Using the Gene Ontology: gene product annotation

Midori Harris

July 18, 2002

The best known goal of the Gene Ontology (GO) project is the development of a dynamic, controlled vocabulary describing important aspects of molecular biology that can be applied to all organisms as knowledge of the roles of gene products is accumulating and changing. The usage of GO terms in database annotations is an indispensable aspect of the project, not only because gene product annotation is the purpose for which the GO vocabularies were originally conceived, but also because application of GO terms helps guide further ontology development.

Terms from the gene ontology are applied in the annotation of gene products. GO annotations are associations made between gene products and the GO terms that describe them. A gene product is an RNA or protein product encoded by a gene. Because a single gene may encode very different products with very different attributes, GO recommends associating GO terms with database objects representing gene products rather than genes. If identifiers are not available to distinguish individual gene products, GO terms may be associated with an identifier for a gene; a gene object is associated with all GO terms applicable to any of its products. A gene product can be annotated to zero or more nodes of each ontology, at any level within each ontology; annotation of a gene product to one ontology is independent of its annotation to other ontologies. Annotations should reflect the normal function, process, or localization (component) of the gene product; an activity or location observed only in a mutant or disease state is therefore not usually included.

The member databases of the GO Consortium use manual and automated methods to annotate genes or gene products using GO terms. Both manual and automated annotations are made according to two principles: first, every annotation must be attributed to a source, which may be a literature reference, another database or a computational analysis; second, the annotation must indicate what kind of evidence is found in the cited source to support the association between the gene product and the GO term. GO uses a simple controlled vocabulary to indicate the type of evidence found in the cited reference to support the annotation.

There are several different methods that can be used to determine which GO terms accurately describe a gene product. The most reliable annotations are those made manually by database curators based on primary and review literature. Manual annotations usually cite peer-reviewed papers that provide

experimental evidence to support the annotation. Several evidence codes represent types of experimental data, such as inferred from mutant phenotype (IMP) or inferred from direct assay (IDA).

Wholly or partially automated methods facilitate the annotation of much larger sets of known or predicted gene products than can be produced manually, but at a cost of less detailed, as well as less reliable annotations. The most commonly used automated methods are based on sequence similarity: annotations can be transferred from a well-characterized gene product identified by running BLAST or InterProScan on a protein of interest. GO also provides a set of external files that map GO terms to entries in other classification systems, including InterPro entries, SWISS-PROT keywords, EC numbers, and others. These mappings can be used to make transitive GO annotations: if the gene product of interest matches an InterPro entry or is already annotated with an EC number or a SWISS-PROT keyword, corresponding GO terms can be applied to the gene product. A single evidence code, 'inferred from electronic analysis' (IEA), is used to denote annotations made by computational methods, the results of which are not usually individually evaluated by a biologist.

Collaborating databases provide tab delimited files of links between database objects and GO terms, with supporting data. In addition to the most critical items – unique identifiers for the database object being annotated, the GO term, and the cited reference, plus the evidence code – the database which contributed the annotation, a symbol for the annotated gene or product, and indication whether the database object represents a gene, a transcript or a protein are also required. Non-mandatory supporting data, such as gene product names and synonyms, can be provided.

High-quality GO annotations, often based on curatorial review of published literature, are now available for gene products in many model organisms. In addition, large sets of annotations made using automated methods cover both model organisms and less experimentally tractable organisms, including human. These annotations are available from a shared central resource, which will permit cross-organism searches based on GO annotations and facilitate the annotation of gene products orthologous to well-characterized proteins or functional RNAs. GO annotation datasets can be obtained from the GO ftp site (`ftp://ftp.geneontology.org/pub/go/`) or searched along with the ontologies themselves using the web-based AmiGO browser (`http://www.godatabase.org/cgi-bin/go.cgi`).

To continue building and maintaining a body of annotation data, the GO Consortium intends to bring in groups that will provide annotations for more organisms; both old and new contributors will work to improve both the quantity and the quality of annotations. An important future direction for GO annotation is to explore approaches to quality control and consistency checking, which are increasing in importance as the use of GO annotation becomes widespread.