

Gene Ontology Users Meeting

McCaw Hall
Arrillaga Alumni Center
326 Galvez Street
Stanford University

January 15, 2004

8:00 - 8:30A Light Breakfast & Coffee

8:30 - 8:45A Welcome, Mike Cherry

8:45 - 10:05A Pathways, Cell types, Physiology (Mike Cherry Chair)

- Ewan Birney, EBI (1)

- Bruce Conklin, UCSF (2)

- Jonathan Bard, Edinburgh University (3)

- David Hill, MGI (4)

10:05 - 10:30A BREAK

10:30 - noon Literature Analysis (Suzanna Lewis, Chair)

- Molly James, Ingenuity Systems (5)

- Colleen Crangle, ConverSpeech (6)

- Lonny Montoya, NCGR (7)

- Nicholas Goncharoff, Reel Two, Inc. (8)

noon - 1:00P LUNCH

1:00 - 2:20P Tools (Judy Blake Chair)

- Gavin Sherlock, Stanford University (9)

- Joel Richardson, MGI (10)

- Sorin Draghici, Wayne State University (11)

- Jill Cheng, Affymetrix, Inc. (12)

2:20 - 2:40P BREAK

2:40 - 4:00P Poster Session

- Paulien Adamse, Plant Research International, Wageningen, NL (13)

- Evelyn B. Camon, EBI (14)

- Mary Dolan, MGI (15)

- Harold J. Drabkin, MGI (16)

- Rebecca Foulger, FlyBase (17)

- Brent Mishler, UC, Berkeley (18)

- Victoria Petri, RGD (19)
- Rama Balakrishnan, SGD (20)
- Mike Bada, University of Manchester (21)
- Tanya Berardini, Carnegie Institution (22)

4:00 - 5:00P Open Forum Discussion

5:00 Close of Meeting

TALKS:

(1)

Use of GenomeKnowledgeBase for pathway storage.

Ewan Birney

EBI, Wellcome Trust Campus, Cambridge, CB10 1SD, United Kingdom
(birney@ebi.ac.uk)

The GenomeKnowledgebase (GK) project (<http://www.genomeknowledge.org>) captures detailed pathway information of known molecular biology. GK has a rich data model of physical entities, complexes and reactions which allows the description of most known molecular events, from small molecule metabolism (eg, TCA cycle) to signalling pathways (eg, Insulin receptor pathway). GK makes extensive use of GO for activity assignment and also uses the compartment and biological process aspects of GO. The GK data model strictly separates species information and allows specific association of references to specific reactions. At GK we have been focusing on populating the database for human pathways. We have now over 600 human proteins references in GK and over 400 complexes participating in over 50 pathways, including some complex molecular biology pathways such as pre-mRNA splicing and the cell cycle. We can project this data 'onto' GO to provide new GO assignments and provide an important quality check for known GO assignments. We believe GK is an excellent mechanism for representing complex pathways leveraging the full use of GO for the controlled vocabulary aspects of biology. We aim to keep increasing the coverage of GK. GK is entirely open (both for our data and software) and we welcome all collaborations

(2)

GenMAPP V2: A Tool for Viewing and Analyzing Microarray Data on Biological Pathways.

Bruce R. Conklin, Kam D. Dahlquist, Nathan Salomonis, Kristina Hanspers, Lynn Ferrante, Karen Vranizan, Jeff Lawlor, Scott W. Doniger, Steven C. Lawlor

Pharmacology, Gladstone-UCSF, PO Box 419100, San Francisco, CA 94141
(bconklin@gladstone.ucsf.edu, <http://www.conklinLab.org/>)

GenMAPP (Gene MicroArray Pathway Profiler) is a free, stand-alone computer program designed for viewing and analyzing gene expression data on MAPPs representing biological pathways or any other functional grouping of genes. A MAPP is a special file format produced with the graphics tools in GenMAPP that depicts the biological relationship between genes or gene products. When a MAPP is linked to an expression dataset, GenMAPP automatically and dynamically color-codes the genes on the MAPP according to criteria supplied by the user. Using the ancillary program MAPPFinder, a user can navigate the GO lists to identify the groups of genes with the most significant gene expression changes. We are in the process of releasing a beta version of GenMAPP 2. This new version is available for several new species, including Drosophila, C.elegans and Zebrafish, in addition to Mouse, Rat, Human and Yeast. It is also possible to create a custom GenMAPP database for a species other than the ones already supported. The program now accepts a variety of gene ID types, such as Unigene, LocusLink, Affymetrix and the MOD ID's. Improvements have been made to many aspects of the program, including a new Zoom function and the ability to export whole sets of MAPPs as interactive HTMLs. The GenMAPP program and accessory files can be downloaded free of charge from <http://www.GenMAPP.org/>

(3)

Linking ontologies with the COBrA ontology editor.

Jonathan Bard and Stuart Aitken

Schools of Biomedical Sciences and Informatics, Edinburgh University, George Square, EH9 9XD, UK

(j.bard@ed.ac.uk)

We are integrating anatomy ontologies for the main model animals on the basis of homology, analogy and common cell-types (<http://www.xspan.org/>). For this, we have made (1) a **cell ontology** (with M Ashburner & D States; <http://obo.sourceforge.net/list.shtml>) with >400 cell types based on morphology, function, species etc. and (2) the COBrA editor (<http://www.xspan.org/applications/cobra/index.html>, details from stuart@aiai.ed.ac.uk).

COBrA runs under Java 1.4, reads two ontologies (in GO flat file, GO XML/RDF, DAGEdit flat file, RDFS, or OWL), and allows a user to make links between their terms (using a third, reference [e.g. cell] ontology if required) that are stored in a new ontology. As COBrA uses a superset of the 5 standard formats, it can save ontologies in any; it thus acts as a format translator. **Engineering:** the formatting has highlighted **3** problems in the RDF and OWL encodings of GO and OBO ontologies due to XML and RDF standards. **1:** Allowable names (Qnames) in URIs forbid ':', but GO IDs include ':'. Existing software libraries (e.g. Jena) require XML Qnames, and OWL-based tools for GO will need these resources. Solutions include replacing ':' with '.' or using the concept names as the URIref. **2:** Existing ontology

tools require files to be downloadable from their stated location so OBO ontologies require namespaces. **3:** The meaning of logical relations in bio-ontologies needs to be made consistent with OWL and RDFS. *Funding: BBSRS (UK).*

(4)

Building the process ontology one branch at a time.

Hill, D.P.¹, Berardini, T.², Foulger, R.³, de la Cruz, N.B.⁴

(1) Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, Maine, (2) The Arabidopsis Information Network, The Carnegie Institute, Palo Alto, California, (3) FlyBase, Cambridge University, Cambridge, U.K., (4) The Rat Genome Database, The Medical College of Wisconsin, Milwaukee, Wisconsin.

(dph@informatics.jax.org, <http://www.informatics.jax.org/>)

One of the most challenging aspects in the Gene Ontology project is the development of a structured vocabulary to represent Biological Process in a large variety of organisms. Particular challenges exist in the areas of the ontology that describe developmental processes and physiological processes. To ensure that the graph is compatible with different model-organism database needs, we have formed a multi-disciplinary working group to address developmental and physiological processes. One of the first steps in creating a vocabulary that satisfies all of our needs was to split out organismal from cellular processes. This split allows us to use the more universal cellular processes as building blocks for more complex and species-specific organismal processes. Once the general cellular and organismal processes are described, they will be used to describe more species-specific processes by combining them with orthogonal vocabularies such as anatomical-structure vocabularies. Eventually, all organismal processes will be broken down into processes that may be species-specific, but will be described by conserved cellular processes that are shared by many organisms. This inclusion of shared cellular processes will provide a level of integration to the vocabularies that may not be evident when viewed at a higher, species-specific level. 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.

(5)

Integrating GO and GO annotations with the Ingenuity ontology.

Molly James

Department of Ontology Modeling, Ingenuity Systems, 1565 Charleston Road, Mountain View, California, 94043

(mjames@ingenuity.com, <http://www.ingenuity.com/>)

One challenge in ontology construction is in aligning and integrating structures and

vocabularies of two different ontologies. The Ingenuity Ontology is a large ontology in the domain of molecular biology focused on the interactions between genes and gene products from humans, mice, and rats. At Ingenuity, we have partially aligned parts of our ontology with the Gene Ontology (GO) by incorporating into our ontology parts of the Gene Ontology structure and associated annotations. Challenges in aligning GO with our ontology structure include different hierarchy semantics (is-a versus part-of) and different approaches for naming and representing classes. In this presentation I will discuss these alignment challenges and the solutions used to overcome them. GO annotations were also incorporated into the Ingenuity ontology to supplement protein functional information. The process for incorporating these annotations into our ontology is discussed and an overview of how this information is ultimately used in the Ingenuity Pathway Analysis software is presented.

(6)

Using the Gene Ontology for text data mining: a case study with human disease genes in yeast.

Colleen Crangle

CoverSpeech & University of Ulster, 60 Kirby Place, Palo Alto, California, 94301
(crangle@converspeech.com, <http://www.converspeech.com/>)

BACKGROUND: As reports of gene-related discoveries increasingly appear in scientific publications, it becomes increasingly important to develop methods to access and analyze this information. In a series of studies, we are using sets of genes identified as new candidate genes for several putative mitochondrial-related disorders such as spastic paraplegia and Friedreich ataxia. We are using these targeted queries to develop improved ways to access and analyze free-text information about gene-related discoveries. **OBJECTIVE:** In this study, we examine the use of GO for this task. **METHOD:** The MEDLINE citation database was searched for all articles relevant to those genes newly identified as candidate genes for spastic paraplegia 5A. We used the ConverSpeech Distiller, a front-end to the PubMed database, for the searches. The Distiller has an integrated biomedical ontology, BioMedPlus, for term expansion and results analysis. BioMedPlus includes GO, which was applied independently in the analysis of the citations found. **RESULTS & CONCLUSIONS:** Full and partial matches were obtained to terms from the three ontologies of molecular function, biological process, and cellular component. We show which matches were made, which failed, and we demonstrate the linguistic processes we added to increase matching. We conclude that GO can improve access to gene-related free-text information and aid in the analysis of text if appropriate language processing is applied to the GO terms.

(7)

ConceptDB: A software system used to establish a link between PubMed and GO.

Lonny Montoya

Department of Bioinformatics, NCGR, 2935 Rodeo Park East, Santa Fe, New Mexico 87505
(lxm@ncgr.org, <http://www.ncgr.org/>)

ConceptDB is a knowledge discovery system developed at the National Center for Genome Resources that translates between different ontologies and provides for the interactive comparison and analysis of data categorized by those ontologies. ConceptDB is composed of a flexible data repository which enables it to relate virtually any type of data from virtually any data source, an interactive query tool with powerful query capabilities and set operations, and a graph visualization tool with graph analysis functions. One of the first implementations we chose for this system was to establish a translation between the PubMed citation database and the GO database. We first imported the PubMed database, which is a database of citations to papers in the area of medical research. PubMed is currently indexed by an ontology called MeSH (Medical Subject Headings). Since PubMed is also very useful to biologists conducting genetic research, we also imported the GO (Gene Ontology) database, which contains gene products annotated by GO. We then related the MeSH indexes from PubMed to the annotations of the gene products from GO using a MeSH to GO mapping provided by UMLS (Unified Medical Language System). Through this mapping we have provided an indirect link to PubMed making it searchable by GO annotation as well as gene product.

(8)

Using text classification software to expand Gene Ontology annotations.

Nicholas Goncharoff

Reel Two, Inc., 2255 Van Ness Avenue, San Francisco, California, 94109
(nicko@reeltwo.com, <http://www.reeltwo.com>)

Gene Ontology annotation includes curation of literature related to specific GO terms designed to give GO users an idea of the gene or protein process through references to experimental research. However, it is not meant to be a comprehensive listing of all literature and research related to each GO term. Researchers could thus benefit from an automated system that substantially increases the amount of GO-related literature. The European Bioinformatics Institute (GOA Project), Flybase, and Reel Two, Inc. have collaborated on a prototype system that automatically classifies Medline abstracts according to GO terms. This system further allows users to submit feedback on classification accuracy and supply additional training examples. During a 2-week evaluation based on the GO SLIM subset, GO annotators from EBI and Flybase found the system to be generally accurate. The main issue identified was, in some cases, that the training data appeared to be biased toward a single gene or protein, sometimes resulting in skewed classifications. EBI, Flybase and Reel Two are examining ways of modifying training data collection to address this problem. Expanding the system to cover the 5000 GO terms for which there exists a reasonable amount of training data could yield a valuable research tool. We present results from the initial evaluation, ideas

on improved training data collection, and what is required to make the GO literature classifications available to the scientific community.

(9)

GO::TermFinder Tool

Elizabeth I. Boyle, Shuai Weng, Jeremy Gollub, Heng Jin, David Botstein, J. Michael Cherry and **Gavin Sherlock**

Department of Genetics, Stanford University, Stanford, California, 94305-5120
(sherlock@genome.stanford.edu, <http://microarray.stanford.edu/>)

GO::TermFinder comprises a set of object-oriented Perl modules for accessing Gene Ontology information, and evaluating the collective annotation of a list of genes to GO terms, and can create an intuitive visual display of the output. It can be used to draw conclusions from microarray and other biological data, calculating the statistical significance of each annotation. GO::TermFinder can be used on any system on which Perl can be run, either as a command line application, or as a web-based CGI script. The full source code and documentation for GO::TermFinder are freely available from <http://search.cpan.org/dist/GO-TermFinder/>, and include several example scripts that can be used to batch process data.

(10)

Vlad: A New GO Tool for Visual Annotation Display

Joel Richardson

Mouse Genome informatics, The Jackson Laboratory, 600 Main Street, Bar Harbor, Maine 04609

(jer@informatics.jax.org, <http://www.informatics.jax.org/>)

Vlad is a new Web-based tool for visualizing the GO annotations of a set of genes. The visualization is intended to help uncover functional similarities and differences, for example, among a collection of upregulated genes from a microarray experiment. Relevant portions of the GO DAGs are rendered directly; node coloring, collapsing, and pruning are used to highlight areas of commonality. Optionally, significance scores can be computed to refine coloring. The data are also presented in tabular form as a supplement to the visual. Vlad uses [GraphViz](#) from AT&T as its graph layout and rendering engine. Vlad itself is a Java framework supporting directed graphs and traversals, reading GO data, interfacing with [GraphViz](#), and other essentials. Vlad is available at <http://www.informatics.jax.org/~jer/vlad>.

(11)

Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate.

Sorin Draghici

Department of Computer Science, Wayne State University, 5143 Cass Avenue, Detroit, Michigan 48084

(sod@cs.wayne.edu, <http://vortex.cs.wayne.edu/>)

Onto-Tools is a set of 4 seamlessly integrated tools based on 4 custom databases: Onto-Express (OE), Onto-Compare (OC), Onto-Design (OD) and Onto-Translate (OT). OE is able to automatically translate lists of genes found to be differentially regulated into functional profiles characterizing the impact of the condition studied upon various biological processes and pathways. OE uses GO to construct functional profiles for the following categories: biological process, cellular component, molecular function and chromosome information. Statistical significance values are calculated for each category. Once the initial exploratory analysis identified a number of relevant biological processes, specific mechanisms of interactions can be hypothesized. Currently, many commercial arrays are available for the investigation of specific mechanisms. Each such array is characterized by a biological bias determined by the extent to which the genes present on the array represent specific pathways. OC is a tool that allows efficient comparisons of any sets of commercial or custom arrays. Using OC, a researcher can determine quickly which array, or set of arrays, covers best the hypotheses studied. In many situations, no commercial arrays are available for specific biological mechanisms. OD is a tool that allows the user to select genes that represent given functional categories. OT allows the user to translate easily lists of accessions, UniGene clusters and Affymetrix probes into one another.

(12)

A Knowledge-Based Clustering Algorithm Driven by Gene Ontology.

Jill Cheng, Melissa Cline, John Martin, David Kulp, and Michael A. Siani-Rose
Affylab, Affymetrix, Inc., 3380 Central Express, Santa Clara, California 95051
(jill_cheng@affymetrix.com)

We have developed an algorithm for inferring the degree of similarity between genes by using the graph-based structure of Gene Ontology (GO). We applied this knowledge-based similarity metric to a clique-finding technique for detecting sets of related genes given a threshold. The extracted GO classes are evaluated by a Bootstrap Test to determine if they are significantly enriched compare to random sampling. A second application involves expression analysis and gene clustering. A co-cluster analysis was developed by combining prior knowledge of biological similarity and expression-based distance measures. This method accentuates genes with both similar expression profiles and similar biological characteristics. These algorithms are demonstrated in the analysis of MPRO cell differentiation time series experiments. Firstly, genes involved in 'transcription regulation'

and are found to be significantly enriched among genes with changed expression levels upon retinoic acid induction. This finding is consistent with previous studies. Secondly, sub-populations of genes sharing functional similarity as well as biological similarity were identified by the co-clustering algorithm. Our results support the notion that a knowledge-guided statistical approach can generate gene clusters that are more informative with respect to both expression profiling and biological meaning.

POSTERS:

(13)

Using ontologies for describing mutants in PlaNet, a network of European plant databases.

Paulien Adamse

Plant Research International, P.O. Box 16, NL-6700 AA Wageningen, The Netherlands.
(paulien.adamse@wur.nl, <http://www.plant.wur.nl/>)

PlaNet is a network of European plant databases for the systematic exploration of the genome of *Arabidopsis thaliana* and other plants. The network can be accessed using BioMOBY (<http://mips.gsf.de/proj/planet/> -> see *News for prototype*), a system through which a client is able to interact with multiple sources of biological data regardless of the underlying format or schema. The data is passed in the form of BioMOBY Objects - lightweight XML documents that conform to BioMOBY object descriptions. Several databases integrated into PlaNet contain information about mutants, among which the Wageningen *Arabidopsis thaliana* Database (**WAtDB**: <http://www.watdb.nl/>). Therefore we have defined new mutant-related BioMOBY objects. To enable the combination of data from different sources, ontologies play an essential role in the design of the databases and these BioMOBY objects. Gene Ontology identifiers are used to describe biological processes, molecular functions and cellular components that are affected in mutants. The phenotype description object uses plant anatomy, development stage and trait terms from a general plant ontology. We plan to use (and contribute to) the ontology being developed by the Plant Ontology ConsortiumTM (<http://www.plantontology.org/>). Special **modifiers** in combination with the GO and PO identifiers will describe how the mutant differs from the wild type. PlaNet is supported by the European Commission - project QLRI-CT-2001-00006.

(14)

The Gene Ontology Annotation (GOA) Database - Sharing Knowledge in UniProt With Gene Ontology

Evelyn Camon, Daniel Barrell, Vivian Lee, Emily Dimmer, Rolf Apweiler

UniProt Knowledgebase & GOA, EBI, Hinxton, Wellcome Trust Campus, Cambridge, CB10 1SD, United Kingdom
(camon@ebi.ac.uk, <http://www.ebi.ac.uk/goa/>)

The GOA database (<http://www.ebi.ac.uk/GOA/>) aims to provide high quality annotations to the UniProt Knowledgebase using the GO standardised vocabulary. As a supplementary archive of GO annotation, GOA promotes a high level of integration of the knowledge represented in UniProt with other databases. GOA provides annotated entries for nearly 60,000 species 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, GOA consolidates specialised knowledge and expertise to ensure the data remains a key reference for up-to-date biological information. Furthermore, the GOA database fully endorses the Human Proteomics Initiative by prioritising 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 proteome and monthly releases of its GO annotation for all species, 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 or downloaded in a parsable format via the EBI/GO FTP websites. The GOA dataset can be used to enhance the annotation of particular model organism or geneexpression datasets, although increasingly it has been used to evaluate GO predictions generated from text mining or protein interaction experiments.

(15)

A procedure for assessing GO annotation consistency.

Mary E. Dolan, Li Ni, and Judith A. Blake, Mouse Genome Informatics
The Jackson Laboratory, 600 Main Street, Bar Harbor, Maine 04609
(mdolan@informatics.jax.org, <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, mutant phenotypes 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 structured, controlled vocabularies in the domain of molecular biology. This has fostered the use of functional annotation standards among model organism database systems. The MGI group is one of the founding members of the Gene Ontology Consortium. A number of software tools have been developed to extend the use of the GO as a research analysis tool. Here we present a new procedure for assessing the consistency of mouse-human GO annotation in the context of curated orthology. For the complete set of mouse-human orthologs as maintained by MGI, we calculate independent GO_Slim categorizations for the mouse genes using MGI and for the human genes using GOA. Based on this categorization, we assess the consistency of the annotations, checking for matches, mismatches and missing annotation for every orthologous

gene pair. We report on the method, the detected inconsistencies, and how the inconsistencies could be resolved. 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.

(16)

Implementation of the Gene Ontology in the Mouse Genome Informatics System.

Harold J. Drabkin, David Hill, Alexander D. Diehl, Li Ni, Mary Dolan, Carol Bult, Janan T. Eppig, Joel E. Richardson, Martin Ringwald, Jim A. Kadin and Judith A. Blake
The Jackson Laboratory, 600 Main Street, Bar Harbor, Maine 04609
(hjd@informatics.jax.org, <http://www.informatics.jax.org/>)

The Mouse Genome Informatics System is a public resource for the integrated representation of relationships between murine gene sequence, function, expression, mutant alleles and their phenotypes, and tumor biology. Information is collected from a variety of sources, including primary literature and large data sets. MGI makes extensive use of controlled vocabularies such as the Gene Ontology (GO) to capture and display this data. The GO is used to describe a gene product's molecular function, cellular localization, and the biological processes that it participates in. Information is obtained from both primary literature curation, and large data sets using keyword translation tables. During literature curation, papers are selected for use in annotation based on standards developed at MGI. Our goal is to provide annotation to mouse genes based on experiments performed in the mouse. Annotations are based on the most current GO available. MGI obtains the GO from the GO FTP site daily, and new annotations are available to the public within 24 hours. Several strategies are used to monitor the quality of the annotations. Quality control reports track obsoleted GO terms from the daily loads and genes annotated to these receive immediate attention. Annotations to GO terms whose definitions have been altered are also logged for curator examination. 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.

(17)

GO data in FlyBase.

R. Foulger, M. Ashburner and the FlyBase Consortium.
FlyBase, Department of Genetics, University of Cambridge, Downing Street, Cambridge, CB2 3EH, United Kingdom.
(ref26@gen.cam.ac.uk)

FlyBase houses molecular and genetic data on all species from the family Drosophilidae, with much of this data stored as controlled vocabularies. This improves data maintenance and also enables efficient searching, both within FlyBase and between other organism databases.

GO (gene ontology) is one of the controlled vocabularies. GO terms are incorporated into FlyBase by a range of methods including curation of primary papers, reviews and conference abstracts, sequence analysis, and input from users. In addition to these manual approaches, collaborations with external groups that have developed tools to electronically assign GO terms to gene products, has led to the bulk annotation of thousands of *Drosophila* genes. Together with the ongoing development of the three ontologies, the combined curation methods ensure that GO data in FlyBase is kept up-to-date for the GO and fly community.

(18)

DEEP GENE: Using green plants to examine relationships between phylogeny, homology, and ontology.

Brent Mishler

Department of Integrative Biology, University of California, VLSB, Berkeley, California 94720-3140

(bmishler@socrates.Berkeley.EDU, <http://ucjeps.berkeley.EDU/bryolab/>)

The green plants represent one of the biggest branches of the tree of life -- more than 1/2 million species -- a clade at least 1 billion years old. Their morphological and chemical diversity, ecological dominance, and importance in human affairs (for food, shelter, and medicines) are paramount among life's lineages. A greatly improved understanding of their phylogeny, coupled with the extensive ongoing genomics projects on Arabidopsis and diverse crop plants, provide an unprecedented opportunity to use the green plants as a model system for all aspects of comparative biology. Systematists have long been interested in the concept of homology, providing names of lineages and parts of organisms. Molecular biologists have also become intensely interested in the naming of genes and phenotypes. There should be many mutual cross-disciplinary insights to be gained. The Deep Gene research coordination network has been funded (NSF DEB-0090227) to explore the ways in which comparative phylogenetic studies can inform genomic studies, and vice-versa. The group is sponsoring a series of professional meetings, workshops, training activities for K-12 teachers, undergraduates, and graduate students, and a web site (<http://ucjeps.berkeley.edu/bryolab/deepgene/index.html>). The fruits of such interdisciplinary collaboration could include new tools for assessing plant relationships as well as new comparative approaches to functional questions combining data from phylogeny and genomics.

(19)

Implementation of multiple ontologies at the Rat Genome Database (RGD).

Norberto de la Cruz¹, Mary Shimoyama¹, Lan Zhao¹, Weiye Wang¹, Simon Twigger¹, **Victoria Petri**¹, Dean Pasko¹, Susan Bromberg¹, Chin-Fu Chen¹, Jiali Chen¹, Chunyu

Fan¹, Aubrey Hughes¹, Jed Mathis¹, Nataliya Nenasheva¹, Rajni Nigam¹, Wenhua Wu¹, Angela Zuniga-Meyer¹, Peter Tonellato¹, Howard Jacob².

(1) Bioinformatics Research Center, and (2) Human Molecular Genetics Center, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, Wisconsin 53226 (vpetri@mcw.edu, <http://www.rgd.mcw.edu/>)

The Rat Genome Database (RGD) has adopted several ontologies to enable scientists to link information through related concepts. We provide search tools that allow users to use the annotations as a way to explore/use the database. Biological ontologies are controlled vocabularies that order concepts in hierarchical fashion. The 'true path rule' governs the way terms relate within directed acyclic graphs (DAGs). Initially, the Gene Ontology (GO) was implemented. However, we determined that additional ontologies were needed in order to capture the richness of available information pertinent to rat research. To this end RGD has adapted and is also developing several other ontologies. The phenotype ontology developed by MGI is used for annotations of QTL data. This is particularly useful in representing QTL linked traits exhibited in rats that relate to diseases. Closely related, yet requiring a separate hierarchy, is the issue of diseases and behavior. We adopted the appropriate branches of the medical subject headings (MeSH) hierarchy used by the National Library of Medicine (NLM) to create a disease ontology (DO) and a behavioral ontology (BO). Efforts are in progress to annotate gene, QTL, and strain data using these ontologies. Paralleling the adaptation and development of ontologies are the efforts to develop rules for linking data items/information and concepts in the ontology. We created annotation systems that more richly captures the attributes of the annotations.

(20)

GO Resources 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, Anand Sethuraman, Barry Starr, Chandra L. Theesfeld, Shuai Weng, Gail Binkley, Qing Dong, Christopher Lane, David Botstein, and J. Michael Cherry
Department of Genetics, Stanford University, Stanford, California, 94305-5120
(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 has created tools and graphical displays: the GO Term Finder, GO Term Mapper and the GO Tree view. The GO Term Mapper and GO Term Finder tools were designed for researchers employing large-scale methods of analysis. Both tools take a list of genes as input (e.g. genes that cluster in a microarray experiment) but each tool analyzes the input list in a different way. The GO Term Mapper determines 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 an upper level GO Term.

Currently, the upper level GO terms used by the GO Term Mapper are pre-defined by SGD and represent a broad slice of the ontologies. Recently a yeast specific GO Slim was developed and provided on the GO sites. In contrast to the GO Term Mapper, the GO Term Finder, rather than mapping the genes to a defined set of GO terms, searches the GO structure to find significant GO terms among the input genes, allowing researchers to identify functions or processes common to members of the set. A generic version of the GO Term Finder that utilizes GO data from the gene_associations file has also been developed in collaboration with Gavin Sherlock. The GO tree view helps users to place selected GO terms into context by graphically illustrating the terms within the GO structure; genes directly or indirectly associated with each term are also shown.

(21)

GOAT: The Gene Ontology Annotation Tool.

Mike Bada

Department of Computer Science, University of Manchester, Oxford Road, Manchester, M1 7AX United Kingdom

As the size of GO continues to grow, finding the terms that users need for the annotation of specific gene products is becoming increasingly difficult. Furthermore, because there are no links among the terms apart from those that form the taxonomic/partonomic hierarchy of GO, current annotation editors do not constrain the combinations of terms users may enter for a given gene product, potentially resulting in inconsistent or even nonsensical descriptions of biological molecules. We have created a formal Description-Logic-based version of GO (specifically, in DAML+OIL) and a database of associations among GO terms that were mined from GOA. Relying upon this version of GO, the database of associations, and the FaCT reasoner, the Gene Ontology Annotation Tool (GOAT; <http://goat.man.ac.uk>) aims to guide the user in the annotation of gene products with GO terms by suggesting the terms that are most likely to be appropriate based on terms previously entered for the given gene product. This can result in a less tedious annotation process by offering the user GO-term subtrees from which large numbers of terms that are probably irrelevant are excluded. In addition, the resulting annotations are likely to be of a higher quality, as the user is encouraged to choose a combination of terms for a specific gene product that have been found to be formally associated with each other in GOA.

(22)

TAIR 2 GO: controlled vocabularies and functional annotation at TAIR.

Tanya Berardini

Plant Biology, Carnegie Institution, 260 Panama St., Stanford, California, 94305
(tberardi@acom.stanford.edu, www.arabidopsis.org)

One of The Arabidopsis Information Resource's (TAIR's) current goals is to associate Arabidopsis genes with structured vocabulary terms developed by the Gene Ontology (GO) Consortium. We will annotate the entire Arabidopsis genome to GO vocabularies that describe the molecular function, biological process, and subcellular component of a gene product. As consortium members since 2000, TAIR has made significant contributions to developing and modifying the controlled vocabularies to accommodate plant gene product annotation. Our first annotations were done using computational methods that provided a general, low resolution overview of the transcriptome. The current strategy uses the experimental evidence in the literature to assign more granular GO terms to the approximately 10% of Arabidopsis genes that have been published in the literature (PubSearch, TAIR's literature curation tool). We have been collaborating closely with The Institute for Genome Research (TIGR) in the annotation effort and display our collective work on TAIR's gene and locus detail pages. Genes annotated with GO terms and expression patterns can now be found using keywords. The GO annotation bulk download interface at <http://www.arabidopsis.org/tools/bulk/go/index.html> allows researchers to obtain GO annotations for any gene or set of genes using locus names. In addition, our new keyword browser allows users to navigate through the ontology structures and explore term relationships and view definitions.