D396–D403 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 27 October 2008*
doi:10.1093/nar/gkn803

The GOA database in 2009—an integrated Gene Ontology Annotation resource

Daniel Barrell*, Emily Dimmer, Rachael P. Huntley, David Binns, Claire O'Donovan and Rolf Apweiler

Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received September 12, 2008; Revised October 9, 2008; Accepted October 10, 2008

ABSTRACT

The Gene Ontology Annotation (GOA) project at the EBI (http://www.ebi.ac.uk/goa) provides high-quality electronic and manual associations (annotations) of Gene Ontology (GO) terms to UniProt Knowledgebase (UniProtKB) entries. Annotations created by the project are collated with annotations from external databases to provide an extensive, publicly available GO annotation resource. Currently covering over 160 000 taxa, with greater than 32 million annotations, GOA remains the largest and most comprehensive open-source contributor to the GO Consortium (GOC) project. Over the last five years, the group has augmented the number and coverage of their electronic pipelines and a number of new manual annotation projects and collaborations now further enhance this resource. A range of files facilitate the download of annotations for particular species, and GO term information and associated annotations can also be viewed and downloaded from the newly developed GOA QuickGO tool (http://www.ebi.ac.uk/QuickGO), which allows users to precisely tailor their annotation set.

INTRODUCTION

The UniProt Knowledgebase (1) (which includes the UniProtKB/Swiss-Prot (2) and UniProtKB/TrEMBL (2) databases) is the world's most comprehensive and highly annotated publicly accessible protein sequence database, having archived more than 6.6 million proteins through a combination of manual and electronic techniques. It is the annotations attached to such sequences that are proving key in enabling researchers to interpret the biology underlying genomic or proteomic investigations. However, efficient management and manipulation of information attached to such large datasets requires standardized vocabularies and ontologies for describing biological function (3).

The Gene Ontology project was founded in 1998 to address the challenges of interpreting functional information attached to gene products. To this end, GO (http://www.geneontology.org/) provides three detailed, structured vocabularies of terms (ontologies) that describe the 'molecular functions' that gene products normally carry out, the 'biological processes' that gene products are involved in and the 'subcellular locations' (cellular components) that gene products are located in (4).

A GO annotation is a specific association between a GO term identifier and a gene or protein and has a distinct evidence source that supports the association. A single well-characterized gene product can be annotated to multiple GO terms at different levels of the three GO hierarchies. GO annotation sets are now widely used in the analysis of large proteomic and genomic datasets, where significantly overrepresented GO terms are measured to determine hypotheses for the biological events behind the data, to validate extraction techniques or to provide a broad overview of the principal characteristics of a proteome (5).

This article will report the improvements and progress of the GOA database since its last description in the database issue in 2004 (6).

OVERVIEW OF THE GOA RESOURCE

The Gene Ontology Annotation (GOA) project was started in 2001 to ensure widespread coverage of UniProtKB with functional descriptions produced by the GO resource. GOA has developed in parallel with the growing number of sequences and annotations available in UniProtKB, and currently contains over 32 million annotations to more than 4.3 million proteins (GOA release, September 2008), an increase of over 10-fold since 2004. This growth has been aided by many of the developments described in this article.

The GOA project applies both manual and electronic methods to associate GO terms to UniProtKB entries;

*To whom correspondence should be addressed. Tel: +44 1223 494444; Fax: +44 1223 494468; Email: goa@ebi.ac.uk

both methods are strictly controlled to produce high-quality GO annotation and both require biologists and software engineers. The electronic annotation techniques exploit existing information present in UniProtKB and Ensembl (7) that have been added by other curation efforts. Such pipelines can quickly produce a large number of annotations for a wide range of species, although in order to be able to maintain the correctness of the generated annotations, associated GO terms from these methods can be quite general and information-poor. Manual annotation, on the other hand, involves curators reading literature for evidence of protein functions and locations and although a slow and expensive process, is able to provide far more accurate and detailed annotations. To ensure that GOA provides a comprehensive resource for all species, the group supplements the manual annotations it creates with those from other databases (Table 1). Both manual and electronic methods are vital in creating a resource that has both large coverage and highly detailed information. GOA provides a range of ways users can access these annotations, including monthly releases of a number of files, as well as a powerful online browser (QuickGO) to enable users to search, browse and tailor their annotation set for download. GOA annotations are also displayed by many gene and protein databases, including UniProtKB, Ensembl and Entrez-Gene, as well as uploaded by a range of tools that provide GO analysis services.

ENHANCEMENTS TO THE ELECTRONIC GO ANNOTATION PIPELINES

The GOA group is the main supplier of electronic GO annotations to the GO Consortium. The group's annotation pipelines primarily use existing cross-references, keywords and Enzyme Commission (EC) numbers in UniProtKB entries and, by using 'translation tables' consisting of mappings between these external vocabularies and their equivalent GO terms, create GO annotations to the protein entries. Such translation tables are manually curated to ensure a high-accuracy is obtained from the created GO annotation set, and are passed over the UniProtKB database every 3 weeks in order to reflect any changes in the annotation work carried out by other annotation groups. Similar electronic methods are applied by UniProtKB/Swiss-Prot for HAMAP (High-quality Automated and Manual Annotation of microbial Proteomes) family rules (8), and by InterPro (9) for InterPro domains, whereby HAMAP rules or InterPro domains are assigned GO terms and the protein entries that either fall within these families or contain a mapped domain are automatically assigned the associated GO term(s).

These four mapping methods (UniProtKB/Swiss-Prot Keywords2GO, EC2GO, HAMAP2GO and InterPro2GO) have been applied to UniProtKB by the GOA group for many years. However, they must be continually maintained and checked to ensure that as changes to both the external resource and GO occur and as the number of proteins annotated by these methods increases, the GO annotation sets generated remain correct.

The nature of electronic annotation means that it can easily be applied to large numbers of protein entries, and because UniProtKB covers such a vast array of species, it follows that some of the less well-studied species can benefit from the addition of GO annotation to their proteins, which may well remain experimentally uncharacterized. In addition to this, there are some species which do not have a dedicated annotation effort and so may never be supplied with manual GO annotation. In UniProtKB there are currently 168 308 species (4 257 090 proteins) for which electronic annotation pipelines are the only source of GO annotation.

In this context the GOA group has recognized the need for additional automatic association of GO terms and so have expanded on their provision of electronic annotation pipelines. One such development has been in collaboration with the Swiss Institute of Bioinformatics, where a new GO mapping table has been created to exploit annotations made in UniProtKB Subcellular Location annotation lines. To date, 92% of subcellular location terms from the Comment (CC) lines of UniProtKB entries have now been manually mapped to GO terms, providing an additional 587 074 new cellular component GO annotations.

The mapping of external resources to equivalent GO terms has been gradually developed so that GOA uses mappings from over 14 450 external terms, which has produced more than 30 million annotations from the UniProtKB over the last four years (a 9-fold increase).

GOA also introduced a complementary electronic GO annotation method in December 2006, in collaboration with the Ensembl Compara group, which applies comparative genomics data to propagate annotations to non-model organism species (7) (Table 1). In this pipeline, one-to-one and apparent one-to-one orthology data from Compara has been used to project manual GO terms from a source species (currently one of human, mouse, rat, Xenopus, Drosophila or zebrafish) into one or more target species. Currently 38 different species benefit from this annotation pipeline (including Xenopus, Macaque and Tetraodon), many of which are non-model organism species with few other annotations available. The Ensembl Compara pipeline currently produces over 147 000 GO annotations for more than 35 000 proteins (GOA release, September 2008).

Electronic annotations are updated every 3 weeks, as part of the release process, and procedures are checked automatically for obsolete GO terms or secondary protein accessions/GO identifiers on a weekly basis as well as at release time. Each electronic annotation can be identified by the 'IEA' (Inferred from Electronic Annotation) evidence code, and the specific methods applied to generate an electronic GO annotation can be identified by their distinct 'GO Reference' identifier, which links to a full description of each method on the GO Consortium's web site (ftp://ftp.geneontology.org/go/doc/GO.references) (see Table 1 for details on each GOA-supplied GO_REF identifiers).

Table 1. Summary of data sources and project inputs to GOA

	Source	Description and/or location of data	Reference
Electronic annotations	Ensembl Compara	Projected annotation based on curated gene orthology data obtained from the Ensembl Compara system: http://www.ensembl.org/info/about/docs/compara/index.html Further details: http://www.ebi.ac.uk/GOA/compara_go_annotations.html	GO_REF:0000019
	EC2GO	Inferred annotation from Enzyme Commission number information using manually curated transitive mapping file available from: http://www.geneontology.org/external2go/ec2go Further details: http://www.ebi.ac.uk/GOA/EC2GO.html	GO_REF:0000003
	InterPro2GO	Inferred annotation from InterPro domain information using manually curated transitive mapping file available from: http://www.geneontology.org/external2go/interpro2go Further details: http://www.ebi.ac.uk/GOA/InterPro2GO.html	GO_REF:0000002
	HAMAP2GO	Inferred annotation from HAMAP family rules information in UniProtKB. A mapping file is available from: http://www.geneontology.org/external2go/hamap2go Further details: http://www.ebi.ac.uk/GOA/HAMAP2GO.html	GO_REF:0000020
	SPKW2GO	Inferred annotation from UniProtKB/Swiss-Prot keyword location information using manually curated transitive mapping file available from: http://www.geneontology.org/external2go/spkw2go Further details: http://www.ebi.ac.uk/GOA/Swiss-ProtKeyword2GO.html	GO_REF:0000004
	SPSL2GO	Inferred annotation from UniProtKB/Swiss-Prot subcellular location information using manually curated transitive mapping file available from: http://www.geneontology.org/external2go/spsl2go Further details: http://www.ebi.ac.uk/GOA/SubcellularLocation2GO.html	GO_REF:0000023
Remote collaborators	AgBase	AgBase, Mississippi State University curators concentrating on GO annotation of agricultural animal genes.	http://www.agbase.msstate.edu
	BHF-UCL	British Heart Foundation, University College London curators, concentrating on GO annotation of cardiovascular genes.	http://www.cardiovasculargeneontology.com
	HGNC	HUGO Gene Nomenclature Committee curators concentrating on human annotation.	http://www.genenames.org
	Roslin Institute	Roslin Institute, University of Edinburgh curators concentrating on chicken annotation.	http://www.roslin.ac.uk
External groups	dictyBase	Dictyostelium manual GO annotation ftp://ftp.geneontology.org/go/gene-associations	http://dictybase.org
	FlyBase	Drosophila manual GO annotation ftp://ftp.geneontology.org/go/gene-associations	http://flybase.org
	GeneDB Spombe	*S. pombe* manual GO annotation ftp://ftp.geneontology.org/go/gene-associations	http://www.genedb.org/genedb/pombe
	Gramene	Oryza manual GO annotation ftp://ftp.geneontology.org/go/gene-associations	http://www.gramene.org
	Human Protein Atlas	Human protein subcellular location experimental data from immunofluorescence studies http://www.proteinatlas.org/data/go_if_loc.php	http://www.proteinatlas.org
	IntAct	High-quality protein–protein interaction annotation data ftp://ftp.ebi.ac.uk/pub/databases/intact/current/various	http://www.ebi.ac.uk/intact
	LIFEdb	Human subcellular location experimental data from fluorescent fusion protein studies http://www.dkfz.de/LIFEdb/LIFEdb.aspx?query=all&output=XML	http://www.dkfz.de/LIFEdb
	MGI	Mouse and non-mouse manual GO annotation ftp://ftp.geneontology.org/go/gene-associations	http://www.informatics.jax.org
	Reactome	Mammalian manual GO annotation ftp://ftp.geneontology.org/go/gene-associations	http://reactome.org
	RGD	Rat manual GO annotation ftp://ftp.geneontology.org/go/gene-associations	http://rgd.mcw.edu
	SGD	*S. cerevisiae* manual GO annotation ftp://ftp.geneontology.org/go/gene-associations	http://www.yeastgenome.org
	TAIR	Arabidopsis manual GO annotation ftp://ftp.geneontology.org/go/gene-associations	http://www.arabidopsis.org

(continued)

Table 1. Continued

Source	Description and/or location of data	Reference
JCBI and IGS, Univ. of Maryland (formerly TIGR)	Bacterial manual GO annotation ftp://ftp.geneontology.org/go/gene-associations	http://www.tigr.org
Wormbase	*C. elegans* and related nematode manual GO annotation ftp://ftp.geneontology.org/go/gene-associations	http://www.wormbase.org

Further details of individual GO_References details can be found at: http://www.geneontology.org/cgi-bin/references.cgi

DEVELOPMENTS TO MANUAL GO ANNOTATION ACTIVITIES

Although electronic GO annotation methods have produced annotations for a large number of proteins in UniProtKB, it is the manual curation of published literature, where a curator reads and extracts information from full-text papers, which enables the creation of associations between very specific, information-rich GO terms and proteins. Curators contributing to the manual GO annotation effort closely follow the guidelines set out by the GO Consortium (http://www.geneontology.org/GO.annotation.conventions.shtml). GOA also provides a sophisticated curation tool, available to both GOA curators and collaborators, called Protein2GO that supports over 40 curators in generating and managing manual GO annotations to UniProtKB accessions by supplying a range of aids and checks to ensure that the submitted annotations are entered in the required format and use valid GO terms and protein accessions.

GOA curators aim to locate and annotate the most recent papers which provide experimental evidence for the unique features of a given protein. All manual GO annotations contain a reference to the publication (such as a PubMed identifier) from where the specific functional information was located, and can make use of a range of evidence codes to indicate further which type of evidence has been found within the cited paper to support the GO term-sequence identifier association [for further details, please see: http://www.ebi.ac.uk/GOA/annotationexample.html and Hill *et al.* (10)]. GOA-supported curators are now able to annotate to specific protein isoforms, as well as prioritising the creation of GO annotations which are directly supported by published experimental evidence (as indicated by the present of the 'IDA','IMP','IPI','IGI' and 'IEP' evidence codes).

Manual annotation is an expensive and time-consuming activity, and the number of manual annotations available for a particular species can be far fewer than those an electronic annotation pipeline can produce. This reflects not only on the amount of functional literature available, but also the amount of resources a curation team has been able to devote to GO annotation. However the GOA group has been able to draw on the expertise of curators from other EBI groups: including the UniProtKB/Swiss-Prot, IntAct and InterPro projects, as well as integrating manual annotations from all GO Consortium groups and other specialist resources. In total GOA integrates annotations from 18 different groups (Table 1), which combined provide an additional 387 897 annotations to the GOA dataset. All integrated annotations are mapped to UniProtKB accessions, enabling users to work with a fully integrated dataset, and the database source of each annotation is acknowledged in column 15 of the gene association file (http://www.geneontology.org/GO.format.annotation.shtml). As of September 2008, GOA was able to provide 467 134 manual annotations and contained citations to over 48 000 unique papers from where functional data has been identified.

It is the manual annotation of the human proteome that is the priority of the GOA group, and to this end, GOA are involved in a number of focused human annotation projects. GOA closely collaborates with the University College London-based Cardiovascular Gene Ontology Annotation Initiative (www.cardiovasculargeneontology.com/), headed by Prof. Philippa Talmud and funded by the British Heart Foundation (11). This project is prioritising the annotation of genes associated with the cardiovascular system and over 4000 cardiovascular-associated genes have been identified as targets for GO annotation over the 5 year duration of the grant. The cardiovascular initiative is complemented by a new three-year project in the GOA group which will focus on the annotation of renal-associated gene products. This work will be funded by Kidney Research UK from January 2009 and aims to manually annotate at least 750 genes over a three year period. Both the heart and the kidney annotation efforts are directly targeted towards benefiting research communities who make considerable use of the GO resource and where much improvement of both the GO terms and annotation are needed for the investigators to fully benefit. Both projects are also being carried out in close consultation with an expert advisory panel, to ensure that the annotation effort maximally benefits the research community.

Curators from the GOA and cardiovascular projects also comprise the human proteome annotation effort for the GO Consortium Reference Genome Annotation project (http://www.geneontology.org/GO.refgenome.shtml), which aims to extend the GO annotation of a representative set of twelve key genomes (known as 'reference genomes') to its broadest and deepest level with the literature available.

To be able to provide the most detailed set of annotations, these three focused manual annotation projects actively seek feedback from the research communities

for their annotation and ontology development efforts, to ensure that current knowledge is comprehensively reviewed and correctly summarized. Interested parties are encouraged to contribute by contacting the groups directly via our mailing lists or wiki pages (see 'Contributing to GOA' section below).

CONTRIBUTING TO GOA

Keeping GO annotation up-to-date and accurate is an enormous task and inevitably there can be inaccuracies in the annotation. The GOA group greatly relies on feedback from users to report such erroneous annotation; this can be done either via e-mail to goa@ebi.ac.uk or by using the GO SourceForge site (http://geneontology.sourceforge.net/). To assist us in improving the annotation, please provide as much detail as possible, including the protein accession number, GO term identifier and the creation date of the annotation.

For further information on the Cardiovascular Annotation Initiative, please see www.ucl.ac.uk/silva/cardiovasculargeneontology. Comments about the project can be posted on the editable wiki at http://wiki.geneontology.org/index.php/Priority_Cardiovascular_genes#Community_Input_Section

For more information on the Renal Annotation Initiative, please see http://www.ebi.ac.uk/GOA/kidney/.

DATA ACCESS

Annotations supplied by the GOA group can be accessed in various ways, which are detailed below.

QUICKGO ONTOLOGY AND ANNOTATION BROWSER

As the GOA annotation set has grown rapidly over the last 10 years, so has its user community grown and diversified. In line with this, QuickGO (http://www.ebi.ac.uk/QuickGO), one of the first web-based GO browsers, has been extensively redeveloped such that users can now query QuickGO with a range of different keywords or identifier types, to find either comprehensive, detailed information on GO terms (Figures 1 and 2) or sets of GO annotation data, which they can filter to their specific needs and download in a range of formats (Figure 3).

The new version of QuickGO has placed GOA's extensive set of GO annotations at the heart of the tool and enables users to easily filter by a number of characteristics, such as species or taxonomic group, evidence type or GO term set, then evaluate and download the resulting set of annotations. Drop-down menus at the top of the annotation table on the 'Annotation Download' page provide a simple way of filtering the GO annotations (Figure 3), whereas more complex queries can be entered into the 'Advanced' search text box on this page, allowing users to apply a combination of Boolean operations (AND, NOT, OR) to their queries.

To support QuickGO users who would like to map GO annotations to different identifier types, an identifier mapping facility for 14 different sequence identifier types (including Ensembl, Entrez Gene, RefSeq, SGD, MGI and IPI identifiers) has been included directly in the tool.

Once users have selected their desired annotation set, QuickGO can provide detailed paginated views of annotations as well as statistics on multiple aspects of the filtered annotations. Statistics are calculated on-the-fly, and provide detailed summaries of the distribution of GO terms, evidence codes and annotation sources within the dataset (Figure 3). Much of the data displayed in the annotation table can be clicked on (including GO terms, taxonomy information and sequence identifiers) to either provide further information, or provide alternative navigation routes. Users can also specify the format that their downloaded data takes—with the tool offering GOA association file, gene2go or customized formats as well as protein FASTA files or identifier lists. These facilities ensure that QuickGO users are now able to download GO annotations that are tailor-made to their requirements. No other GO browser currently provides such extensive identifier mapping or annotation filtering facilities.

QuickGO also now provides users with the ability to view and modify an existing slim GO slim or generate their own. GO slims are subsets of GO terms extracted from the whole Gene Ontology and tend to consist of a limited number of high-level GO terms that have been selected to provide an overview of some or all of the content of GO. GO slims are often used to provide a broad overview of the chief functional characteristics of a set of sequences (12). QuickGO users have direct access to predefined GO slims, which have been extracted from the GO Consortium's OBO file, but can equally easily create a new GO slim set by entering a list of GO identifiers or by selecting terms when browsing QuickGO. QuickGO can provide graphical displays of all slims and GO annotations can be 'mapped-up' to these term sets, along with statistics for number/percentage of annotations associated with each slim term.

WEB SERVICES

All of the data provided by QuickGO can be queried remotely, both for GO term information and annotation data. These web services are fully integrated, so that the filtering options and datasets available are fully synchronized between the browsable and web service interfaces. Web service information is provided at the bottom of each QuickGO page where details on how to construct web queries, format options and sample scripts showing how to query QuickGO in Java, Perl and Bash are provided. The web services have been designed for ease of use; QuickGO provides a REST style query interface in which all information is provided in the URL and the results are in tab separated, OBO or XML formats conforming to well-established standards.

RELEASED FILES

GOA also provides monthly releases of GO annotations, in 15-column tab-delimited 'gene association' file formats

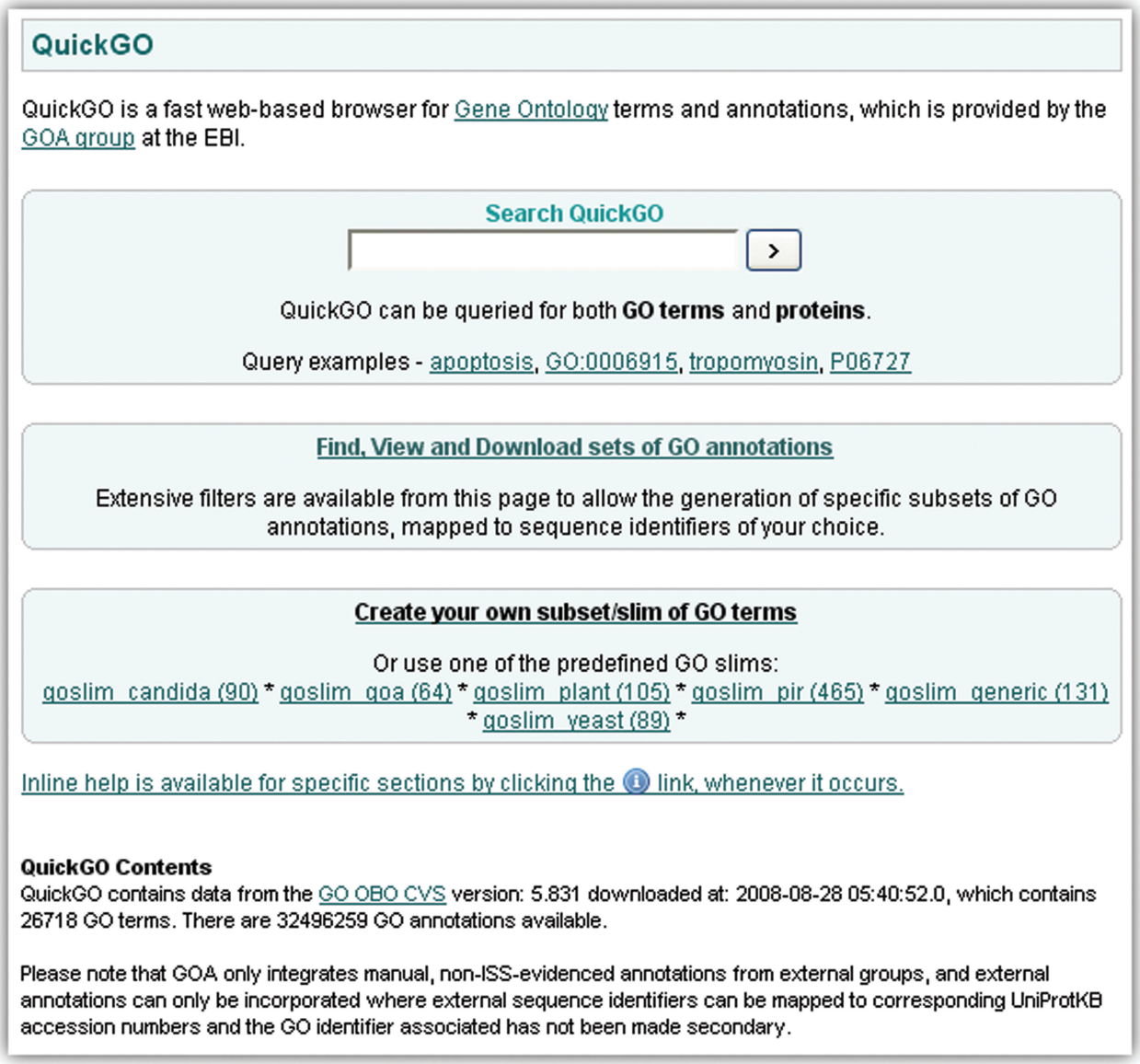

Figure 1. Screenshot of the front page of the QuickGO browser (http://www.ebi.ac.uk/QuickGO).

that provide comprehensive information on each GO annotation, including gene and protein names extracted from UniProtKB. All annotations in the GOA database can be retrieved by downloading the UniProtKB gene association file. This file has rapidly grown and is now 4.3 Gb (363 Mb compressed) in size. However, such a large, unwieldy volume of data is often not required by many researchers who only need annotations to a particular species, therefore GOA also makes available a range of species-specific files. One set of species-specific files available from GOA for the human, mouse, rat, chicken, cow and zebrafish proteomes are created in collaboration with the International Protein Index (IPI) (13) effort, which provides a maximally complete, non-redundant set of identifiers for the main databases that describe proteomes (including UniProtKB, Ensembl, TAIR, Vega, H-Invitational and NCBI's RefSeq databases).

The InterPro2GO method is passed over all non-UniProtKB database entries in these sets, to aid in annotation of objects that may or may not be annotated manually.

At the start of 2005 GOA started to provide Gene Association Proteome sets (http://www.ebi.ac.uk/GOA/proteomes.html), where individual gene association files are created for each strain or species whose genome has been fully sequenced and publicly available and contains more than 25% GO annotation. In 2005 the first release contained 218 species, however this resource has grown in line with the public sequencing projects to cover over 850 proteome sets.

A number of files (gp2protein files) to help users map between sequence identifier types are now also available, including mappings for IPI identifiers between different databases, as well as UniProtKB to GeneID, UniProtKB

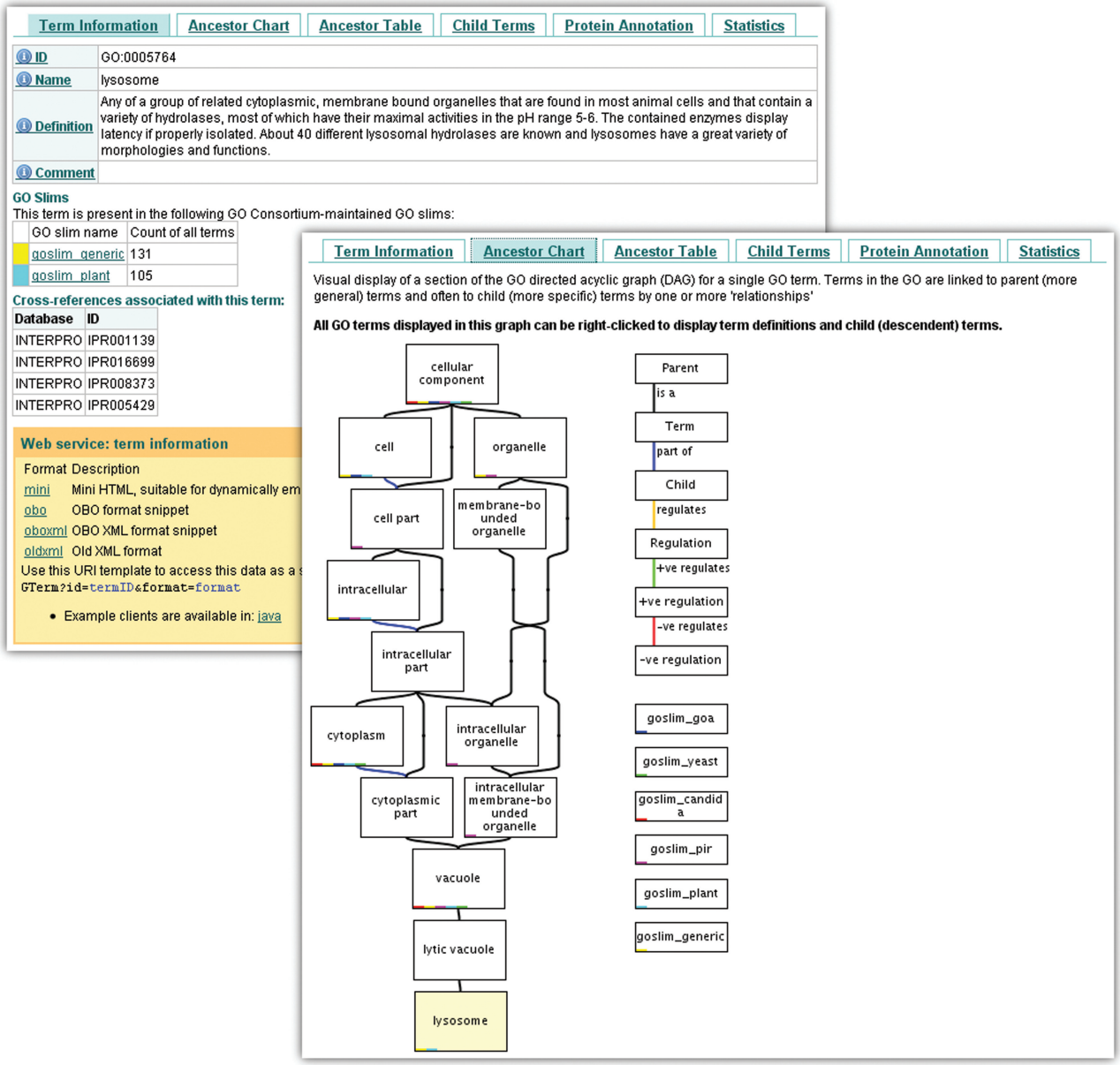

Figure 2. QuickGO GO term information pages. QuickGO screenshots of the tabbed 'Term Information' and 'Ancestor Chart' pages for the GO term lysosome (GO:0005764).

to RefSeq identifiers and UniProtKB to UniGene clusters mappings (http://www.geneontology.org/gp2protein/).

GOA gene association files are available from both the GO Consortium and GOA ftp sites:
ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/
ftp://ftp.geneontology.org/pub/go/gene-associations

EBI MIRROR FOR THE GO MySQL DATABASE

Since 2007, GOA has provided a mirror to the complete GO MySQL database on public servers at the EBI. This is used by the Gene Ontology Online SQL Environment Tool (http://www.berkeleybop.org/goose). However users can also query the database directly using a MySQL client over an internet connection thus negating the need to download, install and maintain your own copy of the database. Full connection details are:

- User: go_select
- Password: amigo
- Host: mysql.ebi.ac.uk
- Port: 4085

An example connection from the command line:

$ mysql -hmysql.ebi.ac.uk -ugo_select -pamigo -P4085

Full connection details are available from http://www.geneontology.org/GO.database.shtml#SQL.

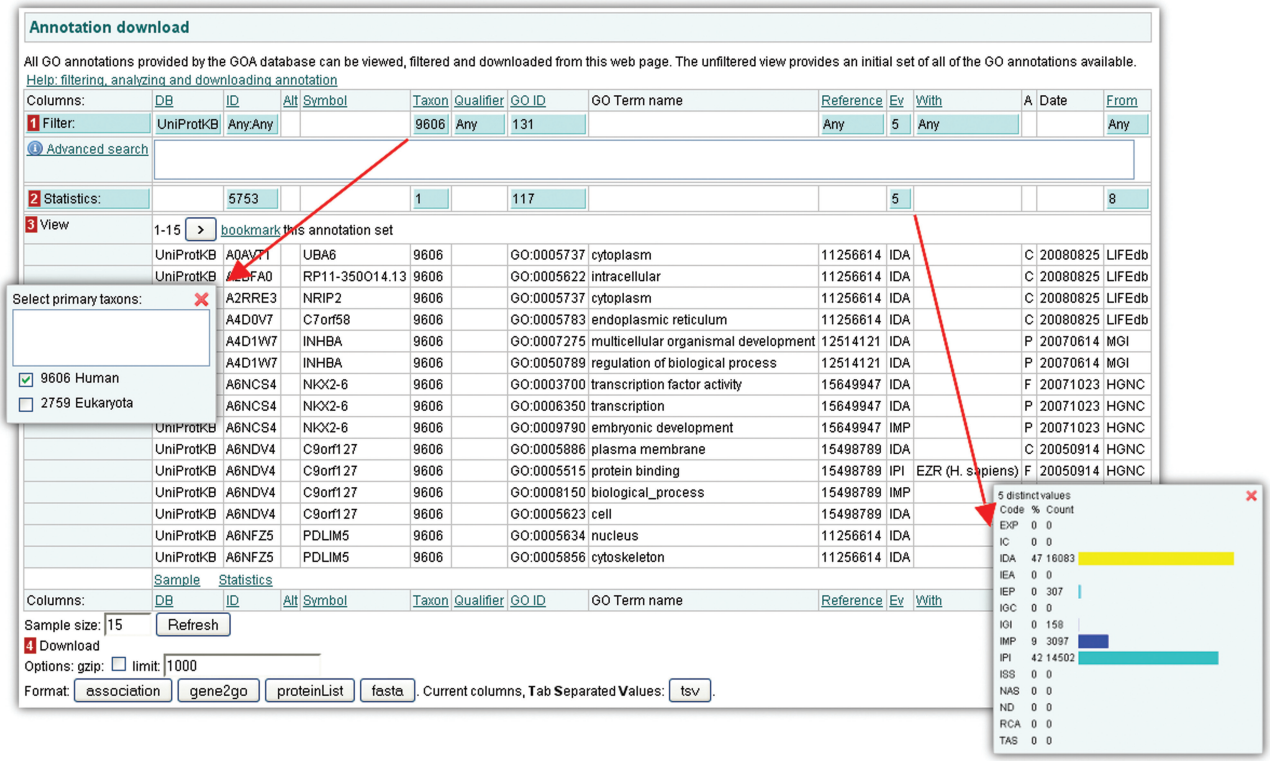

Figure 3. QuickGO GO annotation page screenshot. QuickGO provides a range of options to tailor a GOA annotation set for downloading. The screenshot displays an annotation set which has been filtered so that only those experimentally evidenced annotations to human UniProtKB accessions are selected. This screenshot also shows the taxon filtering option, to demonstrate the ease of filtering annotation sets, as well as the statistics generated for the evidence code usage in the selected set. The bottom of this screenshot also displays the range of download options available to QuickGO.

ACKNOWLEDGEMENTS

The GOA project would like to thank Serenella Ferro Rojas for her work on mapping UniProtKB/Swiss-Prot Subcellular Locations to GO terms and the GO curators from all collaborating groups that help enrich the data set.

FUNDING

National Human Genome Research Institute (HG002273); the British Heart Foundation (SP:07/007/23671); core EMBL funding. Development of the QuickGO browser was possible due to a BBSRC Tools and Resources Fund grant (BB/E023541/1). Funding for open access charge: NIH NHGRI grant number R01HG02273-02: Gene Ontology Consortium.

Conflict of interest statement. None declared.

REFERENCES

1. UniProt Consortium. (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **36**, D190–D195.
2. Boutet,E., Lieberherr,D., Tognolli,M., Schneider,M. and Bairoch,A. (2007) UniProtKB/Swiss-Prot: The Manually Annotated Section of the UniProt KnowledgeBase. *Methods Mol. Biol.*, **406**, 89–112.
3. Thomas,P.D., Huaiyu,M. and Lewis,S. (2007) Ontology annotation: mapping genomic regions to biological function. *Curr. Opin. Chem. Biol.*, **11**, 4–11.
4. Lomax,J. (2005) Get ready to GO! A biologist's guide to the Gene Ontology. *Brief Bioinform.*, **6**, 298–304.
5. Dimmer,E.C., Huntley,R.P., Barrell,D.G., Binns,D., Draghici,S., Camon,E.B., Hubank,M., Talmud,P.J., Apweiler,R. and Lovering,R.C., (2008) *Practical Proteomics*; Jul 17, [Epub ahead of print].
6. Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. and Apweiler,,R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.*, **32**, D262–D266.
7. Flicek,P., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al.* (2008) Ensembl 2008. *Nucleic Acids Res.*, **36**, D707–D714.
8. Gattiker,A., Michoud,K., Rivoire,C., Auchincloss,A.H., Coudert,E., Lima,T., Kersey,P., Pagni,M., Sigrist,C.J., Lachaize,C. *et al.* (2003) Automated annotation of microbial proteomes in SWISS-PROT. *Comput. Biol. Chem.*, **27**, 49–58.
9. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Buillard,V., Cerutti,L., Copley,R. *et al.* (2007) New developments in the InterPro database. *Nucleic Acids Res.*, **35**, D224–D228.
10. Hill,D.P., Smith,B., McAndrews-Hill,M.S. and Blake,J.A. (2008) Gene Ontology annotations: what they mean and where they come from. *BMC Bioinformatics*, **9(Suppl. 5)**, S2.
11. Lovering,R.C., Dimmer,E, Khodiyar,V.K., Barrell,D.G., Scambler,P., Hubank,M., Apweiler,R. and Talmud,P.J. (2008) Cardiovascular GO annotation initiative year 1 report: why cardiovascular GO? *Proteomics*, **8**, 1950–1953.
12. Rhee,S.Y., Wood,V., Dolinski,K. and Draghici,S. (2008) Use and misuse of the gene ontology annotations. *Nat. Rev. Genet.*, **9**, 509–515.
13. Kersey,P.J., Duarte,J., Williams,A., Karavidopoulou,Y., Birney,E. and Apweiler,R. (2004) The International Protein Index: An integrated database for proteomics experiments. *Proteomics*, **4**, 1985–1988.