D440–D444 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 4 November 2007*
doi:10.1093/nar/gkm883

The Gene Ontology project in 2008

The Gene Ontology Consortium*

Received September 14, 2007; Accepted October 1, 2007

ABSTRACT

The Gene Ontology (GO) project (http://www.geneontology.org/) provides a set of structured, controlled vocabularies for community use in annotating genes, gene products and sequences (also see http://www.sequenceontology.org/). The ontologies have been extended and refined for several biological areas, and improvements to the structure of the ontologies have been implemented. To improve the quantity and quality of gene product annotations available from its public repository, the GO Consortium has launched a focused effort to provide comprehensive and detailed annotation of orthologous genes across a number of 'reference' genomes, including human and several key model organisms. Software developments include two releases of the ontology-editing tool OBO-Edit, and improvements to the AmiGO browser interface.

INTRODUCTION

The Gene Ontology (GO) project (http://www.geneontology.org/) is a collaborative effort to develop and use ontologies to support biologically meaningful annotation of genes and their products in a wide variety of organisms. Major model organism databases and other bioinformatics resource centers contribute to the project.

The GO ontologies provide a systematic language, or ontology (1–4), for the consistent description of attributes of genes and gene products, in three key biological domains that are shared by all organisms: molecular function, biological process and cellular component (5–11). A fourth ontology, the Sequence Ontology (SO), covers sequence features (12,13).

GO CONTENT DEVELOPMENT

The branches of the Gene Ontology continue to be dynamic, changing to reflect the current state of biological knowledge and expanding to meet the needs of its user communities. Recently GO has made improvements in both biological content and ontology structure. A summary of the current content of the GO is shown in Table 1.

Biological content improvements

Content-oriented meetings, which bring together GO curators and community experts, have facilitated large-scale changes in specific areas of the ontology. Face-to-face meetings are now supplemented by virtual meetings via teleconferencing utilities and wikis. Interfacing with experts provides GO curators with the most up-to-date views of a given field, thus allowing comprehensive ontological representation of a biological domain; this often entails the addition of many new terms and the refinement of existing terms and relationships. In exchange, GO curators can provide community experts with an in-depth introduction to the GO.

GO curators have recently implemented major changes in the host–pathogen interactions, immunological process (14), central nervous system development and transporter portions of the ontology. The most recent content meetings covered cardiovascular physiology and muscle physiology, and ontology changes agreed upon at these meetings are in progress. Biological topics targeted for improvement include signaling and responses to chemical substances.

Small-scale modifications to the content of the GO are made continuously, in response to individual requests submitted by curators via the GO SourceForge tracker (http://geneontology.sourceforge.net). These requests usually reflect the immediate needs of gene product annotators, and those arising from the Reference Genome annotation project (see below) are given priority.

Structural improvements

A significant improvement to ontology structure was completed at the end of 2006 when all of the GO ontologies became *is_a* complete. This means that all non-root terms in the ontologies now have an *is_a* parent, and a complete path to the root of the ontology that traverses only *is_a* relations. As part of this effort, GOC curators also examined and refined the relationships between many existing terms. As a result, the ontologies are more complete, and are more interoperable with a wider range of ontology tools. This also opens the way

*To whom correspondence should be addressed: Tel: +44(0)1223 494667; Fax: +44(0)1223 494468; Email: midori@ebi.ac.uk

Table 1. Status of GO, September 1, 2007

Biological process terms	13 916
Molecular function terms	7878
Cellular component terms	2007
Sequence ontology terms	1305
Annotation datasets[a]	35
Species with annotation	1 37 454
Annotated gene products	
Total	3 34 7495
Electronic	3 12 8309
Manual	2 19 186

[a]Most datasets represent single species; Gramene, the TIGR gene index, UniProt GOA and UniProt PDB represent multiple species.

to more advanced reasoning, and more flexible ontology visualization.

GO (and SO) curators use OBO-Edit (15), a Java-based ontology editor, to maintain and improve the ontologies. The most recent version of OBO-Edit offers sophisticated filtering, editing, reasoning and error-checking capabilities with both GUI and command-line access. Built-in checks in OBO-Edit help maintain ontology structure, ensuring that curators correctly format definitions and maintain complete *is_a* paths. Other computational checks generate reports that flag possible logical inconsistencies in the ontologies for examination and resolution by GOC curators.

The reasoner in OBO-Edit takes advantage of the increased expressive power of the new OBO Format 1.2 specification (http://www.geneontology.org/GO.format.obo-1_2.shtml). Using OBO Format 1.2, the GO can now provide ontology subsets ('GO slims') and metadata term obsoletion (e.g. alternative terms to update annotations). The format also supports anticipated new features such as cross-products, multi-species interactions (e.g. host–pathogen interactions) and homologous gene sets. The new specification also supports GO's current focus on offering improved support for biochemical pathways by appropriately linking the three branches of the GO. There is also ongoing work specifying cross-products that forge links between the GO and other ontologies in the Open Biomedical Ontologies (OBO) Foundry (http://obofoundry.org/) collection.

THE SEQUENCE ONTOLOGY

The Sequence Ontology continues to provide the terminology needed to describe biological sequences, with the aim of unifying sequence annotation (13). SO has been adopted by many model organism databases as it forms the basis for sequence feature annotation by the Chado database schema (16) and the GFF3 (http://www.sequenceontology.org/gff3.shtml) sequence annotation format.

The ontology now contains over 1300 terms, 161 of which are cross-products; the latter are new terms that are intersections between a sequence feature and a sequence attribute. For example, *low_complexity_region* (SO:0001005) is explicitly defined as a *region* (SO:0000001) that has the characteristic of low sequence complexity. SO held two workshops in 2006 covering topics ranging from mobile genetic elements to RNA editing and mammalian immunology, thereby meeting the needs of diverse users. SO has also actively collaborated with other ontology and database groups to develop and incorporate new terminology. For example, SO recently added 96 protein-based features to the ontology after working with the Biosapiens group (17). New terms can also be requested via a SourceForge tracker or the developers' mailing list.

IMPROVING ANNOTATION COVERAGE

The Reference Genome annotation project

In early 2006, the GO Consortium initiated the Reference Genome project, whose purpose is to offer comprehensive GO annotation for a varied group of organisms: *Arabidopsis thaliana* (thale cress), *Caenorhabditis elegans* (nematode), *Danio rerio* (zebrafish), *Dictyostelium discoideum* (slime mold), *Drosophila melanogaster* (fruit fly), *Escherichia coli*, *Gallus gallus* (chicken), *Homo sapiens* (human), *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Saccharomyces cerevisiae* (baker's yeast) and *Schizosaccharomyces pombe* (fission yeast) (Gene Ontology Reference Genome group, manuscript in preparation). These organisms were chosen because they form an experimentally studied, biologically and evolutionarily diverse group for which expert curators are available to ensure accurate and consistent genome annotation. Comprehensiveness of annotation is measured by how many genes in a genome are annotated (breadth) and how completely those annotations capture known functions (depth). The standards of annotations of the Reference Genome groups are high: measures of comprehensiveness count only annotations based directly on experimental data and those based upon manual inspection of sequence similarity to an experimentally characterized target. The current targets for annotation are human genes associated with heritable diseases and their orthologs (or similar genes) in the other genomes. Figure 1 shows a selection of the molecular function GO terms used to annotate human *MSH2* and its orthologs; the complete graph is available as Supplementary Data.

The Reference Genome project will eventually extend its annotation targets beyond the initial focus on disease genes, and will refine annotation metrics, ortholog identification and public access to annotations.

Annotation of reference genes helps determine ontology development priorities, such that any new terms needed are added to the ontology promptly. A significant expected benefit of the high quality annotation set provided by the Reference Genome group is to assist annotation in other species, especially for newly sequenced genomes.

Supporting new annotation groups

The GO Consortium has continued to actively support the use of GO to annotate emerging genomes, by offering annual annotation camps and one-to-one mentoring to new groups. Several new annotation groups have joined the GO project in the past two years, including AgBase (http://www.agbase.msstate.edu/; (18)), Muscle TRAIT

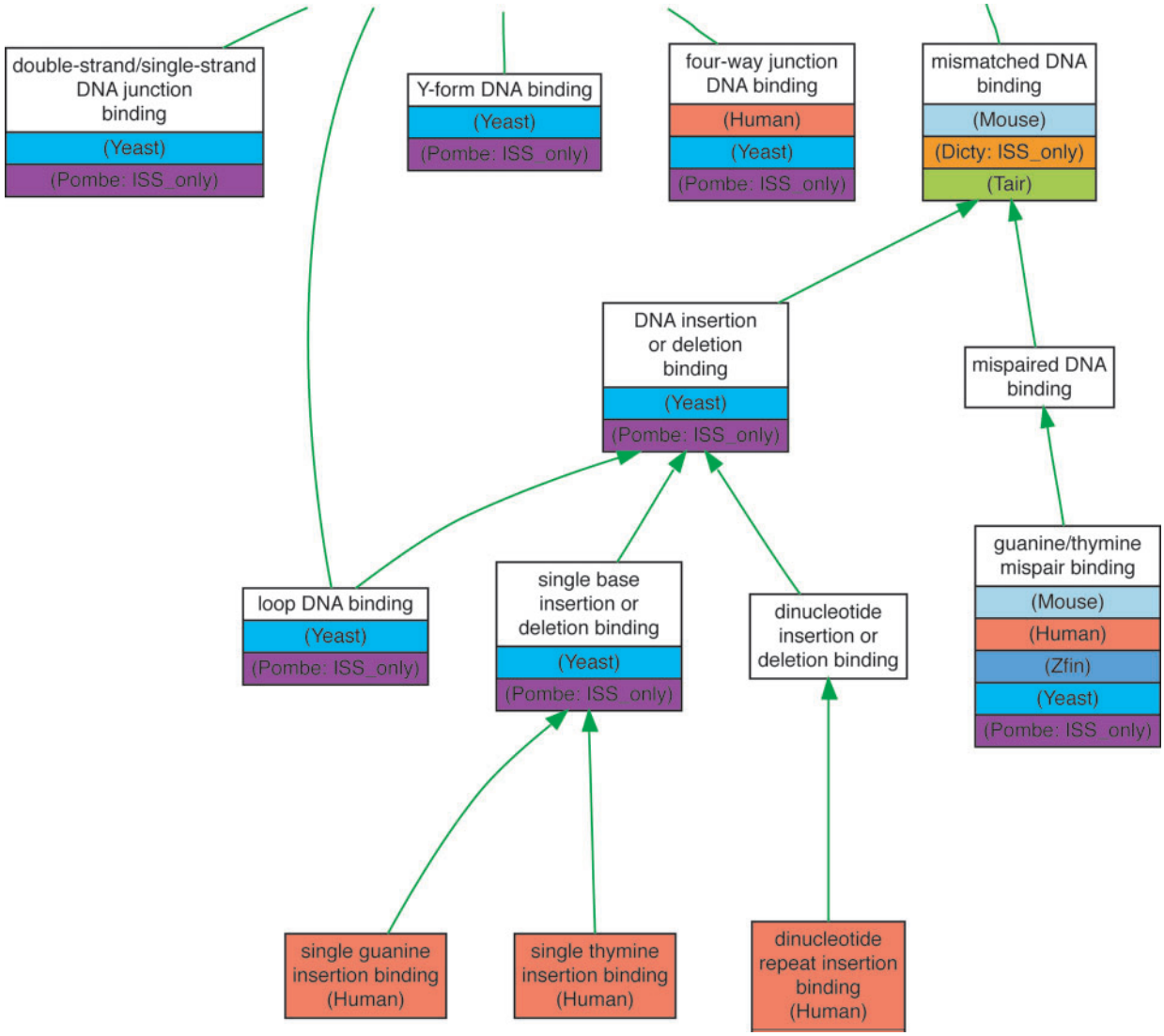

Figure 1. A selection of annotations curated by GO Reference Genome project annotators for human *MSH2* and its predicted orthologs. GO terms are depicted graphically, with arrows representing relationships between terms; each term includes color-coded panels representing the species in which an ortholog has been annotated to the term.

(http://muscle.cribi.unipd.it/), and the Plant-Associated Microbe Gene Ontology (http://pamgo.vbi.vt.edu/) group.

The GO Consortium has also supported a number of new community annotation efforts. In 2006, the Eurofung group (http://www.eurofung.net/) held a community annotation meeting, during which biologists provided manual GO annotation for approximately 500 gene products. GO has, in addition, set up a community annotation wiki to allow experimental immunologists to contribute annotations (see http://www.geneontology.org/GO.immunology.shtml).

USER SUPPORT AND AVAILABILITY

The GO Consortium has initiated a coordinated effort to establish lines of communication between itself and users of the ontologies and annotations in the scientific community to ensure that the Consortium remains aware of, and receptive to, the community's needs. A quarterly newsletter is now produced to communicate current and future developments within the GO Consortium; in addition, the GO Newsletter features uses of GO within the scientific community in its 'Paper Review' and provides tips for using OBO-Edit or AmiGO. The newsletter can be viewed online (http://www.geneontology.org/newsletter/current-newsletter.shtml) or by subscribing to the GO-friends mailing list.

The AmiGO browser (now at http://amigo.geneontology.org), a web-based tool for searching GO terms and annotations, has been enhanced by the addition of new navigation and search options, an improved display of search results, and a simplified user interface. Advanced users have the option to query any of several GO database SQL mirrors programatically or via the GO Online SQL Environment (GOOSE) interface (http://www.berkeleybop.org/goose); the latter includes a collection of

user-adaptable templates for common queries. GO also offers semantic web access to the database via an experimental SPARQL (http://www.w3.org/TR/rdf-sparql-query/) EndPoint. URLs and database access details are available as Supplementary Data.

The GO and SO are available in both OBO and OWL formats and can be browsed online using AmiGO and miSO, respectively. GO annotations are available in their native tab-delimited format and as RDF-XML and a MySQL database dump. URLs for downloads and documentation are available as Supplementary Data.

Please do not hesitate to contact GO at gohelp@geneontology.org with comments, questions or suggestions about any GO resources, or enquiries about contributing GO annotations.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The Gene Ontology Consortium is supported by National Human Genome Research Institute (NHGRI) grant HG02273. GO Consortium member databases receive funding from several National Institutes of Health institutes (NHGRI, National Institute of Child Health and Human Development, National Heart, Lung, and Blood Institute, National Institute of General Medical Sciences) and by the National Science Foundation, the United States Department of Agriculture Cooperative State Research and Education Service and the UK Medical Research Council. The GO Consortium also thanks the community researchers who have participated in content-related meetings or provided valuable feedback on ontology content and annotations. Funding to pay the Open Access publication charges for this article was provided by NHGRI.

Conflict of interest statement. None declared.

REFERENCES

1. Gruber,T.R. (1993) A translation approach to portable ontology specifications. *Knowl. Acq.*, **5**, 199–220.
2. Jones,D.M. and Paton,R. (1999) Toward principles for the representation of hierarchical knowledge in formal ontologies. *Data Knowl. Eng.*, **31**, 99–113.
3. Stevens,R., Goble,C.A. and Bechhofer,S. (2000) Ontology-based knowledge representation for bioinformatics. *Brief. Bioinform.*, **1**, 398–414.
4. Schulze-Kremer,S. (2002) Ontologies for molecular biology and bioinformatics. *In Silico Biol.*, **2**, 179–193.
5. The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
6. The Gene Ontology Consortium (2001) Creating the Gene Ontology resource: design and implementation. *Genome Res.*, **11**, 1425–1433.
7. Blake,J.A. and Harris,M.A. (2003) The Gene Ontology project: structured vocabularies for molecular biology and their application to genome and expression analysis. In Baxevanis,A.D., Davison,D.B., Page,R.D. M., Petsko,G.A., Stein.,L. D. and Stormo,G. (eds), *Current Protocols in Bioinformatics*, John Wiley & Sons, New York.
8. The Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
9. Harris,M.A., Lomax,J., Ireland,A. and Clark,J.I. (2005) The Gene Ontology project. In Subramaniam,S. (ed), *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*, John Wiley & Sons, New York.
10. Lewis,S.E. (2005) Gene Ontology: looking backwards and forwards. *Genome Biol.*, **6**, 103.
11. The Gene Ontology Consortium (2006) The Gene Ontology (GO) project in 2006. *Nucleic Acids Res.*, **34**, D322–D326.
12. Eilbeck,K. and Lewis,S.E. (2004) Sequence Ontology annotation guide. *Compar. Funct. Genom.*, **5**, 642–647.
13. Eilbeck,K., Lewis,S.E., Mungall,C.J., Yandell,M., Stein,L., Durbin,R. and Ashburner,M. (2005) The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.*, **6**, R44.
14. Diehl,A.D., Lee,J.A., Scheuermann,R.H. and Blake,J.A. (2007) Ontology development for biological systems: immunology. *Bioinformatics*, **23**, 913–915.
15. Day-Richter,J., Harris,M.A., Haendel,M. and Lewis,S. (2007) OBO-Edit – an ontology editor for biologists. *Bioinformatics*, **23**, 2198–2200.
16. Mungall,C.J. and Emmert,D.B. and the FlyBase Consortium (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, **23**, i337–i346.
17. Reeves,G.A. and Thornton,J.M. (2006) Integrating biological data through the genome. *Hum. Mol. Genet.*, **15 (Spec No. 1)**, R81–R87.
18. McCarthy,F.M., Bridges,S.M., Wang,N., Magee,G.B., Williams,W.P., Luthe,D.S. and Burgess,S.C. (2007) AgBase: a unified resource for functional analysis in agriculture. *Nucleic Acids Res.*, **35**, D599–D603.

APPENDIX

Midori A. Harris, Jennifer I. Deegan (née Clark), Amelia Ireland, Jane Lomax (GO-EBI, Hinxton, UK); Michael Ashburner, Susan Tweedie (FlyBase, Department of Genetics, University of Cambridge, Cambridge, UK); Seth Carbon, Suzanna Lewis, Chris Mungall, John Day-Richter (BBOP, LBNL, Berkeley, CA, USA), Karen Eilbeck (Eccles Institute of Human Genetics, University of Utah, Salt Lake City, UT, USA), Judith A. Blake, Carol Bult, Alexander D. Diehl, Mary Dolan, Harold Drabkin, Janan T. Eppig, David P. Hill, Li Ni, Martin Ringwald (MGI, The Jackson Laboratory, Bar Harbor, ME, USA); Rama Balakrishnan, Gail Binkley, J. Michael Cherry, Karen R. Christie, Maria C. Costanzo, Qing Dong, Stacia R. Engel, Dianna G. Fisk, Jodi E. Hirschman, Benjamin C. Hitz, Eurie L. Hong, Cynthia J. Krieger, Stuart R. Miyasato, Robert S. Nash, Julie Park, Marek S. Skrzypek, Shuai Weng, Edith D. Wong, Kathy K. Zhu (SGD, Department of Genetics, Stanford University, Stanford, CA, USA); David Botstein, Kara Dolinski, Michael S. Livstone, Rose Oughtred (Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA); Tanya Berardini, Donghui Li, Seung Y. Rhee (TAIR, Carnegie Institution, Department of Plant Biology, Stanford, CA, USA); Rolf Apweiler, Daniel Barrell, Evelyn Camon, Emily Dimmer, Rachael Huntley, Nicola Mulder (GOA Database, UniProt, EBI, Hinxton, UK); Varsha K. Khodiyar, Ruth C. Lovering, Sue Povey (UCL, London, UK); Rex Chisholm, Petra Fey, Pascale

Gaudet, Warren Kibbe (dictyBase, Northwestern University, Chicago, IL, USA); Ranjana Kishore, Erich M. Schwarz, Paul Sternberg, Kimberly Van Auken (WormBase, California Institute of Technology, Pasadena, CA, USA); Michelle Gwinn Giglio, Linda Hannick, Jennifer Wortman (The J. Craig Venter Institute, Rockville, MD, USA); Martin Aslett, Matthew Berriman, Valerie Wood (Wellcome Trust Sanger Institute, Hinxton, UK); Howard Jacob, Stan Laulederkind, Victoria Petri, Mary Shimoyama, Jennifer Smith, Simon Twigger (RGD, Medical College of Wisconsin, Milwaukee, WI, USA); Pankaj Jaiswal (Gramene, Department of Plant Breeding, Cornell University, Ithaca, NY, USA); Trent Seigfried (MaizeGDB, UDSA-ARS, Ames, IA, USA); Doug Howe, Monte Westerfield (ZFIN, University of Oregon, Eugene, OR, USA); Candace Collmer (PAMGO, Wells College, Aurora, NY, USA); Trudy Torto-Alalibo (PAMGO, Virginia Bioinformatics Institute, VA, USA); Erika Feltrin, Giorgio Valle (CRIBI, University of Padua, Italy); Susan Bromberg, Shane Burgess, Fiona McCarthy (AgBase, Mississippi State University; MS, USA).