PANTHER Classification System: Protein subfamily curation, and application to expression data analysis

Anish Kejariwal, Huaiyu Mi, Nan Guo, and Paul D. Thomas

Evolutionary Systems Biology Group, SRI International
333 Ravenswood Ave., Menlo Park CA 94025, USA

The PANTHER (Protein Analysis Through Evolutionary Relationships) Classification System is freely available at http://www.pantherdb.org. PANTHER was designed to model evolutionary sequence-function relationships on a large scale. The current version of PANTHER (6.0) contains trees for over 5000 protein families, divided into over 30,000 functional subfamilies, and has been curated by expert biologists. The demo will cover the curation infrastructure for assigning functional classifications, and will cover one of the major applications for this data: expression data analysis.

The first step in constructing PANTHER families and subfamilies is to cluster proteins into families according to their sequence relationships, and build a sequence similarity tree for each family. A biologist curator reviews the tree structure and protein annotations, and cuts the tree into subtrees (subfamilies) based on divergence in sequence and function. Curators then associate each subfamily with ontology terms that describe the molecular function(s) of the proteins in the subfamily and the biological processes they are known or inferred to participate in (a subset of the terms available in the Gene Ontology). In addition to ontology terms, PANTHER curators can associate protein sequences with components in biological pathways, including both signaling and metabolic pathways. The curation infrastructure is available on the web, and well over one hundred individual biologist curators, located around the world, have worked on PANTHER. In the future, we plan to open up the curation process so that any expert biologist can be involved in the process.

HMMs are built for each curator-defined subfamily, and can be used to categorize genes into functional categories. The functional groupings can be used in statistical analysis of the results of genome-scale experiments, such as gene expression analysis, protein expression analysis, and even analysis of evolutionary selection pressure over large numbers of genes. Two expression analysis tools are available at http://www.pantherdb.org/tools.

The first tool utilizes the binomial test for analysis of gene or protein lists with respect to function. Each input list, as well as a reference list, is divided into groups based on functional classification (molecular function, biological process, or pathway). Then, for each functional category, the binomial test is applied to determine whether there is a statistical over- or under-representation of genes/proteins in the input list relative to the reference list, and P-values for these tests are given as output. From this output page, the user can export the statistics, or follow links to graphically view (as pie charts or bar graphs) the data that were used to compute the P-values, or look at the list of genes/proteins in any functional group. When pathways are chosen as the functional categories, clicking on the pathway name brings up pathway diagrams colored according to preferences specified by the user.

The second tool in this section is for analysis of genes/proteins that have numerical data associated with each gene/protein. The most commonly used numerical data are the fold-change values for each gene in a differential expression experiment, but the statistical test is general enough to handle any numerical data. The user inputs a file containing the gene or protein identifier, and the corresponding numerical value. The statistical tool builds a distribution of values for all input data in the list (this becomes the reference distribution), and then divides the input data into functional categories and builds a distribution of values for each functional category. The probability that the functional category distribution was drawn randomly from the reference distribution is estimated using the Mann-Whitney Rank-Sum Test (U-Test). Each distribution, and how it compares to the reference distribution, can be viewed graphically. The user can export the statistics, or look at the list of genes/proteins in any functional group. For pathways, clicking on the pathway name brings up pathway diagram colored using a “heat map” derived from the input values.

REFERENCES

Thomas, P.D., Kejariwal, A., Guo, N., Campbell, M., Muruganujan, A. and Lazareva-Ulitsky, B. (2006) Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. *Nucleic Acids Res.*, in print

Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremieux, O., Campbell, M.J., Kitano, H., Thomas, P.D. (2005) The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res*. **33**, D284-288.

Thomas, P.D., Campbell, M.C., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. (2003) PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res*., **13**, 2129-2141.