Reviewer's report

Title: Rule based classifier for the analysis of gene-gene and gene-environment interactions in genetic association studies

Version: 1 Date: 27 November 2009

Reviewer: Brett McKinney

Reviewer's report:

This paper compares three rule-based classifier algorithms on 36 simulated data sets based on two genetic association models. The genetic models involve two interacting loci in a background of up to 3000 loci that are unrelated to the phenotype. One of the algorithms (RIPPER) is shown to perform better than the other two rule-based algorithms on these simulations.

1. These methods should be compared with some standard method. MDR would be applicable for model A (XOR model) though perhaps not model B (with a continuous covariate). However, a stepwise or other logistic regression approach should work for either model.

2. Because of the similarity between rule learners and tree learners, another logical comparison method would be Random Forests, which has gained in popularity in genetic association studies.

3. More details are needed for the rule-based methods, or at least for RIPPER, which stood out from the others. It might help to put these methods in the context of the more familiar, Random Forests. Random Forest was recently shown to have limited ability to identify the XOR model (like model A) in genetic studies because the tree node-splitter is based on a main effect measure such as information gain (PLoS Genet 5(3): (2009) e1000432). Are attributes selected in a similar manner with rule-based methods? Should they expected to perform better on an XOR model than Random Forests?

4. What is the criterion for selecting the best rule? Do the authors use a complexity term to properly compare the fitness of a model with more SNPs with one involving fewer SNPs? Can a pure three-way interaction be found with the rule-based system?

5. How exactly are the method settings chosen from the rather large number of possibilities? How would one tune these for a real data set?

6. Are these methods considered filters or modelers? In other words, are they meant to reduce the number of attributes for subsequent modeling or are they mean to identify a final model? Is there a way to estimate the false positive rate?

7. How do rule-based methods scale with number of SNPs? Will they scale to whole genome?
Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Acceptable

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

Declaration of competing interests:
I have no competing interests.