

Scope and Use of the Gene Ontology Vocabularies

Midori Harris

European Bioinformatics Institute, Cambridge, UK

As the amount of available sequence data and associated biological information increases, performing biologically relevant database queries becomes more challenging. The development and application of formal conceptualizations for various domains of biology has great potential to facilitate the organization, retrieval, sharing, and interpretation of biological data.

To address the growing need for meaningful annotation of genes and their products in different organisms, the Gene Ontology (GO) project is developing three independent, structured networks of terms are being developed to describe three key aspects of biology.

"Molecular function" describes the activities or tasks performed by individual gene products at the molecular level; "biological process" describes broad biological goals that are accomplished by ordered assemblies of molecular functions; and "cellular component" encompasses subcellular structures, locations, and macromolecular complexes. At present all three vocabularies are structured as directed acyclic graphs.

In addition to developing the GO vocabularies, the GO Consortium databases are using GO terms to provide gene product annotations, including high-quality annotations, often based on curatorial review of published literature, for gene products in many model organisms, as well as large sets of annotations made using automated methods. 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.

The GO project also develops software for querying, displaying, and manipulating ontologies and annotations. Tools created for use with GO can be applied to other similarly structured ontologies.

Because many important areas of biology are not covered by the three GO vocabularies, the GO consortium supports the development of other open bio-ontologies to complement GO vocabulary development. Examples of ontology domains useful for biology include nucleotide and amino acid sequence features, organism anatomies, phenotypes, and others. The integration of annotations using GO and complementary biological ontologies will facilitate the interconnection of data from disparate sources.

GO Consortium home page: <http://www.geneontology.org>

Biographical notes:

Dr. Midori Harris received her Ph.D from the Section of Biochemistry, Molecular and Cell Biology at Cornell University. Her research, in the laboratory of Prof. Bik-Kwoon Tye, focused on the regulation of DNA replication in the yeast *Saccharomyces cerevisiae*. She joined the *Saccharomyces* Genome Database (SGD) at Stanford University in 1998 as a Curator. Annotating the recently sequenced yeast genome sparked her interest in developing ways to describe the known or predicted functions of gene products. When SGD, the Mouse Genome Database (MGD) and Flybase formed the GO consortium she became involved in the development and use of the nascent GO vocabularies. In 2001 she moved to Cambridge, where she now holds the post

of GO Editor at the European Bioinformatics Institute in Hinxton, working full time on the further development of the GO vocabularies.