Finding Relationships Between Gene Products Using the Gene Ontology Database

By

Patrick C. Bradshaw, PhD

Abstract

The Gene Ontology (GO) database structure provides a controlled vocabulary for annotating gene products with terms. The GO structure is composed of 3 branches: molecular function, biological process, and cellular component. To cluster similar gene products they were first clustered with other gene products with identical annotation. To find other slightly less similar gene products, this constraint was modified in 3 ways. First only evidence in the form of traceable author statements from the scientific literature was considered when clustering. Second, instead of an exact match of terms between 2 gene products one of the gene product's terms may have been a parent or child term of the other gene product's term. Thirdly only the molecular function branch of GO was considered when querying GO. These methods provide a starting point for grouping genes into clusters that may be of interest for biological scientists.

Introduction

Scientific researchers can compare DNA and protein similarity using sequence comparison algorithms such as Blast or FASTA. If no sequence similarity is discovered, it is difficult to classify or group sequences especially if the three-dimensional NMR or crystal structure of the gene product is unknown. However, the Gene Ontology (GO) may be used to group analogous genes that contain no sequence similarity at all. But tools for the common biological researcher to extract information from the GO database are yet to be fully developed.

The Gene Ontology (http://www.geneontology.org) is a controlled vocabulary for describing gene products (Gene Ontology Consortium 2002). It is composed of terms, which may be used as annotations or indices for gene products present in collaborating databases. GO may be represented as a directed acyclic graph (DAG) of terms. A DAG is similar to a hierarchy, but unlike a hierarchy DAG nodes may contain more than one parent node. GO was initiated by scientists at the Saccharomyces Genome Database (SGD), Flybase (the Drosophila Genome database), and the Mouse Genome Informatics Institute (Gene Ontology Consortium 2001). The Arabidopsis Information Resource organism (TAIR) and the Caenorhabditis elegans group later joined. Ontologies facilitate interactions between different systems and increases the ease of communication between groups of people.

GO is composed of three branches of concepts. These branches are molecular function, biological concept, and cellular component. Cellular component is the localization in a cell where a gene product is found. Examples of cellular components are nucleus, plasma membrane, and mitochondria. Examples of molecular functions are enzyme, transporter, and transcription factor. Examples of biological processes include glycolysis, transcription, and cell death. The three branches are completely separate and independent and were chosen so that each gene product from all known organisms may be described by a term from each of the three areas.

Biological researchers strive to find all of the gene products in common metabolic pathways relating to diseases. If they know the gene products that may contribute to the severity of a disease they have more targets in which to develop drugs to combat the disease. Common pathways in which many scientists are interested include cell death, the cell cycle, bacterial cell wall synthesis, and triglyceride synthesis. Unwanted activation of the cell death pathway results in neurodegenerative diseases and ischemia-reperfusion injury. Inhibition of the cell cycle would halt the growth of cancer cells. Antibiotics inhibit cell wall synthesis of pathogenic bacteria to cure infection. Drugs, which would decrease flux through the triglyceride synthesis pathway, would help people suffering from obesity. Because of these large pathways that need to be properly regulated to maintain the health of a person, it is difficult for scientists to discover drugs unless they know all of the complex interactions between gene products in the pathway.

Methods

The GO database was downloaded onto a PC running Windows 2000. GO was loaded into a MYSQL database. Perl scripts, which access the GO database using the DBI and DBD::MYSQL modules were written to cluster and analyze the clusters of gene products.

Five GO tables were used in the SQL queries. The names of the tables are gene product, association, term, evidence, and term2term. The association table links the evidence, term, and gene product tables together. The term2term table is used to find the parent or child of a term.

The first analysis performed clustered the gene products that contain identical terms together. To do this the list of term accession numbers were sorted numerically and then converted to a string. A string comparison was done to determine if any other gene products had an identical annotation. This type of analysis was performed in all of the other clustering methods also.

The second clustering method was similar to the first except that it limited the annotations examined to only those that contain a traceable author statement (TAS) evidence code.

The third clustering method was similar to the first except that it limited the terms examined to be only part of the molecular function branch of GO, not biological process or molecular function.

The fourth clustering method performed was similar to the first except that it contained a condition specifying that a parent or child of a term could substitute for the term itself for one term in the group when comparing the terms for identity.

Combining two clustering methods together was also performed for all combinations of clustering methods.

Results

The results of gene product clustering for individual clustering methods are shown in figure 1.

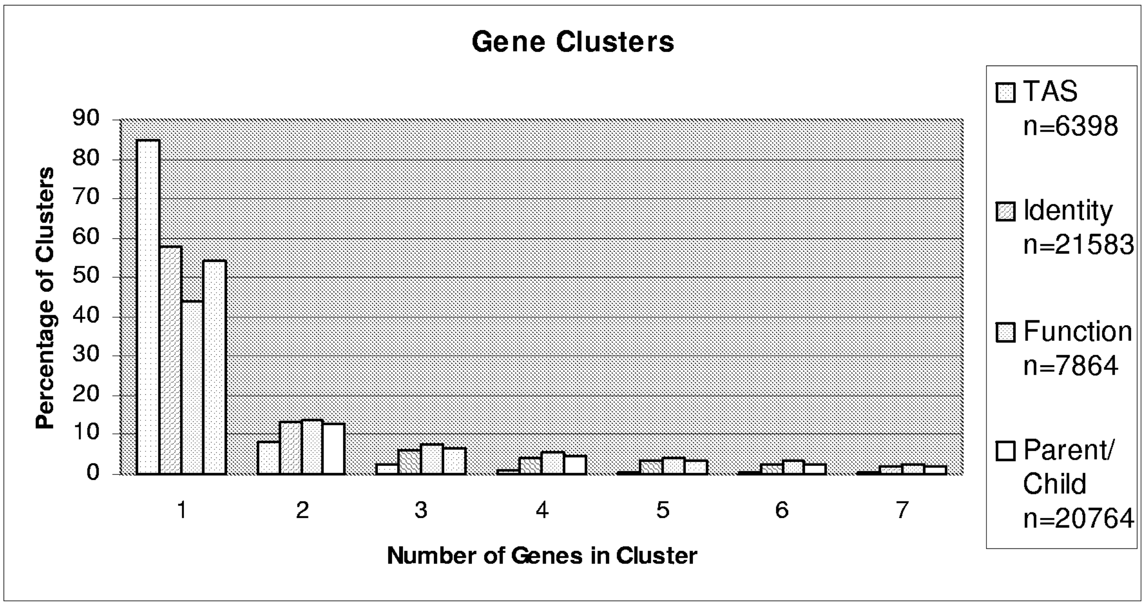

Figure 1. Gene clusters from the different gene product clustering methods.

Much information can be identified by looking at the group of gene products that don't cluster with any others (singularities) (as shown above 1 on the x-axis). The total number of clusters for each method is listed in the graph legend. Using the TAS and molecular function clustering methods led to roughly one third as many total clusters as strict identity or the parent/child method.

Clustering all gene products by strict identity results in 58 percent of the gene products examined being singularities. Using only annotations containing a TAS evidence code resulted in 85 percent of the clusters containing singularities. Relaxing the identity constraints decreased the percentage of total gene products examined that are singularities. Using the parent/child clustering method resulted in 54 % of the clusters containing singularities. Using terms of only molecular function led to 43 percent of the clusters being singularities. Interestingly the methods contain almost the same percentage of clusters having a cluster size of two. The other clusters of slightly larger size containing 3-7 members had a slightly lower percentage for TAS than the other methods.

When combining the different clustering methods, we examined the percentage of clusters containing singularities. Using only the molecular function terms containing TAS annotations lead to singularities being 65 percent of the total number of 3095 clusters. Limiting to molecular function with the parent-child method caused 31 percent of the 7541 clusters to be singularities. Limiting the set of annotations examined to TAS with the parent/child method led to 79 percent of the 6156 clusters being singularities.

Discussion and Conclusion

Modifying the identity constraint to only include TAS annotations greatly increased the percentage of clusters that were singularities while limiting to molecular function greatly reduced the percentage. The reason for this is that the number of possible combinations of clusters is reduced when considering function only. This is because the number of terms used is reduced so the combination of terms (clusters) is also reduced. Limiting to TAS does not reduce the possible combination of clusters while it limits the number of gene products considered by 92 percent (from 123,868 to 9,869).

Limiting strict identity by the parent/child method that we used did not affect the results very much. To reduce the number of singularities further the constraints could be changed to cluster together gene products that also had common sibling terms in addition to a common parent or child. Another way of saying this it to allow the clustering of two gene products when the terms that annotate each share a common parent or child term.

Another way of increasing the cluster size would be to allow one mismatch in the comparison of terms. For example gene products with terms A, B, and C would cluster with gene products with terms A and B. Once the results are expanded, it would be useful to develop a web interface in which a biological researcher could adjust the granularity or specificity of clustering to find a range of biologically related gene products.

When combining two clustering methods the results were numerically in between the results of when each of the individuals was examined. For example, alone, parent/child and TAS yielded percentages of 54 and 85, respectively and together the result was 79 percent. Alone TAS and function only yielded percentages of 85 and 43, respectively and together the result was 65 percent. The parent/child method combined with the function only method decreased the percent singularities from 43 to 31 even though the parent/child method only decreased the percentage using strict similarity from 58 to 54. The parent/child method must become more effective when the number of total terms considered is reduced such as when first applying the function only method.

Can any biologically meaningful results be obtained from these clustering methods? To obtain the result of this question a more detailed evaluation of the data must occur. The data can be compared to results from sequence homology and to results where scientists manually group gene products according to a particular biological interest.

Gene products which contain a nearly identical amino acid sequence from different species were annotated differently the majority of the time. It would be of interest to the scientific community if these gene products could be annotated identically where possible. Therefore clustering by identical annotations would produce more biologically relevant results.

In conclusion, clustering gene products that are functionally similar, in the same localization (component), or part of the same pathway (process) is an important step for data mining and knowledge discovery. However, using the GO to cluster biologically relevant gene products must be undertaken in an intelligent manner and this should be carefully evaluated. Of the methods developed here, the method of restricting to molecular function may be the most globally applicable since it slightly increases the cluster size while maintaining a high level of precision the majority of the time.

References

The Gene Ontology Consortium. (2001) Creating the Gene Ontology Resource: Design and implementation. *Gen. Res.* **11**, 1425-1433.

Gene Ontology Consortium. (online) site URL: http://www.geneontology.org, site visited 7/26/2002.