Inferring Gene Ontology category membership via gene expression and sequence similarity data analysis

Murilo Saraiva Queiroz, Francisco Prosdocimi, Izabela Freire Goertzel, Francisco Pereira Lobo, Cassio Pennachin and Ben Goertzel, Ph.D,
Biomind LLC, Rockville, MD/USA

The Gene Ontology (GO) database annotates a large number of genes according to their functions (the biological processes, molecular functions and cellular components in which they are involved). However, it is far from complete, and so there is a need for techniques that automatically assign GO functional categories to genes based on integration of available data. The present work describes one such technique, that uses a combination of sequence similarity and a similarity measure based on mutual information applied to cross-experiment microarray gene expression analysis.

First of all, in order to test the relevance of sequence similarity for gene function inference, similarity searches of genes belonging to the same GO (from here on we will use "GO" as a shorthand for "GO category", as well as for the "Gene Ontology" as a whole) were done across the human genome. A BLAST attachment value (BAV) for each GO was defined as the sum of the e-value exponents found between pairs of genes in the GO, divided by the sum of all e-value exponents found between genes in the GO and genes outside the GO.

Next, to assess the "expression based similarity" of human genes, we used a dataset (GDS181) from GEO, a gene expression and molecular data repository maintained by the NCBI, providing gene expression profiles from 85 different tissues, organs, and cell lines in the normal physiological state. The dataset contains 12,625 probes, and we used 9,725 of them associated to genes with identifiable GO relationships. For each gene in the dataset, we calculated the Mutual Information (MI) between its expression values measured across all tissues and the corresponding values for the other genes. In order to calculate MI, the gene expression values were discretized, meaning that each one was replaced by one of K symbols. The symbol replacing an expression value was calculated by first normalizing the values into [0,1], and then partitioning this interval into K equally sized subintervals. The normalization was done on a per-gene basis. After experimenting with several different values of K, a value of K=3 was chosen for all further experiments. Using a similar procedure to the one used for calculating the BAV, a MI attachment value (MiAV) was obtained. For each GO, the MiAV was defined as the sum of the MI expression values found for all pairs of genes in the GO, divided by the sum of all MI values between genes in the GO and genes outside the GO.

Then, our gene function inference (GFI) process proceeds as follows. Given a gene for which one wants to know the function, one begins by comparing it with all other genes, using both BLAST and expression data. Then, given a GO, one may calculate the values Bs and Ms, representing the maximum similarity found between the query gene and any gene inside the GO, using BLAST or MI, respectively. Those values plus attachments are used in the following equation for estimating the pertinence of a gene to a GO:

$$\text{(I)} \quad f(Bs,BAV,Ms,MiAV) = x_1(Bs^{y1}BAV^{y2}) + x_2(Ms^{y3}MiAV^{y4}) - z$$

Here, x_1, y_1, y_2, x_2, y_3, y_4, and z represent the weights of the equation. A gene should be classified as belonging to a GO if the equation above gives a value greater than zero when fed similarity and attachment values derived from the GO. A genetic algorithm was used to optimize the weights of this formula, based on assessing the performance of each weight-vector at predicting GO category membership over a training set. The objective function used by the GA was based on the F-measure, which takes into account both precision and recall.

A rigorous testing methodology was utilized. We set aside a subset of the GO and a subset of our overall human genome dataset to train our the genetic algorithm involved in our GFI model. Another subset of the GO and of the human genome dataset was used for testing; and further validation was obtained by applying the parameters learned with human data to the yeast genome. Yeast expression data was composed of the familiar Spellman dataset, and corresponding sequence data from SGD.

In our computational experiments, 2,386 new links were predicted between human genes and GO categories; and 1,111 links between yeast genes and GO categories, spanning the biological process, molecular function and cellular component ontologies. According to tests using the method to replicate already-known GO category assignments, the results are estimated to have precision bounded below by 73% for human data, and 83% for yeast.