

Gene Ontology Users Meeting Abstracts
La Paloma Resort, Cottonwood Room
Tucson, AZ
February 1, 2002

Development of Ontologies

Midori Harris (GO Consortium)

Title: Development of the Gene Ontologies

Abstract:

This is an overview of the process by which new terms are added to GO and existing terms revised. Important considerations include the scope of each ontology, where a new term fits relative to existing terms, and term definitions.

P. Leszek Vincent (Maize Mapping Project, Plant Ontology™ Consortium)

Title: The Plant Ontology™ Consortium and plant ontologies.

Abstract:

Plant databases need to accurately & consistently document features (e.g. gene products, functions, phenotypes, traits, developmental stages, anatomy, morphology), using a syntax & vocabularies which facilitate inter-database searches. This interoperability will enable comparative genomic strategies to elucidate plant functions. The Plant Ontology™ Consortium (POC) is applying & extending the Gene Ontology™ paradigm (www.geneontology.org) to knowledge domains pertinent to plant taxa. The POC aims at providing ontologies and controlled vocabularies for monocot, dicot & other plant taxa - initially maize/corn (*Zea mays*), rice (*Oryza sativa*) and *Arabidopsis thaliana*, but extending to other important taxa in due course. The POC aims to facilitate the communications, productivity & collaborations amongst the core participants of the POC involved in developing ontologies & controlled vocabularies for plant taxa. Further aims: numerical growth of participants; extended collaboration with the research of the GO consortium; provision of educational opportunities in this area of bioinformatics research. It is anticipated that the POC will impact the bioinformatics research of other national & international plant-based research groups/researchers (e.g. Soybean, Sugarcane, Microarray), via the provision of ontology products, community resources & educational inputs. A small sample of ontology & controlled vocabulary for gross morphology - based on *Zea mays* is presented.

Peter E Midford (University of Arizona)

Title: Behavior: the ethogram as an ontology

Abstract:

I will discuss two ontology projects related to behavior. The first is based on a published ethogram (Hailman and Elowson 1992) that describes Loggerhead sea turtle (*Caretta caretta*) nesting behavior. This ontology has been fairly stable since last fall. The second, ongoing, project involves coding videotaped courtship behavior in *Habronattus* jumping spiders. I will also discuss semantic issues pertaining to constructing ontologies for behavior. As time allows, I will share my experiences translating a portion of the Loggerhead ontology to a GO compatible format. This work is supported by an NSF Bioinformatics Postdoctoral Fellowship, DBI-0074524.

Development of Ontologies, continued

Pankaj Jaiswal (Cornell University)

Title: Efforts on development and integration of controlled vocabulary at Gramene.

Abstract:

Gramene (<http://www.gramene.org>) is a comparative genome database for cereal crops and a community resource for rice. We are populating and curating Gramene with annotated sequence data and associated biological information including mutants, phenotypes, polymorphisms and Quantitative Trait Loci. In order to support queries across various data sets as well as across external databases, Gramene will employ three related controlled vocabularies. First, a Trait Ontology (TO) will be implemented across the cereal crops and plants to curate and evaluate phenotype comparisons. An initial vocabulary for TO and definitions for TO terms is available at (http://www.gramene.org/plant_ontology/). Second, a Plant Ontology (PO) will facilitate the curation of morphological and anatomical feature information with respect to expression and localization of gene and gene products. The TO and PO are both in the early stages of development in collaboration with International Rice Research Institute, TAIR, MaizeDB and International Crop Information System. Finally, we have started to classify the confirmed or predicted rice genes by integrating the GO components. The development of plant specific vocabularies is open for community discussion and are encouraged for suggestions, additions or modifications of various ontology terms by web based submission form. The form is available at http://www.gramene.org/plant_ontology/submission/

Ontology Mapping and Applications to the Annotation Pipeline

Richard Belew (UCSD)

Title: Reconciling Hierarchical Taxonomies

Abstract:

GO and related biological ontology efforts, as well as earlier keyword thesauri, all demonstrate that hierarchies are a natural way for people to organize information. However, different people and organizations tend to construct different conceptual hierarchies (e.g., contrast Yahoo! with the UseNet news hierarchy), and while there are often significant commonalities it is in general quite difficult to fully reconcile them. We are particularly interested in the problem of "docking" a narrower, more focused and refined topical hierarchy into a broader one. Two algorithms will be presented that can be used to match or dock hierarchies. The first matches hierarchies based on a weighted bipartite matching algorithm of (textual) features of nodes without consideration of their hierarchic organization, and the second is based on an attributed tree matching algorithm which uses both hierarchic structure and node features. We focus on the field of COMPUTER SCIENCE and several independently developed taxonomies for it, and present experimental results showing the performance of both algorithms.

Ontology Mapping and Applications to the Annotation Pipeline, continued

Evelyn Camon (SWISS-PROT/TrEMBL/InterPro, at EBI)

Title: Integration & Application of GO Annotation at EBI

Abstract:

As a member of the GO Consortium the SWISS-PROT/TrEMBL/InterPro database group at EBI seeks to characterise all held proteins with GO terms and fast track those with completed genomes. The groups' first objective was to create a framework in which to implement and store GO annotation across the various databases in such a way that it complemented the Consortium's format and enabled seamless data transfer. GO term assignments were manually curated to all InterPro entries (interpro2go mapping), EC numbers (ec2go mapping, Michael Ashburner) and, SWISS-PROT keywords (spkw2go mapping) and then transferred electronically to a table of matching SWISS-PROT/TrEMBL proteins. To date this IEA data (inferred from electronic annotation) accounts for 1.5 million GO term assignments in SWISS-PROT/TrEMBL. As part of a Consortium agreement to fast-track GO annotation to human data, the SWISS-PROT curators manually assigned GO terms based on abstracts to a SWISS-PROT/TrEMBL/Ensembl non-redundant proteome data set (approx. 3,042 proteins). The remaining GO annotation of human data (approx. 6,978 proteins) was electronically extracted from LocusLink and represents the manual assignments of Proteome Inc. Together, these assignments represent the first stage of the GOA project at EBI, released in Nov 01. For each association, cross references to SWISS-PROT, TrEMBL, Ensembl, IPI and GO are provided, and evidence for its annotation is presented according to GO's evidence codes. The manual assignment of GO terms to proteins of all organisms in SWISS-PROT and TrEMBL entries is ongoing and will be reflected in subsequent GOA releases. To facilitate data transfer, EBI will continue to update and publicise our GO mappings and associations via the EBI FTP server, QuickGO browser, SRS and GO home page.

Access to GO annotation from the EBI:

InterPro XML flat files: <http://www.ebi.ac.uk/interpro/search.html>

QuickGO Browser: <http://www.ebi.ac.uk/ego/>

GO annotation@EBI(GOA): <http://www.ebi.ac.uk/proteome/goa/goaHelp.html>

Proteome Analysis Database: <http://www.ebi.ac.uk/proteome/>

Sequence Retrieval System, SRS (via GO, GOA InterPro databases):

<http://srs.ebi.ac.uk/>

Public Gene Association files:

http://www.geneontology.org/gene_association.goa

ftp://ftp.geneontology.org/pub/go/gene-associations/gene_association.goa

ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/HUMAN/gene_association.goa

Ontology Mapping and Applications to the Annotation Pipeline, continued

Michelle Gwinn (TIGR)

Title: TIGR's use of the GO system - an update

Abstract:

TIGR has used role category schemes to organize genomic data since it began publishing genomes. Other genome centers have used similar, but distinct, systems of their own. In an effort to facilitate the exchange and usefulness of genomic data, TIGR is adopting the Gene Ontology system as our primary categorization scheme. To that end, we began our efforts with prokaryotic genomes by making associations to GO for the gene products of *Vibrio cholerae* (Heidelberg JF, et al. 2000). The *V. cholerae* associations were done manually resulting in approximately 3800 gene products assigned GO terms. We then used the GO associations for *V. cholerae* as a guide for additional microbial genomes. Currently, in addition to *V. cholerae*, we have completed *Shewanella oneidensis* (~4900 proteins) and have begun *Geobacter sulfurreducens*. To make GO associations for the eukaryote *Arabidopsis thaliana*, we manually assign GO terms to experimentally characterized gene products and then manually propagate those assignments to other related gene products within the same paralogous family. The result is greater than 750 gene products completed for *A. thaliana*. In addition, we have begun associating GO terms with HMMs from both TIGR (TIGRFAMs, Haft D., et al., 2001) and Pfam (Bateman A., et al. 2002) resulting in greater than 1000 HMMs with GO associations in TIGR's database.

The Annotation Pipeline: Use of sequence similarity methods

Antonio Planchard (MGI)

Title: Utilization of orthology relationships in curation

Abstract:

In addition to curating mouse genes at the level of their sequence, chromosome location, embryonic expression, and alleles with associated phenotypes, MGI also curates the orthology relationships between mouse genes and the genes of other mammalian organisms, including humans and rats. In many instances, characterization of the mouse gene product has not been performed. However, in a large fraction of these cases there exists an orthology relationship between the mouse gene and the human or rat genes. Because amino acid sequence similarity is a powerful predictor of functional conservation, this conservation can be used to hypothetically assign function to an uncharacterized protein. The instances where such an orthology exists are captured by MGI GO curators who then determine if the orthologous gene products have been characterized beyond the level available for the mouse gene product. If so, MGI GO curators will transfer this information regarding component, process or function to the mouse gene product. Since this type of annotation is curator-driven and based on sequence similarity, the evidence code used to represent the annotation is ISS (Inferred from Sequence Similarity.)

Finally, at MGI, the assignment of terms from the GO controlled vocabularies to mouse genes is also accomplished by the manual curation of peer-reviewed literature in which mouse genes are explicitly described and their products characterized; and by electronic importation and translation of SWISS-PROT keywords and InterPro domains to GO terms.

The Annotation Pipeline: Use of sequence similarity methods, continued

Jed Mathis (RGD)

Title: Incorporation of GO annotations and future plans for GO within RGD

Abstract:

The primary focus of RGD is to aid Rat researchers in studying the rat as a model organism for human disease. Gene Ontology provides a layer of annotations to the study of comparative genomics and helps in establishing alternative ways to query the database in a more scientifically meaningful fashion. Two different algorithms have been used to obtain initial informatic annotation results of the rat gene data within RGD. In order to avoid redundancy, manual checks have been performed on the annotations that resulted in accomplishing 3778 annotations to 1311 genes in RGD. These have been successfully submitted to the Gene Ontology Consortium and are now available in the gene reports in RGD with appropriate links to AmiGO and other external databases. We are beginning the ongoing manual curation of GO terms to extend and enhance our existing electronic annotation results allowing us to achieve our secondary goal of improving the level of evidence listed for the annotations. The GO annotations are currently stored within RGD as flat files and one of our longer-term goals is the incorporation of the ontology structure into database tables. Integrating the ontology into the database itself will facilitate the use of GO within RGD and also provide an alternative connection to the other model organism databases using the gene ontology. This in turn will allow us to pursue our primary focus of incorporating rat data with that of other organisms, particularly human and mouse.

Matt Berriman (Pathogen Genomes, at Sanger)

Title: GO for the primary annotation of genome sequencing projects

Abstract:

GO for the primary annotation of genome sequencing projects" Gene Ontology is increasingly being used during the primary or first-pass annotation of genome projects. From an annotators perspective, the use of GO evidence codes will be discussed with particular reference to the Malaria Genome Project.

The Annotation Trickle: Hand annotation of individual gene products

Rebecca Foulger (FlyBase)

Title: Assigning GO evidence codes in *Drosophila*

Abstract:

FlyBase currently contains genetic and molecular information on over 31,000 protein coding genes of the species *Drosophilidae*. Since the GO (Gene Ontology) project began, curators at FlyBase have been assigning GO terms to *Drosophila* gene products and currently there are over 2,500 unique GO terms and over 19,000 GO terms in total for *D. melanogaster* alone within the FlyBase database. For GO terms to be useful to the database user, they must be accompanied by a GO evidence code, enabling the user to judge how strongly the evidence supports the assignment of a particular GO term. However in FlyBase there remain approximately 400 GO terms without such an evidence code. To rectify these, the individual papers from which the GO terms were originally annotated are being analysed and the appropriate evidence code(s) assigned to each GO term, in order to create a more complete *Drosophila* research tool.

The Annotation Trickle: Hand annotation of individual gene products, continued

Karen Christie (SGD)

Title: Curation of GO annotations for *Saccharomyces*

Abstract:

To provide the best quality annotations for its users, the *Saccharomyces* Genome Database (SGD) incorporates GO annotations into its Locus Page displays, into Gene Summary Paragraphs, and into displays of microarray expression results. More than half of the approximately 6000 genes in the yeast genome have process and function annotations to specific terms other than "unknown". SGD curators are working to annotate genes which do not yet have GO annotations for one or more ontology and also to examine the existing annotations which bear the IEA evidence code. To do this curators read reviews and primary literature references to assign GO terms with evidence codes to these genes. In this way we are slowly, but steadily, increasing the percentage of the yeast genome which is covered by GO annotations to make a more complete tool for our users.

Uses of GO and GO Annotations

Elizabeth Nickerson (CSHL)

Title: The Human Genome KnowledgeBase

Abstract:

In response to the rapidly expanding body of knowledge concerning human biology we have created the Human Genome KnowledgeBase (GK). The GK serves as an integrative online resource for the scientific community. Taking the form of peer-reviewed, electronic mini-reviews describing processes in human biology, GK summations link researchers to all information relevant to the summation topic including genome and protein databases, literature references and the Gene Ontology. Complete pathways are described comprehensively as text and associated assertions implementing a controlled vocabulary. Assertions are simple statements describing accepted facts, for example *A binds B*, accompanied by all relevant references. All information documented as an assertion is searchable for cross-referencing. We are currently in the pilot phase of this project. The GK, including a portion of our first summation, DNA Replication, can be found at

<http://gkb.cshl.org>

<http://www.genomeknowledge.org>

Uses of GO and GO Annotations, continued

Scott Doniger (Gladstone Institute, UCSF)

Title: GenMAPP and Gene Ontology: Developing new tools for the organization and analysis of DNA microarray data

Abstract:

The large amounts of data produced by DNA microarray experiments have created a need for analytical tools that can quickly identify those biological processes that are differentially changed in an experiment. We have combined the extensive annotations provided by the Gene Ontology Project (GO) with GenMAPP to create tools that address this issue. GenMAPP is a free program designed for viewing and analyzing gene expression data on MAPPs representing biological pathways or any other functional grouping of genes. The GO annotations can be viewed in GenMAPP as a set of MAPP files containing a list of genes from each GO term. In addition, using the GenMAPP database and the GO gene association files, we are developing an additional tool for GenMAPP that will identify those GO processes, functions, and components that show the highest percentage of genes changed in a gene expression dataset. With this tool the user can perform custom queries to generate a list of GO categories ranked by the number of genes changed. The user can then examine specific gene changes on the corresponding GO term MAPP to develop new biological hypotheses. The GenMAPP program, an on-line tutorial, and a growing collection of MAPP files can be obtained from

<http://www.GenMAPP.org>

This work was supported by funding from the Gladstone Institutes and NIH grants HL61689 and HL66621 (Program for Genomics Applications).

Michael Caudy (Cornell)

Title: Using Gene Ontology databases for annotating functional DNA sequence motifs revealed by computational genome analysis

Abstract:

Genes contain combinations of DNA sequence motifs that mediate the function or regulation of the gene and its products. These motifs include transcription regulatory sequences that control the expression of the gene, sequences that encode specific protein domains that mediate protein function, and other functional motifs. We are using computational algorithms that identify clusters of specific combinations of sequence motifs throughout a given genome, followed by functional annotation of those sequences using online annotation databases. We currently are searching the *Drosophila* genome for clusters of specific combinations of transcription factor binding sites that might function as tissue-specific enhancers for promoters of nearby genes. The coordinates and specific combinations of sites within each cluster are saved as GFF files that can be immediately annotated using the GadFly GFF files available online at FlyBase and other GO web sites. By using recently-developed BioPerl-based software such as Bio::DB::GFF, Bio::Graphics, LDAS and the Generic Genome Browser, we can rapidly determine which clusters are positioned near potential genes of interest. However, the annotation information available in the GFFs is limited, and at present we need to search FlyBase or other databases for more comprehensive information. We wish to present our work to the GO community and ask for their input and advice as to how more fully to integrate Gene Ontology data into our analyses.

Uses of GO and GO Annotations, continued

Christopher Hogue (Mt. Sinai Hospital Research Institute)

Title: Large scale use of genome ontology for functional discovery in the Biomolecular Interaction Network Database

Abstract:

The information recently deposited into the Biomolecular Interaction Network Database has come from a diversity of new large-scale experiments with *Saccharomyces cerevisiae*. These include new methods reported in articles in both Science and Nature that report the large-scale discovery of genetic interaction networks, mass-spectrometry based protein complexes, and combined approaches involving yeast-two-hybrid and phage display systems. Taken together there is now approximately 15,143 experimentally determined pairwise interactions among 4825 yeast proteins. Remarkably the networks of interactions can be shown to be sufficient information for complex-finding algorithms to group proteins with known functional relationships. Annotating this information appropriately using GO is part of our challenge, and we have embraced the use of GO for the BIND database and for reporting functional information from large-scale experiments in yeast. Information hiding in interaction network graphics is a problem we are trying to address. We have applied a condensed set of 25 functional annotation groups to simplify a presentation of an interaction graphic with a total of 3617 interacting proteins identified by mass spectrometry. We hope to discuss our use of the GO ontology and work together to determine appropriate means of information hiding for presentation of annotation information on complex interaction network figures. The Biomolecular Interaction Network Database is found at:

<http://www.bind.ca/>

Hong Dang (Incellico)

Title: Evaluation of the accuracy of Gene Ontology (GO) assignments to sequences and integration of GO in a cross-referencing database."

Abstract:

We have incorporated the GO associations as part of a unified cross-referencing database for biological sciences, called the Coded Electronic Life library (CELL'). To evaluate the accuracy of the current GO association assignments, random samples of 100 each of GO function, process, or component associations to GenBank, SWISS-PROT accessions from Compugen, and to LocusLink from NCBI (900 entries total) were examined manually against PubMed references. Accuracies of these assignments were estimated as the lower bounds of 95% confidence, assuming a binomial distribution. The accuracy for Compugen assignments ranges from 0.68 to 0.79, while that for LocusLink ranges from 0.92 to 0.96. The utility of CELL in biological research is demonstrated by cross-clustering of microarray expression clusters to various ontological classifications, such as GO, pathway (KEGG). The GO associations of an elevated expression of a cluster of genes in the melanoma cell lines are consistent with their potential involvements in melanin biosynthesis reflecting the tissue of origin of the tumor cells. Such cross-clustering greatly accelerates the interpretation of experimental results, and the formation of new testable hypotheses.