

Finding Bayesian Difference Networks

Gary Livingston and Thao Nguyen¹

Keywords: Bayesian networks, gene expression analysis

1 Introduction.

Often in biological analysis, identifying the differences in the interactions among attributes between two subpopulations—a *difference network*—is important. For example, identifying the gene interactions that differ between tumor and non-tumor patients would be extremely useful. We have developed a method based upon Bayesian network learning for finding interactions among attributes that appear to differ between two subpopulations. This method is adapted from [3] to identify the edges of a Bayesian network whose “fit” to two subpopulations differs statistically significantly. Merely inducing a network from one subpopulation will not only contain relations unique to the subpopulation, but also relations common to both subpopulations. For example, a network generated from expression data gathered from patients with tumors will not only contain edges for gene interactions that are associated with cancer, but will also contain edges for *non-cancer* interactions. Our approach is general and may be used to generate difference networks from any two subsets of interest from which standard Bayesian network learning may be applied.

The straightforward approach to creating a difference network, creating two edges from each of the subpopulations is not very robust. The two major difficulties with this approach are: (1) aligning networks is difficult—there are omissions to be considered as well as the semantics of the genes themselves that need to be factored into the comparison, and (2) because hard thresholds are used to decide which edges get included in the networks, two networks may be quite dissimilar when the underlying patterns in the data are somewhat similar. For an example of the latter, consider two networks A and B induced from two subpopulations, a and b , respectively. Because the learning algorithms often use threshold tests for determining inclusion of edges in the networks, an edge representing an interaction that is common to both subpopulations might barely pass the threshold test when learning network A from subpopulation a , but might barely fail the threshold test when learning network B from subpopulation b . Thus, the comparing the learned networks would suggest that the interaction is unique to subpopulation a , when the interaction actually occurs in both subpopulations.

2 Our approach.

Our method is straightforward. First, a BN for each population is built. Second, a difference network is constructed for each population. For example, a difference network could contain the edges, i.e. implied causal relationships, which fit population 1 but do not fit population 2. To determine which edges should be included in the difference network, our fitness measure will be a combination of the Diaconis-Efron conditional volume test statistic and the naïve rejection algorithm of Holmes and Jones [2]. This fitness measure will yield a p-value. If the p-value is sufficiently small, one would reject the possibility that the fit difference could have occurred by random variation in the sampling process and therefore accept that the edge probably represents a difference in the subpopulations from which the data were obtained. For example, our method can be used to identify which gene interactions are likely to be particular to cancer by identifying the edges in the network that differ significantly between cancer cases and non-cancer cases. Moreover, the p-values of the edges in the

¹ Department of Computer Science, University of Massachusetts Lowell

difference network could be used to prioritize biological analysis of the interactions represented by the edges. We illustrate our method in Figure 1.

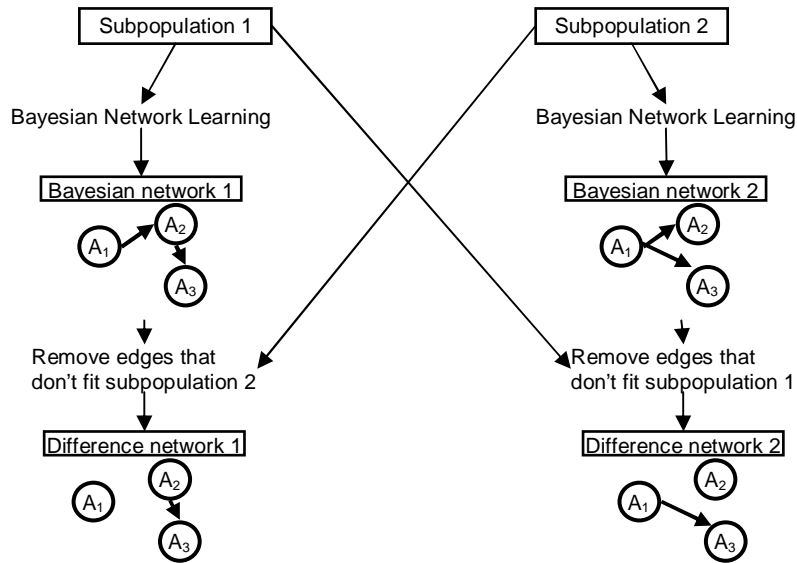


Figure 1. Illustration of our method for generating difference networks

3 Evaluation.

We will evaluate our method using the yeast gene expression data set presented in [1]. Because this dataset is well studied, it will serve as a “gold-standard” against which we can evaluate our method. After evaluation with this dataset, we will apply our method to a cell-growth gene expression dataset to identify gene interactions that differ during different stages of cell growth. Results of these evaluations will be presented.

References

- [1] Cho, R., Campbell, M., Winzeler, E., Steinmetz., L, Conway, A., Wodicka, L., Wolfsberg, T., Gabrielian, A., Landsman, D., Lockhart, D., and Davis, R. 1998. A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle. *Molecular Cell*, Vol. 2, pp. 65-73.
- [2]: Holmes, R. and Jones, L. (1996). On Uniform Generation of Two-way Tables with Fixed Margins and the Conditional Volume Test of Diaconis and Efron. *Annals of Statistics*, 1996, vol 24, no. 1, pp 64-68.
- [3] Livingston, G. R., Li, G., Hao, L., and Li, X. 2003. “The Induction and Analysis of Gene Networks.” *Proceedings of the Fourth International Conference for the Critical Assessment of Microarray Data Analysis (CAMDA 2003)*.