EdgeClipper: an algorithm for Bayesian consensus network refinement and its application in microarray-based pathway refinement

Poster X021

Andrew P. Hodges[1,2], Peter Woolf [1,2,3,4], Yongqun He [1,2,5,6]

1. National Center for Integrative Biomedical Information. 2. Center for Computational Medicine and Bioinformatics. 3. Department of Chemical Engineering. 4. Department of Biomedical Engineering. 5. Unit for

Contact Info (APH):
Mail: 018 ARF, 1150 W Medical Center Dr., Ann Arbor MI 48109
Phone: 734-615-2455
Fax: 734-936-3235
Email: aphodges@umich.edu

Abstract

Bayesian networks (BN) have been used in many studies for reconstructing biological pathways from experimental data. BNs are network models containing the interactions of biological entities (e.g. genes, proteins) in a pathway based on high-throughput gene expression data. However, BNs can show false positive results in predicted associations (i.e., network edges), especially in cases where the available data are few or noisy. To increase the specificity, consensus networks are frequently used to refine the results of BNs out of top predicted networks. In this study, we introduce an algorithm we term EdgeClipper which uses a B-value criterion to integrate the posterior probabilities for all saved BNs when computing consensus networks. The B-value is a cutoff for restricting the number of top networks included in consensus BN generation. Our studies using synthetic and experimental data showed that the decreasing B-value resulted in increased specificity and decreased sensitivity for successive consensus networks. A study with E. coli compendium mRNA microarray data indicated that more restricted consensus BNs with smaller B-values more closely match the manually-curated pathways (e.g., ROS pathway) found in current databases (e.g., EcoCyc, RegulonDB) and literature. In summary, EdgeClipper provide a new method for reconstruction of biological pathways with higher specificity, and provide more focused hypothesis generation for experimental analysis.

Background and Motivation

Bayesian networks are graphical models which can describe causal relationships amongst sets of variables such as genes or proteins when trained using high-throughput datasets. These BNs provide a powerful method to identify both known and unknown interactions amongst variables, by defining conditional probabilities between the variables and graphically presenting them as edges in a network. Given that exploration of all possible network models for a set of variables is NP-hard, the "best" Bayesian networks recovered during a comprehensive search which optimize the overall posterior probability for each network are assumed most likely to represent the underlying or "real" pathway of interest. Unfortunately, the resultant networks often include many spurious or 'false positive' influences predicted between variables which are a challenge to remove computationally when prior knowledge about those edges is unavailable. We recently showed that Bayesian networks could be generated for a selected biological pathway (ROS detoxification in Escherichia coli), combined into a consensus network, and later expanded iteratively to identify novel components of that pathway model [1]. A consensus network is defined here as the set of edges conserved across a set of saved networks meeting some selection criteria.

In our previous study, the consensus network generated from the set of BNs sharing the same best posterior probability (or score) showed low to moderate overlap (29-42%) with existing biological knowledge [1]. Similar findings have been observed in other analyses when benchmarking BNs and consensus networks with a variety of synthetic and biological data [2-5], most specifically in cases where the number of observations are low or the data have abundant noise. Here, an EdgeClipper algorithm is introduced to allow refinement of the top-scoring Bayesian networks within successively-generated consensus networks via a novel B-value parameter and edge selection criterion.

Methods

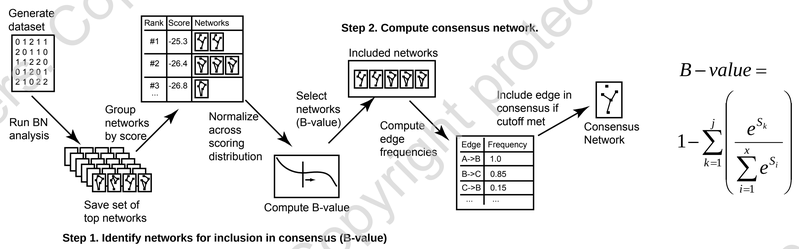

Figure 1. Schema for EdgeClipper algorithm. An initial BN simulation is executed in a parallel simulation architecture, and a set of top-scoring networks are saved for subsequent analysis. A b-value is computed for each unique score (Equation 1) obtained from the saved network set. After selection of a desired b-value, the set of edges to be included in consensus is computed. Currently, edges must appear with 100% frequency in either or both directions to be included in the consensus network.

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

Equation 1. Equation for B-value metric. Each B-value represents a set of networks with score greater than or equal to a unique scoring cutoff.

Results

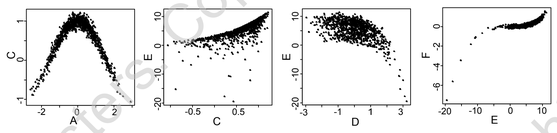

Figure 2. Pair-wise plots of representative data generated using a synthetic network. Data were sampled randomly for various data sizes and used in subsequent simulations (dataset with 1,000 observations shown). The encoded relationships for the synthetic network shown in Fig.3A are as follows: A = N(μ,σ), B = N(μ,σ), C = cosine(A)+ N(μ,σ), D = N(μ,σ), E = 3.5*e$^{(C)}$ - e$^{(D)}$ + N(μ,σ), F = (E/10.0)^3 + N(μ,σ), where N(μ,σ) is normally distributed. A similar sampling approach and synthetic network were recently implemented in a Bayesian network expansion analysis [6]. The functions and corresponding data were generated using custom R code written by APH [7].

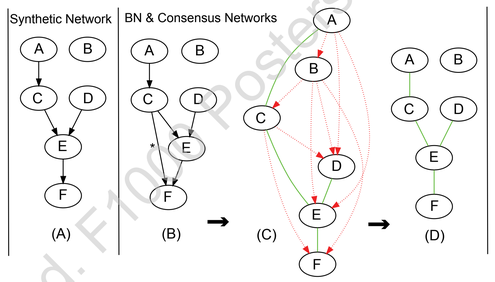

Figure 3. Consensus network refinement of Bayesian networks from a synthetically-sampled dataset using the EdgeClipper algorithm. (A) Synthetic network designed for simulating microarray gene expression data, A set of 1,000 observations were sampled in this example for the six variables (synthetic genes) using functions listed in Fig.2 legend to mathematically model dependencies between variables. (B) Representative network with best log posterior score. (C) Intermediate step in EdgeClipper algorithm. Red edges appear in <100% cumulative directionality and are not included in final consensus network (D). Green edges appear with 100% cumulative frequency in two directions yet <100% frequency in either direction in the network set selected by the B-value. Networks in (A) and (D) show 100% agreement assuming non-directionality. As B-value approaches zero, last supported edge is D-E (not shown).

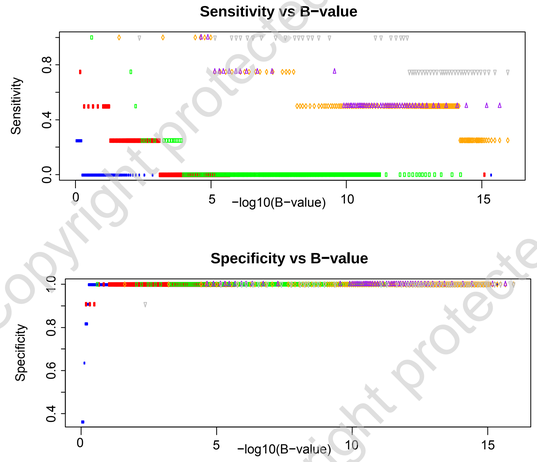

Figure 4. Performance benchmarking of the EdgeClipper Algorithm for Selected Data Sizes. (A) ROC performance for EdgeClipper Algorithm applied to BNs generated from 10 (blue), 50 (red), 100 (green), 250 (orange), 500 (purple), and 1,000 (grey) sampled data points. The B-value shows optimal specificity gains in the 0.1-0.01 range while minimizing sensitivity loss. Sensitivity and specificity were defined as TP/(TP+FN) and TN/(TN+FP), respectively , with TP true positive, TN true negative, FP false positive, FN false negative.

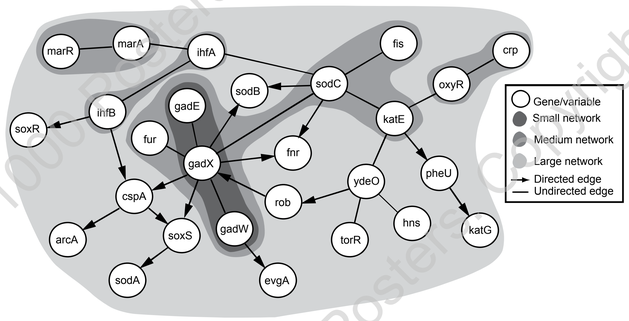

Figure 5. Consensus networks for the *E. coli* ROS detoxification pathway based on gene expression data. Network results originally appearing in [1] for the EcoCyc [8] ROS detoxification pathway were refined using the EdgeClipper algorithm. Each darker shading represents a smaller B-value and more confident prediction as determined by the EdgeClipper algorithm. Gene expression microarray data were obtained from the M3D database [9] and used to construct Bayesian networks (methods and results described in [1]).

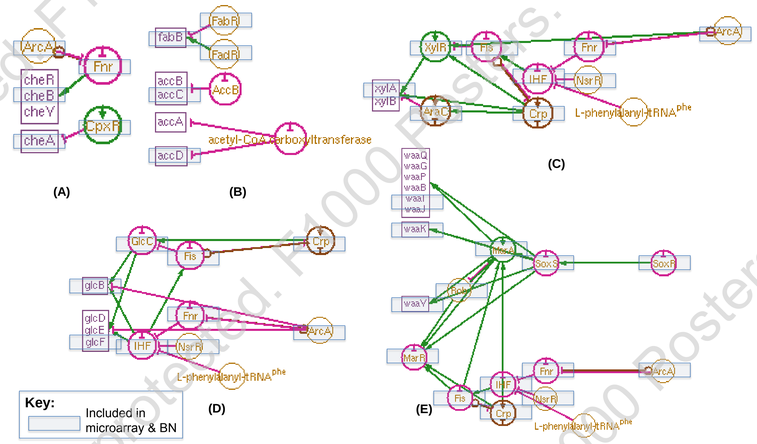

Figure 6. Other EcoCyc pathways [8] analyzed using the EdgeClipper algorithm. (A) Chemotactic signal transduction pathway (CHE-PWY), (B) fatty acid biosynthesis initiation superpathway (FASYN-INITIAL-PWY), (C) xylose degradation I pathway (XYLCAT-PWY), (D) glycolate and glyoxylate degradation II pathway (GLYOXDEG-PWY), and (E) lipid A-core biosynthesis pathway (LIPA-CORESYN-PWY). Results of sensitivity analysis shown below in Table I. Highlighted blue boxes denote those genes or proteins (or synonyms) which match to genes included in the microarray dataset. Proteins such as IHF in EcoCyc can map to two or more genes (e.g. ihfA & ihfB), and are included as separate genes if available.

Pathway	Gene Count	Specificity Top	Specificity B=0.1	Specificity B=0.01	Sensitivity Top	Sensitivity B=0.1	Sensitivity B=0.01
CHE-PWY	5	0.857	0.857	0.857	0.667	0.667	0.667
FASYN-INITIAL-PWY	5	0.500	0.500	0.625	0.000	0.000	0.000
GLYOXDEG-PWY	10	0.826	0.826	0.923	0.357	0.357	0.273
XYLCAT-PWY	10	0.750	0.750	0.833	0.231	0.231	0.154
SALVADEHYPOX-PWY	15	0.829	0.868	0.895	0.125	0.125	0.125
LIPA-CORESYN-PWY	15	0.878	0.946	0.973	0.222	0.222	0.222
FUC-RHAMCAT	20	0.920	0.933	0.939	0.148	0.148	0.111

Table I. Performance assessment for EdgeClipper algorithm using selected EcoCyc pathways and known interactions. Gene count is the number of genes appearing model that appear in the indicated pathway and match to the selected microarray profiles. Interactions were included from the EcoCyc pathway if both parent and child entities appear in the microarray list. Not all pathway members were matched to microarray chips (see Fig.6), possibly explaining some poor behavior observed in the FASYN-PWY pathway. In general, specificity increases at B-values=0.1-0.01 range while sensitivity decreases at B-value = 0.01.

Summary

The EdgeClipper algorithm can be used to refine a Bayesian network to include the most strongly-supported edges given an underlying dataset. The combined B-value and edge clipping steps are able to improve the overall specificity of the consensus networks and can remove model over-fitting. As the B-value decreases, specificity increases and sensitivity decreases. The overall performance of the algorithm was highlighted in small to moderate-sized data sets in synthetic simulations and larger biological pathways. We are currently using the EdgeClipper algorithm to generate network cores for network expansion (BN+1 algorithm) [1,6,10] and identification of novel pathway components in several biological studies. Overall, the EdgeClipper algorithm is a new approach for identifying well-supported interactions from Bayesian network analyses for biological pathways and systems. EdgeClipper and BN+1 are available for use in our web analysis pipeline (http://marimba.hegroup.org).

References

1. Hodges, A.P., D. Dai, et al. (2010). Bayesian network expansion identifies new ROS and biofilm regulators. PLoS One 5(3): e9513.
2. Bansal, M., V. Belcastro, et al. (2007). How to infer gene networks from expression profiles. Mol Syst Biol 3: 78.
3. Steele, E. and A. Tucker (2008). Consensus and Meta-analysis regulatory networks for combining multiple microarray gene expression datasets. J Biomed Inform 41(6): 914-926.
4. Friedman, N., M. Linial, et al. (2000). Using Bayesian networks to analyze expression data. J Comput Biol 7(3-4): 601-620.
5. Bose, R., H. Molina, et al. (2006). Phosphoproteomic analysis of Her2/neu signaling and inhibition. Proc Natl Acad Sci U S A 103(26): 9773-9778.
6. Hodges, A.P., Woolf P., and He Y. (2010). BN+1 Bayesian network expansion for identifying molecular pathway elements. Communicative and Integrative Biology. Accepted for publication, in press.
7. Ihaka R and Gentleman R. (1996). R: A language for data analysis and graphics. J Comput and Graph Stats 5(3):299-314. http://www.r-project.org
8. Keseler, I. M., J. Collado-Vides, et al. (2005). EcoCyc: a comprehensive database resource for Escherichia coli. Nucleic Acids Res 33(Database issue): D334-337.
9. Faith, J. J., M. E. Driscoll, et al. (2008). Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. Nucleic Acids Res 36(Database issue): D866-870.
10. Woolf P., He Y. and Hodges A. (2008). An Automated Method for Building Molecular Pathways Using Incremental Bayesian Learning. Pub. No. US 2009/0105962 A1 (Pub. Date: Apr. 23, 2009). US Patent Application # 12/139,529. Prov. App. Jun 14, 2007.

Acknowledgements

This research was supported in part by NIH Grant U54-DA-021519, NIH Training Grant (5 T32 GM070449-04), 2008 Rackham Spring/Summer Research Grant at the University of Michigan, and the University of Michigan Bioinformatics Program.