Poster J-51

Combined visualization and analysis of Gene Ontology annotations using multivariate representations of annotations and bipartite networks

Authors:
Rohan Williams (*UNSW*)
Kai (Kevin) Xu (*National ICT Australia*)
Jane Huang (*University of Sydney*)
Chris Cotsapas (*UNSW*)
Seokhee Hong (*National ICT Australia*)
Geoff McCaughan (*University of Sydney*)
Mark Gorrell (*University of Sydney*)
Peter Little (*UNSW*)

Short Abstract: We present a "multi-scale" methodology that analyze GO annotations for high-throughput experiments in a network context at various levels of details and employs visualization, statistical analysis, and unsupervised learning techniques. We illustrate this methodology with microarray data from the human liver with hepatitis C virus (HCV) infection.

Long Abstract:
We present a “multi-scale” methodology that analyze GO annotations for high-throughput experiments in a network context at various levels of details and employs visualization, statistical analysis, and unsupervised learning techniques. We illustrate this methodology with microarray data from the human liver with hepatitis C virus (HCV) infection.

Gene Ontology (GO) has become the default means of representing gene-product functional data and is widely used as a basis for analyzing the results of high-throughput experiments. The analysis of GO annotations has focused on the notion of identifying which GO terms are over-represented in a set of genes of interest (relative to an appropriately defined set of background of genes). At an interpretative level, these methods tend to try and characterise the biological functionality of a set of genes with one or a few GO terms. Thus, the diversity of biological function, inherent in any set of gene-products, tends to be neglected by these approaches. Similarly, methods of visualization of ontologies and their annotations have tended to focus on the use of the directed acyclic graph (rDAG) structure of GO itself. We argue that capturing the diversity of biological function should be of fundamental importance when developing methodology to aid the large-scale visualization and analysis of high-throughput biological data.

In this paper, we propose a “gene-GO network” that provides a “multi-scale” and integrated representation of gene identity, their functions (described using GO terms) and the relationships of the functions in the GO hierarchy. Given the large size of GO hierarchy, one of the important choices to build an sensible network is to choose a relevant part of the GO hierarchy. It is reasonable to include all the GO terms describing functions of interested

genes and the relationships between these terms, since these are the majority of available data. Therefore, in the gene-GO network, each gene is connected to GO terms describing its functions and GO terms are connected based on the GO hierarchy. It is also useful to include some of the “ancestor” of these GO terms, i.e., terms that are higher in the GO hierarchy and representing more generic functions, because this could reveal connections between genes and their functions that could be important to explain the functions of the genes as a group. Thus, we extend the concept to k-level gene-GO network, in which each gene is connected to GO term that is the k-th ancestor of the term it directly associates to. While there are multiple ancestors, all are included. This representation provides a “multi-scale” view of the data: a higher k value results in a more abstract network, whereas a lower k value gives a more detailed network.

The k-level gene-GO network can provide a visualisation framework for the exploration of annotations of large sets of genes. To use this construction in a more systematic fashion, we employ independent analysis methods in order to structure the entire visualisation. When dealing with GO annotations, capturing the diversity of functionality across a set of gene-products can be realised by recasting the annotation into a multivariate data matrix, with gene-products indexed by rows (or, as 'samples') and the GO terms indexed by columns. Elements of this matrix are non-zero if the gene annotates to a GO term, either directly or by transitive relation in the ontology (via the 'true path rule'). Here, we choose to use heirarchical clustering of the GO annotation matrix to identify groups of genes that share common functionality under GO Biological Process.

We provide an example to illustrate our approach. The experimental data used in this study was collected as part of a study into the molecular pathogenesis of liver disease in hepatitis C virus infection in humans. The dendrogram generated from gene-GO matrix of the top 100 differentially expressed genes can be summarized as follows: 11 of 19 clusters had >3 gene membership and the remaining 5 clusters had 2 genes per cluster. For example, a subset of the 1-level gene-GO network that arises from the cluster of 20 genes (characterised by immune response; all 20 genes) defined using hierarchical clustering of the gene-GO matrix. Although immune related processes predominate in this cluster, we note that other context-relevant biological functions are present, including inflammation, chemotaxis, cell motility and cell-cell signalling, and signal transduction and cell surface receptor linked signal transduction. Although these relationships are present in the k-level gene-network representation, the fact that genes can be related across multiple functions or processes could confound their identification using purely visual approaches, however they are readily detectable using augmented analysis using clustering methods.

In summary, we introduce a k-level gene-GO network methodology that can be used to explore the GO annotations of large sets of genes, and, in combination with complementary data analysis methods, permit the detection of sets of genes with related biological functionality. Although we have chosen a microarray data set as an example, our methodology is not limited to this setting and could be employed across a diverse range of problems in high-throughput biology.