

An algorithm for reconstruction of Markov blankets in Bayesian networks of gene expression datasets

Catalin Barbacioru, Daniel J. Cowden, Joel Saltz
The Ohio State University, Department of Biomedical Informatics
333 W 10th Graves Hall, Columbus, OH 43210
catalin@bmi.osu.edu

Abstract

This paper presents an efficient algorithm, of polynomial complexity for learning Bayesian belief networks over a dataset of gene expression levels. Given a dataset that is large enough, the algorithm generates a belief network close to the underlying model by recovering the Markov blanket of every node. The time complexity is dependent on the connectivity of the generating graph and not on the size of it, and therefore yields to exponential savings in computational time relative to some previously known algorithms. We use bootstrap and permutation techniques in order to measure confidence in our finding. To evaluate this algorithm, we present experimental results on *S.cerevisiae* cell-cycle measurements of Spellman et al. (1998) [5].

1. Introduction

A Bayesian network $\mathcal{B} = (\mathcal{G}, \theta)$ is a graph-based model of the joint probability distribution that capture properties of conditional dependence and independence between variables represented as nodes. A graph \mathcal{G} encodes the *Markov assumption* if each node X_i is independent of its non-descendants, given its parents. In this case the joint probability distribution can be decomposed into

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | Pa^{\mathcal{G}}(X_i)) \quad (1)$$

where $Pa^{\mathcal{G}}(X_i)$ is the set of parents of X_i in \mathcal{G} . Glymour et al. [3] proposed a greedy algorithm, called **K2** algorithm, to maximize $P(\mathcal{G}|D)$ by finding the parent set of each variable that maximize the “local score”:

$$g(X_i, \pi_i | D) = \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}! \quad (2)$$

Definition 1 A **Markov blanket** of a node X , denoted as $MB(X)$ is a minimal set of variables, such that every other variable is independent of X given $MB(X)$, i.e. $\forall Y \in \{X_1, \dots, X_n\} \setminus \{MB(X), X\}, X \perp Y \mid MB(X)$.

Definition 2 Two variables X_1 and X_2 are **d-separated** given a set of variables S in a BN if and only if there exists no adjacency path between them (i.e. a path ignoring the ordering of the edges) such that (i) every collider (a collider being a node with two incoming edges that belong in the path) is in S or has a descendent in S , and (ii) every non-collider node on the path is in S [3].

Definition 3 The graph \mathcal{G} of some BN $\mathcal{B} = (\mathcal{G}, \theta)$ is **faithful** to a joint probability distribution J over a set a variables V if and only if every dependance entailed by \mathcal{G} is also present in J . We say that a data-generating process K is **faithfully represented** by \mathcal{B} , if K in the sample limit produces data with joint probability distribution J , and \mathcal{B} is faithful to J .

It follows from Markov condition that every conditional independence entailed by \mathcal{G} is also present in J . Thus, together faithfulness and Markov condition establish a close relationship between a graph \mathcal{G} and some probability distribution J and allow us to associate statistical properties of J with graph properties of \mathcal{G} .

Theorem 1 [4] In a faithful BN, *d*-separation captures all conditional dependence and independence relations that are encoded in the graph. Therefore, two nodes are *d*-separated given S , if and only if they are conditionally independent given S .

Theorem 2 [7] (i) The unique $MB(X)$ in a faithful BN is the set of parents, children and parents of children of X .

(ii) If $\mathcal{B}_1 = (\mathcal{G}_1, \theta_1)$ and $\mathcal{B}_2 = (\mathcal{G}_2, \theta_2)$ are two BN both faithful to the same joint distribution, then $MB_{\mathcal{B}_1}(X) = MB_{\mathcal{B}_2}(X)$ for any variable X .

Theorem 3 Let $X, Y, Z = \{Z_1, \dots, Z_n\}$ be a set of discrete variables and D a dataset. Then, if the number of sam-

ples is large enough, $Y \perp X \mid Z$ if and only if

$$g(X, Z \mid D) \geq g(X, Z \cup \{Y\} \mid D). \quad (3)$$

2. Algorithm

Suppose we have a dataset D generated by a process that can be faithfully represented by a Bayesian network $\mathcal{B}_0 = (\mathcal{G}_0, \theta)$. Let $V = \{X_1, \dots, X_n\}$ be the set of observed discrete variables. Suppose D contains m cases, where each case contains a value assignment for each variable in V . Denote $MB_{\mathcal{B}_0}(X_i)$ the Markov blanket of X_i .

Forward phase For every $i \in \{1, \dots, n\}$, and every arbitrary non-negative integer k , let $g_{i,k}(\pi)$, be the restriction of $g(X_i, \pi \mid D)$ introduced in (2), on the space of all subsets of $V \setminus \{X_i\}$ of cardinality k . Let $\pi_{i,k}$ be the subset of variables maximizing $g_{i,k}(\pi)$, and $m_{i,k} = g_{i,k}(\pi_{i,k})$. Let k_0 be the smallest integer such that $m_{i,0} \leq m_{i,1} \leq \dots \leq m_{i,k_0}$ and $m_{i,k_0+1} < m_{i,k_0}$. We will denote $\pi_{i,loc} = \pi_{i,k_0}$. From *Theorem 3*, X and any variable in $V \setminus \pi_{i,loc}$ are d-separated given $\pi_{i,loc}$.

Backward phase Let M_i be initialized with $\pi_{i,loc}$. For every $Y \in M_i$, if $g(X_i, M_i \mid D) < g(X_i, M_i \setminus \{Y\} \mid D)$, then $M_i \leftarrow M_i \setminus \{Y\}$. Repeat this process until M_i cannot be changed any more. Based on *Theorem 3*, it can be shown that M_i is the Markov blanket of X_i .

We make use of bootstrap method in order to estimate confidence in the features of the learned networks, where *confidence* represents the likelihood that a given feature is actually true. The confidence that X_i is in the Markov blanket of X_j will be $conf(m_{i,j}) = \frac{1}{b} \sum_k m_{i,j}(\mathcal{G}_k)$, where \mathcal{G}_k represents the structure obtained from the k^{th} bootstrap step and $m_{i,j}(\mathcal{G})$ is the indicator function in the network structure \mathcal{G} .

3. Results

We applied our approach to cell cycle expression data of Spellman et al. [5], containing 76 gene arrays of 6177 *S. cerevisiae* ORFs. Gene expression levels are discretized into three categories: -1, 0, 1, depending whether the expression value is significantly lower, similar to or greater than the average level of the gene across all experiments. We are making use of all 76 gene array samples for bootstrap estimates of confidence in Markov relations between genes with 100 resampling steps. The analysis shows that we can recover intricate structures even from small data sets. It is important to note that we used no prior biological knowledge or constrains. *Figure1* presents the number of Markov relations with confidence above threshold q . Inspection of the top Markov relations between genes reveals that most of them are functionally related. Among these are the previously known relations between genes coding histones HTB1 and HHT1 (confidence 0.99) presented in

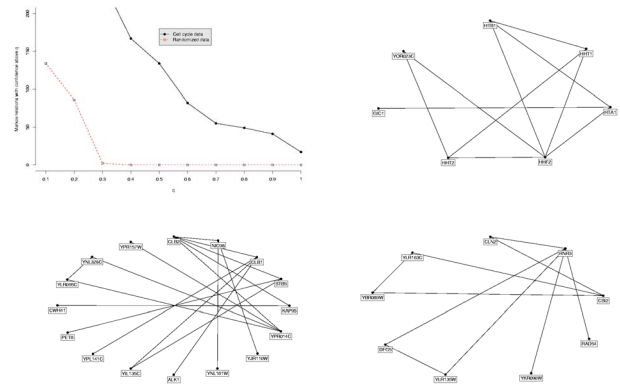


Figure 1. - . 4

Figure2, genes involved in mitosis, CLB1 and CLB2 (confidence 0.97) presented in *Figure3*, both known to be regulated by a combination of two transcription factors Mcm1p and SFF [1] or genes involved in DNA replication, CLN2 and CSI2 (confidence 0.96) presented in *Figure4*.

4. Conclusions

The algorithm presented in this paper, successfully scale up BN-based causal discovery by adopting a local approach, and by reducing the total computational time to $O(n^2)$. We describe how to apply these techniques to gene expression data [5], and uncover many already validated and also not yet validated biological discoveries.

References

- [1] Althoefer H, Schleiffer A, Wassmann K, Nordheim A, Ammerer G., Mcm1 is required to coordinate G2-specific transcription in *Saccharomyces cerevisiae*., *Mol Cell Biol.* 15(11):5917-28, 1995
- [2] Friedman N., Nachman I., Peer D., Learning Bayesian networks structure from masive datasets: The "sparse candidate" algorithm, *Proc.Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI'99)*, 196-205
- [3] Glymour, C., Cooper, G.F. (1999), *Computation, causation and discovery*, AAAI Press/MIT Press
- [4] Pearl, J. , *Probabilistic reasoning in intelligent systems: networks of plausible inference*, Morgan Kaufmann, 1988
- [5] Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B., Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, *Mol Biol Cell.* 9(12):3273-97, 1998
- [6] Tsamardinos I., Aliferis C.F., (2003), *Tword principled features selection:relevancy, filters and wrappers*, *Proc. 16th FLAIRS Conf*