Bioinformatics and Biology Insights

Libertas Academica
FREEDOM TO RESEARCH

OPEN ACCESS
Full open access to this and thousands of other papers at http://www.la-press.com.

ORIGINAL RESEARCH

Mining Gene Ontology Data with AGENDA

Guvanch Ovezmyradov, Qianhao Lu and Martin C. Göpfert

Department of Cellular Neurobiology, Georg-August-University of Göttingen, Schwann-Schleiden Research Centre for Molecular Cell Biology, Julia-Lermontowa-Weg 3, 37077 Göttingen, Germany.
Corresponding author email: govezmu@gwdg.de

Abstract: The Gene Ontology (GO) initiative is a collaborative effort that uses controlled vocabularies for annotating genetic information. We here present AGENDA (Application for mining Gene Ontology Data), a novel web-based tool for accessing the GO database. AGENDA allows the user to simultaneously retrieve and compare gene lists linked to different GO terms in diverse species using batch queries, facilitating comparative approaches to genetic information. The web-based application offers diverse search options and allows the user to bookmark, visualize, and download the results. AGENDA is an open source web-based application that is freely available for non-commercial use at the project homepage. URL: http://sourceforge.net/projects/bioagenda.

Keywords: Gene Ontology, gene annotation, controlled vocabulary, data mining, complex query

Bioinformatics and Biology Insights 2012:6 63–67

doi: 10.4137/BBI.S9101

This article is available from http://www.la-press.com.

© the author(s), publisher and licensee Libertas Academica Ltd.

This is an open access article. Unrestricted non-commercial use is permitted provided the original work is properly cited.

Introduction

The emergence of novel genetic techniques and the exponential accumulation of genomic data have increased the need for bioinformatics tools.[1,2] Biological ontologies facilitate the handling of complex biological data and contribute to the interoperability across multiple data sources.[3,4] The Gene Ontology (GO) database summarizes information about the molecular functions, cellular components, and biological processes related to gene products.[5] Many tools have been created to search, browse, and analyze the GO database.[6] Many of these tools accept only a single gene or GO term as an input, hampering systematic comparisons between GO annotations associated with different GO terms and genes: Complex biological questions that, for example, involve more than one biological process or molecular function cannot be addressed if only one GO term is considered. Similarly, when elucidating a certain biological mechanism, sets of genes rather than single genes are often the focus, raising the need to simultaneously access GO associations of multiple genes. Another limitation in accessing the GO database is that while most programs (eg, EasyGO,[7] GOstat,[8] Onto-Express[9]) produce a short list of significantly enriched GO terms,[10,11] they do not allow to query particular GO terms independent of enrichment, which might be of interest if one wants to know which of the genes that are linked to one GO term are associated with a second, user-defined term.

Here we present AGENDA (Application for mining Gene Ontology data), a novel web-based application for comparing GO annotations associated with multiple GO terms in different species. The program allows for complex queries using GO Slims[17] and Boolean operators. Unlike the programs listed above, with AGENDA it is possible to analyze genes that are annotated to a certain GO term that is defined not by enrichment but by the user. Moreover, using AGENDA, evidences for each annotation can be accessed and the results of the analysis are visualized and can be exported. The usefulness of Boolean operators for mining the GO database had been previously acknowledged.[12] Using Boolean operators to refine queries in a step-by-step manner, AGENDA allows to access the GO database in a more flexible manner than was possible before. By combining GO Slims with Boolean Operators, AGENDA facilitates complex queries of GO data in a step-by-step manner whereby the results of each step can be used principally as a starting point for follow-up steps.

Methods

AGENDA is a web-based application developed using Apache web server and server-side scripting that employs complex SQL queries. The content data of the internal MySQL server is obtained from the GO database archive.[13] HTML pages are dynamically created with PHP and CSS, and supported with JavaScript for the user interactivity. The platform-independent program was successfully tested for cross-browser compatibility on common web browsers. The charts are created dynamically using Google Chart Tools[14] and results of queries are downloadable as CSV files. The application is accessible at http://bioagenda.uni-goettingen.de. The source code and the documentation are freely available under the GNU GPL license for download from the website http://sourceforge.net/projects/bioagenda.

Results

AGENDA offers simple and advanced modes of retrieving GO information that are described below. 12 organisms are supported: *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Danio rerio*, *Dictyostelium discoideum*, *Drosophila melanogaster*, *Escherichia coli*, *Gallus gallus*, *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. The genomes of all these organisms are annotated within the ongoing Gene Ontology's Reference Genome Project.[15] The interface is designed to be intuitive, allowing for an easy navigation to enable convenient data mining. AGENDA provides user-friendly internal and external links for accessing the target data. It is possible to filter the query results by choosing between GO evidence codes such as "IMP" ("Inferred from Mutant Phenotype) and "ISS" ("Inferred from Sequence or Structural Similarity"). Query information can be stored using the URL address, and gene lists can be exported as CSV files. AGENDA also provides access to the homology data generated by the Gene Ontology's Reference Genome Project.[15] Output data obtained with one query can be reused as the input for subsequent queries, allowing to refine searches step-by-step.

Apart from simple queries that focus on only one GO term or gene, two types of batch queries are supported: First, different, user-defined GO terms can be simultaneously queried using the GO Slimmer, a method that uses parent-child relationships between GO terms to compare gene lists of interest with lists that are annotated to GO terms. GO Slimmer identifies overlap between these lists and produces "GO Slims" that quantify the overlap for each GO term. In AGENDA, gene lists of interest are always related with certain GO terms, but GO Slims can be also produced if different gene sets of interests, such as whole genomes of specific organisms, are used as query input.[16,17] Second, queries of different user-defined GO terms can be combined via Boolean operators (AND, OR, NOT). Each query option is represented by a separate web-page in the program. Data from one page can fully be transferred to another so that different types of queries can be linked. Web-pages of the program include input fields, output fields, and charts for visualization of the results.

GO terms can be queried in AGENDA using accession numbers, names or synonyms (if any). When querying apoptotic proteins, for example, "GO:0006915" (accession number), "apoptotic process" (name), or "apoptosis" (synonym) can be typed in as the input. In a similar manner, a gene product can be queried using its symbol, full name or synonyms (if any). For example, "TP53" (symbol), "Cellular tumor antigen p53" (full name) and "P53" (synonym) are all accepted when querying human TP53. A detailed user guide describing this query expansion and other features of AGENDA is available as a web page (http://bioagenda.uni-goettingen.de/userguide.php).

Case study

Many forms of cancer arise from alterations in apoptosis[18] The Gene Ontology database can, for example, be used to find out which genes are implicated in apoptosis (GO:0006915) in humans, and which of the respective gene products localize to mitochondria (GO:0005739), the nucleus (GO:0005634), and the plasma membrane (GO:0005886). Using simple queries only, the cellular localizations of each of the 1771 human genes that are associated with apoptosis would need to be accessed individually. Using the GO Slimmer page of AGENDA, all these GO terms can be simultaneously accessed and the respective information can be obtained with a single mouse click (Fig. 1). Using Boolean queries, in turn, it is, for example, possible to assess which of the human apoptosis genes are associated with mitochondria or the nucleus but not the plasma membrane, linking in one query all the three Boolean operators to delineate genes. By simply exchanging the name of the species, genes of eg, zebrafish or *Drosophila* that satisfy the same query can be displayed. Details about each of the identified genes can be found by clicking on the gene's name: this opens a simple query page for the gene, which includes information about all its GO annotations and links to the supporting evidence.

Discussion

We have presented a novel tool for accessing and mining GO data. While simple search options are similar to the standard services provided by the AmiGO browser,[19] AGENDA employs a new interface for performing complex queries that include different GO terms and species. The GO content of AGENDA is updated regularly using MySQL dump files that are downloaded manually from the GO database archive.[13] To synchronize AGENDA with the latest GO database releases, we plan to implement an automated update. AGENDA undergoes active development to suit the needs of the research community. For example, AGENDA currently does not support the upload and analysis of user-specified gene lists. Our future goal is to enable the uploading of such files for in-depth analysis. Further perspective of development includes the expansion of the queries to all species in the GO database and the implementation of AJAX (Asynchronous JavaScript and XML) to further simplify the usage of AGENDA. AGENDA is an open source application that is freely available for non-commercial use. As the size and value of Gene Ontology is growing steadily together with our understanding of cellular mechanisms, the impact of tools for browsing and mining GO data becomes more apparent. AGENDA complements the existing bioinformatics tools for mining the GO database and provides new functions for accessing GO information.

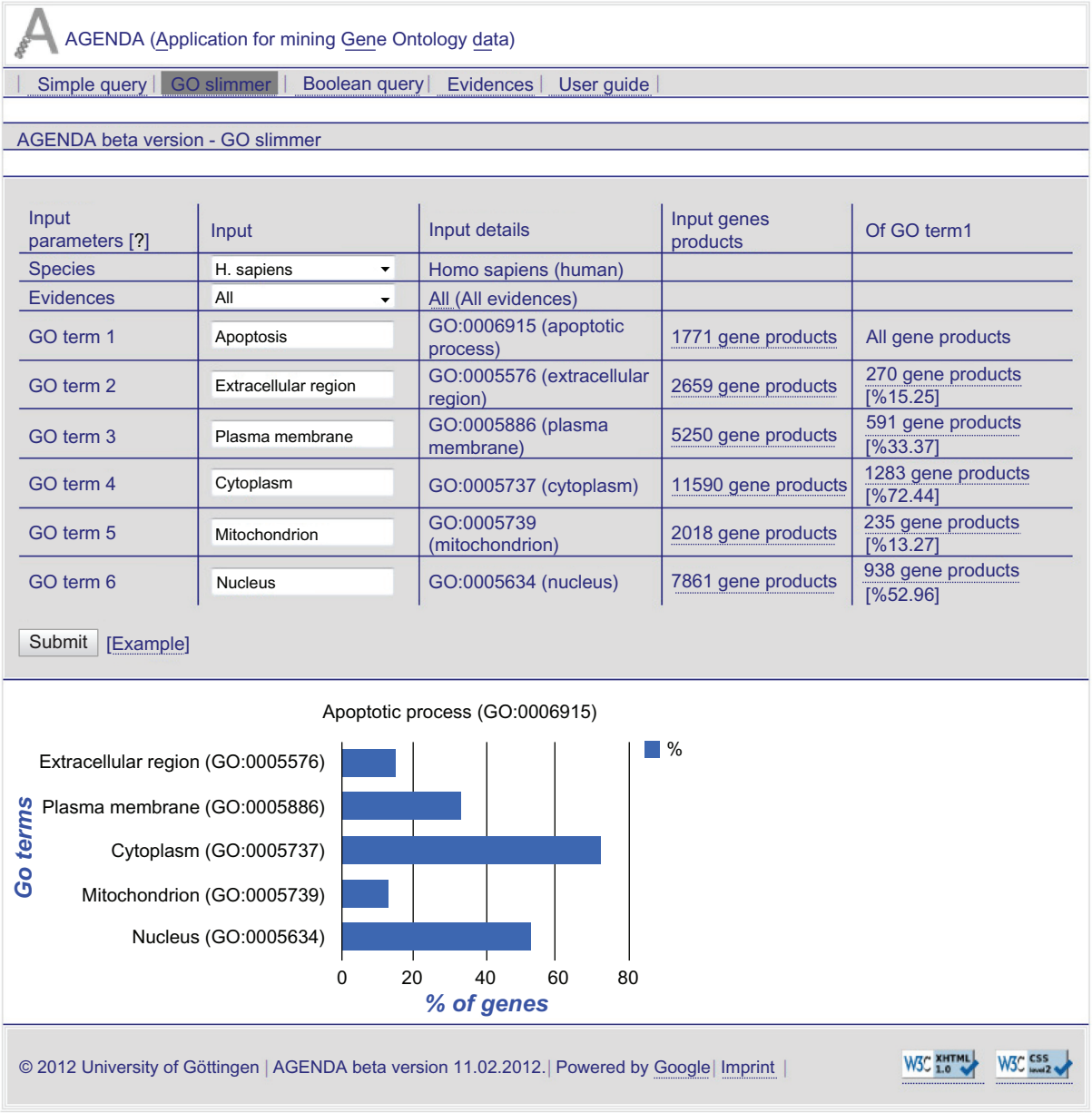

Figure 1. Screenshot of the "GO Slimmer" option in AGENDA. The screenshot displays the cellular localizations of human genes that are implicated in apoptosis.
Notes: 1771 genes are identified that are associated with apoptosis (GO:0006915). 235 of the respective gene products localize, for example, to the mitochondrion (GO:0005739), 591 to the plasma membrane (GO:0005886), and 938 to the nucleus (GO:0005634).

Author Contributions

Conceived and designed the experiments: GO, MCG. Analysed the data: GO, QL. Wrote the first draft of the manuscript: GO. Contributed to the writing of the manuscript: QL, MCG. Agree with manuscript results and conclusions: GO, QL, MCG. Jointly developed the structure and arguments for the paper: GO, QL, MCG. Made critical revisions and approved final version: GO, QL, MCG. All authors reviewed and approved of the final manuscript.

Funding

This work was supported by the NRW IGS GFG (International Graduate School in Genetics and Functional Genomics) (to G.O.) and the DFG (Deutsche Forschungsgemeinschaft) (Go 1092/1-1) (to M.C.G.).

Acknowledgments

We thank the GWDG (Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen) for technical support. We acknowledge the Gene Ontology

Consortium as the source of Gene Ontology data and Google for providing its Charts API infrastructure.

Disclosures and Ethics

As a requirement of publication author(s) have provided to the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.

References

1. Kumar S, Dudley J. Bioinformatics software for biologists in the genomics era. *Bioinformatics*. 2007;23(14):1713–7.
2. Baxevanis AD. The importance of biological databases in biological discovery. *Curr Protoc Bioinformatics*. Chapter 1:Unit 1.1. 2009.
3. Bard JB, Rhee SY. Ontologies in biology: design, applications and future challenges. *Nature Reviews Genetics*. 2004;5(3):213–22.
4. Mi H, Thomas PD. Ontologies and standards in bioscience research: for machine or for human. *Front Physiol*. 2011;2:5.
5. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*. 2000;25(1):25–9.
6. Gene Ontology Tools. Available at: http://www.geneontology.org/GO.tools.shtml. Accessed February 14, 2012.
7. Zhou X, Su Z. EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species. *BMC Genomics*. 2007;8:246.
8. Beissbarth T, Speed TP. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*. 2004;20(9):1464–5.
9. Draghici S, Khatri P, Bhavsar P, Shah A, Krawetz SA, Tainsky MA. Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res*. 2003;31 (13):3775–81.
10. Beissbarth T. Interpreting experimental results using gene ontologies. *Methods Enzymol*. 2006;411:340–52.
11. van den Berg BH, Thanthiriwatte C, Manda P, Bridges SM. Comparing gene annotation enrichment tools for functional modeling of agricultural microarray data. *BMC Bioinformatics*. 2009;10 Suppl 11:S9.
12. Berriz GF, White JV, King OD, Roth FP. GoFish finds genes with combinations of Gene Ontology attributes. *Bioinformatic*. 2003;19(6):788–9.
13. Gene Ontology Database Archive. Available at: http://archive.geneontology.org/latest-full/. Accessed February 14, 2012.
14. Google Chart Tools. Available at: http://code.google.com/intl/de-DE/apis/chart/. Accessed February 14, 2012.
15. Reference Genome Group of the Gene Ontology Consortium. The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. *PLoS Computational Biology*. 2009; 5(7):e1000431.
16. Berardini TZ, Mundodi S, Reiser L, et al. Functional annotation of the Arabidopsis genome using controlled vocabularies. *Plant Physiol*. 2004;135(2):745–55.
17. GO Slim and Subset Guide. Available at: http://www.geneontology.org/GO.slims.shtml. Accessed February 14, 2012.
18. Elmore S. Apoptosis: a review of programmed cell death. *Toxicologic Pathology*. 2007;35(4):495–516.
19. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S; AmiGO Hub. Web Presence Working Group. AmiGO: online access to ontology and annotation data. *Bioinformatics*. 2009;25(2):288–9.