

Slim-o-matic: a semi-automated way to generate Gene Ontology slims

Mélanie Courtot¹, Alex Mitchell¹, Maxim Scheremetjew¹, Janet Piñero², Laura I. Furlong², Robert D. Finn¹ and Helen Parkinson¹

¹ European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Wellcome Genome Campus, Hinxton, United Kingdom, mcourtot@gmail.com

² Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Medical Research Institute (IMIM), Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain

Abstract. The Gene Ontology (GO) currently contains over 40,000 terms describing the locations, activities and processes of gene products. Several millions of gene products have been annotated using the GO, and these annotations are routinely used for multiple applications. However, because of the difference of granularity in the annotations, it is useful to summarize GO annotations using GO slims. GO slims contain a subset of the GO terms, providing a higher-level, broader overview of the ontology while abstracting the finer details. Compiling GO slims is a time consuming process relying on manual human expertise, the process of creating the slims is often poorly documented, and maintaining and updating them can be difficult. In this paper, we present a semi-automated way to generate GO slims based on the annotation data available. We applied the tool to two different use cases, one for data overview in the newly released EBI Metagenomics pipeline, and one for gene-disease enrichment analysis using the DisGeNET platform. The slim-o-matic tool supports choosing the best terms for the slim, ensuring they are representative of the dataset, and have the best coverage using the minimal number of terms.

Keywords: Gene Ontology, slim, visualization, enrichment

1 Introduction

GO slims are subset of the Gene Ontology (GO)[1] which allow grouping at a higher-level of annotations to lower level GO terms. When used for visualization, GO slims employ only the main categories within the annotations, thereby providing an overview of the dataset. When used for enrichment analysis, the statistical power of the signal per slim term is greater than if signals to lower classes were individually counted, which can provide greater insights [2]. While several slims are available on the GO website [3], their development has been ad hoc and based on empirical methods, relying on both an expert GO editor to select the best GO terms and a domain expert to provide guidance and describe

the dataset. To address this, and make the slim generation more transparent and reproducible, the new slim-o-matic methodology was developed.

2 Methods

A 4 step pipeline has been implemented:

1. The GO term identifiers (IDs) and their frequencies in the dataset are mapped to the current GO ontology file. The current label of the terms is retrieved from the GO file, and a new annotation property ‘label with counts’ is populated by concatenating the label and associated frequency for each term. A new Web Ontology Language (OWL)[4] file is generated.
2. The newly generated OWL file is opened in the Protégé OWL editor [5] and manually inspected. The data owner reviews the hierarchy and chooses which higher-level terms maximize coverage of the annotations, using a fixed number of terms and while retaining specificity for the particular dataset. A list of slim term IDs is generated.
3. Based on the slim generated in 2., a script checks which of the original annotations would be included or excluded from the result, to validate whether any large count is falling outside of the chosen slim.
4. After iterations of steps 2 and 3, and once the data owner is satisfied with the terms included in the newly generated slim, a mapping script is run. It is based on map2slim [6], and generates a list of all terms in their dataset mapped to a higher level ontology term from the slim.

All the code and files are available under our GitHub repository, <https://github.com/ebispot/slim-o-matic>.

3 Results

We applied the slim-o-matic tool to two different use cases, one for dataset overview in the newly updated EBI Metagenomics pipeline [7], and one for gene-disease enrichment analysis using the DisGeNET [8] platform.

3.1 EBI metagenomics pipeline

The EBI Metagenomics is a resource for the analysis, archiving and browsing of metagenomic and metatranscriptomic datasets, with the aim of providing understanding of the microbial community composition and functional profile of deposited samples. The number of sequences within these datasets can be potentially vast, running into the 100s of millions, with similar numbers of annotations. Users therefore need to be able to visualize GO terms (assigned by InterProScan [9]) in an easy and compact way. A metagenomics GO slim was first created in 2012, built using 30 million annotations available at the time. Since then, the EBI Metagenomics resource has expanded dramatically and currently contains 10s of billion annotations for taxonomically diverse sequences

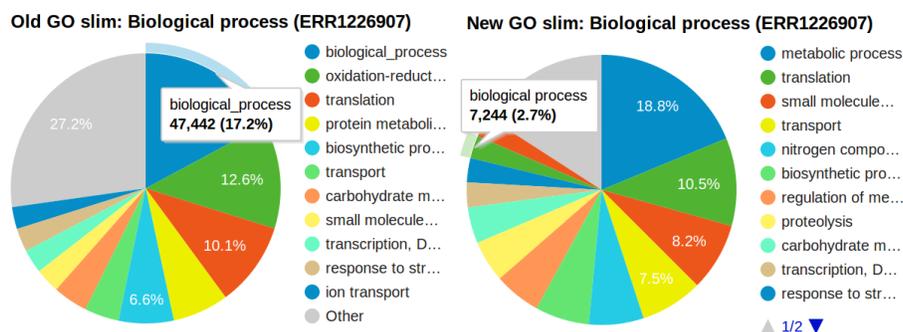


Fig. 1. A comparison of the annotations included when using the old (2012) slim, or the new (2016), slim-o-matic generated one, on a sample from the EBI Metagenomics dataset. Using the old GO slim, over 17% of terms are not mapped to a specific GO term and are instead mapped to the higher level of the hierarchy, `biological_process`. By comparison, with the new GO slim only 2.7% of annotations are not assigned a slim term. **Both slims were run against an unseen before pipeline 3.0 dataset.**

sampled from a wide range of different environments. Given the increase in size and diversity of annotations, and to support the release of an updated v3.0 analysis pipeline, the metagenomics GO slim was rebuilt using the slim-o-matic approach. Compared to the pre-existing slim, the new GO slim contains a few more terms (171 vs 160), and provides vastly improved coverage (98 % vs 80% overall). This increased coverage stems from the fact that using slim-o-matic, the GO terms chosen better reflects the current content of the EBI metagenomics (for example, more eukaryotic-derived sequences) and updates in the GO (such as better representation of viral terms in 2015 [10]). Fig. 1 shows an excerpt of the coverage comparison between the old and new GO slims.

3.2 DisGeNET

DisGeNET [8] is a discovery platform that integrates information about genes and variants associated to human diseases. To facilitate the analysis and interpretation of the data, DisGeNET supplies a variety of annotations describing genes, variants, and diseases. Currently, the genes in DisGeNET are characterized with their Panther protein class, and their top level Reactome pathway. Nevertheless, 46% of genes in DisGeNET have no Panther protein class, and almost 60% have no Reactome pathway. Adding GO information increases the coverage of annotations for protein-coding genes in DisGeNET to over 90%. However, the diverse granularity of the GO terms, and the relatively high number of annotations per gene is a hurdle to straightforward data interpretation. This is why, as a proof of concept, the slim-o-matic tool was applied to the GO cellular component (GO CC) subset of GO terms. As a result, more than 1,400 terms GO CC terms were reduced to 60 slim GO terms, and the median number of annotations per

gene decreased from 8 to 3. Additionally, an enrichment analysis per disease was performed, to test whether the genes associated to each disease showed a preferential distribution of cellular locations. DisGeNET diseases (curated subset) were tested for an over representation of GO CC categories in the complete, and slim set of GO terms. To ease the analysis of results, we grouped diseases by broader categories that correspond to the MeSH classification of diseases [11]. The results for the complete GO set after multiple test correction contained over 500 diseases in 360 GO CC categories, while the slim GO set contained 334 diseases in 47 categories. The results of the GO slim enrichment analysis show that some types of neoplasms and complex cardiovascular diseases are associated to proteins showing enrichment across all cellular compartments, while Mendelian disease proteins tend to be more confined to one specific compartment. For instance, Leigh disease and Coenzyme Q10 deficiency show an enrichment in the mitochondria (both are mitochondrial diseases). Additionally, most nervous system diseases, and mental disorders are enriched in proteins located in the plasma membrane (receptors, channels, and transporters).

4 Discussion

Future development include investigating ways of creating disease-oriented slims, where a term denoting process that might be involved in a disease pathophysiology - such as angiogenesis - is chosen, and co-occurring annotations are fetched from the GOA database with their counts. This can then be used as input to step 1. of the slim-o-matic tool, and allow semi-automated generation of slims focused on specific clinical investigations. While the slim-o-matic method has been developed based on the GO, nothing in the implementation is actually GO-specific. This means it could be applied to other resources, such as the Experimental Factor Ontology (EFO) [12], which currently contains just over 19,000 classes (and increasing), therefore reaching the limits of manual usability, and applied to the NHGRI GWAS catalog [13]. Finally, an interesting idea would be to try and fully automate the slim generation, thereby making it *completely* reproducible. While expert intervention may improve coverage and minimize number of terms, this comes at a cost of both resource and time, and we are aiming at implementing fully automated slim extraction from the Ontology Lookup Service (OLS) [14] hosted resources to provide a one-click slim experience to users.

5 Conclusion

Slim-o-matic allows for easy, fast and semi-automated generation of slims based on the underlying data. Consequently, slims have improved coverage over the existing annotations, and can be regenerated on a regular basis as either the dataset or the ontology evolve. As more and more ontologies reach a large size, the ability to process their hierarchy semi-automatically and summarize their content for visualization or enrichment analysis becomes critical.

References

1. M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–9, may 2000.
2. Seung Yon Rhee, Valerie Wood, Kara Dolinski, and Sorin Draghici. Use and misuse of the gene ontology annotations. *Nat Rev Genet*, 9(7):509–515, jul 2008.
3. Go subsets on the go website. <http://geneontology.org/page/download-ontology#Subsets>. Accessed: 2016-09-23.
4. W3C OWL Working Group. *OWL 2 Web Ontology Language: Document Overview*. W3C Recommendation, 27 October 2009. Available at <http://www.w3.org/TR/owl2-overview/>.
5. Mark A Musen. The protégé project: a look back and a look forward. *AI matters*, 1(4):4–12, 2015.
6. map2slim wiki. <https://github.com/owlcollab/owltools/wiki/Map2Slim>. Accessed: 2016-09-22.
7. Alex Mitchell, Francois Buchini, Guy Cochrane, Hubert Denise, Petra ten Hoopen, Matthew Fraser, Sebastien Pesseat, Simon Potter, Maxim Scheremetjew, Peter Sterk, and Robert D. Finn. Ebi metagenomics in 2016 - an expanding and evolving resource for the analysis and archiving of metagenomic data. *Nucleic Acids Research*, 44(D1):D595–D603, 2016.
8. Janet Piñero, Núria Queralt-Rosinach, Àlex Bravo, Jordi Deu-Pons, Anna Bauer-Mehren, Martin Baron, Ferran Sanz, and Laura I Furlong. Disgenet: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, 2015:bav028, 2015.
9. Philip Jones, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, Sebastien Pesseat, Antony F. Quinn, Amaia Sangrador-Vegas, Maxim Scheremetjew, Siew-Yit Yong, Rodrigo Lopez, and Sarah Hunter. Interproscan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240, 2014.
10. R. E. Foulger, D. Osumi-Sutherland, B. K. McIntosh, C. Hulo, P. Masson, S. Poux, P. Le Mercier, and J. Lomax. Representing virus-host interactions and other multi-organism processes in the Gene Ontology. *BMC Microbiology*, 15(1):146, dec 2015.
11. Medical subject headings (mesh). <https://www.nlm.nih.gov/mesh>. Accessed: 2016-11-18.
12. James Malone, Ele Holloway, Tomasz Adamusiak, Misha Kapushesky, Jie Zheng, Nikolay Kolesnikov, Anna Zhukova, Alvis Brazma, and Helen Parkinson. Modeling sample variables with an experimental factor ontology. *Bioinformatics*, 26(8):1112–1118, 2010.
13. Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, et al. The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic acids research*, 42(D1):D1001–D1006, 2014.
14. Simon Jupp, Tony Burdett, James Malone, Catherine Leroy, Matt Pearce, Julie McMurry, and Helen Parkinson. A New Ontology Lookup Service at EMBL-EBI. In *Proceedings of SWAT4LS International Conference*, 2015.