

Chicago GO Users Meeting - October 2001

Abstracts

Nathan Salomonis

<nsalomonis@gladstone.ucsf.edu>

Phone: (415) 826-7500, FAX:

URL: <http://www.genmapp.org>

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

There is a critical need for analytical approaches that allow for systematic identification of specific biological cascades altered in a gene expression dataset. To address this need, we have utilized an approach that takes advantage of the vast array of annotations currently being developed by the Gene Ontology project. Using the computer application GenMAPP, Gene Micro-Array Pathway Profiler, a freely available program designed to visualize gene expression data upon biological cascades, we have begun to represent Gene Ontology Biological Process, Molecular Function, and Cellular Component groups as MAPP files (GenMAPP format files). Once created, any user can modify the MAPP to illustrate specific biological interactions between genes within this program, thus providing a higher level organization to groups of biologically related genes described within Gene Ontology. In the context of gene expression data, such MAPPs will be amenable to advanced search queries that will enable the identification of those specific MAPPs that contain coordinate changes in gene expression according to the users own statistical criterion. Such an approach will allow researchers to quickly assess which biological cascades contain altered expression patterns and visually identify such changes on two-dimensional maps. The GenMAPP program, a growing collection of MAPPs, and gene databases for other model organisms can be downloaded from www.GenMAPP.org.

Elizabeth Nickerson

<nickerso@cshl.org>

Phone: 516-367-6977, FAX: 516-367-8389

The Human Genome KnowledgeBase

In response to the rapidly expanding body of knowledge concerning human biology we have created the Human Genome KnowledgeBase (GKB). The GKB serves as an integrative online resource for the scientific community. Taking the form of peer-reviewed, electronic mini-reviews describing processes in human biology, GKB 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 GKB, including a portion of our first summation, DNA Replication, can be found at gkb.cshl.org or at www.genomeknowledge.org.

Christopher Larsen

<clarsen@cognia.com>

Phone: 718-777-6558, FAX: 801-459-8850

URL: www.cognia.com

Ontologies and the world of Ubiquitin.

Cognia is a bioinformatics company involved in the aggregation of biological proteomics knowledge, and the creation of databases and tools for their protein analysis. We have created a working database regarding protein turnover, and specifically, the ubiquitin system. With the inclusion of over 1200 individual proteins so far, many novel functions and processes have been uncovered, some not yet in the gene ontology. We find that evolution has paralleled many of the originally discovered functions of the ubiquitin system for new and parallel processes. Thus, many of the new enzymology of sister systems to the ubiquitin system are used to promote entirely new effects. Here we present our database, compare the GO usage to our own ontology, and propose new additions to the current GO version. Our current website (<http://www.cognia.com>) details the focus of our company, and shows the key personnel involved in the project. Contact

Chris Larsen for further details, at clarsen@cognia.com

Michael D. Gonzales

[<mg@ncgr.org>](mailto:mg@ncgr.org)

Phone: 505-995-4436, FAX: 505-995-4432

URL: www.ncgr.org

Interpreting sequence, expression and phenotype data in the context of cellular interaction models.

To facilitate research on biological networks we have developed PathDB 2.0, a curated relational database of information on the mapping or interaction data between cellular building blocks (i.e. genes, proteins and metabolites) in order to build and analyze complex cellular models. This interaction data is derived from the literature, sequence annotation and from analytical methods (i.e. prediction of an ATP binding domain or a regulatory site). The data describing how these building blocks are modified and regulated is also curated. Our content development efforts are currently focused on Arabidopsis, comprised mainly of interactions between proteins and metabolites, as well as yeast which also includes annotation derived from other public databases, protein-protein, protein-DNA and protein-lipid data sets. Our data model includes many features of the Gene Ontology model that allows for easier navigation among biological classes. Information about subcellular location or phenotype information like programmed cell death or disease resistance can be displayed in the context of known biological interactions. Using PathDB, a pathway model can be used to visualize gene, protein and metabolite expression data and its view can then be synchronized with other software programs using NCGR's integration system ISYS. Since our analytical tools will help generate new interactions among building blocks, we allow users the ability to store the information in the database.

Lisa Matthews

[<lrn@proteome.com>](mailto:lrn@proteome.com)

Phone: 978-922-1643, FAX: 978-816-0285

Integration of the Gene Ontology vocabularies into Incyte's model organism and mammalian proteome databases

The BioKnowledge® Library represents a constantly growing collection of searchable databases that integrates knowledge from the research literature with genomic information. The current volumes in the BioKnowledge Library are: YPD™ for *S. cerevisiae*, PombePD™ for *S. pombe*, MycoPathPD™ for a collection of human fungal pathogens, WormPD™ for *C. elegans*, as well as HumanPSD™ and GPCR-PD™ both focusing on human, mouse and rat proteins. Protein report pages are interlinked to allow searching across multiple proteins and species for common protein characteristics. The descriptive terminology of the Gene Ontology™ Consortium (<http://www.geneontology.org/>) has now been integrated into 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.

Simon Twigger

<simont@mcw.edu>

Phone: 414-456-8802, FAX: 414-456-6595

[Rat Genome Database]

The primary focus of RGD is to aid Rat researchers in their work [on] rat as a model organism for human disease. Comparative genomics and the ability to incorporate data from Human and Mouse studies are a key aspect of this focus. To support sequence-based comparative studies we have released VCMMap, a dynamic sequence-based homology tool, which enables Rat, Mouse and Human researchers to view mapped genes and sequences and their locations in the other two organisms. Gene Ontology annotations will provide a further layer of annotation to these comparative studies in addition to creating 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. The analysis of these results and our plans for the further

incorporation of GO into RGD will be presented.

Natalia Maltsev

<maltsev@mcs.anl.gov>

Phone: 630-252-5195

URL: <http://selkov.mcs.anl.gov/WIT2/Natalia>

[Metabolism]

In the past decade the scientific community has witnessed a rapid [increase in the amount] of sequence data and data related to the physiology and biochemistry of organisms. Analysis of the genetic sequences and metabolic processes in phylogenetically diverse set of organisms revealed a substantial degree of similarity between the biochemical pathways in eukaryotic and prokaryotic cells, suggesting their common evolutionary origin. Developing of a dynamic controlled vocabulary of biological function relevant to metabolism and providing a functional context for genomic data is vital to further understanding of biological systems and their evolution. The Computational Biology group at the Mathematics and Computer Science Division of Argonne National Laboratory has substantial expertise in designing integrated systems for sequence analysis and metabolic reconstruction. WIT2 system currently available at Argonne (<http://wit.mcs.anl.gov/WIT2>) contains The General Functional Overview -- a structured dictionary of function that provides a hierarchical representation of major metabolic subsystems in prokaryotic organisms. A talk will concentrate on the possibilities of extending the Gene Ontology approach to the description of metabolic function in prokaryotic organisms. We believe that such development will provide a framework for evolutionary analysis of the biological function and will benefit studies of metabolism both in prokaryotic and eukaryotic organisms.