D322–D327 *Nucleic Acids Research, 2007, Vol. 35, Database issue* *Published online 10 November 2006*
doi:10.1093/nar/gkl799

GO PaD: the Gene Ontology Partition Database

Gil Alterovitz[1,2,3,4,]*, Michael Xiang[5], Mamta Mohan[4] and Marco F. Ramoni[1,3,4]

[1]Division of Health Sciences and Technology, Harvard Medical School and Massachusetts Institute of Technology, Boston, MA, [2]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, [3]Children's Hospital Informatics Program, Boston, MA, [4]Harvard Partners Center for Genetics and Genomics, Harvard Medical School, Boston, MA and [5]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA

Received August 15, 2006; Revised October 1, 2006; Accepted October 2, 2006

ABSTRACT

Gene Ontology (GO) has been widely used to infer functional significance associated with sets of genes in order to automate discoveries within large-scale genetic studies. A level in GO's direct acyclic graph structure is often assumed to be indicative of its terms' specificities, although other work has suggested this assumption does not hold. Unfortunately, quantitative analysis of biological functions based on nodes at the same level (as is common in gene enrichment analysis tools) can lead to incorrect conclusions as well as missed discoveries due to inefficient use of available information. This paper addresses these using an informational theoretic approach encoded in the GO Partition Database that guarantees to maximize information for gene enrichment analysis. The GO Partition Database was designed to feature ontology partitions with GO terms of similar specificity. The GO partitions comprise varying numbers of nodes and present relevant information theoretic statistics, so researchers can choose to analyze datasets at arbitrary levels of specificity. The GO Partition Database, featuring GO partition sets for functional analysis of genes from human and 10 other commonly studied organisms with a total of 131 972 genes, is available on the internet at: bcl.med.harvard.edu/proj/gopart. The site also includes an online tutorial.

INTRODUCTION

The Gene Ontology (GO) (1) is a direct acyclic graph (DAG) with numerous levels and ~20 000 terms. To test biological hypotheses, such as significant functional enrichment (more than by chance) in a set of genes, methods typically assume comparable specificity or rely on the DAG level(s) of the hierarchy (2–4). For example, in some biological experiments, high specificity (e.g. 'terpene metabolism') is needed while in others, more general terms will suffice (e.g. 'metabolism'). DAG structural levels turn out not to be good indicators of specificity. Using the DAG structural levels to identify specificity can cause misleading results or miss biological discoveries altogether (see 'Application' section). Here, the GO Partition Database is presented in order to provide sets of GO terms for hypothesis testing that balances the information content.

GO Partition Database provides three main advantages: the first is that the deliberate selection of a GO term partition containing a relatively small number of GO terms (e.g. 10 or fewer) allows the graph representation to be visually tractable by depicting a manageable set of GO terms. Second, none of the GO terms used in the graph are related by direct ancestor or descendant relationships; thus, if multiple GO terms are over or underrepresented, then each GO term is significant independently and is not due to dependency on other GO term(s) being over or underrepresented. Finally, and most importantly, all the GO terms chosen for a GO term partition are of comparable information content. By ensuring that all GO terms are similar in specificity, GO term partitions allow the GO terms to be compared on equal footing.

The database saves investigators from doing probabilistic analysis as well as presents an advanced query interface that empowers researchers to search for GO partition sets based on specific constraints.

METHODS

Information (in terms of bits) can serve as a proxy for specificity as shown in Figure 1. General GO terms (low specificity) have few bits of information whereas specific terms have more information (higher number of bits). Intuitively, for example, the GO term 'metabolism', which annotates ~40% of human genes, offers little biological insight into the function of a gene; however, the GO term 'carbohydrate metabolism', which annotates ~2% of the human genome, is more specific and informative. Information theory ensures that maximum information can be gained from an analysis

*To whom correspondence should be addressed at: New Research Building, Room 250, 77 Avenue Louis Pasteur, Harvard Medical School, Boston, MA 02115, USA. Tel: +1 617 525 4478; Fax: +1 617 525 4488; Email: gil@mit.edu

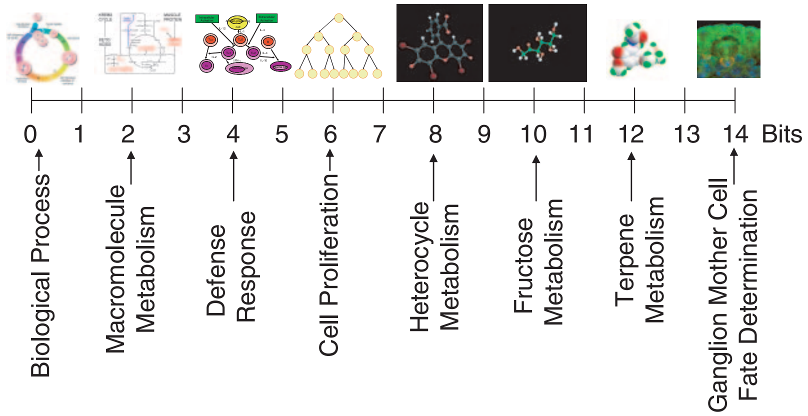


Figure 1. The specificity of GO terms can be captured in terms of bits of information. Heterocycle metabolism image courtesy of: Dr. Brent P. Krueger (14).

of a set of nodes (i.e. GO terms) if the probabilities (of genes corresponding to the analyzed terms) are equivalent. In our method, we use an information-based framework to distribute an ontology, such as GO, into sets of nodes identified by uniform information content (in bits), which represents 'surprisal' or self-information (5,6). Specifically, the probability of observing a gene can be stated in quantitative terms between $p(V_n)$ and $k(V_n)$ and $k(V_m)$ as

$$p(V_n) = \frac{|\, k(V_n)\, |}{|\, \bigcup_{m=1}^{j} k(V_m)\, |} \qquad 1$$

where V_n represents a node, $k(V_n)$ represents the gene set annotated by node V_n, and j is the total number of nodes in the ontology. $p(V_n)$ is the probability of observing a randomly selected gene from the entire genome under study to be annotated by V_n . The information content $I(V_n)$, or Shannon information (7), of a node V_n is denoted as $-\log_2 p(V_n)$ (in bits) as shown below:

$$I(V_n) = -\log_2 p(V_n). \qquad 2$$

This definition of information content implies that an increase in one bit of information relates directly to a 2-fold increase in specificity. Thus, a GO term with 0 bits of information would be expected to annotate all genes; a GO term with 1 bit of information would be expected to annotate 50% of genes; and so forth. A set of nodes with consistent information content represents a partition. A set of GO nodes can be used to partition a set of proteins by function or by enrichment level. The algorithm for selecting j partition nodes is illustrated in Supplementary Figure S1. The result is that information is more evenly distributed across a set of GO terms than by using the GO's graphical structure as a proxy for node specificity. We have found that our approach brings the GO partition terms significantly closer to the optimal information content level compared to GO levels (see Figure S1).

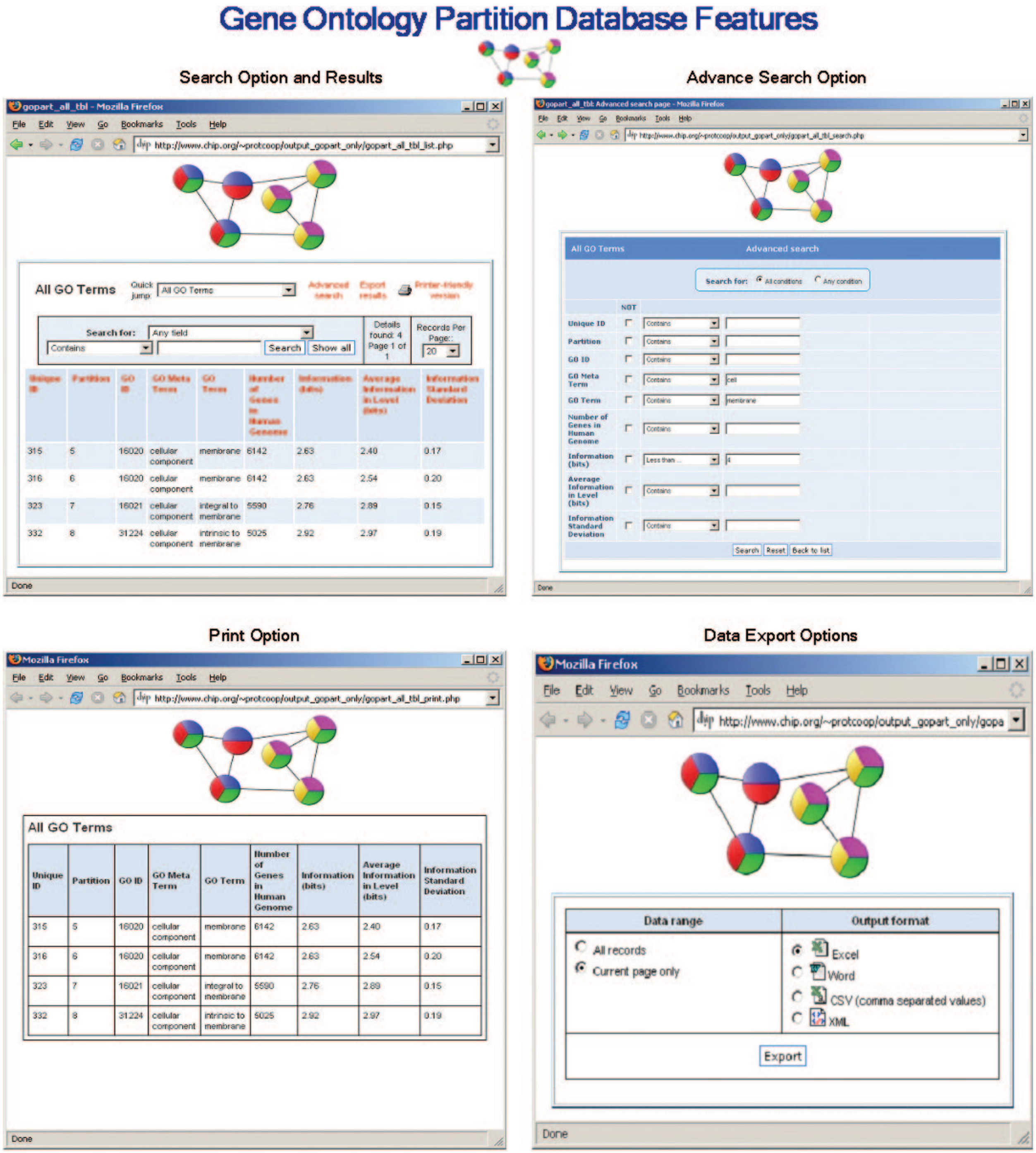
IMPLEMENTATION

The GO Partition Database includes partition information for researchers analyzing data using the entire GO as well as separately for its three main branches. All data for the application is stored in a MySQL database (version 4.1) on a web server with a PHP (version 5.1) interface. Each database schema is created and designed to provide optimum query access. Each GO ID is linked to AmiGO (8) database to provide user annotation and graphical view information of the GO term. Data for the database is generated via MATLAB analysis of 11 genome GO annotations.

The GO Partition Database can be searched using an internet browser (see Figure 2). The search options enable users to input simple as well as complex queries for data access. To facilitate investigations while providing complex search functionality, the database allows users to specify search criteria and ranges from dropdown lists of fields. The advanced search panel allows the user to define specific database queries constructed using Boolean operators connecting multiple fields to fit specific requirements. Each search returns a result list of the matching entries with the partition number, GO ID (linked to AmiGO database), GO term, number of genes, GO term information (in bits), average information content for the associated GO partition and standard deviation information for corresponding GO partition. The user can also print or export query results in multiple formats including Excel, Word and XML. A tutorial (online, supplementary information) has been developed to describe the various features of the database with examples.

APPLICATION

This section presents GO Partition Database applications in gene enrichment. For the functional analysis of human genes, we chose to use a 6-node GO term partition selected within the 'biological process' branch of GO. For optimal analysis, GO term partitions are designed to be organism-specific (see 'Discussion' section). The six GO terms, and their relative placement in the GO graph, are depicted in Figure 3a. The 6-node GO term partition was applied to two human gene sets from the GSEA database (9). Figure 3b contains the proteins from GSEA set 'MAP00190 oxidative phosphorylation' consisting of GenMAPP proteins involved in oxidative phosphorylation. The other GSEA set


Gene Ontology Partition Database Features
Search Option and Results
Advance Search Option
Print Option
Data Export Options
gopart_all_tbl - Mozilla Firefox
gopart_all_tbl: Advanced search page - Mozilla Firefox
Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.chip.org/~protcoop/output_gopart_only/gopart_all_tbl_list.php
http://www.chip.org/~protcoop/output_gopart_only/gopart_all_tbl_search.php
http://www.chip.org/~protcoop/output_gopart_only/gopart_all_tbl_print.php
http://www.chip.org/~protcoop/output_gopart_only/gopa
Done

All GO Terms Quick jump: All GO Terms · Advanced search · Export results · Printer-friendly version

Search for: Any field · Contains · Search · Show all · Details found: 4 Page 1 of 1 · Records Per Page:: 20

Unique ID	Partition	GO ID	GO Meta Term	GO Term	Number of Genes in Human Genome	Information (bits)	Average Information in Level (bits)	Information Standard Deviation
315	5	16020	cellular component	membrane	6142	2.63	2.40	0.17
316	6	16020	cellular component	membrane	6142	2.63	2.54	0.20
323	7	16021	cellular component	integral to membrane	5590	2.76	2.89	0.15
332	8	31224	cellular component	intrinsic to membrane	5025	2.92	2.97	0.19

All GO Terms Advanced search

Search for: ● All conditions ○ Any condition

	NOT		
Unique ID	□	Contains	
Partition	□	Contains	
GO ID	□	Contains	
GO Meta Term	□	Contains	cell
GO Term	□	Contains	membrane
Number of Genes in Human Genome	□	Contains	
Information (bits)	□	Less than ...	4
Average Information in Level (bits)	□	Contains	
Information Standard Deviation	□	Contains	

Search · Reset · Back to list

All GO Terms

Unique ID	Partition	GO ID	GO Meta Term	GO Term	Number of Genes in Human Genome	Information (bits)	Average Information in Level (bits)	Information Standard Deviation
315	5	16020	cellular component	membrane	6142	2.63	2.40	0.17
316	6	16020	cellular component	membrane	6142	2.63	2.54	0.20
323	7	16021	cellular component	integral to membrane	5590	2.76	2.89	0.15
332	8	31224	cellular component	intrinsic to membrane	5025	2.92	2.97	0.19

Data range	Output format
○ All records ● Current page only	● Excel ○ Word ○ CSV (comma separated values) ○ XML

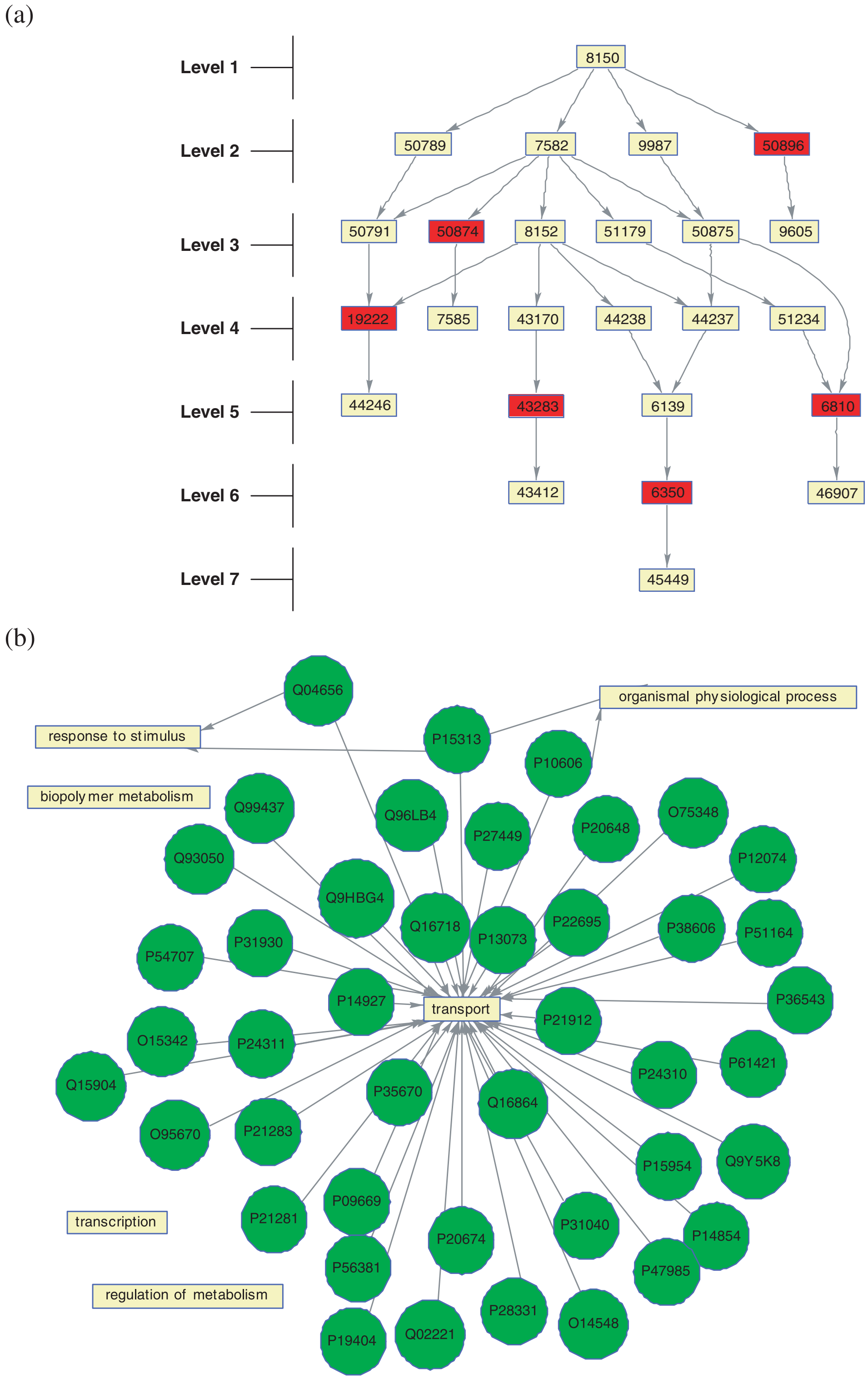
Export

Figure 2. The GO Partition Database has an array of features from customized queries to several export options.

analyzed using GO partitions is the 'HOX_LIST_JP' set, which contains HOX proteins involved in hematopoiesis, and is illustrated in the online tutorial.

From Figure 3b, the enrichment of 'transport' is immediately clear in a visually striking manner. At the same time, other GO terms, such as 'regulation of metabolism', are not represented at all, even though they contain a similar level of information at a genome-wide scope. The enrichment of 'transport' is verified statistically using Fischer's exact test to have a highly significant P-value of 8.55×10^{-33}.

In contrast, using the traditional approach via graphical structure would lead to analysis of functional enrichment with nodes from GO level 2 (as shown in Figure 3a), the closest level with regard to the number of terms. The result using GO level nodes from GO level 2 is shown in Figure 3c and can be compared with the result in Figure 3b using the GO partition method. Several problems with relying on GO level nodes for functional analysis are apparent: the first is the appearance of extremely general nodes such as 'cellular process' and 'physiological process'. These terms are so general that they do not provide much information about the biological processes involved (compared with the much more specific ‘transport’ term found with the 6-node GO partition), and because so many genes share these nodes, the apparent visual enrichment of ‘cellular process’ and ‘physiological process’ is misleading, as it is not statistically significant. Another problem is that these very general nodes are used in the same analysis as inappropriately specific nodes such as ‘viral life cycle’ and ‘pigmentation’. Therefore, the researcher can fall into three traps when relying on GO



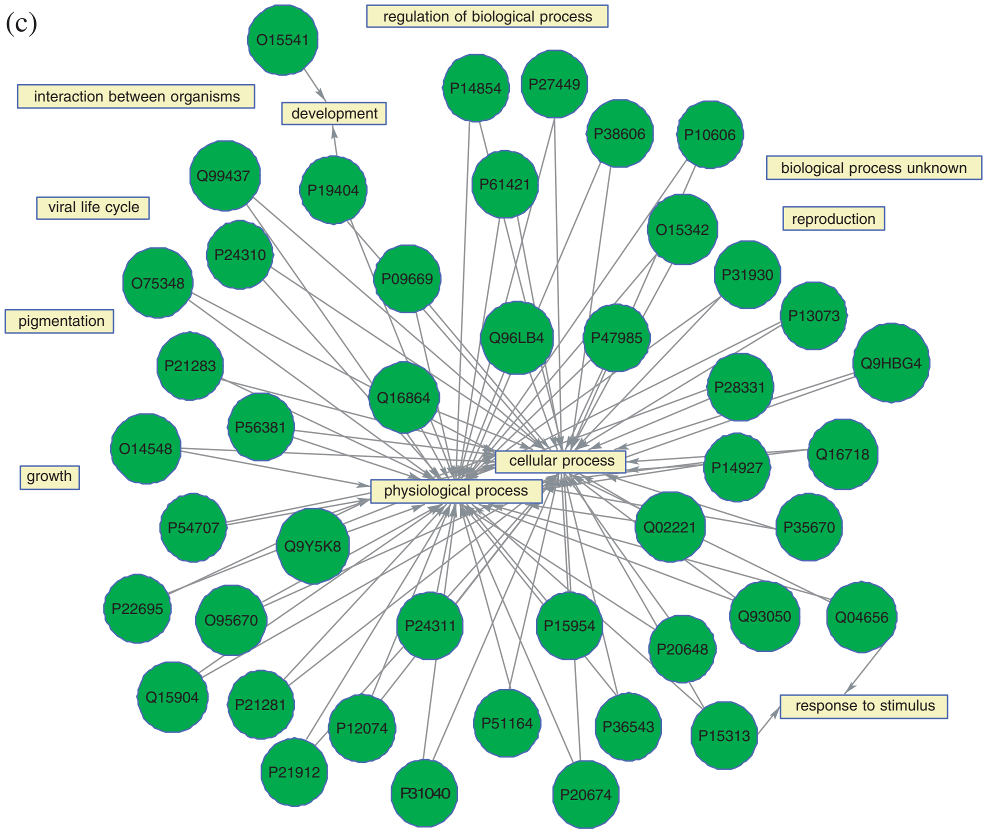


Figure 3. (**a**) GO term partition with six GO terms selected including: regulation of metabolism (19222), response to stimulus (50896), transcription (6350), transport (6810), biopolymer metabolism (43283) and organismal physiological process (50874). (**b**) Visual gene enrichment for transport is evident in these GenMAPP proteins involved in oxidative phosphorylation. Green circles represent proteins (displaying UniProtKB accessions) and rectangles contain the GO terms of the 6-node partition. An arrow going from a protein to a GO term indicates that the protein is annotated by that GO term. (**c**) Visual enrichment is shown based on GO graphical structure—leading to potentially misleading interpretations.

level nodes: some nodes in the analysis may be so general that meaningful, statistically significant enrichment is missed, because too many genes in the background genome are also annotated by those nodes. Second, some nodes may be so specific that no enrichment is found for them, either. Third, the use of nodes that vary widely in information content, as shown in GO level 2, is inconsistent. The researcher, using GO level nodes, might decide to then test additional terms from the next level in the GO hierarchy to get more specific enrichment. However, then the *P*-value would increase due to multiple test correction since more than six tests (the number used in GO Partition Database approach) have to be performed. Thus, significant discoveries made by the GO Partition Database approach can be missed by the traditional methods due to inefficient use of existing information.

DISCUSSION AND CONCLUSION

As illustrated above, the GO Partition Database features sets of GO nodes that offer information consistency for use in functional analysis or visual enrichment, and is a marked improvement over the pitfalls of relying on nodes from a given GO level. Additionally, the GO Partition Database has many other important applications that take advantage of its ability to quantify and compare the specificity of ontology terms. For example, the relatedness of two proteins can be assessed by summing the bitwise information content of GO terms shared by them. This 'information score' can then be compared with that of randomly chosen pairs of proteins to generate a *P*-value. A less mathematical and more visual way to compare pairs of proteins is to assign a color to each term of a GO term partition, and then color proteins based on which terms in a partition annotate it. If the protein is in a node within a protein interaction network, for instance, this approach can be used to show that interacting proteins are more likely to be 'colored' similarly. The protein interaction network involved in the bone morphogenetic pathway (GSEA) on the GO Partition Database home page is colored in this way.

Additionally, the information content of nodes can be used to guide the development of new ontologies. The motivation for GO partitions was due to a lack of correspondence between ontology level and node specificity. As Figure 3a shows, GO terms of similar information content can come from very different levels of GO. However, calculating the information content of ontology terms can help to organize them so that there is greater consistency in information content across each ontology level. Figure 4 shows the uniformity of information content for GO Partition Database relative to a GO level that has the same number of terms.

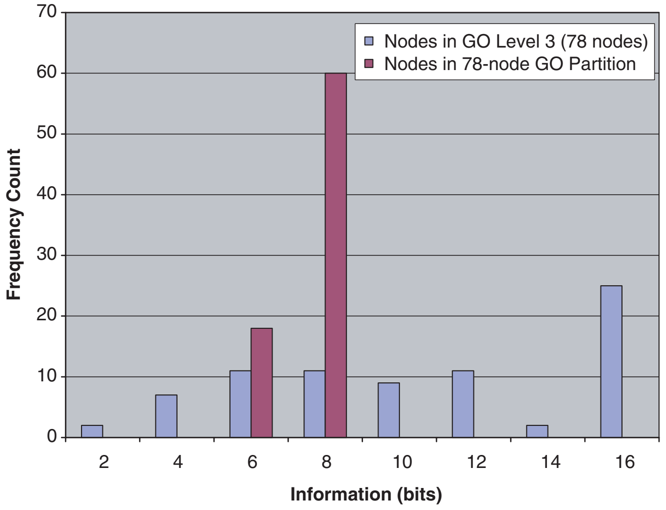


Figure 4. Histogram of GO level 3 versus GO partitions level 3 term information. This shows a tighter distribution for the GO partition-based information compared to that of graphical structure-derived GO level node information.

GO term specificity is also organism specific. Therefore, the composition of GO partitions depends on the context of analysis, i.e. the organism studied. For example, the GO term 'chlorophyll biosynthesis' might not be very informative for plants (for which it is common), but may be highly informative for other organisms. As such, the GO Partition Database includes GO partitions based on GO term specificities within the context of the human genome and of 10 other model or commonly studied organisms, including *Cacnorhabditis elegans, Drosophila melanogaster, Saccharomyces cerevisiae* and *Mus musculus*. Thus, a researcher can access GO partitions suited to a particular organism by simply specifying the organism in the online database.

Information content of GO nodes can also be used to compare and combine related existing ontologies. For instance, GO (10), MIPS (11), YPD (12) and EcoCyc (13) are all biological classification hierarchies, and thus contain overlapping information. The information theoretic approaches used for the GO Partition Database can be used to compare the specificities of terms across ontologies and discover where the ontologies differ or overlap. In addition, such information could be used to create a meta-ontology that combines all terms from related ontologies and to foster ontological portability.

The GO Partition Database empowers researchers with novel ways to do analyses using GO as well as improves existing ones. As our examples have shown, using the information theoretic GO Partition Database, previously obscured information can be utilized to expose significant patterns using fewer statistical tests. Optimization based on GO term information also reduces the need for multiple test corrections. Investigators can thus make significant biological discoveries with less data/tests. The advanced search capabilities of the GO Partition Database allow investigators to do complex Boolean queries in the database without having to learn a query or scripting language. Exporting options allow researchers to easily export to standard formats like XML and to use data in many common applications such as Word and Excel. The GO Partition Database is an important aid to any researcher seeking to determine statistically significant patterns in large-scale genome (or cross-genome) research.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by the National Human Genome Research Institute (NHGRI/NIH) under grant 1R01HG003354-01A2.

REFERENCES

1. The GO Consortium. (2006) The Gene Ontology (GO) project in 2006. *Nucleic Acids Res.*, **34**, D322–D326.
2. Dennis,G., Jr, Sherman,B.T., Hosack,D.A., Yang,J., Gao,W., Lane,H.C. and Lempicki,R.A. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.*, **4**, P3.
3. Al-Shahrour,F., Diaz-Uriarte,R. and Dopazo,J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.
4. Zhou,M. and Cui,Y. (2004) GeneInfoViz: constructing and visualizing gene relation networks. *In Silico Biol.*, **4**, 323–333.
5. MacKay,D.J.C. (2003) *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, Cambridge, UK, New York.
6. Cover,T.M. and Thomas,J.A. (1991) *Elements of Information Theory.* Wiley, New York.
7. Shannon,C.E. (1948) A mathematical theory of communication. *Bell System Technical Journal*, **27**, 623–656.
8. The Gene Ontology Consortium (2006) The Gene Ontology (GO) project in 2006. *Nucleic Acids Res.*, **34**, D322–D326.
9. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
10. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
11. Mewes,H.W., Amid,C., Arnold,R., Frishman,D., Guldener,U., Mannhaupt,G., Munsterkotter,M., Pagel,P., Strack,N., Stumpflen,V. *et al.* (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.*, **32**, D41–D44.
12. Costanzo,M.C., Crawford,M.E., Hirschman,J.E., Kranz,J.E., Olsen,P., Robertson,L.S., Skrzypek,M.S., Braun,B.R., Hopkins,K.L., Kondu,P. *et al.* (2001) YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information. *Nucleic Acids Res.*, **29**, 75–79.
13. Keseler,I.M., Collado-Vides,J., Gama-Castro,S., Ingraham,J., Paley,S., Paulsen,I.T., Peralta-Gil,M. and Karp,P.D. (2005) EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res.*, **33**, D334–D337.
14. Zwier,M.C. and Krueger,B.P. (2003) *Use of Molecular Dynamics Simulations in Analysis of Fluorescence-Detected Resonance Energy Transfer (FRET) Experiments.* Holland, MI.