Author's response to reviews

**Title:** Using Bayesian Networks to discover relations between genes, environment, and disease

**Authors:**

- Chengwei Su (Chengwei.Su.TH@dartmouth.edu)
- Angeline Andrew (Angeline.S.Andrew@dartmouth.edu)
- Margaret R Karagas (Margaret.R.Karagas@dartmouth.edu)
- Mark E Borsuk (mark.borsuk@dartmouth.edu)

**Version:** 3  **Date:** 22 February 2013

**Author's response to reviews:** see over
We found the reviewer comments to be very helpful in sharpening the description of our contribution. Point-by-point responses follow, with reviewer comments given in italics, followed by our response.

*This paper presents an application of Bayesian networks to disease association studies. The paper gives a summary of existing BN algorithms and an example application to bladder cancer data. Overall, I found the summary helpful and the example interesting. But I feel that the example is too limited and does not provide enough motivation for BN to be used in most disease association studies.*

For the bladder cancer data, I am confused about the fact that, after removing the blacklist, some associations reappear in the inferred model. The authors suggest that causal relationships may not be correctly identified due to various reasons. But then why using BN? I think it might be more appropriate, and indeed simpler, to just use Markov blanket without edge directions.

After removing the blacklist, some new associations appear because the blacklist had constrained the allowable associations to those deemed causally plausible. This is not an uncommon practice in BN modeling. However, as we describe in the text, for case-control studies the standard assumptions regarding causality do not apply. Nevertheless, we believe that a directed graph conveys more detailed information on the discovered relations than the corresponding undirected graph (which requires moralization first). Some authors, including Sebastiani and Perls [19], do consider this approach but are ambivalent about its merits. To address the reviewer’s concerns, we have included the moralized undirected graph (called the Markov network) in Figure 5.

*The bladder example has only 11 variables. This seems too small to be realistic in most studies. What is the new "discovery" here? The fact that the authors did not run BN on the entire 1477 SNPs raises concerns of its applicability to larger datasets. Are the problems the authors pointed out about the other methods still be of concerns on such a small dataset?*

Our example is purposefully limited so that the implications of differing assumptions, algorithms, and treatment of missing variables can be easily conveyed and understood by the reader. We do not attempt to report new findings on the bladder cancer dataset, but rather aim to compare the BN approach to the conventional approach previously used to analyze these data. We are currently applying the approach to the full 1477 SNPS and will report on the results separately.

*In fact, there are other Bayesian graphical approaches to identify "causal" relationships in genome-wide scale association studies, but are not reviewed or compared in this paper. A recent method BEAM3 (Zhang 2011 Genetic Epi) seems can do well in the scenarios discussed in this paper, although it does not impute missing data.*

We appreciate the reviewer pointing us to this very recent paper (2012, in fact) and have now included it in our discussion of other approaches.

*Table 5 is a quite interesting result, but is it possible to get p-values for the risks relative to the reference group? Or, can the authors run a logistic regression model with the variables constructed according to figure 4, and check for "interaction" effects? An edge in BN does not necessarily imply interaction, it could be two main effects plus correlation. It would also be useful to use the original continuous values to evaluate the effects on cancer.*
At the reviewer’s suggestion we have added the logistic regression results, which include the interaction effects. As we discuss in the paper, in this paper we prefer to stick with dichotomous variables for consistency with the original study.

Overall, I do not feel that this paper is sufficient to motivate enthusiasms to apply BN in current disease association studies. Its potential is not fully demonstrated, partially because of the limited dataset used and no other popular methods compared on the same studies. There are already tons of methods on large-scale mapping. And for small scale data, at least regression models need to be compared including up to two way interaction terms.

We hope we have addressed the reviewer’s concerns by: (1) explaining why we used a limited dataset in this paper and (2) adding the results of logistic regression modeling, including interaction terms.

Minor: Formula (1) is not shown correctly. Also the formula at the bottom of p7 is not shown properly.

This seems to have been a problem with translation to PDF. We will take care that they are shown correctly in the resubmission.

Fig2, HC and MMHC network, and Fig3, both have a cycle in the network. This seems violating the BN requirement and the corresponding decomposition of conditional probabilities would be incorrect.

This was a graphical error made when we transformed the output of R to a more easily readable figure. It is now corrected.

--- end of response