Reviewer's report

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

Version: 1 Date: 28 October 2012

Reviewer: Yu Zhang

Reviewer's report:

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.

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?

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.

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.

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.
Minor:

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

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.

Level of interest: An article of importance in its field

Quality of written English: Acceptable

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests:

I declare that I have no competing interests