Overview of Bayesian network approaches to model gene-environment interactions and cancer susceptibility

Chengwei Su
Angeline Andrew
Margaret Karagas
M. E. Borsuk

Follow this and additional works at: https://scholarsarchive.byu.edu/iemssconference

Su, Chengwei; Andrew, Angeline; Karagas, Margaret; and Borsuk, M. E., "Overview of Bayesian network approaches to model gene-environment interactions and cancer susceptibility" (2012). International Congress on Environmental Modelling and Software. 124. https://scholarsarchive.byu.edu/iemssconference/2012/Stream-B/124

This Event is brought to you for free and open access by the Civil and Environmental Engineering at BYU ScholarsArchive. It has been accepted for inclusion in International Congress on Environmental Modelling and Software by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.
Overview of Bayesian network approaches to model gene-environment interactions and cancer susceptibility

Chengwei Su\textsuperscript{a}, Angeline Andrew\textsuperscript{b}, Margaret Karagas\textsuperscript{b}, Mark E. Borsuk\textsuperscript{a}

\textsuperscript{a}Thayer School of Engineering, Dartmouth College, Hanover, NH USA. E-mail: chengwei.su.th@dartmouth.edu, mark.borsuk@dartmouth.edu
\textsuperscript{b}Community and Family Medicine, Section of Biostatistics & Epidemiology, Dartmouth Medical School, Lebanon, NH USA. Email: angeline.andrew@dartmouth.edu; margaret.karagas@dartmouth.edu

Abstract: Studies of the relation between genetic traits and cancer susceptibility are often inconclusive or conflicting. This is likely due to the challenges of accommodating multiple genetic and environmental risk factors using traditional analytic models. Each risk factor is likely to contribute to susceptibility through a combination of additive and non-additive interactions with other risk factors, and such interactions are not often addressed by conventional methods. Additionally, data from single studies rarely allow for conclusive identification of causal relationships in such complex systems. Yet, there is often a wealth of knowledge available from previous studies that could be brought to bear on the task of model building. In this paper, we review the potential applicability of Bayesian networks for learning causal relations from gene-environment-cancer data. We first describe the Bayesian network approach, including a variety of algorithms for learning the structure of the causal network from observational data. We then demonstrate application of the approach using a subset of data from a population-based study on bladder cancer in New Hampshire, USA. We find minor differences in the performance or results of different algorithms. However, we expect larger differences when these algorithms are applied to the large number of genes included in the full data set. Incorporation of prior knowledge will thus be a priority.

Keywords: learning causal structure, belief networks, genetic epidemiology, bioinformatics

1. INTRODUCTION

We have found that Bayesian networks (BN) can be effective tools for combining prior knowledge with observational data to infer and model causal relations. In this paper, we review the potential for Bayesian networks to model gene-environment interactions and cancer susceptibility. A Bayesian network is a directed acyclic graph (DAG) in which nodes represent random variables and edges represent probabilistic dependencies between them (Pearl 2000; Spirtes et al. 2000). The strength of the relationship between variables is represented by conditional probability distributions associated with each node (Korb and Nicholson 2010). The absence of directly connecting edges between any two nodes implies that these two variables are independent given the values of any intermediate nodes. A DAG is said to be Bayesian network if it satisfies the Markov condition, which states that each variable is conditionally independent of the set of all its non-descendants.
given the set of all its parents (Neapolitan 2003). Based on the Markov condition, the probability distribution of any variable \( X_i \) in the network can be determined by knowing the values of its immediate predecessors (called its parents, \( PA_i \)), without regard to the values of any other variables. In this way, the joint probability distribution for the entire network can be written as the product of a limited number of conditional distributions using the chain rule of probability calculus: (Borsuk 2008).

\[
P(x_1, ..., x_n) = \prod_{i=1}^{n} P(x_i | pa_i)
\]

### 2. METHODS

#### 2.1. Inference in Bayesian Networks

The edges in BN have casual semantics when we make several assumptions. The first is the Causal Markov Assumption: given the values of a variable’s immediate causes, it is independent of its earlier causes. The second assumption is that there are not latent or hidden variables that affect several of the observable variables (Friedman et al., 2000). Bayesian networks have been applied in various fields for causal study. For examples, in bioinformatics they have been used for the interpretation and discovery of gene regulatory pathways (Friedman et. al., 1999).

Figure 1 illustrates a simple hypothetical BN to demonstrate its application in probabilistic inference. We assume that there are no latent or hidden variables in this particular BN. There is an edge from S to L because Smoking has a direct influence on the presence of Lung Cancer. Smoking also has a direct influence on the occurrence of Bronchitis. Dyspnea (shortness-of-breath) may be due to Lung cancer or Bronchitis or both of them. There is no direct edge from S to D. In probabilistic terms, knowing B and L renders S and D independent. The joint distribution of all variables can be written in the mathematical form of equation [1] as

\[
P(S, L, B, D) = P(S) \cdot P(L | S) \cdot P(B | S) \cdot P(D | L, B)
\]

Say that a patient has Dyspnea, and we want to know what type of disease he or she is likely to have, Lung Cancer or Bronchitis. We can use Bayes’ theorem to compute the posterior probability of each case. Thus, the probabilities of Lung Cancer and Bronchitis conditional on Dyspnea can be computed from the following equation:

\[
P(L | D) = \frac{P(D | L) \cdot P(L)}{P(D)} \quad P(B | D) = \frac{P(D | B) \cdot P(B)}{P(D)}
\]

By comparing these two probabilities, we can infer which disease this patient is more likely to have. Knowing whether the patient is a smoker would provide further evidence on L and B.

#### 2.2. Learning Algorithms
In many real world situations, we do not know the causal structure and would like to learn it from data. Many algorithms have been developed to recover Bayesian networks from observational data to construct Decision Support Systems. Generally, these algorithms can be grouped into two categories:

- **Constraint-based algorithms**: Constraints are typically the conditional independence test statements. These algorithms usually learn the network structure with conditional independence tests, such as the $\chi^2$ test, to determine the lack of edges between variables, and then construct a graph and add directions to edges that satisfy the d-separation criterion. Say that $X$, $Y$, $Z$ are three disjoint sets of variables in a BN, “$Z$ is said to d-separate $X$ from $Y$ if and only if $Z$ blocks every path from a node in $X$ to a node in $Y$” (Pearl 2000).

- **Score-based algorithms**: Algorithms in this category search all possible structures in the space and assign a score that measures how well a BN evaluates the data set $D$. Given a structure $G$, its score is typically defined as:

$$Score(G \mid D) = Pr(G \mid D) = \frac{P(D \mid G) \cdot P(G)}{P(D)}$$  \[4\]

The search returns a structure that optimizes the score (See Copper and Herskovits (1992) and Heckerman et al. (1995) for detailed discussion on score metrics). Most learning algorithms employ heuristic search techniques, such as hill-climbing and simulated annealing.

The constraint-based methods are generally more efficient when the number of variables is large. However, they are sensitive to failures in independence tests. Also, due to their dependence on d-separation criterion to determine the direction of edges, they cannot assign direction to every edge. Thus, the common opinion is that the score-based approach is a better tool for learning structure from observational data (Friedman et al, 1999). Particularly when dealing with small sample size and noisy data, score-based algorithms can give more accurate results as they (potentially) search all possible structures in the space and find the optimal network (Cheng et al, 2001).

In addition to these two methods, there are hybrid algorithms which combine these two methods to maximize their advantages. Generally, they start with the constraint-based method to find the skeleton of the network and then identify the orientation with score-based method.

This section describes some well-known algorithms for learning BN structures using each approach, Grow-shrink (constraint-based), Hill-Climbing (score-based) and Max-Min Hill-Climbing (Hybrid).

- **Grow Shrink (GS) algorithm** consists of a growing and a shrinking phase. It starts from an empty set $S$; the growing phase adds variables to $S$ as long as they are dependent with variable $X$ given the current contents of $S$. In the shrinking process, variables that violate the Markov blanket property of $X$ are removed from $S$. The Markov blanket of a variable $X$ consists of $X$’s parents, children and its children’s other parents. Then, the algorithm identifies the local neighborhood (direct parents and children) of each variable in the network within the Markov blanket to recover the exact structure around each variable. This process can help verify or possibly correct a posteriori the beliefs of an outside expert. Edge directions are determined by examining triples of variables using the d-separation criterion (Margaritis, 2003 Section 3.1).
- Hill-Climbing (HC) search starts from either an empty, full, or possibly random network and existing prior knowledge can also be used to seed the initial candidate network. The main loop consists of attempting every possible single-edge addition, removal, or reversal in a candidate network. The structure that increases the score the most becomes the current candidate. The process iterates until a change in a single-edge does not increase the score (Margaritis, 2003 Section 2.7.2).

- Max-Min Hill-Climbing (MMHC) first learns the skeleton of a Bayesian network using a local discovery algorithm called Max-Min Parents and Children (MMPC). MMPC consists of a forward phase and a backward phase. In the forward phase, all variables with an edge to and from a variable are selected by use of a heuristic function. In the backward phase, all false positives selected in the first phase are removed. After MMPC identifies the skeleton of the BN, a greedy Bayesian-scoring hill-climbing search is used to orient the edges. The algorithm has the advantages of reliably scaling up to thousands of variables in reasonable time (Tsamardinos, et al, 2006).

3. RESULTS

3.1. Preliminary Application of Algorithms

We have available a rich genetic epidemiological data set from a large, population-based, case-control study of bladder cancer in New Hampshire. These data include over 1477 SNPs in cancer-related genes, detailed smoking assessment, gender, age, as well as other risk factors including arsenic exposure. (Karagas et al, 1998, Karagas et al. 2004). As a “proof of concept”, we have begun to analyze a subset of these gene-environment-bladder cancer data using bnlearn (http://cran.r-project.org/web/packages/bnlearn/bnlearn.pdf), an open-source software package for use with the statistical tool R (http://www.r-project.org). We are particularly interested in using the BN method to further assess recent findings of DNA repair genotype interactions with arsenic exposure to increase bladder cancer risk (Andrew et al. 2009a, 2009b). In these recent studies, authors found evidence of increased risk of bladder cancer risk among those with a XRCC3 variant genotype and high toenail arsenic levels.

For this analysis, we focused on the role of genetic variation in the DNA repair genes XRCC3 and ERCC2. Environmental risk factor arsenic exposure is represented by toenail arsenic level, which is the internal biomarker of arsenic exposure. We also include well-accepted risk factors such as gender, age and smoking for analysis. We used the embedded function ‘blacklist’ in bnlearn to preclude unreasonable casual relationships (e.g. smoking influencing gender). The analysis was applied to a subsample consisting of 424 controls and 226 cases without any missing values over the variable of interests.

With the help of the graphics package Rgraphviz in R, we are able to produce Figure 2 showing the relationships among 11 selected variables learned by the Hill-Climbing algorithm. The thickness of the arcs indicates the strength of the dependencies they represent. Results reveal four interaction groups. In one, gender influences smoking status, which further affects cancer. The normally strong influence of age on cancer is mitigated in this case by the fact that controls were chosen in a way that makes than more comparable to cases. Age appear to influence toenail arsenic levels. Figure 2 also shows that the BN captures the interactions among genes, but the arrows are misleading in this case as they represent interactions rather than genuine causal relationships among genetic variants. There is no evidence of genetic influences or gene-environment interactions on bladder cancer risk in this preliminary analysis. Since the complete
data set contains missing values and the previous suggested learning algorithms can only handle complete data, a subset of data without missing values were used for this analysis. This may cause the differences between our preliminary analysis and the study of Andrew et al. (2009a).

Figure 2 Applying score-based Bayesian Network learning algorithm to gene-environment-bladder cancer data of Andrew et al. (2009a).

For comparison, we further employed Grow-Shrink and Max-Min Hill-Climbing algorithms to our data set. Results are shown in Figure 3. The BN derived by MMHC algorithm appears to be consistent with the network derived by HC algorithm. The bold arcs are the extra two arcs learned from GS algorithm and these two arcs are not present in BNs learned from HC and MMHC. In the BN derived by GS algorithm, gender is found to influence cancer directly while XPD_312B appear to influence ERCC2_03B. Besides, the edge directions among XRCC3 genes and ERCC2 genes also differ between BNs learned from GS and other two algorithms.

We constructed a table to compare these three algorithms. From Table 1, we can see that GS algorithm performs slightly fewer independence tests/network comparison than the other two algorithms, thus it is computationally more efficient than them. The major difference among these three algorithms arises from the extra two arcs learned from GS algorithm. Investigating the detailed network learning process of GS shows that the independence test and d-separation criterion conclude that cancer, gender and smoker are neighbors, while in both HC and MMHC the arc from gender to cancer is tested and removed since it does not achieve an optimal score. As stated in the previous section, due to the relatively small data size, scored-based algorithms can give more accurate results as they explore all possible structures in the space and find the optimal network.
Table 1 Comparison of different algorithms

<table>
<thead>
<tr>
<th></th>
<th>Grow-Shrink</th>
<th>Hill Climbing</th>
<th>MMHC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independence tests</td>
<td>92</td>
<td>112</td>
<td>101</td>
</tr>
<tr>
<td>Directed arcs</td>
<td>10</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Undirected arcs</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Alpha threshold</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
</tr>
</tbody>
</table>

4. DISCUSSION, LIMITATIONS AND FUTURE RESEARCH

The intention of this paper is to demonstrate the potential applicability of different Bayesian network learning algorithms to gene-environment-cancer data. The advantage of using Bayesian network is its ability to translate probabilistic dependency and causal relationships into graphic models, thus we can clearly identify the interactions of variables through graphs. However, this analysis has a number of shortcomings due to its preliminary status: (i) We have analyzed only a small subset of the gene data available; (ii) We have only used data without any missing values over any of the variables; (iii) We discretized all continuous variables (e.g. toenail As) using thresholds which may not be optimal; (iv) We have not yet employed formal techniques for eliciting and employing prior judgments about latent variables; (v) We have not yet adopted a reasoned basis for the choice of score thresholds for arc inclusion; and (vi) We have not yet fully investigated alternative learning algorithms or scoring metrics. Our future research will focus on overcoming these limitations. Particularly in dealing with missing values, we will implement the structural EM (SEM) algorithm (Friedman, 1998), to learn BNs from incomplete data.
ACKNOWLEDGEMENTS

This project was supported by grants from the National Center for Research Resources (5P20RR024474-02) and the National Institute of General Medical Sciences (8 P20 GM103534-02) from the National Institutes of Health.

REFERENCES


Cheng, J., Greiner, R., Kelly, J., Bell, D., Liu, W., Learning Bayesian Networks from Data: An Information-Theory Based Approach, Department of Computing Sciences, University of Alberta, Faculty of Informatics, University of Ulster, 2001.


