BIOINFORMATION
Discovery at the interface of physical and biological sciences

Open access

BIOMEDICAL INFORMATICS

www.bioinformation.net
Volume 12(3)

Web Server

Genes2GO: A web application for querying gene sets for specific GO terms

Konika Chawla & Martin Kuiper

Department of Biology, Norwegian University of Science and Technology (NTNU), Høgskoleringen 5, 7491 Trondheim, Norway, Konika Chawla - Email: konika.chawla@ntnu.no

Received May 30, 2016; Accepted June 10, 2016; Published June 21, 2016

Abstract
Gene ontology annotations have become an essential resource for biological interpretations of experimental findings. The process of gathering basic annotation information in tables that link gene sets with specific gene ontology terms can be cumbersome, in particular if it requires above average computer skills or bioinformatics expertise. We have therefore developed Genes2GO, an intuitive R-based web application. Genes2GO uses the biomaRt package of Bioconductor in order to retrieve custom sets of gene ontology annotations for any list of genes from organisms covered by the Ensembl database. Genes2GO produces a binary matrix file, indicating for each gene the presence or absence of specific annotations for a gene. It should be noted that other GO tools do not offer this user-friendly access to annotations.

Availability: Genes2GO is freely available and listed under **http://www.semantic-systems-biology.org/tools/externaltools/**

Keywords: gene sets, GO terms, web server

Background

The interpretation of experimental results from omics experiments often involves the analysis of GO functional annotations. The use of GO terms for analysis of gene lists is in fact an area where considerable intuitive support is already provided, especially concerning the assessment of functional representation within experimental gene set results. GO terms and gene association files containing GO annotation are maintained by the Gene Ontology (GO) Consortium and can be downloaded from http://www.geneontology.org/ **[1]**, whereas GO annotations for single genes or sets of genes can be fetched from various other websites and databases including Amigo **[2]** Ensembl Biomart **[3]**, QuickGO **[4]**, David **[5]**, and GoParGenPy **[6]**. These sources provide comprehensive GO annotation coverage and querying them can provide thousands of annotations for thousands of genes. Subsequently, analysing a result file may require arduous processing and additional parsing to retrieve annotations for specific gene lists with custom sets of GO terms. To provide more flexibility to users we have developed Genes2GO, a flexible, intuitive web tool for fetching custom sets of GO terms for custom sets of genes.

Table 1 shows an overview of popular available GO annotation tools compared with Genes2GO for a number of specific properties, including query input, export formats, annotation file updating, and graphical user interface. Whereas bioinformaticians may use R programming skills in combination with Bioconductor packages such as biomaRt **[7]**, Genes2GO requires no additional programming. Genes2GO is developed for biologists to provide a user-friendly application for creating custom gene annotation matrices in a standard and reusable format, including tab-delimited files that can be further analysed as spreadsheets.

Methodology:

Implementation:

Genes2GO is a web based application implemented in R and constructed with Rwui **[8]**, a web application for running R scripts. Genes2GO uses the biomaRt package of Bioconductor to access the current version of the ENSEMBL GENES (SANGER UK) **[9]** for gene annotations. It uses a table of GO IDs and their term descriptions to fetch the GO IDs matching user-submitted keywords such as 'transcription factor'. Upon entry of these search terms, a matrix is created for specific "Gene IDs" and "GO" terms. In absence of user submitted GO terms or keyword, Genes2GO fetches all the GO terms that correspond to the list of genes. BiomaRt is updated biannually; therefore Genes2GO always retrieves the latest annotations.

Input

The user defines a list of gene IDs in the text input box, or uploads a file. Next, the organism must be specified from a dropdown list with 61 organisms from Ensemble, and the gene identifier type should also be specified (from dropdown list). Five gene identifier types are supported: Entrez Gene, Ensemble, HGNC symbol, UniProt ID and UniProt Accession number. Then, a list of GO IDs from tools such as OLSVis **[10]** or keywords from GO term names can be specified. If left empty, all GO annotations for the input genes will be fetched. Gene ID and GO IDs must be entered as lists separated by commas, spaces or returns.

Table 1: Chart comparing various GO annotation tools for specific properties. The table lists some features to compare Genes2GO with other significant tools that provide gene ontology annotations.

	Genes2GO	biomaRt in R	Biomart Ensembl	QuickGO	AmiGO	GOParGenPy	David
Matrix in text format	✓	-	-	-	-	✓	-
Matrix as excel sheet	✓	-	-	-	-	-	-
Selections of genes	✓	✓	✓	✓	✓	-	✓
Selections of GO terms	✓	-	✓	-	-	-	-
Keywords in GO terms	✓	-	-	-	-	-	-
Most recent annotation files	✓	✓	✓	✓	✓	-	-
Unlimited number of query genes	✓	✓	✓	✓	-	✓	✓
Graphical user Interface	✓	-	✓	✓	✓	-	✓

Output

Genes2GO creates a binary attribute matrix of genes (rows) and GO terms (columns), with values 0 and 1 denoting absence and presence of an annotation for any gene x with GO term y. The processing information and time elapsed after launching a query is displayed in a separate pop-up window. The processing time is essentially determined by the speed of data retrieval from biomaRt. This means that for small data sets (10 genes/10 terms) it may take between 20 to 150 seconds, whereas for large data sets (3000 genes/100 terms) it may take from 1 to 5 minutes. After query processing is completed, a link to the result file appears at the bottom of the query page and results can be downloaded either as a tab-delimited or an excel file.

Discussion

Although many tools are available for retrieval of GO terms for single or multiple genes, none of them returns annotations with only a specific set of GO terms, in a user-friendly format that allows further filtering and processing **(Table 1).** Genes2GO produces well-structured attribute matrices that can be easily analysed in spreadsheets, and exported for use is subsequent tools, without having to process large annotation files produced by other tools. Genes2GO always fetches the latest annotations of the genes. Genes2GO is available as a web tool requiring no download and installations, and requires no scripting knowledge.

Although, Genes2GO does not differentiate between the GO subtrees 'cellular compartment', 'molecular function' and 'biological processes', it can retrieve annotations with GO terms of any hierarchical level exactly as available via biomaRt **[3]**, and lists of GO IDs with a particular hierarchical relation may be easily procured using tools such as OLSVis **[10].** In addition to numerical GO IDs, Genes2GO allows users also to select for GO terms based on a text string that is part of the term name.

As Genes2GO depends on the BioMart server, it performs best with limited numbers of genes and GO terms, a condition that is compatible with many biological use cases. It currently caters to vertebrate gene annotations only and does not fetch the evidence code for individual annotations. These improvements could be made in the later version of the application.

Acknowledgements:

We wish to thank Astrid Lægreid for helpful suggestions and motivation for the development of Genes2GO. We also acknowledge Richard Newton and Norstore for helping with technical issues.

References

[1] Ashburner M *et al. Nature genetics*. 2000 **25(1):**25 [PMID: 10802651]
[2] Carbon S *et al. Bioinformatics*. 2009 **25(2):**288 [PMID: 19033274]
[3] Smedley D *et al. BMC genomics*. 2009 **10:**22 [PMID: 19144180]
[4] Binns D *et al. Bioinformatics*. 2009 **25(22):**3045 [PMID: 19744993]
[5] Dennis G, Jr. *et al. Genome biology*. 2003 **4(5):**P3 [PMID: 12734009].
[6] Kumar AA *et al. BMC bioinformatics*. 2013 **14:**242 [PMID: 23927037]
[7] Durinck S *et al. Nature protocols*. 2009 **4(8):**1184 [PMID: 19617889]
[8] Newton R & Wernisch L, R News 2007. p. 32.
[9] Yates A *et al. Nucleic acids research*. 2016 **44(D1)**:D710 [PMID: 26687719]
[10] Vercruysse S *et al. BMC bioinformatics*. 2012 **13:116** [PMID: 22646023]

Edited by P Kangueane
Citation: **Chawla & Kuiper,** Bioinformation 12(3): 231-232 (2016)