subunit protein S22); S=-3850.29	**hdeA** (stress response protein acid-resistance protein); S=-954.57
3	**uspE** (stress-induced protein); S=-8310.63	**yodD** (predicted protein); S=-3850.30	**hdeB** (acid-resistance protein) S=-958.11
4	**yohF** (predicted oxidoreductase with NAD(P)-binding Rossman-fold domain); S=-8312.24	**fbaB** (fructose-bisphosphate aldolase class I); S=-3860.69	**gadA** (glutamate decarboxylase A, PLP-dependent); S=-968.53,
5	**yncG** (predicted enzyme); S=-8313.04	**slp** (outer membrane lipoprotein); S=-3865.13	**gadB** (glutamate decarboxylase B, PLP-dependent); S=-972.15
6	**msyB** (predicted protein); S= -8318.20	**hdeA** (stress response protein acid-resistance protein); S=-3870.05	**hdeD** (acid-resistance membrane protein); S=-973.65
7	**yedP** (conserved protein); S=-8320.30	**msyB** (predicted protein); S=-3871.68	**yhiD** (predicted Mg(2+) transport ATPase inner membrane protein); S=-975.68
8	**sra** (30S ribosomal subunit protein S22); S=-8323.97	**hdeB** (acid-resistance protein); S=-3873.59	**dctR** (predicted DNA-binding transcriptional regulator); S=-993.91
9	**ydcK** (predicted enzyme); S=-8325.91	**erfK** (conserved protein with NAD(P)-binding Rossmann-fold domain); S=-3877.97	**ybaS** (predicted glutaminase); S=-996.20
10	**ynhG** (conserved protein); S=-8326.20	**ynhG** (conserved protein); S=-3878.40	**mdtE** (multidrug resistance efflux transporter); S=-1017.59

Table 1. Top 10 genes with best log posterior scores predicted from BN+1 expansion based on the large, medium, and small consensus networks. Numbers shown after gene names are negative logs of posterior probabilities (S) for each top network containing the respective predicted gene. Highlighted cells represent known acid fitness genes, which are important for ROS stress response.

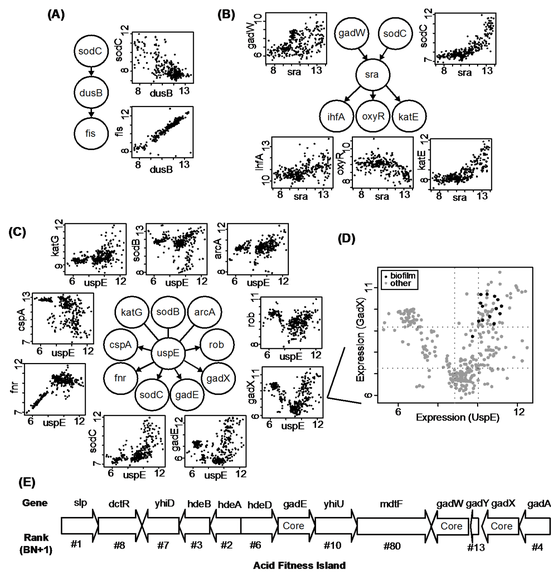

Fig. 3. Top predicted BN+1 genes from three consensus network expansions. The genes *dusB* (A) and *sra* (B) are predicted from the medium network expansion. The genes *dusB*(A), *uspE* (C) were the top results for the large network expansion. (D) Scatter plot for *uspE* versus *gadX* highlighting experiments with the word "biofilm" in the experiment title and/or description. High levels of *uspE* and *gadX* were observed for all conditions mapped to 'biofilm', suggesting possible roles of these two genes in biofilm activities. The dotted lines indicate boundaries for binning used in network learning. Many nonlinear expression patterns were predicted by our BN+1 simulations. (E) The entire acid fitness island (4) was recovered from the expansion using the small network consisting of three core genes (*gadE*, *gadW*, and *gadX*).

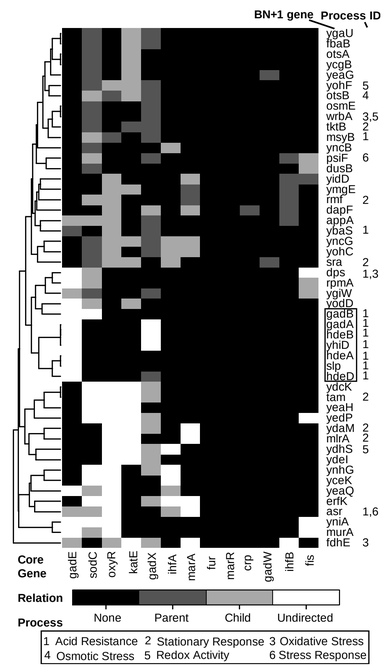

Fig 4. Novel heatmap representation of consensus neighborhoods for the top fifty BN+1 genes predicted using medium network. This representation was designed to test whether certain core genes had a preferential role in the recovery of the top BN+1 genes. Each cell represents a relationship between a BN core gene (x-axis) and a particular BN+1 gene (y-axis) with selected grayscale shading representing predicted relationships of core genes respective to the predicted genes. The heatmap demonstrates preferential connectivity of BN+1 genes to those core genes on the left side of the figure versus those on the right. Manually-curated biological functions and localization (Entrez Gene and literature) are indicated in margin of vertical axis. Roughly half of the top 50 genes identified for the medium network expansion have relevant ROS and/or stress-related activities, whereas many are yet un-annotated or with unknown functions. Boxed gene names identify those genes from the acid fitness island which clustered together.

Discussion

Our study addressed the two questions described in the Introduction. Regarding the first question, We find that the agreement between these network types is only partial, indicating that the widespread usage of literature networks as a gold standard for expression networks and their expansions is not recommended. Our case study on the ROS pathway analysis indicates that when this network model based on transcriptional gene expression data is further constrained (low *B*-value), the consensus model more closely matches the known regulatory ROS pathway. Regarding the second question, here we show how pathways can be intelligently expanded based on experimental data. In the study, we describe a novel Bayesian method called BN+1 that systematically searches for new genes to add to a pathway description. By searching for new participants in a pathway, we automatically detect connections between current pathway definitions and also uncover new genes that play a central role in existing pathways. In this study, genes selected for inclusion in the ROS pathway showed clear biological relevance to ROS in all three models, supporting the premise that the network expansion approach employed in this study are valid. The BN+1 algorithm recovered genes (e.g. *gadX* and *uspE*) that would be very difficult to identify using methods such as clustering and Pearson correlation. Overall, the consensus network and BN+1 approach is a generalized method that is applicable to the investigation of various biological pathways in living systems.

Acknowledgements

This research was supported in part by NIH Grant U54-DA-021519. A.P.H. was also supported by a NIH Training Grant (5 T32 GM070449-04) and a 2008 Rackham Spring/Summer Research Grant at the University of Michigan. Additional support for A.P.H. was provided by the University of Michigan Bioinformatics Program. We gratefully acknowledge the critical review and editing of this manuscript by Dr. George W. Jourdian, University of Michigan Medical School.

References

1. Faith JJ, et al. (2008) Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. Nucleic Acids Res 36(Database issue):D866-870.

2. Cooper GF & Herskovits E (1992) A Bayesian method for the induction of probabilistic networks from data. Machine Learning 9:309-347.

3. Smith VA, Yu J, Smulders TV, Hartemink AJ, & Jarvis ED (2006) Computational inference of neural information flow networks. PLoS Comput Biol 2(11):e161.

4. Mates AK, Sayed AK, & Foster JW (2007) Products of the Escherichia coli acid fitness island attenuate metabolite stress at extremely low pH and mediate a cell density-dependent acid resistance. J Bacteriol 189(7):2759-2768.