A Scoring Algorithm for Ontology Information Extraction

David.Outteridge@uchsc.edu
Department of Pharmacology
University of Colorado Health Sciences Center

Interpreting the results of a high-throughput gene expression experiment can create a list of genes which can be associated with the established collective knowledge base by using some publicly available mapping to a knowledge bank. For example, Locuslink can be used to map between GenBank and Gene Ontology identifiers.

A scoring algorithm has been developed at UCHSC which identifies the ontology entries which describe the largest numbers of genes. This enables rapid extraction of the potentially most useful information about the genes.

The procedure is to reverse the original mapping, i.e., establish which genes are described by each ontology entry, which results in some ontology entries becoming more interesting than others by virtue of the number of experimentally significant genes they describe. In addition, there is information contained within the ontology hierarchy(ies), since the entries have ancestral relationships (e.g., is-a). For example, if, in the set of entries, A is-a B, C is-a B, and D is-a C, and A, C, & D all describe many genes, then an investigator may be interested in the role of B in his experiment, even if B does not show especially well by itself.

This concept is used as the basis for the scoring algorithm of the entries connected by the hierarchical relationships of the ontology. Thus a hierarchical ancestor gets credit for the genes described by its descendents. However, a satisfactory algorithm is more complex than simple addition, since being simplistic can lead to misleading results. For instance, in the example above, under simple addition it is always true that B will describe more genes than any of its descendants, which is not an interesting fact. Or it could be that A and C are of interest in their own right, and their, low scoring, common ancestor B is not of special interest.

A score is a gene density, i.e., a number of unique genes divided by a number of unique nodes. Two scores are used in the algorithm. The first score is the total number of genes described by a node and its hierarchical descendents divided by the number of hierarchical descendents plus one (for the node itself). This score decreases for a node scoring low relative to its descendents and draws attention to the higher scores of the descendents. The second score is the gene density of a node without any other consideration. The algorithm simply selects the higher score of these two for any given node. If the first score is the higher then attention is drawn both to the subject node and its descendents. If the second score is the higher then the effectsof uninteresting descendents are eliminated and attention is focussed on the subject node only.

Presentation of the resultant scores for (a few hundred) sets of entries also is demanding because of the quantity of data and because human judgement may be important for

resolving difficulties. Gradation of colour in a graphical presentation of the ontology hierachy is used, together with the ability to expand areas of interest for more detailed study. The overall effect is that parts of a graphical display are highlit.