The Journal of Open Source Software

GeneScape: A Python package for gene ontology visualization

Istvan Albert [1,2]

1 Bioinformatics Consulting Center, Pennsylvania State University, United States of America **2** Department of Biochemistry and Molecular Biology, Pennsylvania State University, United States of America

DOI: 10.21105/joss.06624

Software

- Review
- Repository
- Archive

Editor: AHM Mahfuzur Rahman

Reviewers:

- @sridhar0605
- @j-andrews7

Submitted: 29 March 2024
Published: 05 June 2024

Summary

The Gene Ontology (GO) (Ashburner et al., 2000; Consortium et al., 2023) is a structured vocabulary that describes gene products in the context of their associated functions. The ontology takes the form of a directed graph, where each node defines a term, and each edge represents a hierarchical relationship between the terms (the words of the vocabulary).

For example, in the GO data, `GO:0090630` defines *activation of GTPase activity* and is a child of `GO:0043547`, defined as *positive regulation of GTPase activity* which in turn is a child of `GO:0051345` representing *positive regulation of hydrolase activity*.

Gene association files (GAF) are text files used to annotate an organism's gene products with Gene Ontology terms, associating functions to gene products. For example, a GAF file connects a gene product label, such as `ZC3H11B`, with multiple GO terms, such as `GO:0046872` or `GO:0016973`. The complete human genome GAF representation contains 288,575 associations of 19,606 gene symbols with over 18,680 GO terms.

The Gene Ontology Consortium maintains GAF files for various organisms. Typical genomic analysis protocols generate gene lists that must be placed in a functional context.

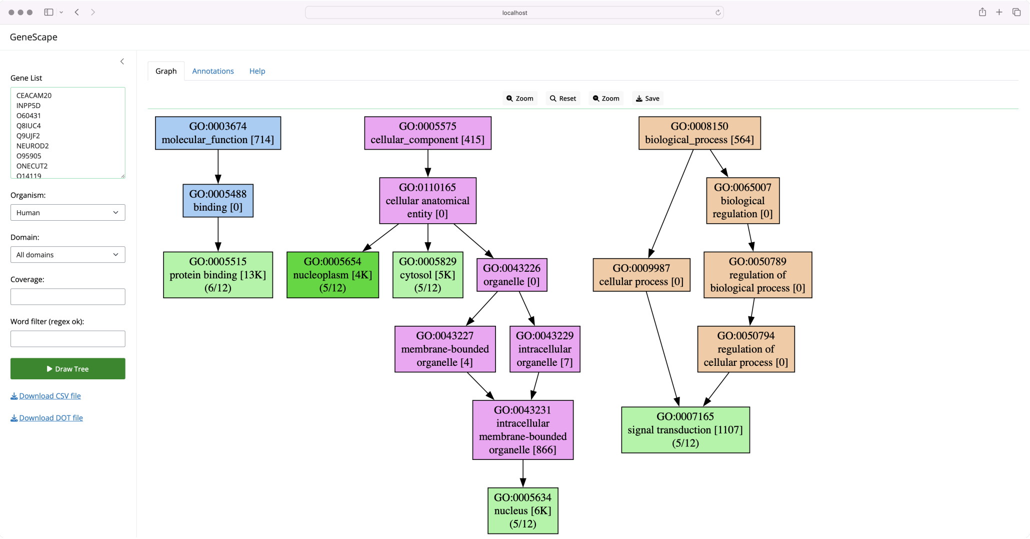
Statement of need

The most annotated gene in the human genome, `HTT`, currently has 1100 annotations. Thus, even small lists of genes may have a large number of annotations presenting an extraordinary challenge for interpretation. There is a clear need to visualize shared gene functions in an informative manner.

Web-based tools designed to visualize and filter gene ontology data include `AmiGO` (Carbon et al., 2008) and `QuickGO` (Binns et al., 2009). Command line tools like `goatools` (Klopfenstein et al., 2018) support GO term lineage visualization. R packages like `topGO` (Alexa & Rahnenfuhrer, 2023) implement GO structure visualizations of enriched GO terms. We are unaware of locally installable software that allows for interactive filtering and visualization of gene ontology terms derived from gene lists.

GeneScape is a Python package that allows users to visualize a list of genes in the functional context represented by the Gene Ontology

GeneScape is distributed both as a command-line tool and as GUI-enabled standalone software via the Shiny platform (Chang et al., 2024), making it accessible to a wide range of users.



GeneScape

Gene List

CEACAM20
INPP5D
O60431
Q8IUC4
Q9UJF2
NEUROD2
O95905
ONECUT2
O14119

Organism: Human

Domain: All domains

Coverage:

Word filter (regex ok):

localhost
Graph
Annotations
Help
Zoom
Reset
Zoom
Save
Draw Tree
Download CSV file
Download DOT file
GO:0003674 molecular_function [714]
GO:0005488 binding [0]
GO:0005515 protein binding [13K] (6/12)
GO:0005575 cellular_component [415]
GO:0110165 cellular anatomical entity [0]
GO:0005654 nucleoplasm [4K] (5/12)
GO:0005829 cytosol [5K] (5/12)
GO:0043226 organelle [0]
GO:0043227 membrane-bounded organelle [4]
GO:0043229 intracellular organelle [7]
GO:0043231 intracellular membrane-bounded organelle [866]
GO:0005634 nucleus [6K] (5/12)
GO:0008150 biological_process [564]
GO:0065007 biological regulation [0]
GO:0009987 cellular process [0]
GO:0050789 regulation of biological process [0]
GO:0050794 regulation of cellular process [0]
GO:0007165 signal transduction [1107] (5/12)

Figure 1: GeneScape as a Shiny App

GeneScape is distributed with several prebuilt databases for model organisms including the human, mouse, rat, fruitfly and zebrafish genomes. To study additional organisms, users must download GAF files from the Gene Ontology website and create custom databases using the `build` subcommand:

```
genescape build --gaf mydata.gaf.gz --index mydata.index.gz
```

For detailed instructions on using the software, users should refer to the GeneScape documentation. A Q&A discussion board is also available on the GeneScape GitHub page.

Typical usage

A typical usage starts with a gene list such as:

```
ABTB3
BCAS4
C3P1
GRTP1
```

Users can process the list above via the command line or the Shiny interface. A command line invocation might look like:

```
genescape tree genes1.txt -o output.pdf
```

The command above will produce the image:

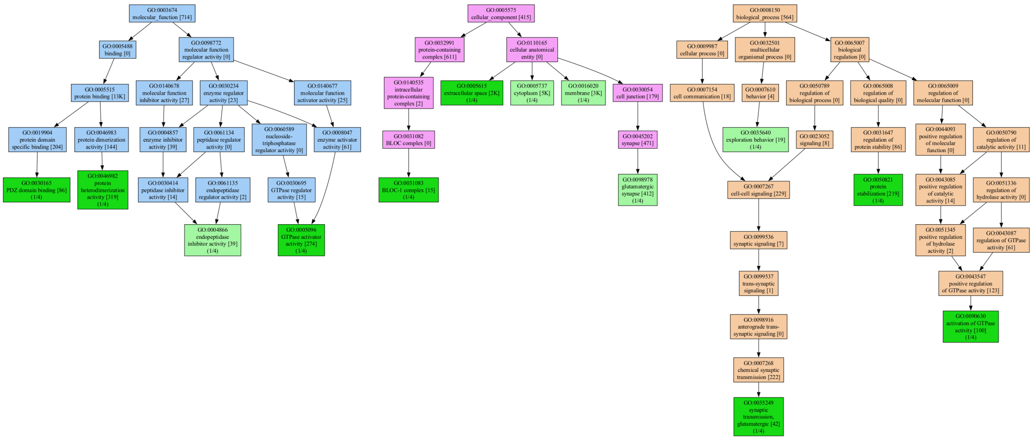

GO:0003674 molecular_function [714]
GO:0005488 binding [0]
GO:0098772 molecular function regulator activity [0]
GO:0005515 protein binding [13K]
GO:0140678 molecular function inhibitor activity [27]
GO:0030234 enzyme regulator activity [23]
GO:0140677 molecular function activator activity [25]
GO:0019904 protein domain specific binding [204]
GO:0046983 protein dimerization activity [144]
GO:0004857 enzyme inhibitor activity [39]
GO:0061134 peptidase regulator activity [0]
GO:0060589 nucleoside-triphosphatase regulator activity [0]
GO:0008047 enzyme activator activity [61]
GO:0030165 PDZ domain binding [86] (1/4)
GO:0046982 protein heterodimerization activity [319] (1/4)
GO:0030414 peptidase inhibitor activity [14]
GO:0061135 endopeptidase regulator activity [2]
GO:0030695 GTPase regulator activity [15]
GO:0004866 endopeptidase inhibitor activity [39] (1/4)
GO:0005096 GTPase activator activity [274] (1/4)
GO:0005575 cellular_component [415]
GO:0032991 protein-containing complex [611]
GO:0110165 cellular anatomical entity [0]
GO:0140535 intracellular protein-containing complex [2]
GO:0005615 extracellular space [2K] (1/4)
GO:0005737 cytoplasm [5K] (1/4)
GO:0016020 membrane [3K] (1/4)
GO:0030054 cell junction [179]
GO:0031082 BLOC complex [0]
GO:0045202 synapse [471]
GO:0031083 BLOC-1 complex [15] (1/4)
GO:0098978 glutamatergic synapse [412] (1/4)
GO:0008150 biological_process [564]
GO:0009987 cellular process [0]
GO:0032501 multicellular organismal process [0]
GO:0065007 biological regulation [0]
GO:0007154 cell communication [18]
GO:0007610 behavior [4]
GO:0050789 regulation of biological process [0]
GO:0065008 regulation of biological quality [0]
GO:0065009 regulation of molecular function [0]
GO:0035640 exploration behavior [19] (1/4)
GO:0023052 signaling [8]
GO:0031647 regulation of protein stability [86]
GO:0044093 positive regulation of molecular function [0]
GO:0050790 regulation of catalytic activity [11]
GO:0007267 cell-cell signaling [229]
GO:0050821 protein stabilization [219] (1/4)
GO:0043085 positive regulation of catalytic activity [14]
GO:0051336 regulation of hydrolase activity [0]
GO:0099536 synaptic signaling [7]
GO:0051345 positive regulation of hydrolase activity [2]
GO:0043087 regulation of GTPase activity [61]
GO:0099537 trans-synaptic signaling [1]
GO:0043547 positive regulation of GTPase activity [123]
GO:0098916 anterograde trans-synaptic signaling [0]
GO:0090630 activation of GTPase activity [100] (1/4)
GO:0007268 chemical synaptic transmission [222]
GO:0035249 synaptic transmission, glutamatergic [42] (1/4)

Figure 2: Ontology subgraph for a gene list

Internally, GeneScape first transforms the input gene list into a GO term list, where additional information is added to each term:

```
Coverage,Function,Domain,GO,Genes
1,endopeptidase inhibitor activity,MF,GO:0004866,C3P1
1,GTPase activator activity,MF,GO:0005096,GRTP1
1,extracellular space,CC,GO:0005615,C3P1
1,cytoplasm,CC,GO:0005737,BCAS4
1,membrane,CC,GO:0016020,ABTB3
1,PDZ domain binding,MF,GO:0030165,ABTB3
1,BLOC-1 complex,CC,GO:0031083,BCAS4
1,"synaptic transmission, glutamatergic",BP,GO:0035249,ABTB3
1,exploration behavior,BP,GO:0035640,ABTB3
1,protein heterodimerization activity,MF,GO:0046982,ABTB3
1,protein stabilization,BP,GO:0050821,ABTB3
1,activation of GTPase activity,BP,GO:0090630,GRTP1
1,glutamatergic synapse,CC,GO:0098978,ABTB3
```

In the next step, GeneScape draws the GO terms as the graph structure using the Networkx package (Hagberg et al., 2008), helping users visualize the functional context of the genes relative to the larger Gene Ontology.

Various colors and labels are used to provide additional context to the nodes in the graph; for example, functions present in the input genes are colored green. Intermediate nodes are colored by their category. Node labels display the total annotations and the number of genes that carry that function.

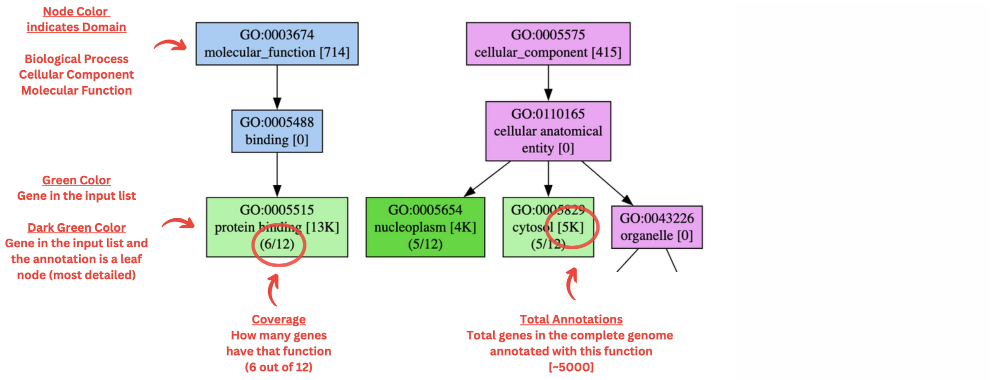


Figure 3: Filtering a large graph for a specific term

In the web interface, users can zoom in and out of the tree. The software's command-line version supports generating outputs in various formats, such as PDF or PNG.

Since the resulting graphs may also be large, with thousands of nodes, the main interface provides input widgets that allow users to interactively reduce the subgraph to nodes for which:

1. The function definitions match certain patterns.
2. A minimum number of genes share a function.
3. Nodes belong to a specific GO subtree: Biological Process (BP), Molecular Function (MF), Cellular Component (CC).

As an example, take the input gene list of just four genes:

```
Cyp1a1
Sphk2
Sptlc2
Smpd3
```

the resulting functional ontology graph is large with 641 nodes and 1,007 edges:

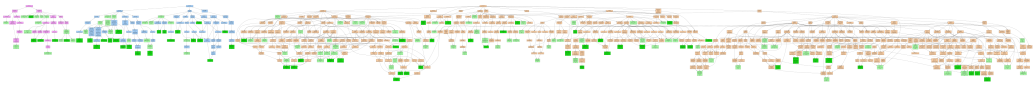


Figure 4: Very few genes can produce a large ontology tree

Users can reduce the tree to show only terms that match the word `lipid` and with at least two genes supporting the function via the graphical user interface or the command line:

```
genescape tree -m lipid --mincov 2 genes2.txt -o output.pdf
```

The filtering process will result in a smaller tree with 18 nodes and 29 edges, focused on the functions that contain the word "lipid":

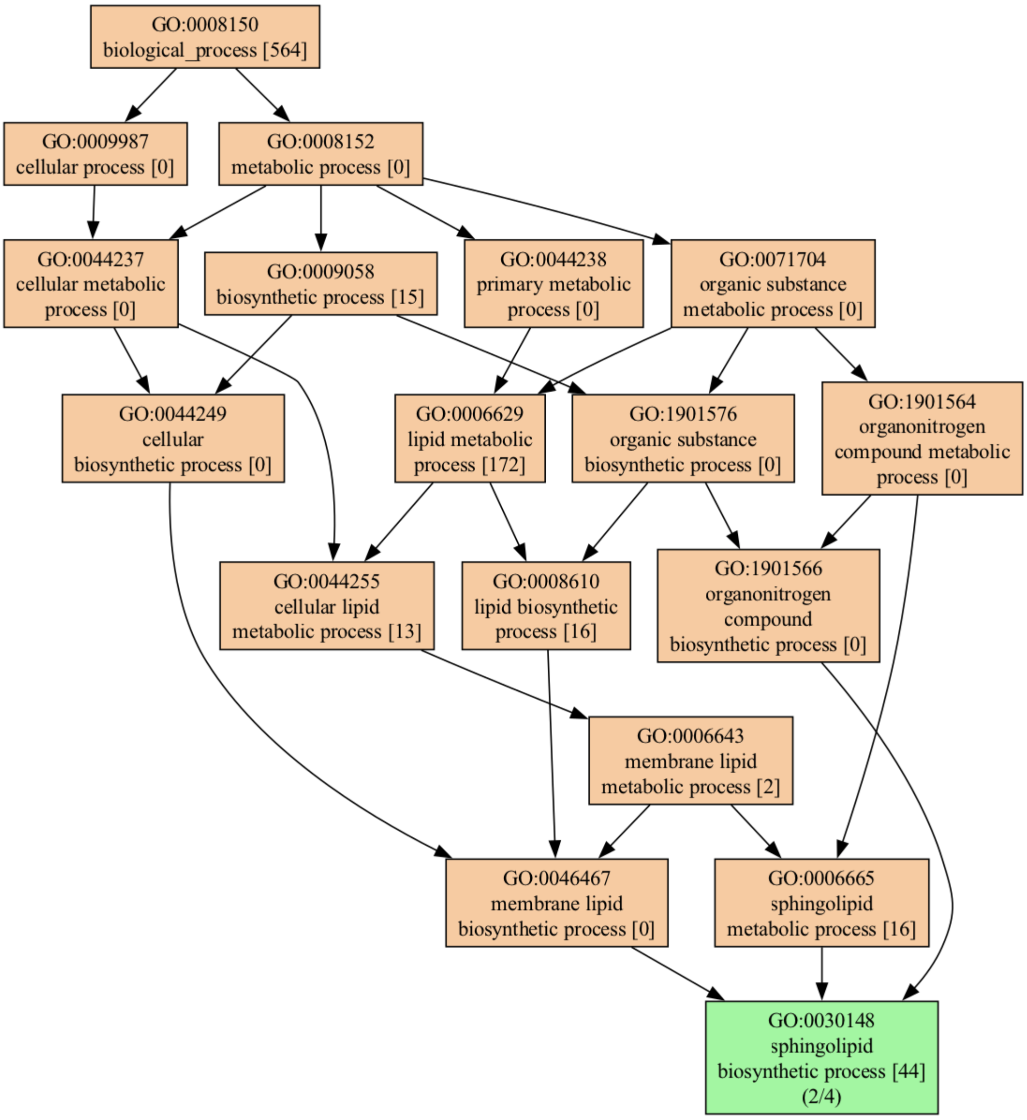

GO:0008150
biological_process [564]
GO:0009987
cellular process [0]
GO:0008152
metabolic process [0]
GO:0044237
cellular metabolic process [0]
GO:0009058
biosynthetic process [15]
GO:0044238
primary metabolic process [0]
GO:0071704
organic substance metabolic process [0]
GO:0044249
cellular biosynthetic process [0]
GO:0006629
lipid metabolic process [172]
GO:1901576
organic substance biosynthetic process [0]
GO:1901564
organonitrogen compound metabolic process [0]
GO:0044255
cellular lipid metabolic process [13]
GO:0008610
lipid biosynthetic process [16]
GO:1901566
organonitrogen compound biosynthetic process [0]
GO:0006643
membrane lipid metabolic process [2]
GO:0046467
membrane lipid biosynthetic process [0]
GO:0006665
sphingolipid metabolic process [16]
GO:0030148
sphingolipid biosynthetic process [44]
(2/4)

Figure 5: Filtering a large graph for a specific term

The software's primary purpose is to allow users to assess the functional depth of genes and identify commonalities and differences in the functional context of these genes.

Acknowledgments

We acknowledge support from the Huck Institutes for the Life Sciences at the Pennsylvania State University.

References

Alexa, A., & Rahnenfuhrer, J. (2023). topGO: Enrichment analysis for gene ontology. In *Bioconductor*. Bioconductor. https://doi.org/10.18129/B9.bioc.topGO

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., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'Donovan, C., & Apweiler, R. (2009). QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics*, *25*(22), 3045–3046. https://doi.org/10.1093/bioinformatics/btp536

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., AmiGO Hub, the, & Web Presence Working Group, the. (2008). AmiGO: online access to ontology and annotation data. *Bioinformatics*, *25*(2), 288–289. https://doi.org/10.1093/bioinformatics/btn615

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2024). *Shiny: Web application framework for r*. https://shiny.posit.co/

Consortium, T. G. O., Aleksander, S. A., Balhoff, J., Carbon, S., Cherry, J. M., Drabkin, H. J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N. L., Hill, D. P., Lee, R., Mi, H., Moxon, S., Mungall, C. J., Muruganugan, A., Mushayahama, T., Sternberg, P. W., Thomas, P. D., … Westerfield, M. (2023). The Gene Ontology knowledgebase in 2023. *Genetics*, *224*(1), iyad031. https://doi.org/10.1093/genetics/iyad031

Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In G. Varoquaux, T. Vaught, & J. Millman (Eds.), *Proceedings of the 7th python in science conference* (pp. 11–15).

Klopfenstein, D., Zhang, L., Pedersen, B. S., Ramírez, F., Warwick Vesztrocy, A., Naldi, A., Mungall, C. J., Yunes, J. M., Botvinnik, O., Weigel, M., & others. (2018). GOATOOLS: A python library for gene ontology analyses. *Scientific Reports*, *8*(1), 1–17. https://doi.org/10.1038/s41598-018-28948-z