GOArray: Interpreting microarrays with GODB

Michael V. Osier[1], David Tuck[2]. Kevin P. White[3], Christopher E. Mason[3], Hongyu Zhao[4], Kei-Hoi Cheung[1]

Keywords: microarray interpretation, Gene Ontology Database, inference, permutation

1 Introduction.

Once a researcher has performed a microarray experiment and determined which genes are differentially expressed, the process of interpretation begins. For the rare case where only a few genes appear relevant, this can be an easy task. When dozens or hundreds of genes are differentially transcribed, however, it might be best to allow computers to reduce the effort of elucidation. Several analysis tools have been implemented to do this in the context of the terms in the Gene Ontology Database (GODB, http://www.godatabase.org/dev/database/). GODB is a rooted directed acyclic graph (rooted-DAG) of terms which represent a biological concept. Some of the tools that interpret arrays in the context of GODB score by "strict association" in which a GO term is scored purely by which genes are directly associated with it. Other tools score by "inclusive association" in which a term is scored by which genes are directly associated with it or one of its child terms.

A problem faced by both scoring methods is how to correct for the large number of statistical tests that must be performed in the analysis. The Bonferoni correction, multiplying the p-value by the number of tests performed, is overly-conservative to the point of being an impediment since few, if any, terms will ever be significant. Untangling the many interdependencies of the rooted-DAG of terms in GO to find an appropriate correction, however, is impractical.

As a way of assessing confidence in the results of the analysis from our tool GOArray, formerly named GOMine [1], which uses an inclusive association analysis of microarray data in the context of GODB, we have implemented two tests based on a permutation of the microarray data: a False Detection Rate to estimate how many significant terms we would expect on average, and a Confidence Test of how frequently we observe permuted arrays to have at least as many significant terms as the observed data.

2 Methods.

GOArray uses an inclusive association algorithm to identify genes associated with a term in GODB. The statistical analysis for each term tests if, among the genes associated with this term or one of its descendants, there is an overrepresentation of genes considered of interest (Genes Of Interest, GOI) relative to none GOI (NGOI). Note that determination of GOI and NGOI is made by an outside source, allowing the researcher to answer the question of "what is a gene that is differentially expressed?" in a manner appropriate for their experiment. A z-score and a p-value are

[1] Yale Center for Medical Informatics, Yale University School of Medicine, New Haven, Connecticut. E-mail: michael.osier@yale.edu

[2] Dept of Pathology, Yale University School of Medicine, New Haven, Connecticut

[3] Dept of Genetics, Yale University, New Haven, Connecticut

[4] Dept of Epidemiology and Public Health, Yale University, New Haven, Connecticut

calculated for each term based on the overall frequency of GOI [1]. Terms with a p-value less-than-or-equal-to a user-defined cutoff value are reported.

Permutations of the GOI are then generated, and statistics calculated for each permutation. For each permutation, the number of terms with a p-value less than or equal to the cutoff are counted (T_f). For the real dataset, the same statistic is calculated (T_r). The False Discovery Rate (FDR) is determined as the mean T_f divided by the T_r. The Confidence Test (CT) is the number of permutations where the T_f is greater than or equal to the T_r divided by the number of permutations. Together, these two tests give a feel for how confident one can be in the results of the analysis.

In the output of GOArray, the list of significant terms, FDR, and CT are reported as are two tree representations of the significant terms and an archive of how all statistics are calculated. An HTML format is used so that a) results are viewable in the future, minimizing the chance of data loss when application software is retired, and b) GOArray can be easily harnessed by a web interface in the future. GOArray is available from the web site "http://ycmi.med.yale.edu/gomine/".

3 Results and Discussion.

We have used GOArray to analyze a Drosophila expression dataset using a cutoff p-value of 0.001 and 1000 permutations. The data was taken from the Arbeitman et al. [2] stage 0-1 hours results, available from the NCBI database GEO (http://www.ncbi.nlm.nih.gov/geo/) with the accession GSM3612. All genes with at least a five-fold increase in expression were marked as GOI (1374 spots), and all other genes as NGOI (7425 spots). On a 2.4 GHz Xeon processor, the analysis took ~10 min with permutations and ~0.5 sec without. The FDR was high (51.4%) and the confidence test was marginal (10.7%). In contrast to what one would expect based purely on the FDR and CT results, however, the terms determined to be significant appear to be biologically appropriate. Multiple terms were related to rapid cell replication (e.g. "DNA replication and chromosome cycle", "S phase of mitotic cell cycle", "pre-replicative complex", etc.). The only unexpected term was "leucyl aminopeptidase activity" (p=0.00007). The biological role of this metalloexopeptidase activity is uncertain. It may be involved in the modification of signaling peptides, or it could be a false positive. A closer examination of the specific GOI that resulted in this high z-score would be interesting. Given the high percentage of significant terms that seemed biologically appropriate, it may be that the FDR and CT are themselves overly conservative measures. Alternatively, it may be that other, more statistically robust, means of separating GOI from NGOI may result in a higher confidence in the results. Indeed, we are examining methods to maximize the power of GOArray. The HTML formatted results of this analysis are available from the web site "http://microarray.yale.edu/ymd_public/white_science_ratio5.html".

References

[1] Osier, M.V., Tuck, D., White, K.P., Mason, C.E., Zhao, H., and Cheung, K.H. GOMine – a model for microarray interpretation. *Submitted.*

[2] Arbeitman, M.N., Furlong, E.E., Imam, F., Johnson, E., Null, B.H., Baker, B.S., Krasnow, M.A., Scott, M.P., Davis, R.W., White, K.P. 2002. Gene expression during the life cycle of Drosophila melanogaster. *Science* 297:2270-2275.