Gene Ontology Users Meeting

Montgomery Ward Building, R 4-075
303 E Chicago Ave
Northwestern University
Chicago, IL

October 14, 2004

ABSTRACTS

Ontological Visualization of Protein-Protein Interactions

Harold J. Drabkin[1], Chris Hollenbeck[2], David P. Hill[1] and Judith A. Blake[1]

[1]Mouse Genome Informatics, The Jackson Laboratory, 600 Main St., Bar Harbor, ME 04609 USA and [2]Department of Computer Science, Rensselaer Polytechnic Institute, 110 Eighth St., Troy, NY 12180-3590

Cellular processes require the interaction of many proteins across several cellular compartments. Determining the collective network of such interactions should allow insight into which processes the individual members participate, and how they may be regulated. The Gene Ontology is used by many model organism databases to provide functional annotation of gene products. The annotation process provides a mechanism to document the binding of one gene product with another. The Mouse Genome Informatics system (MGI) currently provides annotation for over 900 genes using this method. Using the basic structure of these annotations, we have constructed protein interaction networks for various mouse proteins represented in the database. GO Annotation at MGI captures over 1300 potential interactions. These include 129 binary interactions and 125 interaction sets of three or greater. Three sets involve over 30 proteins, the largest involving 109 proteins. Several tools such as GraphViz, MGI Gene Ontology Term Finder, and the MGI Gene Ontology GO_Slim Chart Tool, and Vlad (VisuaL Annotation Display) are available to visualize and analyze these data. MGI integrates not only data used for GO annotation, but also data other aspects of mouse biology including embryonic gene expression, chromosome location, and alleles and their phenotypes. The integration of protein-protein network visualization into this system will be useful in determining the significance of more complex interaction networks.

MGI database resources are funded by NHGRI (HG00330, HG02273), NIH/NICHD (HD33745), and NCI (CA89713).

GO Slim Mapper Tool at SGD

Rama Balakrishnan, Karen Christie, Maria C. Costanzo, Kara Dolinski, Selina S. Dwight, Stacia R. Engel, Dianna G. Fisk, Jodi E. Hirschman, Eurie L. Hong, Robert Nash, Rose Oughtred, Marek Skrzypek, Barry Starr, Chandra L. Theesfeld, Gail Binkley, Qing Dong, Christopher Lane, Stuart Miyasato, Mark Schroeder, Anand Sethuraman, Mayank Thanawala, Shuai Weng, David Botstein, and J. Michael Cherry
(rama@genome.stanford.edu, www.yeastgenome.org)

In order to aid biologists in the utilization of GO annotations for *S. cerevisiae*, the *Saccharomyces* Genome Database (SGD) has recently added new options to the SGD GO Slim Mapper tool (also known as the GO Term Mapper). The GO Slim Mapper was designed for researchers employing large-scale methods of analysis. This tool can take a list of genes as input (e.g., genes that cluster in a microarray experiment) and can determine the upper level GO terms associated with the input genes by tracing the ontologies from the granular, specific term associated directly with a gene to upper level, more general parent GO Terms (GO Slim terms). Previously only one set of GO Slim terms was included, but now users have the option of choosing from three different sets. The Macromolecular Complex Terms set contains granular protein complex terms from the cellular component ontology and is useful for determining whether a protein of interest is a member of a particular complex. The Super GO-Slim set is a small set of very broad, high-level GO terms from the process, function, and component ontologies; it is useful for binning groups of genes in general categories. Finally, the Yeast GO-Slim set contains high-level GO terms that best represent the major biological processes, molecular functions, and cellular components that are found in *S. cerevisiae*. Together with other GO tools such as the GO Term Finder and GO Tree View, the GO Slim Mapper tool helps SGD users to analyze genome-wide or large-scale data in an efficient manner.

Tools for assessing GO annotation consistency

Mary E. Dolan, Li Ni, Megan Campbell, Chris Hollenbeck and Judith A. Blake
Mouse Genome Informatics, The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609 USA, http://www.informatics.jax.org

The Mouse Genome Informatics (MGI) system provides a comprehensive public resource about the laboratory mouse that integrates information on sequences, genes, expression, and mammalian orthology. The integration of such diverse data depends upon quality determinations of object identities and relationships and upon the use of defined, structured vocabularies (ontologies). The Gene Ontology (GO) project provides vocabularies in the domain of molecular biology that have fostered the use of functional annotation standards among model organism database systems. But a crucial requirement in making use of comparative analyses is that annotations be consistent.
We extend our work of assessing mouse-human GO annotation consistency in the context of curated orthology using GO annotations for the mouse genes using MGI and for the human genes using GOA (GO Annotation@EBI) to include mouse-rat comparisons based on GO annotations for rat genes from RGD (Rat Genome Database). We have augmented the comparison to include a three-way comparison in cases where curated mouse-human-rat orthologs triplets are available. In addition, we have developed a visualization tool to explore detailed annotation information to facilitate inconsistency resolution.

MGI is funded by grants from NIH/NHGRI, NIH/NICH and NCI. The GO project is funded by NIH/NHGRI and by the European Union RTD program.

Gene Ontology and the development of tools for improved data search, navigation, mining and visualization at Rat Genome Database (RGD)

Dean Pasko, Lan Zhao, Mary Shimoyama, Victoria Petri, Susan Bromberg, Simon Twigger, Norberto de la Cruz, Jiali Chen, Chunyu Fan, Cindy Foote, Glen Harris, Yuan Ji, Weihong Jin, Dawei Li, Jedidah Mathis, Natalya Nenasheva, Jeff Nie, Rajni Nigam, Dorothy Silver-Reilly, Weiye Wang, Wenhua Wu, Angela Zuniga-Meyer, Anne Kwitek, Peter Tonellato, and Howard Jacob
Bioinformatics Program and Human and Molecular Genetics Center, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226. http://rgd.mcw.edu

The use of Gene Ontology (GO) to annotate gene products has spurred the development of new tools to allow for more comprehensive and robust data integration, navigation and visualization. GO searches can be done both in the Ontology Browser and the Advanced Search tool - a new and unique feature of data integration. GO terms provide easy means of navigation around RGD via the Ontology Search and Browser, the Ontology and Object Reports. Each can be a starting point for searches, their wealth of information. Users can start from a Gene Report and the annotations within to link to the Ontology Report with all the objects annotated to the same term or link to other Object Reports. From the Ontology Report or Browser the user can choose to see the tree, other terms and their annotations, link to other Ontology or Object Reports. The Ontology Reports provide a genome-wide view of the chromosomal distribution of genes annotated to the term through the Gviewer tool. GO can be used to analyze very large sets of genes using another recently developed tool – the Gene Annotation that allows for a broad range of input and output choices. Finally, GO can provide a means to attach biology to the genome via the GBrowse. The tool allows users to view genes on a chosen chromosomal region. We added the Ontology and confidence tracks: genes are displayed along with their GO terms and the color-coded evidence codes. GO terms link to their respective Ontology Reports in all tools.

GO Chicken: From MudPIT to Swamp

Fiona M. McCarthy, Amanda M. Cooksey, G. Todd Pharr and Shane C. Burgess.
College of Veterinary Medicine, P.O. Box 6100, Mississippi State University, MS 39762.
fmccarthy@cvm.msstate.edu

The chicken genome project deposited 17 040 previously unidentified proteins into the NCBI non-redundant protein database (6/30/04), approximately doubling chicken entries. The NCBI database prefaces the names of these proteins with "PREDICTED"; meaning they have been electronically-inferred. During experiments that combined differential detergent fractionation with multidimensional protein identification technology (DDF MudPIT) to analyze the chicken Bursa of Fabricus proteome we identified 3 602 proteins, 2 142 (59.4 %) of which are PREDICTED; most thus have no GO classification. Currently only 3 091 out of 27 015 (11.4 %) chicken proteins have any GO annotation at all. We confirm for the first time that each gene with the PREDICTED preface produces at least one protein in vivo. Our plan for data-analysis was to rely first on GO annotation. GO was invaluable to assigning protein function, but only 8% our proteins currently are GO annotated. Because DDF MudPIT separates proteins by cellular component and physico-chemistry we can assist chicken GO annotation. We have generated a figurative "swamp" of data that will be electronically-annotated using InterPro2GO followed by continued in-depth human data-searches to identify direct experimental evidence.

The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology.

Evelyn Camon, Daniel Barrell D, Vivian Lee, Emily Dimmer and Rolf Apweiler
European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. goa@ebi.ac.uk

The Gene Ontology Annotation (GOA) database (http://www.ebi.ac.uk/GOA) aims to provide high-quality electronic and manual annotations to the UniProt Knowledgebase (Swiss-Prot, TrEMBL and PIR-PSD) using the standardized vocabulary of the Gene Ontology (GO). As a supplementary archive of GO annotation, GOA promotes a high level of integration of the knowledge represented in UniProt with other databases. This is achieved by converting UniProt annotation into a recognized computational format. GOA provides annotated entries for over 70,000 species (GOA-UniProt) and is the largest and most comprehensive open-source contributor of annotations to the GO Consortium annotation effort. By integrating GO annotations from other model organism groups (FlyBase, SGD, MGD, GeneDB, RGD, TAIR), GOA consolidates specialized knowledge and expertise to ensure the data remain a key reference for up-to-date biological information. Furthermore, the GOA database fully endorses the Human Proteomics Initiative by prioritizing the annotation of proteins likely to benefit human health and disease. In addition to a non-redundant set of annotations to the human, mouse and rat proteomes (GOA-Human, GOA-Mouse, GOA-Rat) and monthly releases of its GO annotation for all species (GOA-UniProt), a series of GO mapping files and specific cross-references in other databases are also regularly distributed. GOA can be queried through a simple user-friendly web interface (QuickGO) or downloaded in a parsable format via the EBI and GO FTP websites. The GOA data set can be used to enhance the annotation of particular model organism or gene expression data sets, although increasingly it has been used to evaluate GO predictions generated from text mining or protein interaction experiments. In 2005, the GOA team will build on its success and will continue to supplement the functional annotation of UniProt and work towards enhancing the ability of scientists to access all available biological information. Researchers wishing to query or contribute to the GOA project are encouraged to email: goa@ebi.ac.uk.

Interactive functional analysis from the Biomedical Literature

Bruce Schatz, ChengXiang Zhai, Gene Robinson
University of Illinois at Urbana-Champaign

The BeeSpace Environment (http://www.beespace.uiuc.edu/) project is building an information system for functional analysis, focusing on social behavior of honey bee. We have just been awarded $5M for this project from the NSF Frontiers in Integrative Biological Research program. An earlier project, the Interspace Prototype (http://www.canis.uiuc.edu/), showed it was feasible to index the entire biomedical literature to support conceptual navigation. This computation extracted noun phrases from all of Medline, then computed co-occurrence frequency -- abstract by abstract, collection by collection. A user could specify a general phrase, then select related phrases from automatically generated lists. The semantic indexing thus proceeds by computing the statistical context. In BeeSpace, we will construct an Interspace spanning biology (Biosis), medicine (Medline), agriculture (Agricola), by extracting out the conceptual phrases necessary to functionally analyze genome data. The statistical computation requires partitioning the entire discipline collection into small community collections, where narrow topics are consistently discussed. A typical scientific community might be 10,000 articles and 100 members. Previously, for a medical application, we used MeSH for the partitioning. For this biological application, we are planning to use GO, the Gene Ontology. Currently, the GO classifications are much stronger in molecular function than in biological process, reflecting the current state of biological knowledge. We hope that our effort might push the classification scheme at the higher levels towards social behavior. As we are just at the beginning of this project, comments and collaborators are most welcome from all viewpoints.

Using Cross-Products to Represent Dictyostelium Development

Pascale Gaudet, Petra Fey, Karen Pilcher, Warren Kibbe and Rex Chisholm
dictyBase, Northwestern University, Chicago, IL

The GO describes biological processes, cellular components and molecular functions in a species-independent manner. Other ontologies are currently being developed that describe chemicals, phenotypes, anatomy, etc. At dictyBase, we have produced an ontology that describes the anatomy of the eukaryote Dictyostelium discoideum (available at OBO http://obo.sourceforge.net/).

The existence of several ontologies allows the of creation 'cross-products' that maximize the utility of each ontology while avoiding redundancy. Moreover, it facilitates the maintenance of both ontologies, since they are much smaller then the composite ontology. This method has been used to depict the high-levels terms of heart development in mouse (Hill *et al*. Genome Res. 2002 12:1982).

We have taken advantage of the anatomy of Dictyostelium to test this method to represent the development of a simple multicellular organism. Forty-six terms were chosen from the development node of the GO process ontology and crossed with the 45 Dictyostelium anatomy ontology terms. The plain cross product results in 660 terms. Manual inspection was done to remove terms that do not apply to specific structures or cell types. The possibility of expanding the Dictyostelium development ontology to include terms from other nodes of the process ontology, including cell communication and cell death, will be discussed.

Creating compound annotations to represent biological detail in MGI

Hill, D.P., Drabkin, H.J., Diehl, A., Ni, L., Ringwald, M., and Blake, J.
Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, ME

Mouse Genome Informatics (MGI) is a database resource that represents biological data about the genetics and genomics of the laboratory mouse. MGI uses a variety of standardized structured vocabularies to describe the biology of the mouse. We have developed an annotation format that allows us to track and integrate disparate vocabulary-based information in a single annotation. MGI curators enter data from direct interpretation of experiments described in primary literature. Often, the manuscript results involve details that go beyond the use of any single vocabulary. Curators have taken an approach of using a structured text format to create private annotations that include information from several vocabularies. Vocabularies included in this system are anatomical dictionaries, the cell-type ontology and the GO vocabularies. Constructing compound annotations using a combination of vocabularies allows for precise descriptions of exactly when and where a gene product is involved in a biological process. Such structured annotations:
- allow for accurate representation of experimental conclusions.
- integrate annotations of different types of experimental data.
- guide the creation of new cross-product terms for a vocabulary.
- place composite information in the context of external ontologies, thus facilitating recovery and analysis of complex biological knowledge.

This project is funded by NIH/NICHD grant HD33745, NIH/NHGRI grants HG002273 and HG00330.