Use and misuse of the gene ontology annotations

Seung Yon Rhee, Valerie Wood[‡], Kara Dolinski[§] and Sorin Draghici[||]*

Abstract | The Gene Ontology (GO) project is a collaboration among model organism databases to describe gene products from all organisms using a consistent and computable language. GO produces sets of explicitly defined, structured vocabularies that describe biological processes, molecular functions and cellular components of gene products in both a computer- and human-readable manner. Here we describe key aspects of GO, which, when overlooked, can cause erroneous results, and address how these pitfalls can be avoided.

The accumulation of data produced by genome-scale research requires explicitly defined vocabularies to describe the biological attributes of genes in order to allow integration, retrieval and computation of the data[1]. Arguably, the most successful example of systematic description of biology is the Gene Ontology (GO) project[2]. GO is widely used in biological databases, annotation projects and computational analyses (there are 2,960 citations for GO in version 3.0 of the ISI Web of Knowledge) for annotating newly sequenced genomes[3], text mining[4,5], network modelling[6] and clinical applications[7], among others. GO has two components: the ontologies themselves, which are the defined terms and the structured relationships between them (GO ontology); and the associations between gene products and the terms (GO annotations). GO provides both ontologies and annotations for three distinct areas of cell biology: molecular function, biological process, and cellular component or location.

Researchers who use GO should understand how the ontologies are structured and how genes are annotated so that they can avoid errors in interpretation. Here we describe the frequently overlooked aspects of GO and discuss their consequences through examples from common applications.

Ontology structure

Ontologies are formal representations of a specific knowledge domain, in this case, cell biology. The GO ontology is represented as a directed acyclic graph (DAG) in which the terms are nodes and the relationships among them are edges. Key characteristics of a DAG in the context of GO are that: parent–child relationships are defined (FIG. 1), with parent terms representing more general entities than their child terms; and, unlike a simple tree (FIG. 1a), a term in a DAG can have multiple parents (red node or grey edge in FIG. 1b). These characteristics of the GO structure enable powerful grouping, searching and analysis of genes.

Fundamental aspects of GO annotations

A GO annotation associates a gene with terms in the ontologies and is generated either by a curator or automatically through predictive methods. Genes are associated with as many terms as appropriate as well as with the most specific terms available to reflect what is currently known about a gene. When a gene is annotated to a term, associations between the gene and the terms' parents are implicitly inferred. Because GO annotations to a term inherit all the properties of the ancestors of those terms, every path from any term back to its root(s) must be biologically accurate or the ontology must be revised[8]. For example, if a gene is known to be specifically involved in 'vesicle fusion', it will be annotated directly to that term, and it is implicitly annotated (indirectly) to all of its parents' terms, including 'membrane fusion', 'membrane organization and biogenesis', 'vesicle-mediated transport', 'transport' and so on, back to the root node (FIG. 1c). Thus, a gene annotated to vesicle fusion can be retrieved not only with this term, but also with all of its parent terms, increasing flexibility and power when searching for and making inferences about genes.

Evidence codes — not all annotations are created equal. All GO annotations include an evidence code to record the type of information on which the annotation is based. Evidence codes can be broadly divided into four categories: experimental, computational, indirectly derived from either of the first two categories, or unknown. TABLE 1 describes the 14 evidence codes that are used by GO. Annotations derived from direct experimental evidence are generally thought to be of higher quality than those inferred from computational or indirect evidence, although this has

**Carnegie Institution for Science, Department of Plant Biology, 260 Panama Street, Stanford, California 94305, USA.*
[‡]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK.
[§]Lewis–Sigler Institute for Integrative Genomics, Carl Icahn Laboratory, Princeton University, Princeton, New Jersey 08544, USA.
[||]Wayne State University, Department of Computer Science, 5,143 Cass Ave, Room 431 State Hall, Detroit, Michigan, 48202, USA.
All authors contributed equally to this work.
Correspondence to S.Y.R. or S.D.
e-mails: rhee@acoma.stanford.edu; sod@cs.wayne.edu
doi:10.1038/nrg2363
Published online 13 May 2008

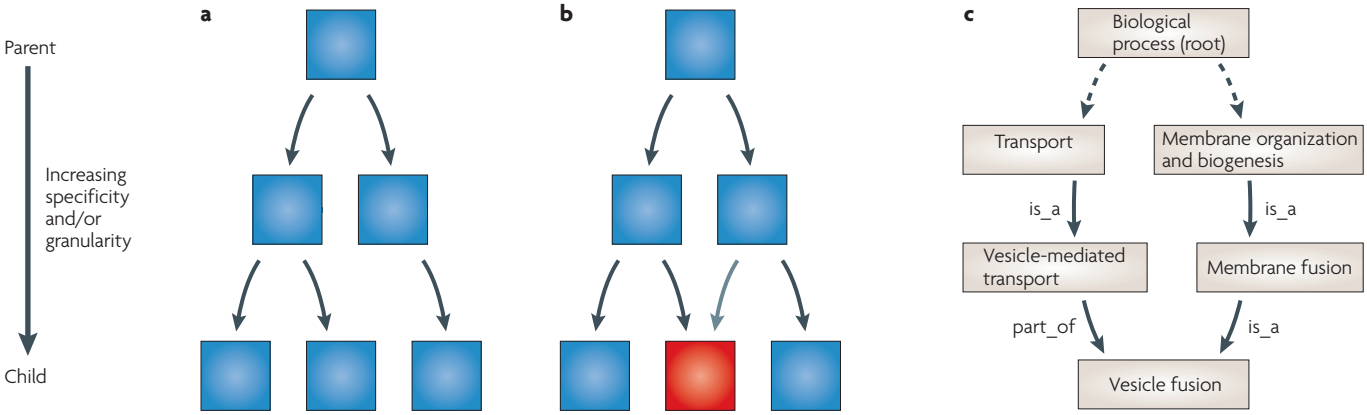

Figure 1 | **Simple trees versus directed acyclic graphs.** Boxes represent nodes and arrows represent edges. **a** | An example of a simple tree, in which each child has only one parent and the edges are directed, that is, there is a source (parent) and a destination (child) for each edge. **b** | A directed acyclic graph (DAG), in which each child can have one or more parents. The node with multiple parents is coloured red and the additional edge is coloured grey. **c** | An example of a node, vesicle fusion, in the biological process ontology with multiple parentage. The dashed edges indicate that there are other nodes not shown between the nodes and the root node (biological process). A root is a node with no incoming edges, and at least one leaf (also called a sink). A leaf node is a node with no outgoing edges, that is, a terminal node with no children (vesicle fusion). Similar to a simple tree, A DAG has directed edges and does not have cycles, that is, no path starts and ends at the same node, and will always have at least one root node. The depth of a node is the length of the longest path from the root to that node, whereas the height is the length of the longest path from that node to a leaf[41]. is_a and part_of are types of relationships that link the terms in the GO ontology. More information about the relationships between GO terms are found online (An Introduction to the Gene Ontology).

not been shown robustly. As of October 2007, there are over 16 million GO annotations. Strikingly, over 95% of these annotations are computationally derived and have not been manually curated; these are associated with the 'inferred from electronic annotation' (IEA) evidence code. Most of these annotations come from the GO annotation project at the European Bioinformatics Institute (GOA[9]). In addition to the GOA set, individual model organisms also have a substantial portion of annotations not derived from direct experimental evidence (TABLE 2). Among the 27 organisms with more than 5,000 annotations, the portion of genes with at least one experimentally derived annotation varies widely from 1.1% to 90.9%. Although computational and indirectly derived annotations increase coverage significantly, they probably contain a higher portion of false positives. Researchers who use GO annotations should be cognizant of the differences between annotations associated with different evidence codes.

Table 1 | **Evidence codes used by GO**

Evidence code	Evidence code description	Source of evidence	Manually checked	Current number of annotations*
IDA	Inferred from direct assay	Experimental	Yes	71,050
IEP	Inferred from expression pattern	Experimental	Yes	4,598
IGI	Inferred from genetic interaction	Experimental	Yes	8,311
IMP	Inferred from mutant phenotype	Experimental	Yes	61,549
IPI	Inferred from physical interaction	Experimental	Yes	17,043
ISS	Inferred from sequence or structural similarity	Computational	Yes	196,643
RCA	Inferred from reviewed computational analysis	Computational	Yes	103,792
IGC	Inferred from genomic context	Computational	Yes	4
IEA	Inferred from electronic annotation	Computational	No	15,687,382
IC	Inferred by curator	Indirectly derived from experimental or computational evidence made by a curator	Yes	5,167
TAS	Traceable author statement	Indirectly derived from experimental or computational evidence made by the author of the published article	Yes	44,564
NAS	Non-traceable author statement	No 'source of evidence' statement given	Yes	25,656
ND	No biological data available	No information available	Yes	132,192
NR	Not recorded	Unknown	Yes	1,185

*October 2007 release

Table 2 | **Distribution of gene ontology (GO) annotations for species with more than 5,000 annotations**

Species (NCBI taxon ID)	Genes* with experimental annotations‡	Total annotated genes*	Percentage of genes* with at least one experimental annotation	Total genes*	Percentage annotated§	Percentage known in genome‖
Schizosaccharomyces pombe (4896)	4,482	4,930	90.9%	4,930	100%	90.9%
Saccharomyces cerevisiae (4932)	4,947	5,794	85.4%	5,794	100%	85.4%
Mouse (10090)	10,621	18,386	57.8%	27,289	67.4%	38.9%
Caenorhabditis elegans (6239)	4,614	14,154	32.6%	20,163	70.2%	22.9%
Human¶ (9606)	4,780	17,021	28.1%	20,887	81.5%	22.9%
Arabidopsis thaliana# (3702)	5,530	26,637	20.8%	27,029	98.5%	20.5%
Rat (10116)	3,566	17,243	20.7%	17,993	95.8%	19.8%
Fruitfly (7227)**	2,790	9,563	29.2%	14,141	67.6%	19.7%
Candida albicans (5476)	806	3,756	21.4%	6,166	60.9%	13.0%
Pseudomonas aeruginosa PAO1 (208964)	491	2,506	19.6%	5,568	45.0%	8.82%
Slime mold (44689)	797	6,892	11.6%	13,625	50.6%	5.9%
Trypanosoma brucei (5691)	449	3,914	11.5%	9,154	42.8%	4.92%
Zebrafish (7955)	1,235	13,574	5.8%	21,322	63.7%	3.7%
Plasmodium falciparum (5833)	188	3,243	5.8%	5,420	59.8%	3.47%
Rice (39947)	654	29,877	2.2%	41,908	71.3%	1.57%
Chicken¶ (9031)	75	6,063	1.2%	16,737	36.2%	0.4%
Cow¶ (9913)	96	8,536	1.1%	21,756	39.2%	0.4%

*Total genes in genomes include only those that encode proteins. These numbers were obtained from the databases that contribute annotations to GO and are listed on the GO annotations download page (http://www.geneontology.org/GO.current.annotations.shtml). ‡Experimental annotations include those only with the following evidence codes: IDA (inferred from direct assay), IEP (inferred from expression pattern), IGI (inferred from genetic interaction), IMP (inferred from mutant phenotype) and IPI (inferred from physical interaction). §Percentage annotated is determined by dividing the number of genes annotated by total genes. ‖Percentage known in genome is determined by multiplying the percentage of experimentally derived annotations by the percentage of the genome annotated. This is an approximation of the extent of knowledge about the portion of the genome that encodes proteins in an organism with a complete genome sequence that is captured by annotation. ¶Numbers are from the GO annotation project at the European Bioinformatics Institute, human data last updated 14 September 2007, cow data last updated 17 January 2007, chicken data last updated 10 July 2007. #Numbers are from The *Arabidopsis* Information Resource (TAIR), last updated 14 December 2007. **Numbers are based on release 5.4 of the *Drosophila melanogaster* genome and GO annotations from FlyBase release FB2007_03 (dated 11 January 2007). NCBI, National Center for Biotechnology Information.

Annotation of 'unknowns' — we know what we don't know. If the process, function or location of a gene is unknown, then it is annotated to the root node of the respective ontology with the evidence code 'no biological data available' (ND). This indicates that a curator has exhaustively checked the literature and could find no data for the gene in question. The ND annotations provide a way to distinguish genes that are unannotated from those that are uncharacterized. Some databases also annotate genes to the root node with the evidence code ND when no orthologues with known function are identified on the basis of computational analyses. However, this is not performed consistently among databases (S.Y.R., unpublished observations). Procedures are being developed, as part of the reference genome annotation project at GO, to perform these analyses more consistently among participating databases.

Annotation qualifiers — to be or not to be is crucial for GO. GO uses three qualifiers, contributes_to, colocalizes_with and NOT, to further refine annotations (see the GO annotation conventions). The NOT qualifier, which indicates the lack of a property, is most vital in data interpretation. This is used judiciously, only when there is potential for confusion or contradiction. For example, a gene product might have sequence similarity to protein kinases, but the curator can apply the NOT qualifier to indicate that, contrary to expectation, the gene product does not exhibit kinase activity based on published results. Although the total number of NOT annotations is minor, several databases have hundreds of these annotations (TABLE 3). Therefore, any analyses using GO annotations should consider the NOT qualifier and exclude them as appropriate. A quick survey of several GO profiling tools shows that many fail to properly consider the NOT annotations (Supplementary information S1 (table)).

Ontology and annotation changes. The GO ontology and annotations are continually updated to reflect current knowledge, to correct errors and to improve logical consistency. The GO ontology is updated daily and most of the annotation files are released weekly. GOA's mappings between GO terms and other descriptors (for example, domains from the Interpro database or Enzyme Commission numbers), which are the major sources for IEA annotations, are updated monthly. Each change to the ontology is tracked, and when terms are deleted they retain their identifier but are flagged as 'obsolete'. These practices allow for versioning of the ontologies and annotations. Although annotations are robust to changes in the ontology because they are made

Table 3 | **NOT annotations in the gene ontology (GO) database***

Contributing database	Number of NOT annotations
CGD	11
Dictybase	76
FlyBase	246
GeneDB_Spombe	83
UniProt	148
AgBase	3
HGNC	41
MGI	217
RGD	21
SGD	88
TAIR	127
ZFIN	37

*As of 12 November 2007. CGD, *Candida* Genome Database; HGNC, HUGO Gene Nomenclature Committee; MGI, Mouse Genome Informatics; RGD, Rat Genome Database; SGD, *Saccharomyces* Genome Database; TAIR, The *Arabidopsis* Information Resource; ZFIN, Zebrafish Information Network.

to the definition of the term and not to the term name or its position in the graph, it is nonetheless vital that researchers use the latest versions of the ontology and annotations (available at the GO downloads web page), and cite the version downloaded, so that their results can be reproduced.

Applications using GO annotations

In one of its most common applications, GO is used to analyse results from high-throughput studies. These studies typically produce sets of genes, and researchers often use the GO annotations to determine which biological processes, functions, and/or locations are significantly over- or under-represented in a group of genes, what new gene functions can be inferred on the basis of the data, and how the given genes are distributed across a pre-defined set of biological categories. Here, we discuss approaches to address these issues using GO, as well as the pitfalls to be avoided.

Functional profiling — seeing the forest from the trees. In many cases, the result of a high-throughput experiment is a set of genes that are differentially expressed between different conditions (for example, cancerous versus healthy). The goal of functional profiling is to determine which processes might be different in particular sets of genes, a process that is often conducted by determining which GO terms are represented differently (for example, significantly more or less often than expected by chance) within the gene set[10–18].

The simplest approach is to calculate 'enrichment/depletion' for each GO term (that is, a higher proportion of genes with certain annotations among the differentially expressed genes than among all of the genes in the study). The main problem here is that any enrichment value can occur just by chance. Therefore, enrichment alone should not be interpreted as unequivocal evidence implicating the GO term in the phenomenon studied without an appropriate statistical test.

More sophisticated approaches calculate the probability of observing a particular enrichment value just by chance using a binomial model. For example, the probability of picking a gene annotated to 'apoptosis' is fixed, and is equal to the proportion of apoptosis genes in the reference set (the set of genes under study). Here, the binomial distribution provides the probability of obtaining a particular proportion of apoptosis genes among the differentially expressed genes by chance[19]. This is a good approximation for large reference sets (for example, whole-genome microarrays) because the probability of picking an apoptosis gene from the reference set hardly changes after each gene is picked. However, once a gene or protein is picked from a smaller reference set (for example, an antibody array that might screen only hundreds of proteins), the probability that the next picked gene is annotated to apoptosis is substantially influenced by whether the previously picked genes were annotated to apoptosis. Under these circumstances, better suited models are the hypergeometric distribution (a discrete probability distribution that describes the number of successes in a sequence of n draws from a finite population without replacement)[19] or the chi squared[20] distribution, both of which take into consideration how the probabilities change when a gene is picked. More recent approaches perform the analysis while considering information about the position of the GO terms in the hierarchy[21–23].

When applying a statistical test in functional profiling, several things should be considered. First, all of the statistical models described above calculate the probability of having the observed number of genes annotated to a given GO term when a random draw is performed from the same reference set. Therefore, it is crucial that an appropriate reference set be used. The reference set should only include the genes that were monitored in the experiment. This is often distinct from the background (that is, total) set of genes in a genome, yet many of the functional profiling tools use an incorrect reference set and hence produce incorrect results. Second, many tools only measure enrichment and ignore depleted GO terms, which could lead to only partial answers to the given biological problem.

Two types of questions can be addressed when performing functional profiling: a hypothesis-generating query, such as which GO terms are significant in a particular set of genes; or a hypothesis-driven query[24], such as whether apoptosis is significantly enriched or depleted in a particular set of genes. In the hypothesis-driven query, one can include all of the genes that are annotated both directly to apoptosis and indirectly to all of its children, and calculate the p-value, maximizing the statistical power because no correction for multiple comparisons is required[25]. The hypothesis-generating approach can also be valuable. An unbiased search for significant GO associations can be done with a bottom-up approach as follows: for every leaf term, calculate p-values with the genes directly associated to it. If any term is significant, do not propagate its genes above. This would provide the most specific node that is significant in that particular branch. If a term is not

significant, propagate the annotations to its parent and re-calculate with the parent term. The genes will propagate upwards until a significant node is found or until the root is reached. A careful analysis is still necessary to properly correct for multiple comparisons.

It is important to correct for the fact that many tests performed in parallel will greatly increase the number of false positives in the entire set of tests. This is a general problem that is not specific to GO. The simplest correction, Bonferroni, would multiply the p-values of all terms with the number of parallel tests performed. If genes are propagated all the way up, to the root node, the number of tests is equal to the number of terms in the GO hierarchy — 25,473 in the October 2007 release. In practice, a term would need to have a raw p-value less than 4×10^{-7} for it to be significant at the 1% significance level. Other corrections, such as Holm's[26] and false discovery rate[27], are less conservative but loss of power cannot be completely avoided (see REFS 28,29 for further reviews). Hence, as a general rule, one can increase the power of the statistical analysis by performing the fewest possible number of tests. One way to do this is to ask specific biological questions by collapsing terms in different regions of the GO structure, before any p-values are calculated, on the basis of the biological hypotheses tested. Genes are then propagated up to the collapsed nodes and the multiple comparison correction needs to use only the number of nodes in this custom cut of the GO. Most tools that are currently available are limited to performing analysis either at a fixed depth or with all nodes, thus preventing the customized collapsing of the GO that can improve significance in most circumstances.

Another issue stemming from the propagation of genes to parent terms is that the parallel tests performed for nodes in a given path will be clearly correlated because the same genes can appear several times on each path. Not all correction methods perform well under such circumstances. GO's structure is important because the lack of independence arises from clear inheritance phenomena that could be used to decorrelate the analysis of various terms[21–23].

A thorough comparison of 14 tools, which discusses their scope, statistical models, visualization capabilities, corrections for multiple comparisons and reference sets, is available elsewhere[25]. Interestingly, submitting the same data set to 10 different ontological analysis programs resulted in p-values ranging over several orders of magnitude for some GO terms (S.D., unpublished observations). Factors that can cause such wildly different results for the same input data include: the method used to map gene and sequence identifiers; the sources and versions of the annotation files (for example, GOA, GO or model organism database); the method of annotation propagation (for example, direct annotations only versus propagated to parents); the statistical testing method (for example, one-sided versus two-sided tests); the actual mathematical formula for the calculation; and the multiple hypothesis correction method. These and other relevant variables should be made explicit in software distributions, as well as in reports of the results. At the very least, researchers should try a few different functional profiling programs before interpreting the results of an experiment.

Using GO to predict gene function, and assessing the results. The GO is often used to infer gene function computationally[30–32]. Typical approaches tend to be variations of the same theme: genes are grouped together on the basis of some criteria such as similar gene expression or through a protein–protein interaction network. Enrichment of GO terms is detected by methods such as those described above, and the uncharacterized genes are presumed to be involved in the same biological processes as the genes with which they are grouped. Therefore, these uncharacterized genes can be putatively associated with the enriched GO terms. It is imperative that annotation practices are considered when taking this approach (for example, the NOT annotations described in the previous section). In a similar manner, propagating gene function on the basis of annotations that are neither manually checked nor experimentally verified (TABLE 1) is likely to result in a substantial number of false positives. Also, the aspect of the GO ontology (biological process, molecular function and cellular component or location) should be considered when making inferences about a gene on the basis of the GO annotations. For example, inferences that are based on correlated gene expression might be reasonable for the biological process ontology terms, but less so for the molecular function and cellular component or location terms. Gene functions can also be inferred from GO annotations without the need for a prior gene grouping, for instance, on the basis of a semantic analysis of the gene function association matrix[33]. This type of analysis relies on capturing the implicit dependencies that might be present between genes.

Computational studies that analyse high-throughput data often benchmark their results to assess their performance. Advantages of using GO for as a tool for evaluation include: comprehensiveness, in which GO encompasses all cellular biological processes as opposed to classification schemes that are specific to a certain domain (for example, metabolic pathways) and thus should exhibit less bias; manual annotation, which is generally thought to be higher quality than computationally generated annotation; and consistent annotation standards across species.

There are some important issues to consider when using GO as a tool for evaluation. Consideration of the DAG structure of the ontology and the information content and source of each annotation are crucial to accurate benchmarking. For example, computational biologists could benchmark their analysis techniques by using precision-recall curves; if they happen to use high-level GO terms that are so general that most gene products are annotated to them from propagations to parent terms they would get artificially inflated performance scores. Alternatively, if only direct annotations are considered, the performance would be artificially poor. One could use a particular level of the GO ontology (for example, terms that are three levels below the root node), but this is problematic because GO's structure is not uniform in its

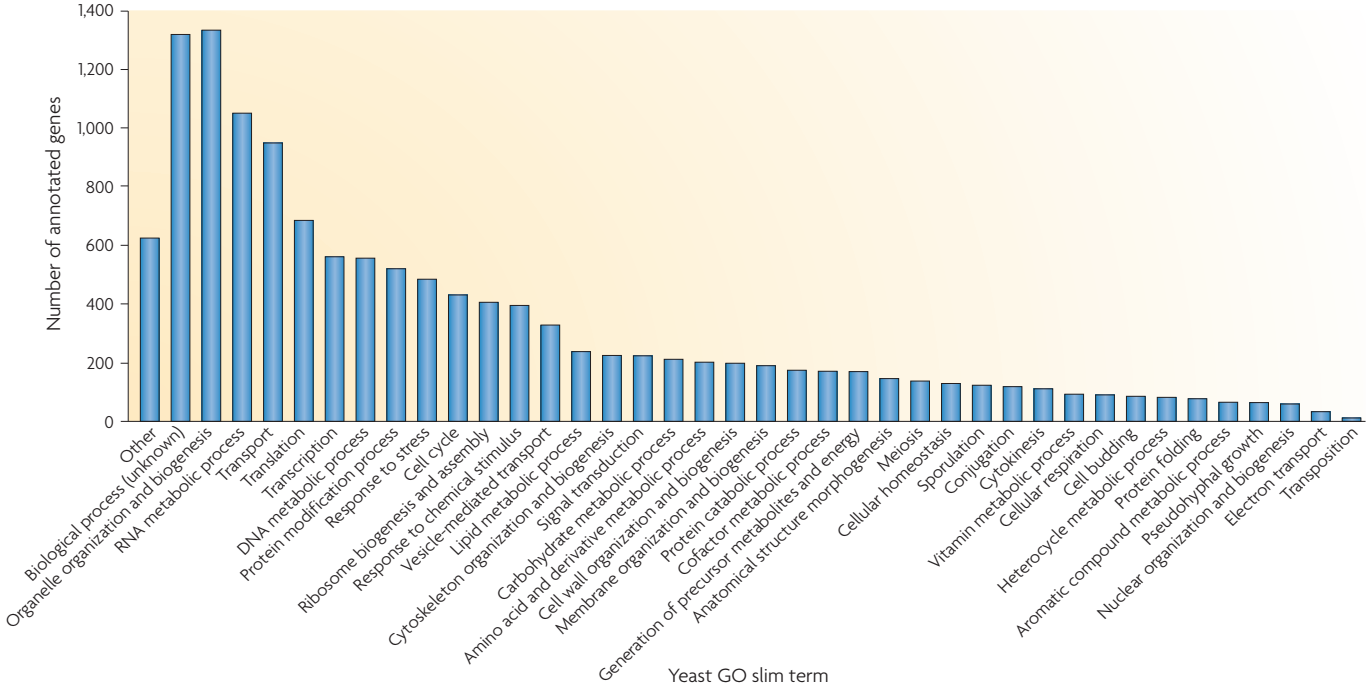

Figure 2 | **Using gene ontology (GO) to bin the yeast genome into broad biological process categories.** This example was generated by downloading the go_slim_mapping.tab file from the *Saccharomyces* genome database ftp site (dated 19 January 2008). This file maps every gene in the yeast genome to the yeast GO slim ontology available from the GO website. The number of genes (6,200 in total, including RNAs but excluding 'dubious' genes) annotated to a particular term in the yeast GO slim ontology is indicated on the graph. Dubious genes are those that were originally predicted to exist, on the basis of ORF length, but that are now thought to be unlikely to encode an expressed protein, on the basis of functional and comparative genomics data. The 'other' term is used when genes are annotated to terms other than those included in the GO slim ontology, and the 'biological process' term, the root node in the biological process ontology, indicates that genes annotated to it are not yet characterized. Note that because genes can be binned to more than one category, there are more annotations (13,074) than total genes (6,200) with annotations.

information content. A better approach is to determine the information content of GO terms. One recent paper chooses informative terms on the basis of the number of gene products annotated to them[34] and another uses a manually derived set[35]. Another common issue with using GO for benchmarking is that some of the data that is used as an input for a particular analysis might also be used as a source for GO annotations, which might cause a circularity problem. For instance, an algorithm that predicts gene function on the basis of expression data should not include GO annotations based on microarray expression from those same experimental studies. Evidence codes and citations provided with each annotation can be used to filter annotations appropriately.

Functional categorization using GO. Another common application of GO is to categorize genes on the basis of a relatively small set of high-level GO terms. Results of the functional categorization are frequently shown as pie charts or bar charts (FIG. 2). This involves the mapping of a set of annotations for the genes of interest to a specified subset of high-level GO terms called a GO slim ontology. This is a typical way of providing an overview of the broad biology encoded by a genome[3,36], EST or cDNA collection[37,38], or of differential expression patterns[18,39,40].

Four GO slim ontologies (generic, plant, yeast and human) are updated regularly and are available from the GO website (see the GO Slim and Subset Guide); previously published GO slims are also archived. The GO website provides an algorithm, map2slim, to map annotations to a slim ontology, which is used by several tools (for example, Princeton University's GO Term Mapper. When categorizing using GO, it is important to choose (or create) both the GO slim and the binning algorithm carefully to generate results that are applicable for a particular analysis. A GO slim file should be chosen or created on the basis of the type of analysis (that is, the specific biological processes and organism of interest).

Regardless of the GO slim chosen, the structure of the GO and the nature of the annotations must be considered. Because the GO is a DAG, a GO term used for a specific annotation might be a child of multiple terms in the slim set. Also, individual gene products often have several annotations to different terms to reflect their multiple functions, roles or locations. Because one gene's annotations frequently map to many slim terms, pie charts — traditionally used to illustrate functional distribution of genes — are not a good representation of the data because the sum of the annotations is larger than 100%, that is, genes are found in more than one slice of the pie. Bar charts are more appropriate here (FIG. 2).

Annotation coverage is also a vital consideration when categorizing a gene set from one organism, or when categorization of gene sets from different organisms are compared (TABLE 2). Taking these caveats together, any GO categorization should provide an indication of how many genes are not mapped to any slim term, how many genes are unknown (that is, mapped directly to the root node), and how many genes are unannotated.

Conclusion

Although GO is a powerful tool, researchers who use it should be cognizant of the features of the ontologies and annotations to avoid common pitfalls. Available annotation for a given organism might affect results and conclusions. Therefore, care should be taken when choosing an analysis method; it might be essential to include or exclude certain types of annotations for certain types of analysis. In addition, it is crucial for any analysis to cite data sources (including the version of ontology, date of annotation files, numbers and types of annotations used, versions and parameters of software, and so on) to ensure that results are fully reproducible. The GO is a tool that will become increasingly powerful for data analysis and functional predictions as the ontologies and annotations continue to evolve. Our hope is that researchers fully understand and thus can take full advantage of this vital resource.

1. Bard, J. B. & Rhee, S. Y. Ontologies in biology: design, applications and future challenges. *Nature Rev. Genet.* **5**, 213–222 (2004).
This paper provides a more detailed overview of types and uses of ontologies in biology, with an emphasis on GO.
2. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25**, 25–29 (2000).
This paper includes more details about the Gene Ontology.
3. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
4. Hirschman, L., Yeh, A., Blaschke, C. & Valencia, A. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics* **6** (Suppl. 1), S1 (2005).
5. Camon, E. B. *et al.* An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC Bioinformatics* **6** (Suppl. 1), S17 (2005).
6. Liu, M. *et al.* Network-based analysis of affected biological processes in type 2 diabetes models. *PLoS Genet.* **3**, e96 (2007).
7. Dressman, H. K. *et al.* Gene expression signatures that predict radiation exposure in mice and humans. *PLoS Med.* **4**, e106 (2007).
8. The Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Res.* **11**, 1425–1433 (2001).
This paper describes in more detail how the GO ontology is built and maintained in more detail.
9. Camon, E., Barrell, D., Lee, V., Dimmer, E. & Apweiler, R. The Gene Ontology Annotation (GOA) Database — an integrated resource of GO annotations to the UniProt Knowledgebase. *In Silico Biol.* **4**, 5–6 (2004).
10. Cai, S. & Lashbrook, C. C. Stamen abscission zone transcriptome profiling reveals new candidates for abscission control: enhanced retention of floral organs in transgenic plants overexpressing *Arabidopsis zinc finger protein 2*. *Plant Physiol.* **146**, 1305–1321 (2008).
11. Datu, B. J. *et al.* Transcriptional changes in the hookworm, *Ancylostoma caninum*, during the transition from a free-living to a parasitic larva. *PLoS Negl. Trop. Dis.* **2**, e130 (2008).
12. Faustino, R. S., Behfar, A., Perez-Terzic, C. & Terzic, A. Genomic chart guiding embryonic stem cell cardiopoiesis. *Genome Biol.* **9**, R6 (2008).
13. Ginos, M. A. *et al.* Identification of a gene expression signature associated with recurrent disease in squamous cell carcinoma of the head and neck. *Cancer Res.* **64**, 55–63 (2004).
14. Li, Y. & Sarkar, F. H. Gene expression profiles of genistein-treated PC3 prostate cancer cells. *J. Nutr.* **132**, 3623–3631 (2002).
15. Okada, H. *et al.* Genome-wide expression of azoospermia testes demonstrates a specific profile and implicates ART3 in genetic susceptibility. *PLoS Genet.* **4**, e26 (2008).
16. Uddin, M. *et al.* Sister grouping of chimpanzees and humans as revealed by genome-wide phylogenetic analysis of brain gene expression profiles. *Proc. Natl Acad. Sci. USA* **101**, 2957–2962 (2004).
17. van der Pouw Kraan, T. C. *et al.* Expression of a pathogen-response program in peripheral blood cells defines a subgroup of rheumatoid arthritis patients. *Genes Immun.* **9**, 16–22 (2008).
18. Zhang, X. *et al.* Whole-genome analysis of histone H3 lysine 27 trimethylation in *Arabidopsis*. *PLoS Biol.* **5**, e129 (2007).
19. Draghici, S., Khatri, P., Martins, R. P., Ostermeier, G. C. & Krawetz, S. A. Global functional profiling of gene expression. *Genomics* **81**, 98–104 (2003).
This paper describes how the significance of enriched or depleted terms is calculated using a number of alternative models in GO profiling.
20. Man, M. Z., Wang, X. & Wang, Y. POWER_SAGE: comparing statistical tests for SAGE experiments. *Bioinformatics* **16**, 953–959 (2000).
21. Alexa, A., Rahnenfuhrer, J. & Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**, 1600–1607 (2006).
This paper explains some of the problems related to the structure of GO and proposes an approach that can be used to address them.
22. Grossmann, S., Bauer, S., Robinson, P. N. & Vingron, M. Improved detection of overrepresentation of Gene Ontology annotations with parent child analysis. *Bioinformatics* **23**, 3024–3031 (2007).
23. Schlicker, A., Rahnenfuhrer, J., Albrecht, M., Lengauer, T. & Domingues, F. S. GOTax: investigating biological processes and biochemical activities along the taxonomic tree. *Genome Biol.* **8**, R33 (2007).
24. McCarthy, F. M., Bridges, S. M. & Burgess, S. C. GOing from functional genomics to biological significance. *Cytogenet. Genome Res.* **117**, 278–287 (2007).
25. Khatri, P. & Draghici, S. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* **21**, 3587–3595 (2005).
This includes a detailed comparison of 14 functional profiling tools using a number of different criteria, including scope of the analysis, visualization capabilities, statistical model(s) used, correction for multiple comparisons, reference microarrays available, installation issues and sources of annotation data.
26. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65–70 (1979).
27. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. (Ser. B)* **57**, 289–300 (1995).
28. Draghici, S. *Data Analysis Tools for DNA Microarrays* (Chapman & Hall/CRC, Boca Raton, Florida, 2003).
29. Farcomeni, A. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Stat. Methods Med. Res.* 14 Aug 2007 (doi:101177/0962280206079046).
30. Marcotte, E. & Date, S. Exploiting big biology: integrating large-scale biological data for function inference. *Brief. Bioinform.* **2**, 363–374 (2001).
31. Markowetz, F. & Troyanskaya, O. G. Computational identification of cellular networks and pathways. *Mol. Biosyst.* **3**, 478–482 (2007).
32. Srinivasan, B. S. *et al.* Current progress in network research: toward reference networks for key model organisms. *Brief. Bioinform.* **8**, 318–332 (2007).
33. Khatri, P., Done, B., Rao, A., Done, A. & Draghici, S. A semantic analysis of the annotations of the human genome. *Bioinformatics* **21**, 3416–3421 (2005).
34. Wong, S. L., Zhang, L. V. & Roth, F. P. Discovering functional relationships: biochemistry versus genetics. *Trends Genet.* **21**, 424–427 (2005).
35. Myers, C. L., Barrett, D. R., Hibbs, M. A., Huttenhower, C. & Troyanskaya, O. G. Finding function: evaluation methods for functional genomic data. *BMC Genomics* **7**, 187 (2006).
36. Yu, J. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**, 79–92 (2002).
37. Kawai, J. *et al.* Functional annotation of a full-length mouse cDNA collection. *Nature* **409**, 685–690 (2001).
38. Whitfield, C. W. *et al.* Annotated expressed sequence tags and cDNA microarrays for studies of brain and behavior in the honey bee. *Genome Res.* **12**, 555–566 (2002).
39. Perrin, R. M. *et al.* Transcriptional regulation of chemical diversity in *Aspergillus fumigatus* by LaeA. *PLoS Pathog.* **3**, e50 (2007).
40. Qin, X., Ahn, S., Speed, T. P. & Rubin, G. M. Global analyses of mRNA translational control during early *Drosophila* embryogenesis. *Genome Biol.* **8**, R63 (2007).
41. Bender, M. A., Farach-Colton, M., Pemmasani, G., Skiena, S. & Sumazin, P. Lowest common ancestors in trees and directed acyclic graphs. *J. Algorithms* **57**, 75–94 (2005).

Acknowledgements

We are grateful to the GO Consortium for their efforts in developing, maintaining and making accessible the GO ontology and annotations. We thank S. Carbon and C. Mungall for their help with SQL queries to the GO database and the following individuals for feedback on this manuscript: M. Ashburner, E. Camon, P. D'Eustachio, E. Dimmer, P. Gaudet, R. Huntley, R. Lovering, C. Mungall, S. Twigger, and K. Van Auken.

FURTHER INFORMATION

Seung Yon Rhee's homepage: http://carnegiedpb.stanford.edu/research/research_rhee.php
Sorin Draghici's homepage: http://vortex.cs.wayne.edu
An Introduction to the Gene Ontology: http://www.geneontology.org/GO.doc.shtml#term-term-relationships
Gene Ontology (GO) project: http://www.geneontology.org
GO annotation conventions: http://www.geneontology.org/GO.annotation.conventions.shtml#qual
GO annotation project at the European Bioinformatics Institute (GOA): http://www.ebi.ac.uk/GOA
GO downloads: http://www.geneontology.org/GO.downloads.shtml
GO Slim and Subset Guide: http://www.geneontology.org/GO.slims.shtml?all
Interpro database: http://www.ebi.ac.uk/interpro
ISI Web of Knowledge: http://apps.isiknowledge.com
Map2slim: http://search.cpan.org/~cmungall/go-perl/scripts/map2slim
Princeton University's GO Term Mapper: http://go.princeton.edu/cgi-bin/GOTermMapper/GOTermMapper
Reference genome annotation project at GO: http://www.geneontology.org/GO.refgenome.shtml

SUPPLEMENTARY INFORMATION

See online article: S1 (table)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF