Summary of Discussion for
Predicting Gene Function From Patterns of Annotation
by Oliver D. King, Rebecca E. Foulger, Selina S. Dwight, James V. White, and Frederick P. Roth
Genome Research 13(5), 896-904, May 2003
http://www.genome.org/cgi/doi/10.1101/gr.440803

presented by Christopher Maier for
INLS 279: Bioinformatics Research Review
2006-02-22

This article presents a supervised machine learning technique for predicting Gene Ontology annotations for a gene, based on existing annotations of that gene. As such, it is similar to recommending purchases to customers based on prior purchase history. The group was generally pleased with the article. However, two main issues were raised, concerning the consequences of various data reduction methods, as well as the significance of the machine learning techniques used.

Throughout the article, the authors detail steps taken to reduce the amount of data in their datasets. Removing all annotations of "unknown" is likely a sound approach to take, but this simple act removes a significant number of annotations (a fact not shown in the article, but raised in the discussion). What effect does this have? What kind of information is being left out? The authors do not say.

Similarly, the authors speak of using the A10 set of attributes to improve their statistics. It would be nice to have an explanation as to how they decided on 10 as the cutoff. An investigation of the effects of alternate attribute sets (e.g. A9, A8, A15, A20) on classifier performance could be instructive as well.

The other main concern revolved around the choice of classifiers used. The authors create and evaluate both a decision tree and Bayesian network. Additionally, they use a model that treats all GO terms as independent; the predictive statistic $q(i,j)$ is the proportion of genes tagged with annotation j. The group was not sure that the independent model contributed much to the article. Creating alternative models, such as support vector machines or artificial neural networks, would have been more helpful.

The lack of any real discussion of the relative strengths and weaknesses of the two methods used (decision tree or Bayesian network) was the group's biggest concern. It was mentioned that the Bayesian classifier performed best overall, while the decision tree was better at low false-positive rates. We wondered what aspects (if any) of the data might contribute to this, as well as what aspects of the learning mechanism influence their performance on this data set. This is not necessarily a mark against this particular article, as similar articles omit such a discussion as well. We would be interested to see a more in-depth investigation of these machine learning methods in order to more fully understand their interaction with the kinds of biological data discussed.

Pedagogically, the article served as an introduction to the use of ROC analysis to assess classifier performance.

Overall, the group found the article to be well-done and informative. More discussion on the effects data reduction would be helpful, as would a discussion of the rationale behind the choice and implications of the machine learning methods used.