BIOINFORMATICS APPLICATIONS NOTE doi:10.1093/bioinformatics/bti106

GOChase: correcting errors from gene ontology-based annotations for gene products

Yu Rang Park[1], Chan Hee Park[1], Ju Han Kim[*1,2]

[1]SNUBI: Seoul National University Biomedical Informatics, [2]Human Genome Research Institute, Seoul National University College of Medicine, Seoul 110-799, Korea

Received on April 11, 2004; revised on July 10, 2004; accepted on September 14, 2004
Advanced Access Publication October 28, 2004

ABSTRACT

Summary: The Gene Ontology (GO) is a controlled biological vocabulary that provides three structured networks of terms to describe biological processes, cellular components, and molecular functions. Many databases of gene products are annotated using the GO vocabularies. We found that some GO-updating operations are not easily traceable by the current biological databases and GO browsers. Consequently, numerous annotation errors arise and are propagated throughout biological databases and GO-based high-level analyses. GOChase is a set of web-based utilities to detect and correct the errors in GO-based annotations.

Availability: *http://www.snubi.org/software/GOChase/*

Contact: *juhan@snu.ac.kr*

INTRODUCTION

The Gene Ontology (GO) provides structured controlled biological vocabulary for describing genes and gene products in terms of their associated biological processes, cellular components, and molecular functions (Ashburner *et al.*, 2000). As more and more biological databases are using GO terms to annotate their gene products and many high-level methods analyzing GO annotations are being developed (Dennis *et al.,* 2003; Doniger *et al.,* 2003; Zeeberg *et al.,* 2003), it is essential to provide a mechanism for preventing GO-based annotations from inconsistencies, errors, or error propagations.

The structural foundation of GO is formally a directed acyclic graph (DAG) wherein the terms are equivalent to nodes and the relationships edges of the graph (Aho *et al.,* 1983). The GO consortium provides DAG-Edit for editing GO. Monthly reports (http://www.geneontology.org/MonthlyReports/) are generated by a set of Perl scripts to describe what has happened to the ontologies each month. They report six different types of change that may have happened to a term; 'new terms', 'new obsoletions', 'term name changes', 'new definitions', 'new term merges', and 'term movements'.

We found that the two operations, 'new obsoletions' and 'new term merges', are not easily traceable by the current biological databases and GO browsers and hence cause errors in GO-based annotations. These errors may have already created systematic errors in biological databases, GO browsers, and GO-based high-level data analyses. Table 1 shows the numbers of gene products annotated to invalid GO terms (i.e., 'merged' and 'obsolete') in various databases. It seems evident that the errors are widespread. For a fair comparison, we tested each database version against the corresponding latest GO version that might have been used for the annotation process.

PROGRAM OVERVIEW

Different databases and methods use different GO versions. Without an error-proof mechanism, it is non-trivial to correct the widespread errors and error propagations. Although powerful ontology-management tools are available (Klein *et al.*, 2002; Noy *et al.*, 2002), these 'general-purpose' tools use heuristic algorithms that do not guarantee 100% exact matches. For the purpose of illustration, we applied PromptDiff (Noy *et al.*, 2002) to compare the 2004 January and February versions. PromptDiff correctly detected more than 95% of most of the GO-updating operations. It missed one for "new term" (87 out of 88) and one for "new term merge" (3 out of 4). It exhibited three false positives for 'term name change' (239 calls for 236 true positives) and perfect matches for 'new obsoletion' (60 out of 60). For the 201 'term movement' operations, however, only 24 out of the 246 PromptDiff predictions were correct.

On the other hand, the monthly report generated by the Perl scripts captures all GO-update operations applied to the previous version. Therefore, if we integrate all GO-update information in the monthly reports in sequence, it can serve as the gold standard for GO versioning information. GOChase is a set of web-based utilities available at http://www.snubi.org/software/GOChase/ to detect and correct the possible errors in GO-based annotations. GOChase integrates all monthly reports with major

* To whom correspondence should be addressed

Table 1. Errors in Gene Ontology-based annotations for gene products in selected databases.

Databases	DB Version	GO version*	No. of gene products annotated with GO terms			No. of GO annotations applied to gene products			No. of GO terms used in gene-product annotations		
	mm/dd/yy	mm/dd/yy	Merged term	Obsolete term	Total gene products	Merged term	Obsolete term	Total GO annotations	Merged term	Obsolete term	Total GO terms
NetAffyx[a]	10/03/03	10/01/03	921	7,637	153,369	1,178	15,416	861,349	29	216	6,311
	12/10/03	12/01/03	1,613	8,587	166,651	2,641	16,723	1,149,348	45	230	6,414
1.0	02/08/01	02/01/01	0	154	6,621	0	156	23,089	0	3	1,008
1.1	10/04/01	10/01/01	2	259	7,746	2	275	28,765	1	11	1,216
FANTOM 1.2[b]	12/25/01	12/01/01	3	280	8,031	3	308	30,876	2	12	1,289
2.0	10/04/02	10/01/02	450	2,288	25,130	462	2,709	116,967	27	81	3,034
2.1	11/27/02	11/01/02	528	2,329	25,130	540	2,753	116,967	29	84	3,034
LocusLink[c]	06/02/03	06/01/03	184	1,800	33,118	190	2,184	136,657	22	196	5,256
	08/26/03	08/01/03	124	1,260	33,225	130	1,521	137,114	22	181	5,265
	10/01/03	10/01/03	117	1,233	33,596	123	1,321	141,195	21	162	5,614
	11/09/03	11/01/03	117	1,273	33,731	123	1,363	141,616	21	167	5,648
	12/08/03	12/01/03	110	738	35,329	115	832	155,291	20	129	5,685
	01/24/04	01/01/04	112	795	36,282	117	893	157,734	22	129	5,783
	02/13/04	02/01/04	121	1,707	36,333	135	2,893	157,749	23	163	5,801
SGD[d]	02/26/04	02/01/04	0	0	6,450	0	0	28,865	0	0	2,383
FlyBase	08/29/03	08/01/03	0	0	7,938	0	0	33,153	0	0	3,356
MGI	02/20/04	02/01/04	0	0	12,848	0	0	66,980	0	0	3,382
WormBase	02/04/04	02/01/04	57	407	7,023	57	573	24,700	12	58	1,179
RGD	02/19/04	02/01/04	13	159	3,657	19	366	19,607	7	18	2,412
Gramene	01/05/04	01/01/04	0	5	19,640	0	5	57,376	0	1	966
ZFIN	02/26/04	02/01/04	0	40	1,548	0	84	7,334	0	9	732
DictyBase	02/03/04	02/01/04	18	34	3,459	29	50	9,114	6	13	1,753
TAIR	02/26/04	02/01/04	0	319	23,663	0	348	63,102	0	12	2,341

* The latest GO version for to the database version was used for each comparison for fair analyses.
[a] http://www.affymetrix.com/analysis/index.affx
[b] http://www.gsc.riken.go.jp/e/FANTOM/
[c] http://www.ncbi.nih.gov/LocusLink/
[d] http://www.geneontology.org/GO.current.annotations.shtml

biological databases containing GO annotations (Table 1) and parses them into relational tables, which are then integrated into the GO DB schema (http://www.godatabase.org/dev/database/schema_diagram.html).

GOChase provides four web-based interfaces. (1) GOChase-History resolves the whole evolution history of a GO ID. As an example, the GO term, GO:0006489 (dolichyl-diphosphate biosynthesis), has repeatedly swung back and forth among the seven GO terms (i.e., metabolism, catabolism, biosynthesis, lipid metabolism, protein biosynthesis, protein metabolism, protein modification) by the 16 GO operations in the six updates between March 2001 and August 2003. (2) GOChase-Correct highlights a 'merged term' and redirects it to the correct 'target term' into which the 'merged term' has been merged. For a discarded (or 'obsolete') term, GO consortium provides suggested alternative terms in the comments field of the obsolete term (http://www.geneontology.org/ontology/gene_ontology.obo), which is decided on by a curator. As of May 2004, there are 805 suggested alternative terms for 871 'obsolete' terms. For an obsolete term, GOChase recommends the nearest non-discarded parent term as well as the alternative terms whenever available. The databases that create GO annotations may well find this feature useful to fix the broken hyperlinks for the 'merged' and 'obsolete' terms. (3) A whole database like LocusLink can be input to GOChase in a flat-file format. The annotation errors will be reported with GOChase corrections. (4) When one inputs a GO ID, GOChase will resolve all gene products annotated with the ID across all databases in Table 1. Of course, one can resolve the GO annotations for each gene product.

The annotation errors, i.e. annotations to the merged and obsolete GO terms, may exist in databases simply due to a time lag, as many databases update the annotations only periodically. We learned, however, that certain GO-update processes should be carefully traced to prevent error propagation. An error-conscious mechanism can help GO-based high-level analysis tools like clustering microarray data with GO annotations. Functionalities like showing the evolution history and redirecting to the correct target term may benefit GO Browsers. When a database containing GO annotations is being updated, inconsistencies and errors should be checked against the latest version of GO, for which GOChase can help. Otherwise, the errors may be propagated to the secondary users.

ACKNOWLEDGEMENTS

This study was supported by a grant from Korea Health 21 R&D Project, Ministry of Health & Welfare, Republic of Korea (0405-BC02-0604-0004).

REFERENCES

Aho, A.V., Hopcroft, J.E., and Ullman, J.D. (1983) Directed graphs. In *Data structures and algorithms*, pp. 219-221. Addison-Wesley, Reading, MA.

Ashburner, M., Ball, C.A., *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**(1):25-29.

Dennis, G.Jr., Sherman, B.T., *et al.* (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* **4**(9):R60.

Doniger, S.W., Salomonis, N., *et al.* (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol.* **4**(1):R7.

Klein, M., Kiryakov, D., *et al.* (2002) Ontology versioning and change detection on the web. In *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02)*, Sigüenza, Spain, 2002.

Noy, N.F. and Musen, M.A. (2002) PromptDiff: a fixed point algorithm for comparing ontology versions. In *18th National Conference on Artificial Intelligence (AAAI-2002)*, Edmonton, Canada, 2002.

Zeeberg, B.R., Feng, W., *et al.* (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.* **4**(4):R28.