

## Talk Abstracts

### Literature-based annotation tools and strategies at TAIR

Suparna Mundodi, Tanya Berardini, Aisling Doyle, Margarita Garcia-Hernandez, Eva Huala, Lukas Mueller, Leonore Reiser, Julie Tacklind, Iris Xu, Daniel Yoo, Jungwon Yoon, Peifen Zhang and Sue Rhee

*The Arabidopsis Information Resource, The Carnegie Institution of Washington, 260 Panama St, Stanford, CA, USA*

The *Arabidopsis* Information Resources (TAIR) is a comprehensive data resource for the plant *Arabidopsis thaliana*. As members of the GO consortium since 2000, TAIR has made significant contributions to developing controlled vocabularies. In collaboration with TIGR (The Institute of Genomic Research), TAIR focuses on annotating the entire *Arabidopsis* genome. In order to facilitate the annotation process, TAIR has designed a powerful in-house software tool with Gene Ontology standard in mind, known as Pubsearch. Pubsearch incorporates literature, gene, functional annotations and keyword data and provides associations between various objects such as genes and go annotation to that gene. Further, Pubsearch provides a link to the local GO browser, AmiGO, that reads the local go database, which updates the latest version of the ontologies daily. Pubsearch software is available on Source Forge at <https://sourceforge.net/> for public download. Our first-pass GO annotations were obtained using several computational strategies. However, our current emphasis uses the existing literature to assign GO terms to gene products. These annotations will replace existing electronic annotations as well as provide new annotations for previously un-annotated genes. Some of our literature-based annotation strategies include gene family annotations and cellular component annotations. Our current search interface allows researchers to obtain GO annotations for any gene using locus names. Bulk annotations are available for download at <http://www.arabidopsis.org/tools/bulk/go/index.html>

## **Practical application of GO to RIKEN mouse cDNA clones.**

Hidemasa Bono

*Genome Exploration Research Group, RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi, Yokohama, 230-0045 Japan*

We have pursued RIKEN mouse encyclopedia project, which attempts to collect full-length enriched cDNAs. We held the FANTOM meeting in August 2000, which aimed to annotate functional description to 21,076 fully sequenced RIKEN mouse cDNA clones. In that meeting, the experts from genomics and bioinformatics extensively and substantially worked not only for the functional annotations, but for GO term annotations. These data are available with web-based viewer (FANTOM-DB, <http://fantom.gsc.riken.go.jp/db/>). GO annotation pipeline developed from the experience after much discussion in the FANTOM meeting was used for the FANTOM2 set, comprising 60,770 fully sequenced cDNA clones. Making full use of GO hierarchy and in-house gene association of these, mouse metabolome and tissue specific genes were analyzed. Metabolome was analyzed by reconstructing mouse metabolic pathways using KEGG by matching EC numbers. And tissue specific genes determined by microarray analysis were investigated using associated GO terms to get biological insights in that tissue. In order to look into the feature of biological functions, GO hierarchy was explored in detail with utilizing GeneAround GO viewer. These results will be presented in the meeting.

## **A general implementation of ontologies in model organism databases, the Mouse Genome Informatics model.**

Judith A. Blake

*Mouse Genome Informatics, The Jackson Laboratory, 600 Main St., Bar Harbor ME 04609 USA*

The Gene Ontology (GO) project provides a model for practical development of ontologies in the domain of molecular biology. Participating in the GO project since its inception, the Mouse Genome Informatics (MGI) system incorporates the GO and GO annotations in its representation of mouse genes. MGI ([www.informatics.jax.org](http://www.informatics.jax.org)) is a comprehensive, public bioinformatics resource about the laboratory mouse from sequence (genotype) to phenotype. MGI provides a consensus view of mouse biology as well as extensive integration of small and large experimental data sets. In addition to the GO, MGI incorporates structured vocabularies for mouse anatomy, phenotypes, and other semantic standards. New data structures in MGI support a generic DAG model that facilitates edits, views, browsing, and queries of ontologies and the associated knowledgebase. The MGI-GO group has also defined a GO-Slim view for the overall summation of mouse-to-GO information.

The Gene Ontology work is supported by NHGRI grant HG-02273. The Mouse Genome Database (MGD) and the Gene Expression Database (GXD) are two components of the MGI database system. MGD is supported by NHGRI grant HG-00330. GXD is supported by National Institute of Child Health and Human Development grant HD33745.

## **Semantic similarity measurements and the Gene Ontology.**

Phillip Lord, Robert Stevens, Andy Brass, and Carole Goble

*Department of Computer Science, University of Manchester, Oxford Road, Manchester M13 9PL UK*

The Gene Ontology (GO) is fast becoming the de facto standard for the annotation of gene products. One of the claims made, is that it should allow improved querying of databases. Different resources queried the same term should recover all and only entities conforming to that notion. One obvious way to query a database would be to ask for all proteins with 'semantically similar' annotation to a query protein. We have investigated using an information content based measure for semantic similarity. This uses the notion that less frequently occurring terms are more informative. Such information content based measures were initially developed to operate over the WordNet dictionary / thesaurus for such purposes as word sense disambiguation. In these cases the measures were validated over small humanly generated data sets. We have applied these measure to GO, and validated them over SWISS-PROT, by comparing sequence similarity, as defined by BLAST bit score, with the semantic similarity of the annotation. A highly significant correlation was found, serving to validate the measure. We have also used the measure to investigate the use of GO annotation within GO. Finally we discuss the use of this measure to develop a prototype search tool, which we believe represents the first step toward the application of semantic similarity as a valuable tool for the researcher.

## **Predicting gene function using patterns of annotation**

Oliver D. King<sup>1</sup>, James V. White<sup>2</sup>, Frederick P. Roth<sup>1</sup>

*<sup>1</sup>Dept. of Biol. Chem. and Mol. Pharmacology, Harvard Medical School; <sup>2</sup>JVWhite.Com*

The Gene Ontology (GO) Consortium has produced a controlled vocabulary for annotation of gene function that is used to annotate genes in several model organism databases. Probabilistic relationships between GO attributes allow the prediction of gene function based on partial annotation. For example, if two attributes tend to co-occur in a database, then a gene holding one attribute is likely to hold the other as well. We predicted which genes are likely to hold which attributes by modeling the relationships between attributes using two approaches: Bayesian networks, and optimal linear estimation. The fifty strongest resulting hypotheses for each method were inspected; those for the Bayesian network were found to be significantly more likely to be true than would be expected by chance. In more extensive cross-validation, the Bayesian network approach made 2703 correct predictions (64% sensitivity) with 361 'false' positives (99.93% specificity). This approach holds promise in assisting database curators and in generating strong testable hypotheses.

## **Beauty and the beast: GONG reconciles biological beauty with the computer science beast.**

Robert Stevens<sup>1</sup>, Chris Wroe<sup>1</sup>, Carole Goble<sup>1</sup>, Michael Ashburner<sup>2</sup>

<sup>1</sup>*University of Manchester, Oxford Road, Manchester M13 9PL UK;* <sup>2</sup>*Department of Genetics, Cambridge University, Cambridge CB2 3EH UK*

GO now captures a considerable amount of beautiful biological knowledge in its 11,000 terms; yet many computer scientists berate GO's representation. Surely the two communities should co-operate? Experience from other communities has shown that hand crafted terminologies are prone to errors. Many of these are errors of omission — in a graph that allows multiple inheritance, it is easy to omit some of the is-a relationships. The GONG project aims to develop a methodology to migrate a handcrafted, phrase-based ontology, to a descriptive, property-based form expressed in the description logic (DL) DAML+OIL and show that Computer Science can aid biologists in capturing knowledge. The end-point of the migration is a collection of logical expressions used to describe a concept's properties. DL's turn object orientated modelling of terminology on its head. Rather than manually classifying concepts and then describing their properties; the properties form a concept's logical definition from which a classification and logical consistency can be computed using a reasoner. We present early results from the project, which has focused on the enzyme and metabolism sections of GO. Use of the FaCT reasoner has produced a suggested set of additional is-a relationships; the majority of which have been added to the published version of GO, whilst detecting a reassuringly small number of logical inconsistencies.

GONG is a subcontract to Stanford University funded through the DAML program.

## Poster Abstracts

### Gene ontology annotations and tools in SGD.

Karen Christie

*Department of Genetics, Stanford University, Stanford CA 94305-5120*

The Gene Ontology (GO) Consortium is developing a species independent common language, a controlled vocabulary, to aid in the annotation of gene products. This system is useful for annotations produced by either curators or via computational analysis. To aid those performing large scale analysis, each model organism database generates a file, available for download, containing all of the associations between that organism's gene products and GO terms. To further help the analysis of multigene data sets, the *Saccharomyces* Genome Database (SGD) is developing new tools to analyze the GO annotations of groups of genes, whether derived from clustered expression results or based on some other selection criteria of the researcher's choosing. The SGD GO pages have incorporated links to the AmiGO browser, a useful tool created by the GO Consortium, where one can view all the genes, from many species, that have been annotated to the same term. Along with the incorporation of GO annotations into gene specific pages and into displays of microarray expression results, GO and these new tools to utilize it become even more useful as our GO annotations of the *S. cerevisiae* genome becomes more complete, an ongoing process at SGD. For information on the current progress of GO annotations at SGD or the Gene Ontology project, please visit the GO Consortium page at <http://www.geneontology.org>. SGD is found at <http://genome-www.stanford.edu/Saccharomyces/>.

### COMe, the ontology of bioinorganic proteins

Kirill Degtyarenko

*EMBL - EBI Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD UK*

The ontology is a structured vocabulary in the form of directed acyclic graph or a network in which each term may be a 'child' of one or more than one 'parent'. COMe (Co-Ordination of Metals) represents the ontology for bioinorganic centres in complex proteins. Therefore, COMe is orthogonal to GO as well as to structural or sequence classifications used for annotation of biological databases. COMe consists of three types of entities: bioinorganic motif (BIM), molecule (MOL), and complex proteins (PRX), each entity is assigned a unique identifier. BIM consists of at least one centre (metal atom, inorganic cluster, organic molecule) and two or more endogenous and/or exogenous ligands. MOL entity represents 'small molecule' which, in complex with polypeptide(s), forms a functional protein. The PRX entity refers to the functional protein as well as separate protein domains and subunits. The main groups of complex proteins in COMe are (1) metalloproteins, (2) organic prosthetic group proteins and (3) modified amino acid proteins. *IsKindOf* (ISA) relationship occurs between entities of the same class: PRX->PRX, BIM->BIM, MOL->MOL. *IsPartOf* (aggregation) relationship can occur between entities of the same class (BIM->BIM, MOL->MOL) or different classes (MOL->BIM, BIM->PRX). *IsBoundTo* relationship occurs only in the case (MOL->PRX). The data are currently stored in both XML format and relational database and are available at <http://www.ebi.ac.uk/~kirill/come/>.

## **GO Annotation in FlyBase.**

Rebecca Foulger

*FlyBase, Department of Genetics, Cambridge University, Cambridge CB2 3EH UK*

FlyBase is a database of genetic and molecular data for *Drosophila*, with *D. melanogaster* being the primary species represented. Gene Ontology (GO) annotations in FlyBase come from a variety of sources including scientific literature, sequence records, computer analyses and the original annotation of the *Drosophila* genome sequence. However, FlyBase was founded in 1992, six years before the GO Consortium began, and thus the annotation of genes with GO terms is inevitably incomplete. Several approaches are being taken to rectify this. SWISS-PROT records that are attributed to a *Drosophila* gene record are being curated. In addition, I am currently curating recent scientific reviews to increase the number of genes that have attributed GO data. Since 1999, FlyBase curators at Cambridge have been adding GO terms on a daily basis as they curate papers, and work is ongoing to incorporate much of the free text in the database, added before the GO Consortium was founded, into relevant GO data. Furthermore, these methods have been complimented with computer analyses, which have been used in conjunction with human review for GO annotation. Release 3 data from the *Drosophila* genome annotation is also emerging, and many of the resulting gene annotations will need to be evaluated and GO data added. All these approaches will increase the coverage of GO annotation in FlyBase, to provide a more complete information tool for *Drosophila* researchers.

## **Gene Ontology assignments to the *Arabidopsis thaliana* genome at TIGR.**

Hannick, L.I. , Maiti, R., Chan, A.P., Smith, R.K., Haas, B.J. Wortman, J., Whitelaw, C., White, O., Town, C.D., Fraser, C.M.

*The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850 USA .*

The Arabidopsis Genome Initiative produced the sequence of Arabidopsis in December 2000, using a BAC-based sequencing strategy. The effort resulted in a heterogeneous annotation. The Institute for Genomic Research (TIGR) is funded to reannotate the entire genome. Consistent naming of gene products is being applied and paralogous families of proteins have been constructed. Gene products in the genome have been organized into paralogous families. Paralogous proteins exist due to gene duplications that evolved from an ancestral gene. The grouping of these related proteins allows annotators to better evaluate the function of the predicted genes and to understand the gene duplication. We have identified these families based on Arabidopsis Paralogous Domains and Pfam HMM domain organization. To define the functions of encoded proteins systematically, consistent naming of gene products is being applied to the paralogous families. The Gene Ontology (GO) system of classification is being applied to Arabidopsis. The TIGR GO annotation is organized through the paralogous families. Many gene products now have GO assignments, each curated and inspected manually. The resulting annotation will be more consistent and complete, providing the research community with a unified resource for rapid progress plant research.

This work is supported by the National Science Foundation.

## **Non-homology approaches in functional and structural genomics**

Michael Lappe

*EMBL - EBI Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD UK*

Protein Interaction data have taken center stage in the experimental elucidation of protein function, describing their cooperativity in the context of dynamic complexes and pathways as interaction maps. We present a quantitative comparison of the two known methods for functional and structural prediction based on interaction patterns, namely 'guilt-by-association' and 'EMBed'. Using a comprehensive compilation of a number of different high-throughput datasets, we investigate how useful and accurate these methods are for the annotation of proteins. This analysis, based on the available protein interaction networks and GO, should provide some insight into the current state of automated annotation. Looking ahead, we suggest ways forward to estimate and eventually enhance the quality of interaction data based on orthogonal functional and structural annotations. Vice versa, we will discuss approaches to detect inconsistencies in the existing annotations based on established interactions. Along these lines, we describe a statistical approach to text-mining which allows to tap into the vast amount of information about protein function accumulated in the biological literature. Using the resulting structured information, we have developed an algorithm that is capable of accurate reconstruction of known pathways in the form of easily interpretable graphical summaries.

## **Integration and use of the GO<sup>TM</sup> vocabularies in the Bioknowledge<sup>®</sup> Library**

Lisa R. Matthews, Burk Braun, Angela Russell Winnier

*Incyte Genomics, 100 Cummings Center, Suite 420B, Beverly, MA 01915*

The BioKnowledge<sup>®</sup> Library (<http://www.incyte.com>) is a collection of literature-based databases that integrates genomic and proteomic data to provide information about biological function. The complete database volume set, now available through subscription, includes WormPD<sup>TM</sup>, YPD<sup>TM</sup> (*Saccharomyces cerevisiae*), PombePD<sup>TM</sup> (*Schizosaccharomyces pombe*), and for the first time to academic users, MycoPathPD<sup>TM</sup>, HumanPSD<sup>TM</sup>, and GPCR-PD<sup>TM</sup>. MycoPathPD is an annotated database of 17 human fungal pathogens. HumanPSD is a survey database of over 29,000 human, mouse, and rat proteins. Also a mammalian protein database, GPCR-PD provides literature-based annotations for G protein-coupled receptors, their ligands, and downstream signaling proteins. These databases are fully interconnected and searchable, and are accessible in a Web-based format. □

The descriptive terminology of the Gene Ontology<sup>TM</sup> Consortium (<http://www.geneontology.org/>) has now become fully integrated into all volumes of the BioKnowledge Library. Ph.D.-level scientific curators use the Gene Ontology (GO) system of controlled vocabulary to record structured textual annotations. Examples will be shown revealing how curation with GO terms alone creates a detailed picture of protein function, based solely on GO molecular function, biological process and cellular component properties.

The information in our databases enables transfer of knowledge about known proteins to unknown but related proteins. The common vocabulary provided by GO can also be used to describe the predicted properties of these uncharacterized proteins. Examples will be shown describing how knowledge incorporated into the BioKnowledge Library combined with GO terms predicts functional features of proteins of interest.

### **Linking GO annotation to the RZPD.**

Kathrin Meissner

*Bioinformatics, RZPD, Heubnerweg 6, Berlin 14059 Germany*

The Resource Center and Primary Database (RZPD) was initially established within the German Human Genome Project (DHGP). Its main tasks are the supply of researchers with high quality, standardized experimental material and clones and the collection and integration of data created with these materials. The RZPD harbors one of the most comprehensive clone collections world-wide. 225 cDNA libraries and 126 genomic libraries from 32 organisms contain about 30 million clones. The clone-related structure of the Primary Database provides the possibility to link and integrate various quality of data to physically existent clones at the RZPD. The RZPD started to link GeneOntology Identifier to the clones. Annotations that are provided by the EBI, Compugen, MGD and RGD for genes and proteins, respectively, will be used to link this information to the materials stored at and administered by the RZPD. Pipelines to link the EBI(human), Compugen Swissprot (human) and MGD (mouse) GO associations have been set up. Work to include information from the other sources is in progress. A first conception of a user interface has been designed and implemented enabling users to search for GO information linked to certain clones, genes or Unigene Clusters. Furthermore, it is possible to search for clones that match certain GO Identifiers. We are currently working on extending these query possibilities.

## **ProToGO - Evaluating biological features for a set of proteins using GO annotations.**

Hagit Ulanovsky<sup>1</sup>, Shany Ron<sup>1</sup> and Ilona Kifer<sup>2</sup>

<sup>1</sup>*Department of Biological Chemistry, Life Science Institute and* <sup>2</sup>*The School of Computer Science and Engineering, The Hebrew University, Jerusalem, 91904, Israel*

The accelerated rate of sequence accumulation raised a critical demand for assigning biological significance to large sets of sequences. In DNA microarray experiments, methods for clustering differentially expressed genes and for dealing with experimental pitfalls are established, while tools for quantifying the likelihood that a set of genes participates in a certain biological process are missing. We attempt to fill this gap by applying a new tool- ProToGO, that offers an on-line analysis on the biological features of a set of proteins. The ProToGO server receives a set of entries from the user (Swissprot, TrEMBL or Genbank accessions) and uses the EBI and Compugen GO annotations. Most entries are associated with multiple terms according to the GO partitions: Molecular function, Biological process and Cellular component. ProToGO produces a graph that includes the statistically significant GO terms that are associated with the query set, consistent with the complete GO graph. Each node is assigned a P-value to reflect the significance of obtaining such a node in the graph. ProToGO results are presented in a graphical or textual format according to the user preference. Another application of ProToGO is to assess the quality of clusters created automatically by ProtoNet. ProtoNet provides a classification for all proteins in Swissprot at different levels of granularity. ProToGO results can be used to assign a score for a cluster according to its purity and uniformity.