

Abstracts

The Gene Ontology Annotation (GOA) database and enhancement of GO annotations through InterPro-to-GO

Nicola Mulder^{}, Evelyn Camon and Rolf Apweiler
EMBL Outstation—European Bioinformatics Institute, Hinxton, UK*

GOA (<http://www.ebi.ac.uk/GOA/>) is a project managed by the European Bioinformatics Institute (EBI) and aims to provide high quality manual and electronic GO annotations to proteins in the UniProt Knowledgebase (UniProtKB). The GOA group prioritises the annotation of the human proteome and focuses on proteins involved in health and disease. However because of the multi-species nature of the UniProtKB, and because many UniProt curators at the EBI also annotate to GO during their curation procedure, GOA provides electronic and/or manual annotations to proteins from over 100,000 species. The group implements large-scale electronic annotations from mapping such as InterPro2GO, SPKW2GO, EC2GO etc., and also integrates high quality GO annotations from many model organism groups (AgBase, FlyBase, GDB, GeneDB, Gramene, HGNC, MGI, RGD, SGD, TAIR, TIGR and ZFIN) as well as the Reactome pathways and IntAct protein-protein interaction databases. This ensures that the GOA data set remains a key reference and a comprehensive source of GO annotations. The interpro2go mappings are done by the InterPro team, which now also includes the GOA group. InterPro2GO mappings account for 93% of all distinct proteins that have GO annotations, and therefore form an important component of the GOA project. This presentation will describe the GOA project at EBI and the process and impact of electronic GO mappings such as InterPro2GO.

Distinguishing manually curated GO annotations from computationally predicted GO annotations at the *Saccharomyces* Genome Database

*Eurie L. Hong^{*1}, Rama Balakrishnan¹, Karen R. Christie¹, Maria C. Costanzo¹, Selina S. Dwight¹, Stacia R. Engel¹, Dianna G. Fisk¹, Jodi E. Hirschman¹, Michael S. Livstone², Rob Nash¹, Rose Oughtred², Julie Park¹, Marek Skrzypek¹, Chandra L. Theesfeld¹, Rey Andrada¹, Gail Binkley¹, Stan Dong¹, Stuart Miyasato¹, Anand Sethuraman¹, Shuai Weng¹, Mark Schroeder², Kara Dolinski², David Botstein², and J. Michael Cherry¹*

¹ Department of Genetics, School of Medicine, Stanford University, Stanford, CA 94305, USA;

² Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA

The *Saccharomyces* Genome Database (SGD; <http://www.yeastgenome.org/>) annotates *S. cerevisiae* gene products to Gene Ontology (GO) in order to provide information about their biological roles in the cell. GO annotations can be generated by manual curation of the published literature by curators; by automated prediction of functions by sequence similarity methods such as InterProScan; and by computational prediction from analysis of large-scale interaction and expression datasets. Since 2003, every gene that produces an RNA or protein product has been assigned at least one annotation in each of the three ontologies (Biological Process, Molecular Function, and Cellular Component) based on experimental evidence in the published literature. To complement these ~34,000 manually curated GO annotations, SGD is incorporating ~24,000 GO annotations that are "Inferred from Electronic Annotation." These annotations, provided by the UniProt GOA project, provide valuable information for genes that have not been experimentally characterized. Here we present revisions to the user interfaces and GO analysis tools that distinguish between the sources of data and the methods used to generate the GO annotations. All data are publicly available via web interfaces and in downloadable files. SGD is funded by the US National Human Genome Research Institute.

Increasing GO annotation through community involvement

Fiona M. McCarthy^{*†1,2}, *Susan M. Bridges*^{†1,3} and *Shane C. Burgess*^{1,2}

¹*Institute for Digital Biology*, ²*Department of Basic Sciences, College of Veterinary Medicine and*

³*Department of Computer Science and Engineering, Bagley College of Engineering,*

Mississippi State University, MS 39762, USA

[†]*These authors contributed equally to this work*

Analysis of functional genomics experiments is hampered by lack of GO annotation. Moreover, the overall quality of GO annotation is not only based on the number of gene products that have GO annotations (breadth) but also the level of detail of these GO annotations (depth) and the quality of evidence. Manual curation of published literature remains the “gold standard” for GO annotation yet there are few annotators trained to do GO biocuration and experimentalists who are experts in their field do not have the resources to commit to becoming active trained GO annotators. We have developed a two-tier system of GO annotations to allow users to contribute their expert knowledge. The AgBase download page provides a “GO Consortium” gene association file containing only fully quality-checked annotations that are submitted to the central GO database and a more comprehensive “Community” gene association file containing additional GO annotations including GO annotations for electronically predicted proteins, comprehensive annotations based on sequence homology and annotations from community researchers that have not yet been quality checked. This system gives researchers the initial breadth required for functional modeling, leading to experiments that test the function of these gene products, which leads to higher quality GO annotations. We envision that as the overall annotation quality improves, the GO annotations in the community gene association file will be superseded.

The AgBase GO annotation tools

S. M. Bridges^{*}, *F. M. McCarthy*, *N. Wang*, *G.B. Magee*, *B. Nanduri* and *S. C. Burgess*

Mississippi State University, Mississippi State, MS, USA

AgBase (<http://www.agbase.msstate.edu>) has been created to provide a coordinated, directed approach for functional annotation in agricultural species. In addition to providing Gene Ontology annotations of gene products from agricultural species, AgBase also provides a suite of tools for functional analysis of proteomics and gene expression datasets. The tools are available on-line and can be used individually or as components of a pipeline. The AgBase GO tools include GProfiler, GORetriever, GOanna and GOSlimViewer. GProfiler is designed to provide a summary of the number and type of GO annotations that are available for gene products from a particular species. GORetriever finds existing GO annotations for a set of proteins. GOanna accepts a list of protein or gene accession numbers or a FASTA file and returns the results of the appropriate blast search in a Gene Association file format. Each entry also has a link to the corresponding alignment that enables the researcher to judge the quality of the match. GOSlimViewer can be used with the output of GORetriever or GOanna (or both) and provides GO Slim summaries of a dataset that can be displayed visually. The AgBase suite of tools has been used for functional modeling of both proteomics and gene expression datasets from chicken, maize, cow, and a number of microbes.

Potential errors in protein GO function annotations returned by AmiGO: How to find them and what to do about them

*Carson Andorf, Drena Dobbs, and Vasant Honavar**

*Artificial Intelligence Research Laboratory, Department of Computer Science,
Bioinformatics and Computational Biology Graduate Program, Center for Computational
Intelligence, Learning, and Discovery, Iowa State University, Ames, IA 50011-1040, USA*

Incorrectly annotated sequence data are becoming more commonplace as databases increasingly rely on automated techniques for annotation. Hence, there is an urgent need for computational methods for checking consistency of such annotations against independent sources of evidence and detecting potential annotation errors. We show how a machine learning approach designed to automatically predict a protein's Gene Ontology (GO) functional class can be employed to identify potential gene annotation errors. In a set of 211 previously annotated mouse protein kinases, we found that greater than 95% of the GO annotations returned by AmiGO appear to be inconsistent with the UniProt functions assigned to their human counterparts. In contrast, 97% of the predicted annotations generated using a machine learning approach are consistent with the UniProt annotations of the human counterparts, as well as with available annotations for the mouse protein kinases in the Mouse Kinome database. We conjecture that most of annotations of mouse kinases predicted by the classifier trained using machine learning are, most likely, correct. Our results raise the possibility that machine learning approaches might be useful for detecting potential errors in GO annotations generated by high-throughput gene annotation projects.

NCBO, the OBO-Foundry, and you

*Suzanna Lewis^{*1}, Barry Smith², Chris Mungall¹, Michael Ashburner³, and Mark Musen⁴*

¹Berkeley Bioinformatics and Ontology Project, Lawrence Berkeley National Laboratory, Berkeley, CA, USA; , ²Department of Philosophy, University at Buffalo, Buffalo, NY, USA;

³Department of Genetics, Cambridge University, Cambridge, UK; ⁴Stanford Medical Informatics, Stanford University, Stanford, CA, USA

The National Center for Biomedical Ontology has been founded as one of seven National Centers for Biomedical Computing established under the NIH Roadmap for Medical Research. The creation of the Center represents a milestone in the history of ontology as a vital tool of biomedical science. The scores of biomedical ontologies now in existence will not solve the problem of biomedical data and knowledge management by themselves. Current solutions, based on terminological coding, are frustrated by the uncontrolled heterogeneity even of terminologies dealing with identical subject-matters. We propose a new type of solution, which is to create a family of interoperable gold standard reference ontologies, one for each core domain of biomedical science. We shall describe how this solution is being realized, and show how it can not only address the problems of data retrieval and reuse, but also lead to enhancements in currently available terminology resources.

OBO-Edit: The Browser, The Editor

*John Day-Richter^{*1}, The OBO-Edit Working Group²*

¹Berkeley Bioinformatics and Ontology Project, Lawrence Berkeley National Laboratory, Berkeley, CA, USA; ²The Gene Ontology Consortium, <http://www.geneontology.org/>

This brief talk will introduce users to OBO-Edit, an editor for OBO ontologies. This talk will focus on OBO-Edit's many useful ontology browsing features, with some brief discussion of editing and ontology checking tools.

From genes to functional blocks in the study of biological systems

*Fatima Al-Shahrour¹, Leonardo Arbiza¹, Hernan Dopazo¹, Jaime Huerta-Cepas^{1,2},
David Montaner^{1,2}, Pablo Minguez¹, and Joaquin Dopazo^{*1}*

*¹Bioinformatics Department, Centro de Investigacion Principe Felipe (CIPF) and
²Functional Genomics Node, INB-CIPF, Autopista del Saler 16, E-46013, Valencia, Spain*

With the popularisation of high-throughput techniques, the need for procedures that help in the biological interpretation of the results has increased enormously. The strategies most commonly used are based on thresholds derived exclusively from experimental values, that assume an independent behaviour for the genes and ignore the functional correlations existing among them. Recently, it has been noted that these procedures are inefficient because of the lack of information caused upon the application of such stringent thresholds. New procedures more inspired in systems biology criteria are under development. Here we present an implementation of a threshold-independent test for the functional interpretation of large-scale experiments, that does not depend on the pre-selection of genes based on the multiple application of independent tests to each gene. The test implemented aims to directly test the behaviour of blocks of functionally related genes, instead of focusing on single genes. In addition, the test does not depend on the type of the data for obtaining significance values and consequently can be applied to different types of genome-scale studies. We exemplify its application in evolution, microarray gene expression data and interactomics data. A web server that performs the test described and other similar can be found at: <http://www.babelomics.org>

GUI GoMiner and High-Throughput GoMiner analysis of alternative splice variants

Barry R Zeeberg^{*1}, David W. Kane², Ari B. Kahn^{1,3}, Michael C. Ryan³, D. Curtis Jamison^{3,4},
Hongfang Liu^{1,5}, Alessandro Ferrucci^{1,6}, William C. Reinhold¹, 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;

³Department of Bioinformatics, George Mason University, Fairfax, Virginia, USA;

⁴College of Informatics, Northern Kentucky University, Highlands Heights, KY;

⁵Department of Biostatistics, Bioinformatics, and Biomathematics,

Georgetown University Medical Center, 4000 Reservoir Road, NW, Washington, DC 20007, USA;

⁶Department of Computer Science and Electrical Engineering,

University of Maryland, Baltimore County, 1000 Hilltop Circle, MD 21050, USA

GUI GoMiner and High-Throughput GoMiner leverage GO for analysis and interpretation of results from microarrays or other high-throughput molecular profiling platforms. We have now developed the functionality to analyze differential splice variant expression as well as “exon expression.” To achieve that goal, we have: (1) introduced a splice variant annotation mechanism into GUI GoMiner and High-Throughput GoMiner (<http://discover.nci.nih.gov/gominer/faq.jsp>; Ule *et al.*, *Nat Genet* **37**: 844, 2005); (2) developed software for customization of the traditional GO database (GOD); (3) developed comprehensive tools for curation of all Affymetrix microarrays (<http://www.affymetrix.com>); and (4) developed a database for exhaustive curation of transcripts associated with all HGCN entries in the NCBI Evidence Viewer (<http://www.ncbi.nlm.nih.gov/sutils/static/evvdoc.html>). We describe the implementation of those resources and their application to microarray results obtained from studies of the DU145 human prostate cancer cell line and its camptothecin-resistant derivative, RC0.1 (Reinhold *et al.*, *Cancer Res* **63**: 1000, 2003). Acknowledgements: This research was supported in part by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research. This research was supported in part by grant IIS-0430743 from the National Science Foundation.

Are your gene products of interest annotated to the GO?

Jennifer I. Clark^{*1} and the Gene Ontology Consortium²

¹EMBL Outstation—European Bioinformatics Institute, Hinxton, UK;

²<http://www.geneontology.org/>

The GO Consortium conducts GO Annotation Camps to introduce new groups to the GO system. We also pair new groups learning to use GO terms to annotate a genome with more experienced ‘mentor’ groups. To further aid annotating groups, the GO project provides links to free, open source annotation software that can be used by any group, and mailing lists where advice can be sought on any GO related matter. For any questions on the GO project, or for assistance using the system to annotate your genome or gene products of interest, please contact the GO mailing list on gohelp@geneontology.org.

Using The Gene Ontology (GO) to Annotate Mouse Genes at Mouse Genome Informatics

David Hill, Mary Dolan, Harold Drabkin, Alex Diehl, Li Ni and Judith Blake
The Jackson Laboratory, Bar Harbor, ME, USA*

The Mouse Genome Informatics Consortium uses the GO to record and display information about the function and cellular location of mouse gene products. Curators use the standards adopted by the GO Consortium to curate primary biomedical literature articles that link GO terms with mouse gene products. MGI uses several formats to visualize GO annotations for users and to support complex queries and data access. These include:

1. Computational methods to generate textual descriptions of gene products.
2. Tabular formats of GO annotations
3. Graphical formats based on the GO DAG structure
4. Browser-based views permitting entry into the data from the GO itself
5. Comparative graphs of functional annotations for mammalian orthologs

We also have several prototype tools available that extend the mammalian-centric view of functional annotations based on the biomedical literature. One of these tools represents all annotation made to mouse gene products and their orthologs. This tool allows Users to retrieve all information about a 'gene' in a species-independent manner. Future plans include the representation of GO curation status of a gene in our database and the representation of anatomical and cell type information that is co-curated with GO annotations. The integration of the functional annotations for mouse genes using the GO system complements the comprehensive integration of mouse genetics, genomic and phenotypic information represented in MGI.

Improving AmiGO, the Gene Ontology Browser

*Jane Lomax*¹, The AmiGO Working Group²,
¹EMBL Outstation—European Bioinformatics Institute, Hinxton, UK;
²The Gene Ontology Consortium, <http://www.geneontology.org/>*

AmiGO is an HTML based application (<http://www.godatabase.org/>) developed and supported by the Gene Ontology (GO) project. AmiGO allows users to browse, query and visualize GO and related data. Users can search by GO category or by gene product name/symbol, and a tree view of the three ontologies is also available to help users browse and view the relationships between the GO terms. AmiGO can also be used to BLAST multiple sequences against the GO database to retrieve best match GO annotations.

With the vast number of new genomes being sequenced and more gene products being annotated using GO terms, the GO project is making attempts to improve the AmiGO web application so that searching, retrieving and visualizing data can be made more efficient and meaningful. Here we present some options for redesigning the AmiGO web interface and we would like to solicit your feedback for an optimal representation of GO data.

Plant Associated Microbe Gene Ontology (PAMGO): A community resource of gene ontology terms describing gene products involved in microbe-host interactions

*Trudy Torto-Alalibo*¹, *Bryan Bieh*², *David Bird*², *Marcus Chibucos*¹, *Allan Collmer*³, *Candace Collmer*⁴, *Ralph Dear*², *Michelle Gwinn Giglio*⁵, *Jeremy D. Glasner*⁶, *Amelia Ireland*⁷, *Magdalen Lindeberg*³, *Jane Lomax*⁷, *Thomas K. Mitchell*², *Nicole Perna*⁶, *Joao Setubal*¹, *Brett Tyler*¹ and *Owen White*⁵

¹Virginia Bioinformatics Institute, Blacksburg, VA, USA; ²North Carolina State University, Raleigh, NC, USA; ³Cornell University, Ithaca, NY, USA; ⁴Wells College, Aurora, NY, USA; ⁵The Institute for Genome Research, Rockville, MD, USA; ⁶University of Wisconsin, Madison, WI, USA; ⁷EMBL Outstation—European Bioinformatics Institute, Hinxton, UK

The PAMGO interest group was formed to develop new gene ontology (GO) terms describing the various processes, functions and cellular components related to microbe-host interactions. Plant-associated microbes have evolved similar mechanisms to evade, neutralize or suppress defense systems of their plant hosts and obtain nutrients. Such similarities can only be discovered if a controlled vocabulary is set in place to describe these processes amongst diverse microbe-host interactions. In a multi-institutional collaborative effort, we are currently working on developing new GO terms and relationships for gene products implicated in plant interactions in the bacterial pathogens *Erwinia chrysanthemi*, *Pseudomonas syringae* pv tomato and *Agrobacterium tumefaciens*, the fungus *Magnaporthe grisea*, the oomycetes *Phytophthora sojae* and *Phytophthora ramorum* and the nematode *Meloidogyne hapla*. Most terms developed are housed under the “interaction between organisms” (IBO) node. This collaborative effort has since led to the establishment of 291 terms in the IBO node, of which 113 were added directly by the PAMGO group. At a recently held PAMGO ontology development meeting, we presented 190 more terms, which are currently being processed for integration into the GO. Annotations are being done concurrently with ontology development but the number of annotations will increase when the new terms are placed in the ontology. In the future, we hope to develop and evaluate an automated system to transfer PAMGO terms from the species being worked on now to related genomes of plant-associated microbes. Researchers willing to contribute to the ontology development and other discussions can subscribe to the PAMGO discussion list at the PAMGO website (<http://pamgo.vbi.vt.edu/>). This project is supported by grants to BT from the National Research Initiative of the USDA and the US National Science Foundation.

Using GO terms as entities in phenotype annotations

Leyla Bayraktaroglu, *Tom Conlin*, *David Fashena*, *Ken Frazer*, *Melissa Haendel*, *Doug Howe*, *Prita Mani*, *Christian Pich*, *Srihar Ramachandran*, *Kevin Schaper*, *Erik Segerdell*, *Xiang Shao*, *Peiran Song*, *Judy Sprague*, *Brock Sprunger*, *Sierra Taylor*, *Ceri Van Slyke*^{*}, and *Monte Westerfield*
The Zebrafish Information Network (ZFIN), University of Oregon, Eugene, OR, USA

Phenotypic characteristics are often described in terms of biological processes, cellular components, or molecular functions. Consequently, terms from the Gene Ontology are well suited to serve as primary entities in phenotype annotations to which qualities from the Phenotype and Trait Ontology (PATO) are added to describe how the entities are changed by mutation or disease. This poster explores the proposed use of GO terms as an integral part of phenotype annotations. A series of use cases is presented, representing a spectrum of annotation complexities that illustrate difficult query and curation issues associated with complex and detailed phenotype annotations involving GO. ZFIN is supported by the NIH (P41 HG002659).