

The Gene Ontology Annotation database (GOA): A community resource of GO annotations to the UniProt Knowledgebase

Evelyn Camon^{}, Emily Dimmer, Daniel Barrell, Nicola Mulder and Rolf Apweiler
European Bioinformatics Institute, Hinxton, UK*

The Gene Ontology Annotation (GOA) database (<http://www.ebi.ac.uk/GOA>) provides high-quality electronic and manual annotations to the UniProt Knowledgebase 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 92,000 species (GOA-UniProt) and is one of 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 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 proteome (GOA-Human) 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 web interface or downloaded from 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. Researchers wishing to query or contribute to the GOA project are encouraged to email: goa@ebi.ac.uk.

ChickGO takes flight

Fiona M. McCarthy^{†}, N. Wang[†], S. M. Bridges[‡] and S. C. Burgess[‡]
Mississippi State University, Mississippi State, MS, USA*

^{†,‡} These authors contributed equally to this work

The chicken is the first agricultural organism sequenced but presently its genome is poorly structurally and functionally annotated. Currently 12,854 (43%) of chicken proteins in the NCBI protein database are predicted based solely on homology to known genes. A further 30% of chicken proteins are predicted to have no mammalian homolog. Currently, only 5,062 proteins (17%) have any gene ontology (GO) annotation and >99% of these proteins are ascribed function inferred by electronic annotation (IEA). We have established ChickGO as part of AgBase, a curated, open-source, Web-accessible resource for functional analysis of agricultural plant and animal genomes. Our long-term goal is to serve the needs of the agricultural research communities by facilitating post-genome biology for agricultural species. ChickGO has three goals. 1. Working with GOA, ChickGO provides GO functional annotations for chicken gene products. GOA Chicken 3.0 has just been released. 2. ChickGO will aid structural annotation of the chicken genome using proteomics approaches. We have confirmed the *in vitro* expression of 8% of the 77,600 *ab initio* ORFs frames predicted by Ensembl and submitted these to the PRIDE database. 3. We will develop and support tools for functional genomics, including tools for analysis of GO data and proteomics. ChickGO accepts annotations from the research community and will provide an intermediary service for agricultural researchers interested in functional genomics and ontologies. ChickGO has served as a model for CowGO and MaizeGO, and for GO annotation of other agriculturally important genomes with small research communities and limited funds.

Uses of the GO in the Reactome Knowledgebase of Biological Processes

*Lisa Matthews^{*1}, Ewan Birney², David Croft², Bernard De Bono², Peter D'Eustachio¹, Marc Gillespie¹, Gopal Gopinath¹, Bijay Jassa², Geeta Joshi-Tope¹, Suzanna Lewis³, Esther Schmidt², Lincoln Stein¹, Imre Vastrik² and Guanming Wu¹*

¹Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA;

²European Bioinformatics Institute, Hinxton, UK;

³Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Reactome is an on-line database of human biological reactions and pathways. The information contained within Reactome is contributed by expert research scientists, interpreted and integrated in the context of the Reactome data model by Reactome curators, and reviewed using a formal peer-review system. This process is facilitated by new graphical author and curator tools. Reactome's event-centric data model offers a computationally formal description of the ordered relationships among biological reactions and their molecular components. Reactome molecules and concepts are cross-referenced to their counterparts in numerous databases including GO facilitating the interpretation and use of Reactome data in the context of functional annotations available from other biological databases. In addition to cross-referencing the GO, Reactome compares and exchanges GO annotations with the human GOA project. While Reactome focuses primarily on the description of human processes, we provide orthology-based electronic annotations for numerous other species. The web interface to Reactome can be browsed like an online textbook, used as a "quick reference" guide to gene function or used to identify the composition of macromolecular complexes. In addition, new tools and a new event page display introduced this year facilitate navigation and orientation within Reactome and aid in the interpretation of microarray expression studies and other large-scale data sets. Here we focus on specific features of Reactome that could be useful to Gene Ontology users, including the new interface, tools and applications.

Recent additions and improvements to the Onto-Tools

Purvesh Khatri^{}, Sivakumar Sellamuthu, Pooja Malhotra, Kashyap Amin, Arina Done and Sorin Draghici*

Department of Computer Science, Wayne State University, Detroit, MI, USA

The Onto-Tools suite is composed of an annotation database and 6 seamlessly integrated, web-accessible data mining tools: Onto-Express, Onto-Compare, Onto-Design, Onto-Translate, Onto-Miner, and Pathway-Express. The Onto-Tools database has been expanded to include various types of data from 12 new databases. Our database now integrates different types of genomic data from 19 sequence, gene, protein, and annotation databases. Additionally, our database is also expanded to include complete Gene Ontology (GO) annotations. Using the enhanced database and GO annotations, Onto-Express now allows functional profiling for 24 organisms and supports 17 different types of input IDs. Onto-Translate is also enhanced to fully utilize the capabilities of the new Onto-Tools database with an ultimate goal of providing the users with a non-redundant and complete mapping from any type of identification system to any other type. Currently, Onto-Translate allows arbitrary mappings between 29 types of IDs. Pathway-Express is a new tool that helps the users find the most interesting pathways for their input list of genes. Onto-Tools are freely available at <http://vortex.cs.wayne.edu/Projects.html>.

High-Throughput GoMiner: leveraging GO for integrative interpretation of multiple microarray studies

Barry R Zeeberg^{*†}, Haiying Qin^{2†}, Sudarshan Narasimhan³, Hong Cao³, Hongfang Liu^{1,4}, Robert S Sfeir³, David W Kane³, Andre Muller⁵, Hans A Kestler^{5,6} and John N Weinstein¹

¹Genomics and Bioinformatics Group, Laboratory of Molecular Pharmacology, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA;

²Metabolism Branch, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

³SRA International, Fairfax, VA, USA

⁴Department of Information Systems, University of Maryland at Baltimore County, USA

⁵Internal Medicine I - Gastroenterology, University Hospital Ulm, Germany

⁶Neuroinformatics, University of Ulm, Germany

† These authors contributed equally to this work

High-Throughput GoMiner (BMC Bioinformatics 2005, 6:168; <http://discover.nci.nih.gov/gominer/htgm.jsp>) is a web (or optionally command-line) resource that leverages GO for integrative interpretation of data from multiple microarrays. For example, assume that the input represents lists of genes with altered expression level at a series of time points. Then one of the output file types would be a clustered image map (CIM) visualization of the time course of the statistical significance of the important biological process categories. Alternatively, assume that the input represents knock-outs of different genes related to a disease state. Then the CIM provides an intuitively simple picture of the complex underlying relationship of altered biological processes and particular knock-outs.

High-Throughput GoMiner may prove particularly attractive in an industrial setting. Imagine using microarray technology for screening dozens of derivatives of a promising lead compound. Few, if any, of the screened compounds will provide interesting results. It would be extremely tedious, or perhaps not even possible in practice, to use the original GUI GoMiner to examine each individual microarray result. In contrast, High-Throughput GoMiner not only automates this laborious and human-error prone process, but also prepares a report that flags the compounds most likely to be of interest.

Integration of GoMiner with VennMaster (Bioinformatics 2005, 21:1592; <http://www.informatik.uni-ulm.de/ni/mitarbeiter/HKestler/vennm/doc.html>), a program package that generates sophisticated Venn diagrams, provides a novel approach to visualization of the GO categorization of genes. GoMiner and High-Throughput GoMiner are being integrated into the Cancer Biomedical Informatics Grid (caBIG) project and into caWorkbench. Acknowledgments

This research was supported [in part] by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

Tutorial on using the command line version of High-Throughput GoMiner

*Barry R Zeeberg¹, Sudarshan Narasimhan², Margot Sunshine², Hong Cao²,
David W Kane² and John N Weinstein¹*

*¹Genomics and Bioinformatics Group, Laboratory of Molecular Pharmacology,
National Cancer Institute, National Institutes of Health, Bethesda, MD, USA;*

²SRA International, Fairfax, VA, USA

High-Throughput GoMiner (htgm) is a resource that leverages GO for integrative interpretation of data from multiple microarrays. Given a list of genes with altered expression, htgm produces a number of export files that help the user to interpret the microarray results by illuminating the significant GO categories. The set of export files provides the potential for both quantitation and visualization.

Although htgm is primarily intended as a web application (<http://discover.nci.nih.gov/gominer/GoCommandWebInterface.jsp>) it is also available in a command line version (<http://discover.nci.nih.gov/gominer/htgmcommand.jsp>) that runs on a Unix platform (including Mac OS X) for users who prefer a greater degree of customization than is afforded by the web application. This tutorial focuses on using the command line version.

The tutorial will cover:

- How to install htgm
- How to prepare data files to run in htgm
- How to interpret htgm results

The performance of the command line version will be greatly enhanced by installation of a local database (<http://discover.nci.nih.gov/gominer/enhance.jsp>).

In summary, installing and running the command line version of htgm is straightforward. It is expected to be particularly attractive to those users who seek customization or who are performing intensive number crunching that would significantly downgrade the performance of our web server.

Acknowledgements: This research was supported by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

**GOALIE: an application that describes a time course microarray experiment
in terms of the GO “biological processes” taxonomy**

Marco Antoniotti¹, Naren Ramakrishnan² and Bud Mishra¹

¹ Bioinformatics Group, Courant Institute of Mathematical Sciences, New York University;

² Virginia Tech

Current micro-array data analysis techniques draw the biologist's attention to targeted sets of genes e.g., those that vary in a well correlated manner, are under similar regulatory control, or that have consistent functional annotation or ontological categorizations. Yet, such methods do not provide dynamic perspectives inferred collectively over a dataset. To address these problems we have started pursuing an approach and a system — GOALIE — based on a statistical analysis of time course micro-array data in terms of the annotations available in the Gene Ontology and other publicly available databases. Several researchers looked into the significance of “describing” a given micro-array dataset in terms of the GO contents. The question asked is “what is the set of terms that conveys the most information about an experiment”? This question is asked after statistical inference tests have been applied (e.g. a Fisher Exact Tests). Our approach adds one more dimension to the analysis, by introducing a breakdown of the time course experiment (usually 5 to 50 time points) into “overlapping windows”. Clustering is performed within each window and the “biological processes” are tracked as they “move” from window to window. This yields valuable information about the locality of phenomena to a biologist, who may better target new experiments. In our presentation we will discuss several examples we looked into: SEB host-pathogen interaction, *S. cerevisiae* cell cycle, *P. falciparum* development, Fibroblast serum responses. Finally we will discuss how the interactions reconstructed from the datasets can be rendered in terms of a formal temporal logic and hence into Natural Language. We wish to acknowledge several colleagues, especially Dr. Marti Jett of Walter Reed Army Institute of Research and the DARPA BioCOMP program that funded our research.

eGOn, explore Gene Ontology, V2.0

*Vidar Beisvåg^{*1}, Frode Romstad Krohn Jünge^{1,3}, Hallgeir Bergum^{1,3}, Clara-Cecilie Günther²,
Mette Langaas² and Astrid Lægrid^{1,3}*

¹Dept. of Cancer Research and Molecular Medicine and

²Dept. of Mathematical Sciences, Norwegian University of Science and Technology;

³Norwegian Microarray Consortium (NMC), National Plan for Functional Genomics

The mantra of the post-genomic era is gene function and functional genomics. Datasets generated with technologies such as DNA microarrays have created a critical need for resources that facilitate the interpretation of large-scale biological data. GENETOOLS is a collection of web-based tools on top of a database that brings together information from a broad range of resources, and provides this in a manner particularly useful for genome-wide analyses. Today, the two main tools connected to this database are the NMC Annotation Database V2.0 and eGOn V2.0 (explore Gene Ontology). eGOn V2.0 facilitates interpretation of GO annotation. GO terms are retrieved in batch modus from Entrez Gene and the GO database and displayed in the GO di-acyclic hierarchical graph (DAG). Essential features of eGOn V2.0 are: 1. Visualization: gene annotations are visualized in the GO DAG or as a table view. The granularity of the GO DAG can be edited freely by the user. 2. Filtering: GO annotations can be filtered on evidence codes. 3. Include user defined GO annotations: previously added to the Annotation database. 4. Statistical analysis: Several gene lists are analyzed simultaneously to compare the distribution of the annotated genes over the GO hierarchy. Statistical tests are implemented to allow the user to compute GO annotation dissimilarities within or between gene lists. 5. Connection to Annotation database: Links to Annotation database gene and protein information are offered directly from the GO DAG or in exported data. 6. Export: GO DAG information, statistical results and gene and protein information can be exported in excel, text or XML format. GENETOOLS are freely available at www.genetools.no.

**Blast2GO: a universal annotation, visualization and analysis tool
for functional genomics research**

*Ana Conesa^{*1}, Stefan Goetz², Juan Miguel García², Javier Terol¹, Manuel Talón¹
and Montserrat Robles²*

¹Centro de Genómica, Instituto Valenciano de Investigaciones Agrarias, Apartado Oficial 46113, Moncada, Valencia; ²BET-ITACA, Universidad Politécnica de Valencia, con. Vera s/n, Valencia

The recent advances and the cost reduction in new high throughput technologies have made possible to extend genome-wide research to many field specific plant and animal species. To obtain the maximal research benefit that these powerful technologies can provide is also necessary to have analysis tools adequate to the type of organisms of study. One very important aspect in mining genomics data is to associate individual sequences and related expression information to biological function. Functional annotation allows categorization of genes in functional classes, which can be very useful to understand the physiological meaning of large amounts of genes and to assess functional differences between subgroups of sequences. Gene Ontology has imposed in the last years as a standard for functional gene annotation and as common framework for function based statistical analysis of genomics data. Different tools have been developed that provide this type of GO-based analysis. However, when trying to apply these data mining approaches to poorly characterized organisms we encountered that most tools fail to combine the tasks of high-throughput annotation, data exploration and function based analysis that would be required in this case. Looking for a suitable solution to this problem we have developed Blast2GO (B2G), a universal GO annotation, visualization and statistics framework that brings advanced functional analysis to the genomics research of non-model species. Briefly, B2G uses BLAST to find homologs to fasta formatted input sequences. The program extracts GO terms to each obtained hit by mapping to existent annotation associations. An annotation rule finally assigns GO terms to the query sequence. Genomics experimental data can then be analyzed based on this annotation with a various descriptive and hypothesis testing statistical tools. Annotation and functional analysis can be visualized in graph form reconstructing the GO relationships and color-highlighting the most significant areas. B2G was conceived to be an attractive tool for research environments where genetic and/or computational resources are limited and where much work is still done in an explorative fashion. It allows monitoring and interaction at different steps of the analysis, and emphasizes visualization as an important component of knowledge acquisition. B2G is a Java application made available by Java Web Start. It is platform independent and has no further requirements than an Internet connection.

Monitoring dynamics of biological systems via changes in GO-term annotations

*Stefan Enroth^{*1}, Adam Ameer¹, Astrid Lægrid² and Jan Komorowski¹*

¹The Linnaeus Centre for Bioinformatics, Uppsala University, Sweden;

²Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

A summary of which genes are involved in which process or function over a sequence of time points of gene expressions may provide an attractive approach to understanding the dynamics of a biological system. To this end we compare the number of GO-annotations in a selected subset of genes to a predefined background distribution. This allows us to test, statistically, the significance of each GO term individually. We have made this kind of analysis available in The Linnaeus Center for Bioinformatics microarray data warehouse. A GO-based toolkit consisting of three major parts — importing GO annotation to previously in-house un-annotated sets via Entrez Gene IDs, visualization of terms through graphs and histograms and, finally, significance testing of terms using a hypergeometric selection assumption — has been implemented in R. The results of the tests are presented in tables and images as well as expandable graphs, allowing the user to browse through the results by clicking on and expanding the GO-terms in the GO-DAG. The toolkit can handle time course experiments allowing the user to monitor change in function/process over time and not just snapshot gene expression. This feature can also be used to compare results of any partition of the data such as different treatments or differentially expressed genes. Our tool has been tested on numerous datasets including time course experiments and even ChIP-chip data with good results.

BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks

Steven Maere

Department of Plant Systems Biology, Ghent University / VIB, Ghent, Belgium

BiNGO is a Java-based tool to determine which Gene Ontology (GO) categories are statistically overrepresented in a set of genes or a subgraph of a biological network. BiNGO is implemented as a plugin for Cytoscape, which is an open source bioinformatics software platform for visualizing and integrating molecular interaction networks. Gene sets can either be selected or computed from a Cytoscape network or compiled from sources other than Cytoscape (e.g. a list of genes that are significantly upregulated in a microarray experiment). BiNGO maps the predominant functional themes of the tested gene set on the GO hierarchy, and takes advantage of Cytoscape's versatile visualization environment to produce an intuitive and customizable visual representation of the results. The main advantage of BiNGO over other GO tools is the fact that it can be used directly and interactively on molecular interaction graphs.

GOdist — a new tool and approach for microarray data analysis

Lilach Soreq^{}, Yoram Ben-Shaul, Hagai Bergman, Nissim Ben-Arie and Hermona Soreq
The Hebrew University of Jerusalem, Jerusalem, Israel*

In microarray experiment, an arbitrary 2-fold threshold of change is conventionally employed to define up/down regulated genes. Often, Gene Ontology (GO) classification is used to identify over-represented functions or processes undergoing modulation. Conventional methods (the “discrete” approach) define a GO term as changed if the number of above-threshold modulated transcripts is significantly larger than expected. Changes among specific terms are compared to the entire set of transcripts, using chi-square or an exact test employing the hypergeometric distribution. However, the implicit assumption underlying these analyses, namely that gene expression is a continuous rather than a discrete (binary) process is unwarranted. Setting an arbitrary threshold discounts meaningful, subtler changes in individual genes and in the distribution of expression level changes within terms. We developed GOdist, to analyze microarray expression data implementing the continuous approach by Kolmogorov-Smirnov statistics, enabling identification of a much wider spectrum of distributional differences between each specific term and the entire set of transcripts. For each GO term, GOdist tests whether it is differentially distributed compared to the entire set of transcripts. For comparison, GOdist also implements the discrete approach using an exact hypergeometric distribution. Additionally, GOdist allows in-depth inspection of specific GO terms, comparing them to their parents in the GO hierarchy and allowing identification of the key nodes in which a change occurred. It also plots the distribution functions of specific terms in comparison to their direct parent and/or child terms and to the global tested population and provides various displays of several statistical measures. We applied GOdist to Affymetrix data derived from mouse models of Parkinsonism and of impaired development of cerebellar neurons and for studying the effects of cholinergic signaling on alternative splicing configurations in transfected cells in culture. In all of these uses, GOdist indicated both kinds of changes in many terms represented on the array and served to analyze the change of specific GO terms.

New Gene Ontology Structure for Improved Biological Reasoning

Henrik Tveit

Norwegian University of Science and Technology (NTNU), Trondheim, Norway

Gene Ontology (GO) consists of three independent sub-ontologies: molecular function (MF), biological process (BP) and cellular component (CC). We generated a new structure, the Second Gene Ontology Layer, using data mining and biological information. This model captures biological relationships not reflected in the present ontology structure. Given a molecular function, the Second Layer paths identify biological processes where the MF is involved and cellular components where the MF is active. The Second Layer is a validated biological model, which currently consists of 7237 paths. The Second Layer model can be used to improve consistency and completeness of existing gene annotation sets as it utilises the MF annotations to provide new annotations in the sub-ontologies of BP and CC. Applying the Second Layer paths to a set of 4223 human genes increased BP and CC annotations by 30% and 41%, respectively. Furthermore, the enlarged annotation set improved sensitivity and specificity of a model predicting gene biological process annotations from gene expression profiles. The Second Gene Ontology Layer co-exists with the original GO-structure and enhances the ontology to reflect a wider range of biological concepts. This creates new opportunities for computer-assisted reasoning with biological information. The new structure and ANNEX, an application complementing existing annotation sets, are publicly available through the GO Annotation Toolbox: www.goat.no.

Using Gene Ontology information for biologically constrained variable selection in modeling large scale functional genomics data

*Victor Trevino and Francesco Falciani**

School of Biosciences, The University of Birmingham, Birmingham, UK

One of the most challenging aspects in the analysis of microarray data is to identify genes that are associated to relevant aspects of cell physiology. Algorithms to perform unsupervised and supervised classification have been used to this effect and have proven to be useful to improve our understanding of biology. One of the advantages of statistical models over unsupervised methods is that they can easily incorporate non-linear relationships. Because of the large number of variables sets to explore in large datasets a comprehensive analysis of models built on all possible combinations of n genes is computationally impossible. For this reason variable selection strategies [1,2,3] have been applied. Current methodologies however, explore variable space in an unbiased manner. When mining genomics data however, we may be interested in exploring particular sets of variable combinations that may reflect biologically plausible scenarios. Biologically driven variable selection can be achieved in the context of stochastic search methodologies by defining the prior probability of inclusion of a gene in a model as conditional to its function, for example, as defined in the context of a Gene Ontology system (GO). This communication reports our initial results on biologically driven model selection using a genetic algorithm search strategy where the probability of gene mutations is a function of the position of the gene in the GO hierarchy.

Bibliography

[1] N. Sha *et al.* *Comp Funct Genom* 4: 171–181, 2003

[2] N. Sha *et al.* *Biometrics* 60, 812–819, 2004

[3] Liu JJ *et al.* *Bioinformatics* 11: 2691–7, 2005

Recent developments in OBO

Michael Ashburner

Department of Genetics, Cambridge University, Cambridge, UK

I will discuss recent developments in the OBO project and, in particular, how the Gene Ontology will use orthogonal ontologies from OBO for the maintenance and development of the GO.

AgBase: Targeted gene ontology annotation databases for agriculture

Susan Bridges^{}, F. M. McCarthy, D.S. Luthe, N. Wang, B. Nanduri and S. C. Burgess
Mississippi State University, Mississippi State, MS, USA*

The genomes of several agricultural species have been sequenced and more are scheduled for sequencing in the near future. These genomes must be both structurally and functionally annotated to derive benefit from the sequencing effort. However agricultural research communities interested in using this data are small, diverse and have limited funding. AgBase (www.agbase.msstate.edu) has been created to provide a coordinated, directed approach for functional annotation in agricultural species. The three specific aims of AgBase are 1) to provide highly curated, functional information about agriculturally important gene products following the Gene Ontology Consortium guidelines, 2) to provide direct experimental evidence for further structural and functional annotation of agricultural genomes, and 3) to supply tools that assist in analysis and visualization of large scale data. AgBase is a confederation of curated open-source databases that provides a Web-accessible resource for functional analysis gene products for agricultural species. Current components of AgBase include ChickGO, CCatfishGO, CowGO, MaizeGO, and MheamGO and several others are scheduled to come online within the next year. Unlike the global annotation approach used by the large well-funded model organism communities, we target our annotations based on experimental and modeling needs. AgBase provides community access for the annotations produced by the research and also serves as a repository for other researchers working with agricultural species. The current database is implemented in MySQL although we plan to move to Oracle and to integrate our work with the implementation of the CHADO schema developed by dictyBase in the near future. Apache is used as the web server and the web interface is implemented in HTML and Perl CGI.

Does your species show in GO?

*Jennifer I. Clark
European Bioinformatics Institute, Hinxton, UK*

Currently the model organism database resources are the main contributors of annotation. We are also developing systems to enable domain experts to contribute annotations for their area of expertise.

The Consortium provides support for groups beginning the process of annotation. We run regular courses on manual annotation, including those specific to GO annotation, and those covering more general topics run by database resource groups such as TIGR. We can also pair new groups with more experienced 'mentor' groups for help in the initial stages. For example mentoring by Mouse Genome Informatics (MGI) enabled the new ChickGO group to start annotation with GO. The mentor groups can provide advice on all aspects of annotation. Topics covered may include such things as the meaning of evidence codes, the running of scripts to verify the syntax of the gene_association file, and the requesting of new terms for the GO. The consortium provides links to free open source annotation software that can be used by any group, and mailing lists where advice can be found on use of the software.

To discuss the possibility of having your model species or gene products of interest annotated, please contact the GO mailing list.

A semantic analysis of the annotations of the human genome

*Purvesh Khatri, Bogdan Done, Archana Rao, Arina Done and Sorin Draghici**
Department of Computer Science, Wayne State University, Detroit, MI, USA

The correct interpretation of any biological experiment depends in an essential way on the accuracy and consistency of the existing annotation databases. Such databases are ubiquitous and used by all life scientists in most experiments. However, it is well known that such databases are incomplete and many annotations may also be incorrect. In this paper we describe a technique that can be used to analyze the semantic content of such annotation databases. Our approach is able to extract implicit semantic relationships between genes and functions. This ability allows us to discover novel functions for known genes. This approach is able to identify missing and inaccurate annotations in existing annotation databases, and thus help improve their accuracy. We used our technique to analyze the current annotations of the human genome. From this body of annotations, we were able to predict 212 additional gene–function assignments. A subsequent literature search found that 138 of these gene–functions assignments are supported by existing peer-reviewed papers. An additional 23 assignments have been confirmed in the meantime by the addition of the respective annotations in later releases of the Gene Ontology database. Overall, the 161 confirmed assignments represent 75.95% of the proposed gene–function assignments. Only one of our predictions (0.4%) was contradicted by the existing literature. We could not find any relevant articles for 50 of our predictions (23.58%). The method is independent of the organism and can be used to analyze and improve the quality of the data of any public or private annotation database.

The Lab of Milk and Honey: Plans for Annotating Bovine and Bee

Christine G. Elsik
Department of Animal Science, Texas A&M University, College Station, TX, USA

BeeBase (http://racerx00.tamu.edu/bee_resources.html) and the Bovine Genome Database (<http://bovinegenome.org>) are model organism databases currently under development at Texas A&M University. The immediate objectives are to support the community-based genome annotation efforts, which are coordinated by the Baylor College of Medicine Human Genome Sequencing Center. The long-term objectives are to develop community informatics resources that will integrate and attach biological meaning to genomic data. Following release of the official predicted gene sets, BeeBase and the Bovine Genome Database will begin GO annotation by ISS and literature curation. The 7.5X coverage honey bee genome assembly was released in January 2005. A consensus gene set was created using gene predictions generated by several groups, and was the starting point for ongoing manual gene family curation by honey bee research community members. The 6.2X bovine genome assembly was released in June 2005, with anticipation of the 7-8X release in January 2006. Bovine gene prediction analysis is ongoing at ENSEMBL and NCBI, and we expect the bovine research community's manual curation effort to begin in October. The community-based efforts include both structural and functional annotation, using a variety of methods, such as sequence homology, multiple alignment and phylogenetics. Upon submission, manually annotated genes are checked and assigned identifiers by database curators. We will continue to support community input after conclusion of the sequencing projects, and wish to develop a system that will allow GO annotation to benefit from community expertise.

Refining the content of the Gene Ontology

Midori A. Harris

European Bioinformatics Institute, Hinxton, UK

As part of the Gene Ontology (GO) project (<http://www.geneontology.org>) the GO Consortium is committed to the continued refinement of its ontologies to keep abreast of advances in biology and to meet the needs of database curators and many other users, both within and outside the GO Consortium. Suggestions for additions or other changes in GO vocabulary content come from a variety of sources. The model organism database curators who use GO terms for gene product annotation are the most active contributors, playing a key role in guiding the development of GO. More recently, members of the research community have begun proposing changes as they incorporate GO data into their research programs. To coordinate the efforts of Consortium members and outside experts, the GO Consortium has established Curator Interest Groups to focus on areas within the ontologies that are likely to require extensive additions or revisions. Complementing input from biologists, a third source of suggested changes is computational analysis of existing GO terms and relationships, which can identify missing relationships and missing or misplaced terms. These computational efforts, most notably the OBOL project (see <http://www.fruitfly.org/~cjm/obol/>), improve the logical consistency of GO, and will eventually enable GO to adopt more formal computational representations for its ontologies.

To ensure consistency and promote communication, changes to the ontologies are centrally coordinated by the GO Editorial Office. We have adapted the online tracking system provided by SourceForge to manage suggestions for changes to the ontologies (see <http://geneontology.sourceforge.net/>). The GO mailing list and meetings devoted to GO biological content provide additional mechanisms by which the GO Consortium continually refines its ontologies. The GO Consortium welcomes feedback from the biology and bioinformatics communities.

Gene nomenclature and Gene Ontology — the perfect partnership?

Varsha K. Khodiyar^{}, Ruth Lovering and Sue Povey*

HUGO Gene Nomenclature Committee (HGNC), Department of Biology, University College London, London, UK

The HUGO Gene Nomenclature Committee (HGNC) works to provide a unique approved name and short-form symbol for every human gene. To date we have provided unique gene symbols for over 21,000 human genes and genomic features. There are two ways in which we are contributing to the Gene Ontology (GO) effort.

The increasingly widespread use of approved nomenclature across genomic browsers, gene databases, journals and publications greatly enhances the retrieval of genetic information. The use of approved nomenclature in Gene Ontology (GO) databases is also becoming more prevalent; promoting this can only serve to augment the use of GO in scientific research.

The process of providing approved gene symbols entails correspondence with authors, reading the literature and performing data analyses. This accumulation of information could also be used for GO annotation. Consequently, we are assessing how we could implement GO annotation within the framework of gene nomenclature. Since 2004 we have established a good relationship with UniProt and have added a limited number of GO annotations to human, mouse and rat proteins.

The HGNC is supported by the NIH, the UK MRC and the Wellcome Trust.

Protein Ontology Development: 2005 updates

Amandeep S. Sidhu and Tharam S. Dillon*

Faculty of Information Technology, University of Technology Sydney, Australia

We defined a Protein Ontology (PO) that provides a common structured vocabulary for researchers who need to share knowledge in proteomics domain. It consists of concepts (or classes), which are data descriptors for protein data and the relations among these concepts. PO provides a structured vocabulary description for protein domains that can be used to describe cellular products in any organism. Protein Ontology (PO) Framework describes: (1) Protein Sequence and Structure Information, (2) Protein Folding Process, (3) Cellular Functions of Proteins, (4) Molecular Bindings internal and external to Proteins and (5) Constraints affecting the Final Protein Conformation. PO uses all relevant protein data sources of information. The sources include new proteome information resources like PDB and SCOP as well as classical sources of information where information is maintained in a knowledge base of scientific text files like OMIM and from various published scientific literature in various journals. PO is available online at <http://www.proteinontology.info/>. PO is defined by Web Ontology Language (OWL) and the complete OWL file is also available online. Database of 10 out all the 57 Major Prion Proteins in various Protein data sources based on the vocabulary provided by PO is also available on. The XML Representation of the Database is available on Protein Ontology (PO) Website. Various User Interfaces for the Database will be made available soon. PO currently contains 92 concepts or classes, 261 attributes or properties and 17550 instances. The Proposed Protein Ontology is the first ever work to integrate protein data based on data semantics describing various phases of protein structure. PO helps to understand structure, cellular function and the constraints that affect protein in a cellular environment. The attribute values in PO are not defined as text strings or as set of keywords. Most of the attribute values entered in PO are instances of concepts defined in Generic Classes.