Reconstructing Condition-Specific Signal Transduction Hierarchy using Bayesian Networks

Dongxiao Zhu[1], Hua Li[2], Alfred O Hero[3]

Keywords: Bayesian Network, Gene Regulation Network, Hierarchy, Microarray

1 Introduction.

Bayesian networks (BNs) can simultaneously predict probabilistic causal relationships among biomolecules from noisy genome-wide measurements. However, optimal inference of Bayesian Network on the given data is a NP-hard problem. Consequently, existing BN learning algorithms use heuristics, such as constraining the search space by pre-screening the candidate causal relationships [1], [2], [3] or using heuristic search of good BN structure with randomly start, genetic algorithm or simulated annealing [4], but do not guarantee an optimal solution to the problem. Motivated by the recent findings that gene regulatory networks are likely to be hierarchical organized [5], we propose a novel BNs learning algorithm that is able to find the best BN structure(s) from the condition/tissue-specific genomics data in polynomial time. The algorithm takes advantage of a biological constraint, namely that signal is transmitted sequentially and hierarchically from the master regulators on the "top layer" to the regulators in the subsequent layers. The layer constraint makes the NP-hard problem of learning optimal BN structure(s) solvable in a heuristic manner.

2 The Algorithm

A BN is a directed acyclic graph (DAG) representation of the joint probability distribution of a set of random variables, $\mathbf{X} = \mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n}$. Based on the assumption of first-order Markovian dependency, a random variable X_i is independent of the other variables in the network given its immediate parents π_i, the joint probability distribution can be decomposed into a particularly simple form of conditional distribution, in which the π_i varies following BN structures:

$$P(\mathbf{X}) = \prod_{\mathbf{i}} \mathbf{P}(\mathbf{X_i}|\pi_\mathbf{i}). \tag{1}$$

A score function can be applied to score conditional distributions (Eq. 1) corresponding to a given BN structure. We use discrete score function that is suitable for noisy genomics data as follows:

$$Score(B_s|D) \propto \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{r_i - 1}{N_{ij} + r_i - 1} \prod_{k=1}^{r_i} N_{ijk}, \tag{2}$$

in which B_s represents a BN structure, D represents genomics data, r_i is the number of discrete outcomes of each variable (i.e. 3) and N_{ijk} equals the total number of cases in the dataset where dataset variable i in state k and the parent nodes are in stage j. A number of existing literature discussed the advantages of discretizing data in both supervised

[1]Stowers Institute for Medical Research. E-mail: doz@stowers-institute.org
[2]Stowers Institute for Medical Research. E-mail: hul@doz@stowers-institute.org
[3]Dept. of EECS, University of Michigan, Ann Arbor. E-mail: hero@umich.edu

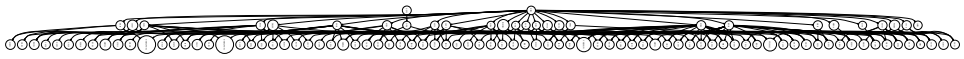

Figure 1: Optimally reconstructed MMS-specific signaling hierarchy. Only layers 1,2 & 3 are shown.

learning (e.g. tissue classification) and unsupervised learning (networking and clustering). The proposed BN learning algorithm is presented as the following:

Algorithm 1 Hierarchical BNs Learning Algorithm

1: Initialize hierarchical layers $L_1, L_2, \ldots, L_n$. j represents layer index, G_{ji} represents ith gene in the jth layer, and π_{ji} represents the parent(s) of G_{ji}.
2: **for all** layer j, $j = 2, 3, \ldots, n$ **do**
3: **for all** gene i in layer j **do**
$B_s = \text{argmax} Score(G_{ji}|\pi_{ji})$
4: **end for**
5: **end for**

3 Learning Methyl-MethaneSulfonate (MMS) Responsive Signaling Hierarchy

Based on the yeast signaling hierarchy constructed from manually curated datasets [5], the MMS responsive hierarchy was optimally inferred (Fig. 1), ie, the highest-score structure by Eq. 2 given the MMS response microarray data. Many predictions made in this analysis are supported by biological literature. For example, two out of eight top-layer master regulators in [5] (HIR3, GZF3) were predicted to be involved in the MMS response. After checking references, HIR3 was shown to be able to repress transcription and cause MMS sensitivity [6], and GZF3 is a transcription represser, which is a good indication of its role in DNA damage response.

References

[1] Friedman, N., Nachman I and Peer D. 1999. Learning Bayesian Network Structure from Massive Dataset: The "Sparse Candiate" Algorithm. In: *Proc. Fifteenth Conf. on Uncertainty in Artificial Intelligence (UAI 99).*

[2] Heyssier M and Koller D. 2005. Ordering-Based Search: A Simple and Effective Algorithm for Learning Bayesian Networks. *Proc. of the Twenty-first Conference on Uncertainty in AI (UAI) (pp. 584-590)*

[3] Pe'er D, Tanay A and Regev 2006. MinReg: A Scalable Algorithm for Learning Parsimonious Regulatory Networks in Yeast and Mammals. *Journal of Machine Learning Research* 7:167-189.

[4] Heckerman DA 1996. A Tutorial on Learning with Bayesian Networks. *Technical Report MSR-TR-95-06, Microsfot Research*

[5] Yu, H and Gerstain M. 2006. Genomic Analysis of the Hierarchical Structure of Regulatory Netowrks. *Proceedings of the National Academy of Sciences USA* 103:14724–14731.

[6] Qian, Z, Huang H, Hong JY 1998. Yeast Ty1 Retrotransposition Is Stimulated by a Synergistic Interaction between Mutations in Chromatin Assembly Factor I and Histone Regulatory Proteins. *Mol Cell Biol.*, 18(8): 47834792.