

Bar Harbor GO Users Meeting - July 2001

Abstracts

Natalia Maltsev

<maltsev@mcs.anl.gov>

Phone: (630)252-5195, FAX: (630) 252-5986

URL: <http://www-unix.mcs.anl.gov/~maltsev/>

Representation of metabolism

In the past decade the scientific community has witnessed a rapid accumulation 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.

Simon Twigger

<simont@mcw.edu>

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

URL: <http://rgd.mcw.edu>

Rat Genome Database - Data integration to support comparative genomics in the rat, mouse, and human

The primary focus of RGD is to aid Rat researchers in their work studying the rat as a model organism for human disease. To support these studies we have integrated a large amount of rat genetic and genomic resources in RGD and these are constantly being expanded through ongoing literature curation. One of the major features of RGD version 1.1, released in January of this year, is incorporation of QTL data to support physiological genomics studies relating disease with the genome. In addition VMap, a dynamic sequence-based homology tool, has been developed to enable Rat, Mouse and Human researchers to view mapped genes and sequences and their locations in the other two organisms. This will facilitate the application of results in one species to experiments in another species. In collaboration with the Mouse Genome Database and NCBI, close links are being created between RGD and MGD, LocusLink and UniGene to increase access to each set of data. In keeping with this goal of integrating data across databases RGD plans to incorporate GO annotations into its reports to further enhance the user's ability to make connections between organism databases. By providing these links we aim to provide a means for researchers to harness the physiological genomics studies performed in the rat and apply these results to their own work in Human or Mouse.

Carl Schmidt

<schmidtc@udel.edu>

Phone: 302-831-1334, FAX: 302 831-2282

URL: <http://udel.edu/~schmidtc/>

A multi-agent system for integration of gene ontologies with sequence annotation

We are using DECAF, a multi-agent system toolkit, to construct a prototype multi-agent system for automated annotation and storage of DNA sequencing data. As part of this effort, we intend to incorporate existing Gene Ontology efforts and use software agents to propose annotation for genes. An Ontology Agent contains both the GO ontologies and the mappings of other symbologies to GO terms. This agent, based roughly on the FIPA standard for ontology agent services, provides term translation services and answers queries about how terms are related in the ontology. A set of agents provide wrappers to BLAST

services and GO annotations at MGD, SGD, and Flybase. Working from a set of local unannotated gene sequences, an Ontology Reasoning Agent uses queries to all of the above agents to build a minimum spanning tree between all terms returned for sequence homologies from these GO organism databases. This information is then used to suggest likely annotation for new sequences and to display this information graphically for the biologist. An additional effort aims at extending the GO ontology to incorporate additional terms related to virology.

Robert Ireland

<rci@proteome.com>

Phone: 203-421-0097

URL: www.proteome.com

Application of GO to the Proteome Protein Survey Database

Proteome has developed the BioKnowledge™ Library (BKL™) as an information resource that brings biological knowledge to genomics and proteomics. The BKL™ comprises species-specific protein reports derived from comprehensive curation of research literature about mammals, including human, mouse and rat, to *C. elegans* (WormPD™), and through to the single-cell microorganisms *S. cerevisiae* (YPD™), *S. pombe* (PombePD™) and *C. albicans* (CalPD™). Protein report pages are linked to allow searching across multiple proteins and species for common characteristics. We have begun applying the Gene Ontology system to curation of our Protein Survey Database (HumanPD™). The first stage of GO incorporation involves conversion of Proteome Properties to GO terms for over 18,000 human, mouse and rat proteins. Stage two is built into the curation process. Information about each protein is obtained by curation of the scientific literature by a staff of Ph.D. curators who, in addition to collecting protein properties, writing annotations for each protein page and collecting additional information not resident in GO, also apply GO terms and collate information on expression patterns for each protein. Consequently, our curators are in a unique position to collect new terms to be consideration of inclusion in the GO hierarchies. Proteome has established a GO update team and is implementing a protocol for updating the GO information in HumanPD in accordance with changes adopted by the Consortium.

Gregory Harhay

<harhay@email.marc.usda.gov>

Phone: 402 762 4250, FAX: 402 762 4155

Integration and use of Gene Ontology terms in a livestock genomics database

Gene ontology (GO) terms have the potential to provide a wide variety of scientists an efficient means to access and exploit information from livestock expressed sequence tags (EST). GO provides a controlled vocabulary that can be understood by a broad range of scientists and end-users. This is important in livestock genomics and improvement where molecular biologists, geneticists, veterinarians, physiologists and microbiologists collaborate. GO facilitates the exploitation of functional information from other species. This is especially important in the case of livestock genomics, as investigators have a relatively large number of EST contigs with relatively few assigned functions. We have integrated GO terms into our database to provide scientists with a means of selecting EST for further study. Our approach is to relate publicly available information about genes and gene products associated with man and other organisms with similar gene and gene products in cows and pigs. TIGR cow and pig EST contig sequences are BLASTed against the NCBI nt database and the results are parsed into the database with a Perl script. Another Perl script subsequently extracts the Ref Seq gids of these BLAST hits from our database, parses the corresponding Ref Seq, Locus Link, GO, GeneCard and OMIM terms from hypertext pages on the Web, and incorporates these terms back into our database. Implementation of these Perl scripts and how livestock scientists are using GO terms is described.

Michelle Gwinn

<mlgwinn@tigr.org>

Phone: 301-315-2536

TIGR's transition to the GO ontology system

To organize proteins from sequencing projects, TIGR has used in-house categorization schemes for both Prokaryotic and Eukaryotic data. To

facilitate data exchange and data consistency among genome centers, TIGR is switching to the Gene Ontology (GO) system. We have tables in our database to house the GO terms, ids and the relationships between them. We have downloaded the protein annotations from the model organisms available on the GO web site and have created a database of those proteins for searches. In collaboration with Michael Ashburner, we have generated a map of TIGR microbial roles to GO terms which allows some GO terms to be assigned automatically while others will need to be assigned manually. Once GO terms have been assigned with high confidence to a few diverse microbial species, we intend to propagate those assignments to other bacteria by sequence similarity. Assignments to the Eukaryotic proteins will be made both manually and through searches to the database of organisms whose proteins are already assigned. The assignment of Arabidopsis genes to GO terms will greatly facilitate the comparison of the model plant genome to other GO-compliant genomes of great biological interest. GO assignment has been incorporated into the Eukaryotic annotation software and a web-based tool allows TIGR users to search for, and examine, GO terms and their position in the GO hierarchy. Both role systems will exist in our database until the transition is complete.

Leszek Vincent

<Leszek@missouri.edu>

Phone: 573 884-3716, FAX: 573 884-7850

URL: <http://www.cafnr.missouri.edu/mmp>

The development of ontologies and structured controlled vocabularies for plants

Plant genomic databases are expanding in number and complexity. These information-rich databases need to accurately and consistently document features e.g. gene structures, products, functions, phenotypes, traits, developmental stages, anatomical parts, besides other information. Inter-database queries will enable comparative genomic strategies to elucidate plant functions. However, terms for comparable objects in each database are variable, limiting successful querying of information in and across different databases. One solution involves the development and application of ontologies of structured common Controlled Vocabularies (CVs). The CVs in ontologies would be generic enough to facilitate inter-database queries for related organisms (e.g. monocots and dicots) and customizable to

accommodate taxon diversity. We summarize our efforts to develop a model for these structured CVs for plants that expands upon the protocols and tools developed by the Gene Ontology Consortium. This work is a collaborative effort between MMP, IRRI, TAIR (The Arabidopsis Information Resource) and Gramene: A Comparative Mapping Resource for Grains. *Partially supported by NSF award DBI-9872655.*

Lei Liu

<leiliu@uiuc.edu>

Phone: 217-265-5061, FAX: 217-265-5066

URL: keck1.biotec.uiuc.edu

An EST informatic framework incorporating Gene Ontology

Many EST (Expression Sequence Tag) projects have been carried out for a wide range of organisms and generated millions of sequences. There are more than 8 million EST entries in NCBI's dbEST. This is a rich resource for the life science research, especially for functional genomics using microarray technology. But most of the sequences are poorly or not annotated and have little use without annotation. We have developed a web based EST informatic framework which provides a preliminary annotation using Gene Ontology (GO) terms. EST sequences are first assembled by CAP3 program. The resulting putative unique sequences are then searched against a well-annotated model genome (e.g. Drosophila, Mouse) using BLAST. A gene association table can then be generated from the BLAST result through GO association for the model genome. In the framework, we have ported the MySQL schema of GO Database to an Oracle schema and built a GO term query interface for browsing the GO terms and retrieve the GO annotations for a particular EST data set. We have implemented the framework for two EST projects. One is Honey Bee Brain EST Project (http://keck1.biotec.uiuc.edu/bee/honeybee_project.htm). The other is Bovine Placenta EST Project (http://keck1.biotec.uiuc.edu/cattle/cattle_project.htm).

Keith Decker

<decker@cis.udel.edu>

Phone: 302-831-1959, FAX: 302-831-409
URL: <http://www.cis.udel.edu/~decker>

Agent-based support for functional annotation

Today biological information and algorithms for the analysis of biological data are available on the Internet in many different locations with overlapping content, different structure, and varied amounts of curation. Our approach to these problems, called multi-agent information gathering, is to apply multi-agent systems technologies to create software agents for information retrieval, filtering, integration, analysis and display. Currently we have developed a prototype system for the automated annotation of herpesvirus sequences with homologs, motifs, domains, and sub-cellular location predictions. The system automatically produces a searchable database of this information (at <http://udgenome.ags.udel.edu/herpes/> [soon]). Our presentation would discuss the goals and design of this system, and in particular the new subsystem for assisting biologists in functional annotation using the GO ontologies. This subsystem includes agents that wrap MGD, SGD, and Flybase (or local copies) and perform BLAST searches on these databases for each unknown gene, collecting the GO annotations and evidence codes from close homologues. An ontology agent (based on the FIPA standard) computes a minimal spanning tree of the annotations and the resulting graphs are displayed in a web browser to the biologist in order to assist in choosing an initial annotation. This architecture allows us to also experiment with arbitrary mechanisms for automated functional annotation as well.