3712–3715 *Nucleic Acids Research, 2003, Vol. 31, No. 13*
DOI: 10.1093/nar/gkg582

Automated Gene Ontology annotation for anonymous sequence data

Steffen Hennig*, Detlef Groth and Hans Lehrach

Max-Planck Institute for Molecular Genetics, Ihnestrasse 73, D-14195 Berlin, Germany

Received February 16, 2003; Revised and Accepted April 7, 2003

ABSTRACT

Gene Ontology (GO) is the most widely accepted attempt to construct a unified and structured vocabulary for the description of genes and their products in any organism. Annotation by GO terms is performed in most of the current genome projects, which besides generality has the advantage of being very convenient for computer based classification methods. However, direct use of GO in small sequencing projects is not easy, especially for species not commonly represented in public databases. We present a software package (GOblet), which performs annotation based on GO terms for anonymous cDNA or protein sequences. It uses the species independent GO structure and vocabulary together with a series of protein databases collected from various sites, to perform a detailed GO annotation by sequence similarity searches. The sensitivity and the reference protein sets can be selected by the user. GOblet runs automatically and is available as a public service on our web server. The paper also addresses the reliability of automated GO annotations by using a reference set of more than 6000 human proteins. The GOblet server is accessible at http://goblet.molgen.mpg.de.

INTRODUCTION

The last 5 years have seen the completed sequences of several eukaryotic genomes with the human (1,2) and mouse (3) genome as the most prominent examples. Comparative analysis between various species [e.g. yeast and *Caenorhabditis elegans* (4) or human and mouse (5)] has shown an extensive number of orthologous genes with highly conserved function. As an example, the present annotation of human and mouse genes by ENSEMBL (http://www.ensembl.org) shows a clear synteny relationship for almost all genes in both species, and even more distant species like yeast and *C.elegans*, separated by >200 myr of evolution, display a remarkable percentage of orthologous genes on the order of 20–40% (4). Orthology in its original definition means the direct relationship between genes in different species by their descent from a common ancestor and is usually detected by sequence similarity. While sequence similarity search tools (6–9) are well established and can be run in highly automated fashion, functional annotations cannot easily be compared or exchanged between species, since this requires a universal terminology. The Gene Ontology (GO) consortium, since its start a few years ago, has focussed on the development of a structured and universal vocabulary describing the molecular function, biological process and cellular location of gene products (10). Currently GO contains ~9000 terms in total. GO itself provides the species independent vocabulary and hierarchy of terms and several groups have used the vocabulary for annotating various genomes and protein sets (11–14). A comprehensive overview with links to respective addresses can be found at http://www.geneontology.org. There one can also access extensive documentation about data formats and quality codes, which specify how the annotation by GO terms was actually achieved. The hierarchy of GO terms can be regarded as a complex tree structure, with the exception that one GO term can have multiple parent terms. Since the complex hierarchy is difficult to survey several GO browsers have been constructed, e.g. AmiGO, QuickGO etc. (see references at the GO web site), which offer tools for text searching within the GO vocabulary and associated datasets and for graphical display of the hierarchy of target terms. However, to our knowledge direct sequence annotation based on GO terms is not supported by any of the GO related sites. Here we present a public web service (GOblet) which allows annotation of anonymous sequence (cDNA, protein) data based on similarity searches versus a collection of specially designed protein databases.

CONSTRUCTION OF THE ANNOTATION SYSTEM

The way we designed our GO annotation system was mainly inspired by personal experience with the use of GO terms for annotation of in-house EST projects for model organisms like amphioxus or sea urchin. A common problem in these projects is how to compare large transcript libraries represented by ESTs (e.g. from different embryonic stages or different species) with respect to global functional classes like transcription regulation, energy metabolism, in order to find fundamental differences.

*To whom correspondence should be addressed. Tel: +49 30 8413 1612; Fax: +49 30 8413 1380; Email: hennig@molgen.mpg.de

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Published by Oxford University Press 2003

Typically in cases where no GO annotation is available it is imported by sequence similarity searches against data sets with existing links to GO terms. This procedure generates a wealth of information of high specificity, which is not always convenient for a more general classification. Here the well defined hierarchical structure of GO is an excellent resource, since all parent terms for a specific GO-Id can be traced up to the more general ontology classes like binding, enzyme, transcription regulation, cell communication, which are more suitable for a survey annotation of large data sets like whole cDNA libraries or complete genomes (1,4,15). An optimal automated annotation system should therefore use a broad data set of protein and gene sequences connected with GO terms and it must contain parsers, which allow effective screening of the GO hierarchy up to any level of specificity.

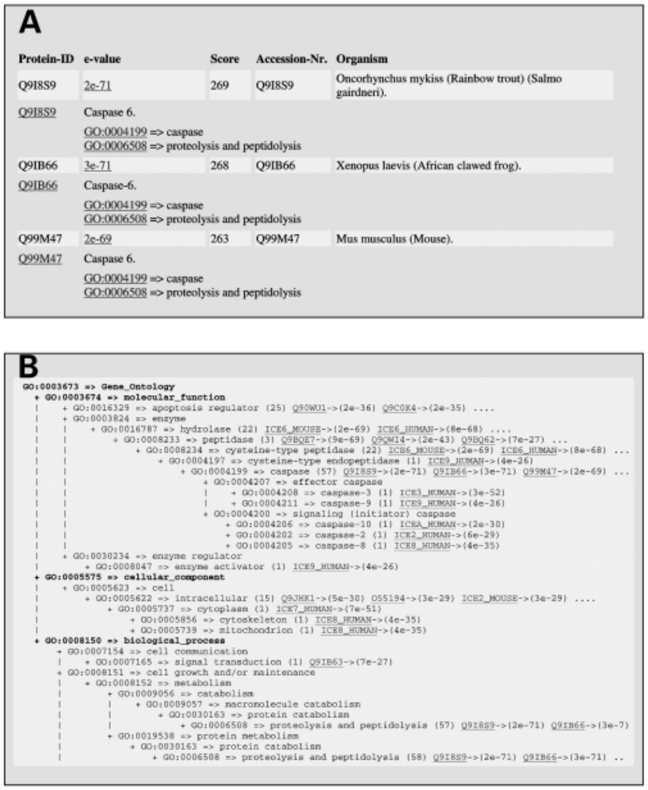
While the vast majority of data sets published on the GO web site has evidence code IEA (inferred from electronic annotation), which normally means that the annotation was based on sequence similarity searches without inspection by a curator, there is also a significant amount with more confident evidence codes (indicating that annotation was controlled by a curator). Especially the Gene Ontology annotations (GOA) for yeast (11), *C.elegans* (16), *Drosophila* (17), mouse (18) and the human GOA (www.ebi.ac.uk/GOA/) maintained by the EBI have a large number of high-confidence evidence codes. The largest GOA set (as of February 2003) covers 566 342 protein IDs from SWISS-PROT (19) and TrEMBL from almost 50 000 taxa and is also provided by the EBI, but here the majority of entries only has evidence code IEA. We imported all these data sets into local protein databases (1 per GOA set), which contain the respective GO terms, and made them accessible by a local BLAST (6,7) server. Similarity searches can now be performed with any level of sensitivity, with DNA or protein sequences as query. The protein databases are regularly updated and build the core of the annotation system. Once a BLAST run is finished the relevant GO terms are extracted from the BLAST output files together with the functional description of the respective database proteins (Fig. 1A).

The complete gene ontology (i.e. the hierarchy of GO identifiers and their description) is available in various formats from the GO consortium (www.geneontology.org). Since we wanted to set up a fully integrated local analysis system we developed a GO parser, which maps the GO hierarchy onto a set of linearised trees, with terms and nodes connected by hash tables, so that any partial hierarchy (starting backwards from a single GO-Id) can be easily reconstructed. For each query sequence the complete set of relevant GO-Ids (obtained from the BLAST output) is used then for construction of a summary tree (Fig. 1B), that lists all the single proteins leading to a specific leaf of the tree. Furthermore, the total counts per GO-Id are given, which allows easy identification of the most significant GO terms.

All the procedures described above are combined in a package of perl scripts. The web server handles the queries via Perl-CGI modules.

USING THE GOblet SYSTEM

Our web server is publicly accessible under http://goblet.molgen.mpg.de. On the front page the available protein databases are listed and described. All of them were constructed by using the GOA information made available by the GO consortium: protein identifiers were used to download the respective protein sequences from their source databases (SWISS-PROT/TrEMBL, ENSEMBL, Flybase, Wormbase, SGD) and the sequences were connected with the respective GO identifiers.



Protein-ID	e-value	Score	Accession-Nr.	Organism
Q9I8S9	2e-71	269	Q9I8S9	Oncorhynchus mykiss (Rainbow trout) (Salmo gairdneri).
Q9I8S9	Caspase 6. GO:0004199 => caspase GO:0006508 => proteolysis and peptidolysis			
Q9IB66	3e-71	268	Q9IB66	Xenopus laevis (African clawed frog).
Q9IB66	Caspase-6. GO:0004199 => caspase GO:0006508 => proteolysis and peptidolysis			
Q99M47	2e-69	263	Q99M47	Mus musculus (Mouse).
Q99M47	Caspase 6. GO:0004199 => caspase GO:0006508 => proteolysis and peptidolysis			

```
GO:0003673 => Gene_Ontology
 + GO:0003674 => molecular_function
 |    + GO:0016329 => apoptosis regulator (25) Q90WU1->(2e-36) Q9C0K4->(2e-35) ....
 |    + GO:0003824 => enzyme
 |    |    + GO:0016787 => hydrolase (22) ICE6_MOUSE->(2e-69) ICE6_HUMAN->(8e-68) ....
 |    |        + GO:0008233 => peptidase (3) Q9BQE7->(9e-69) Q9QWI4->(2e-43) Q9BQ62->(7e-27) ...
 |    |            + GO:0008234 => cysteine-type peptidase (22) ICE6_MOUSE->(2e-69) ICE6_HUMAN->(8e-68) ...
 |    |                + GO:0004197 => cysteine-type endopeptidase (1) ICE9_HUMAN->(4e-26)
 |    |                    + GO:0004199 => caspase (57) Q9I8S9->(2e-71) Q9IB66->(3e-71) Q99M47->(2e-69) ...
 |    |                        + GO:0004207 => effector caspase
 |    |                        |    + GO:0004208 => caspase-3 (1) ICE3_HUMAN->(3e-52)
 |    |                        |    + GO:0004211 => caspase-9 (1) ICE9_HUMAN->(4e-26)
 |    |                        + GO:0004200 => signaling (initiator) caspase
 |    |                            + GO:0004206 => caspase-10 (1) ICEA_HUMAN->(2e-30)
 |    |                            + GO:0004202 => caspase-2 (1) ICE2_HUMAN->(6e-29)
 |    |                            + GO:0004205 => caspase-8 (1) ICE8_HUMAN->(4e-35)
 |    + GO:0030234 => enzyme regulator
 |        + GO:0008047 => enzyme activator (1) ICE9_HUMAN->(4e-26)
 + GO:0005575 => cellular_component
 |    + GO:0005623 => cell
 |        + GO:0005622 => intracellular (15) Q9JHK1->(5e-30) O55194->(3e-29) ICE2_MOUSE->(3e-29) ....
 |            + GO:0005737 => cytoplasm (1) ICE7_HUMAN->(7e-51)
 |                + GO:0005856 => cytoskeleton (1) ICE8_HUMAN->(4e-35)
 |                + GO:0005739 => mitochondrion (1) ICE8_HUMAN->(4e-35)
 + GO:0008150 => biological_process
     + GO:0007154 => cell communication
     |    + GO:0007165 => signal transduction (1) Q9IB63->(7e-27)
     + GO:0008151 => cell growth and/or maintenance
     |    + GO:0008152 => metabolism
     |        + GO:0009056 => catabolism
     |        |    + GO:0009057 => macromolecule catabolism
     |        |        + GO:0030163 => protein catabolism
     |        |            + GO:0006508 => proteolysis and peptidolysis (57) Q9I8S9->(2e-71) Q9IB66->(3e-7)
     |        + GO:0019538 => protein metabolism
     |            + GO:0030163 => protein catabolism
     |                + GO:0006508 => proteolysis and peptidolysis (58) Q9I8S9->(2e-71) Q9IB66->(3e-71) ..
```

Figure 1. (**A** and **B**) Excerpts of a GOblet result web page for an Amphioxus RNA for caspase-6. Note that in the figure the original output is truncated for easier display. (A) Upper part of result page. The protein matches are shown in the order of their significance. Links to external databases and to the BLAST alignments are provided and the GO-Ids associated with the respective target protein are displayed. (B) Bottom of result page. All GO-Ids positive with the query sequence are condensed into a summary tree. Contributions of single database proteins are displayed. The numbers in brackets give the amount of distinct protein contributions for that branch.

The input of the query sequence is simply performed by cut and paste, then the user has to specify the type of query (DNA or protein), the target database from a popup menu and the threshold (cut-off e-value). Once the job has started a unique URL is created on the fly, which can be bookmarked to find the results later on. Although at the moment we allow only one query sequence per search during a session many searches can be started and will be accessible by their unique URL. Currently result files are stored for at least 1 week on our system.

At the top of the main result page the query sequence is displayed again (which is useful if a user has started several jobs) and the hits are tabulated sorted by significance. In addition, the description of the target protein as given by the respective source database and the species is shown, and there are links to the original BLAST output file as well as to source documents in SWISS-PROT, FlyBase, ENSEMBL, etc. For each hit the complete list of GO-Ids associated with it is shown as well and links to the QuickGO browser at the EBI are provided. Figure 1A gives an excerpt of a GOblet run with an amphioxus (*Branchiostoma floridae*) mRNA for the CASP-6 gene, which is an apoptosis-related cysteine protease (at least in *Homo sapiens* and higher vertebrates). Several homologs in other species were found, which indicates the high degree of conservation.

At the bottom of the result page a summary tree is drawn constructed from all distinct GO terms as they were imported from the matching proteins. To save space the tree displayed in Figure 1B is truncated. The advantage of our summary presentation is that the most significant branches are easily detected and the contributing proteins are listed as well, so that any single piece of information can easily be inspected.

TESTING THE VALIDITY OF GO-BASED ANNOTATIONS

Although orthology between genes from different species is frequently detected a central question is, how far GO terms derived for a specific organism [e.g. *Drosophila* (17), *C.elegans* (16)] can be used for annotating distant species like amphioxus (a project we are actually working on) or *H.sapiens*. Although a detailed inspection of the correlation between sequence similarity and protein function is beyond the scope of this article at least a rough estimate can be tried for protein sets with existing GO-based annotations. We made explicit use of the evidence codes (attached to all data sets published by the GO consortium) by using a reference set of 6544 human proteins in SWISS-PROT/TrEMBL, where the GO annotation was checked and verified by a curator, as reflected by the evidence code TAS (traceable author statement). The reference set was then annotated alternatively by running sequence similarity searches (BLAST, e-value <e-20) against (a) GOA sets of mammalian proteins, with all human proteins excluded; and (b) non-mammalian (mainly invertebrate) proteins, so that finally the reference GO annotation could be compared to two independent GO annotations. The aim of our analysis was to get a global picture as to what extent annotations from one set are reproduced by other sets. In Figure 2 the results for two branches of the class ‘biological process’ are depicted. There are pronounced differences in some cases (e.g. ‘cell–cell signaling’ and ‘cell proliferation’) but the shape of the distributions is quite similar in all three sets. Qualitatively the same results are found in all major branches of the GO hierarchy. We therefore conclude that GO-based annotation of sequence data in the majority of cases will give the correct result. Nevertheless, the pure electronic annotations as done by our GOblet system have to be taken as tentative, but can inspire interpretation of experimental results as, for example, gene expression studies, *in situ* hybridisations, protein–protein interactions.

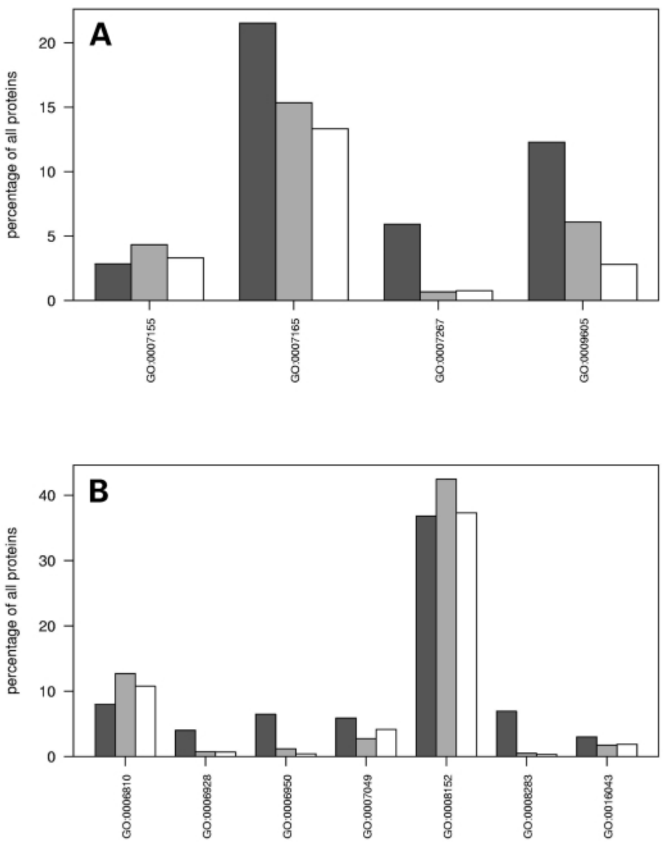


Figure 2. Comparison of three alternative GO annotations (GOA) for the same reference set of 6544 human proteins: (dark grey) original annotation (EBI) with high evidence code (TAS) only, (light grey) GOA imported by sequence similarity from a set of mammalian, non-human proteins, (white) GOA imported from a purely non-mammalian protein set. Given are the percentages of all proteins in the query set matching GO terms of the respective classes. (**A**) Biological process: cell communication: GO:0007155: cell adhesion, GO:0007165: signal transduction, GO:0007267: cell–cell signaling, GO:0009605: response to external stimulus. (**B**) Biological process: cell growth/maintenance: GO:0006810: transport, GO:0006928: cell motility, GO:0006950: response to stress, GO:0007049: cell cycle, GO:0008152: metabolism, GO:0008283: cell proliferation, GO:0016043: cell organization and biogenesis.

ACKNOWLEDGEMENTS

S.H and D.G. thank the BMBF (German Ministry for Education and Research) for funding in the frame of the National Genome Research Network (NGFN).

REFERENCES

1. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
2. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
3. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

4. Chervitz,S.A., Aravind,L., Sherlock,G., Ball,C.A., Koonin,E.V., Dwight,S.S., Harris,M.A., Dolinski,K., Mohr,S., Smith,T. *et al.* (1998) Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science*, **282**, 2022–2028.
5. Mural,R.J., Adams,M.D., Myers,E.W., Smith,H.O., Miklos,G.L., Wides,R., Halpern,A., Li,P.W., Sutton,G.G., Nadeau,J. *et al.* (2002) A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science*, **296**, 1661–1671.
6. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
7. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
8. Korf,I. and Gish,W. (2000) MPBLAST: improved BLAST performance with multiplexed queries. *Bioinformatics*, **16**, 1052–1053.
9. Pearson,W.R. (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.*, **132**, 185–219.
10. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
11. Dwight,S.S., Harris,M.A., Dolinski,K., Ball,C.A., Binkley,G., Christie,K.R., Fisk,D.G., Issel-Tarver,L., Schroeder,M., Sherlock,G. *et al.* (2002) Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.*, **30**, 69–72.
12. Bono,H., Kasukawa,T., Furuno,M., Hayashizaki,Y. and Okazaki,Y. (2002) FANTOM DB: database of Functional Annotation of RIKEN Mouse cDNA Clones. *Nucleic Acids Res.*, **30**, 116–118.
13. GO-Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res.*, **11**, 1425–1433.
14. Xie,H., Wasserman,A., Levine,Z., Novik,A., Grebinskiy,V., Shoshan,A. and Mintz,L. (2002) Large-scale protein annotation through gene ontology. *Genome Res.*, **12**, 785–794.
15. Schug,J., Diskin,S., Mazzarelli,J., Brunk,B.P. and Stoeckert,C.J., Jr (2002) Predicting gene ontology functions from ProDom and CDD protein domains. *Genome Res.*, **12**, 648–655.
16. Harris,T.W., Lee,R., Schwarz,E., Bradnam,K., Lawson,D., Chen,W., Blasier,D., Kenny,E., Cunningham,F., Kishore,R. *et al.* (2003) WormBase: a cross-species database for comparative genomics. *Nucleic Acids Res.*, **31**, 133–137.
17. FlyBase-consortium. (2003) The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res.*, **31**, 172–175.
18. Hill,D.P., Davis,A.P., Richardson,J.E., Corradi,J.P., Ringwald,M., Eppig,J.T. and Blake,J.A. (2001) Program description: strategies for biological annotation of mammalian systems: implementing gene ontologies in mouse genome informatics. *Genomics*, **74**, 121–128.
19. Bairoch,A. and Apweiler,R. (1997) The SWISS-PROT protein sequence database: its relevance to human molecular medical research. *J. Mol. Med.*, **75**, 312–316.