Chapter 11

Get GO! Retrieving GO Data Using AmiGO, QuickGO, API, Files, and Tools

Monica Munoz-Torres and Seth Carbon

Abstract

The Gene Ontology Consortium (GOC) produces a wealth of resources widely used throughout the scientific community. In this chapter, we discuss the different ways in which researchers can access the resources of the GOC. We here share details about the mechanics of obtaining GO annotations, both by manually browsing, querying, and downloading data from the GO website, as well as computationally accessing the resources from the command line, including the ability to restrict the data being retrieved to subsets with only certain attributes.

Key words Gene ontology, Ontology, Annotation resources, Annotation, Genomics, Transcriptomics, Bioinformatics, Biocuration, Curation, Access, AmiGO, QuickGO

1 Introduction

The efforts of the Gene Ontology Consortium (GOC) are focused on three major subjects: (1) the development and maintenance of the ontologies; (2) the annotation of gene products, which includes making associations between the ontologies and the genes and gene products in all collaborating databases; and (3) the development of tools that facilitate the creation, maintenance, and use of the ontologies. This chapter is focused on the mechanics of obtaining GO annotations, both directly and computationally, including the ability to restrict the data being retrieved to subsets with only certain attributes.

GO data is the culmination of various forms of curation, made accessible through a variety of interfaces and downloadable in different forms, depending on your intended use. Because the data and software landscape are constantly changing, it is hard to cover with any permanence the best way to access the data; this inherent limitation should be kept in mind as we navigate through this section. This chapter is intended as an overview of the different ways users can access GO data (via web portals, downloadable files, and API)

Christophe Dessimoz and Nives Škunca (eds.), *The Gene Ontology Handbook*, Methods in Molecular Biology, vol. 1446,
DOI 10.1007/978-1-4939-3743-1_11,

a quick description of basic software used by GO, and as a reference for where to find more detailed and up-to-date information about these subjects.

2 Web Interfaces to Access the GO

This section covers the online interfaces for accessing and interacting with the data using standard web browsers. Most consumers of the GO can make use of data browsers such as AmiGO, QuickGO, and data browsers embedded within more specific databases.

2.1 AmiGO

AmiGO ([1] http://amigo.geneontology.org; Fig. 1a) is the official web-based open-source tool for querying, browsing, and visualizing the Gene Ontology and annotations collected from the MODs (model organism databases), UniProtKB, and other sources (complete list of member institutions currently contributing to the GOC at http://geneontology.org/page/go-consortium-contributors-list). Notable features include: basic searching, browsing, the ability to download custom data sets, and a common question "wizard" interface. Recent changes have brought improvements both in speed and the variety of search modes, as well as the availability of additional data types, such as the display of annotation extensions (*see* Chap. 17 [2]) and display of protein forms (splice variants and proteins with post translational modifications). More details about the latest improvements on the AmiGO browser can also be found at GOC—Munoz-Torres (CA), 2015 [3].

2.2 QuickGO

The Gene Ontology Annotation (GOA) project at the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) also makes available the QuickGO browser ([4]; http://www.ebi.ac.uk/QuickGO; Fig. 1b), a web-based tool that allows easy browsing of the Gene Ontology (GO) and all associated electronic and manual GO annotations provided by the GO Consortium annotation groups. Included in its many features are extensive search and filter capabilities for GO annotations, a powerful integrated subset/slim interface, as well as an integrated historical view of the terms. For data consumption, QuickGO provides broad-ranging web services and cart functionality (a way of persisting abstract elements, like term IDs, between parts of the QuickGO web application).

AmiGO and QuickGO make use of the same GO data sets, with somewhat different implementations according to the requirements of funding sources and respective users. AmiGO, in its entirety, is a product of the GO Consortium and is the official channel for dissemination of the GO data sets, adhering to funding recommendations from NHGRI-NIH. QuickGO is produced, managed, and funded by EMBL-EBI; the members of QuickGO's managing team are also members of the GOC.

Fig. 1 Landing pages for the AmiGO (**a**) and QuickGO (**b**) browsers. A few features are highlighted for each browser

2.3 Other Browsers

The ontology component of the GO is also searchable and browsable from various third party generic ontology browsers such as OntoBee (http://ontobee.org), the EMBL-EBI Ontology Lookup Service (OLS) (http://www.ebi.ac.uk/ontology-lookup), OLSVis (http://ols.wordvis.com), and BioPortal (http://bioportal.bioontology.org). Each of these systems has their own particular strengths—for example OntoBee is aimed at the semantic web community, and provides the ontology as part of a linked data platform [5], whereas OLSVis is geared towards visualization. However, none of these browsers currently provide access to the annotations.

2.4 Term Enrichment Tool

One of the main uses of the GO is to perform enrichment analysis on gene sets. For example, given a set of genes that are up regulated under certain conditions, an enrichment analysis will find which GO terms are overrepresented (or underrepresented) using the available annotations for that gene set. The GO website offers a service that directly connects users with the enrichment analysis tool from the PANTHER Classification System [6]. The PANTHER database is up-to-date with GO annotations, and their enrichment tool is driven by GO data. Further details about this enrichment tool, as well as a list of supported gene IDs, are available from the PANTHER website at http://www.pantherdb.org/ and at http://www.pantherdb.org/tips/tips_batchIdSearch_supportedId.jsp. More information on enrichment analysis using the GO is available Chap. 13 [7] on "*Gene-Category Analysis.*"

2.5 A Simple Example of Data Exploration Using AmiGO (See Fig. 2)

To give a concrete example of the type of easy GO data exploration that can be accomplished using a web interface, we here provide an example where a user on the AmiGO annotation search interface (http://amigo.geneontology.org/amigo/search/annotation) is trying to find associations between genes/gene products and epithelial processes, while searching only data outside those available for human, and which have experimental evidence.

The user could:

- Type "*epithel*" into the text filter box (*Free-text filtering*) to the left of the results area.
- Open the "*Taxon*" facet and select the [−] next to "*Homo sapiens.*"
- Open the "*Evidence type*" facet and select the [+] next to "*experimental evidence.*"

The remaining results would fit the initial search criteria. However, suppose that the user wants to further refine their search to strictly look at all GO annotations that are directly or indirectly annotated to the GO term "*epithelial cell differentiation*" (GO:0009913). Following the steps above, they could:

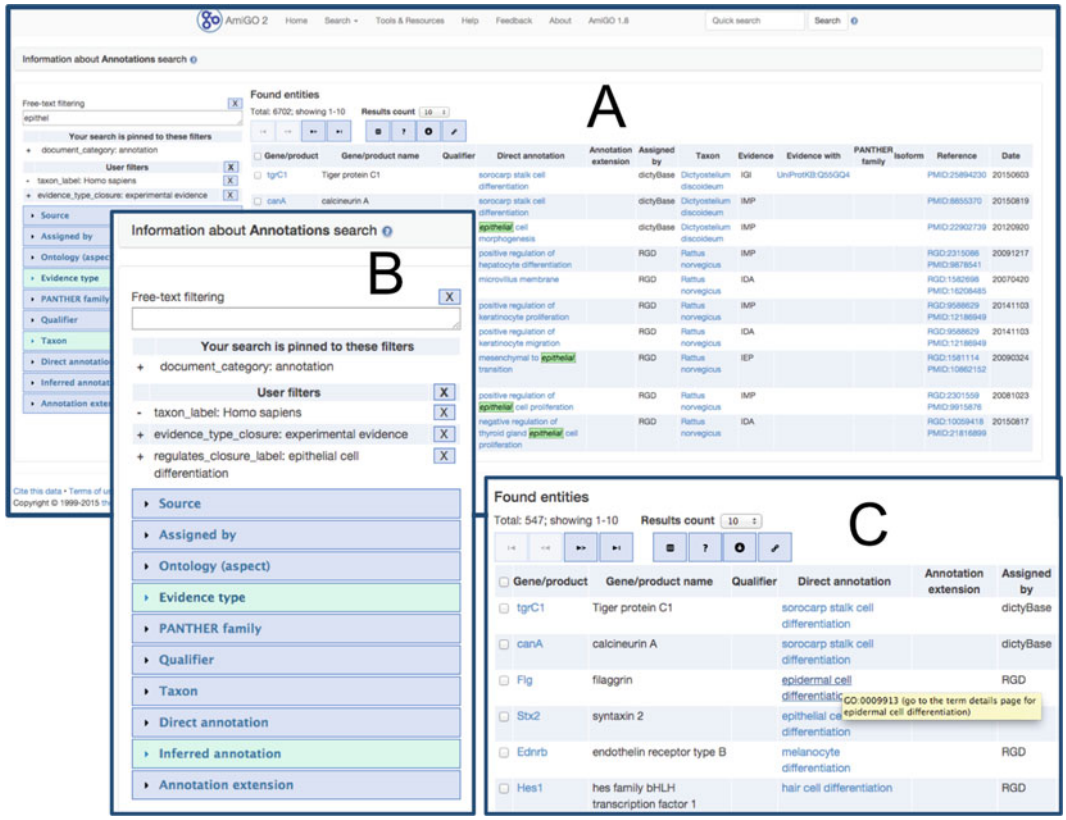

Fig. 2 Data exploration using the AmiGO annotation search interface. All results from this example are listed in *panel* (**a**). (**b**) Shows a detail about the filters applied throughout the search, listed under "**User filters**." An example of the details that appear for each gene or gene product is visible in (**c**): note that the information about the GO term ID for "*epithelial cell differentiation*" (GO:0009913) appears when users hover over the "Direct annotation" details

– Open the "*Inferred annotation*" facet and select the [+] next to "*epithelial cell differentiation*," then
– Remove the text filter by clicking the [x] next to the text entry.

This would leave the user with all GO annotations directly or indirectly annotated with "*epithelial cell differentiation*" (GO:0009913), that are not from human data, and have some kind of experimental evidence associated with them.

3 GO Files: Description and Availability

GO data files contain the current and long-term output of ontology and annotation efforts that are used for exchanging data across various systems. There are several use cases where it may be easier to mine the data directly from the files using a variety of tools. The most commonly used raw data files can be broken down into two categories: ontology and association files.

3.1 Ontology

In the context of GO, ontologies are graph structures comprised of classes for molecular functions, the biological processes they contribute to, the cellular locations where they occur, and the relationships connecting them all, in a species-independent manner [3]. Each term in the GO has defined relationships to one or more other terms in the same domain, and sometimes to other domains. Additional information about ontologies in general is also available from Chap. 1 [8].

GO ontology data are available from the GO website at http://geneontology.org/page/download-ontology. There are three different editions of the GO, in increasing order of complexity: go-basic, go, and go-plus.

go-basic: This basic edition of the GO is filtered such that annotations can be propagated up the graph. The relations included are *is_a*, *part_of*, *regulates*, *negatively_regulates*, and *positively_regulates*. It is important to note that this version excludes relationships that cross the three main GO hierarchies. Many legacy tools that use the GO make these assumptions about the GO, so we make this version available in order to support these tools. This version of the GO ontology is available in OBO format only.

go: This core edition of the GO includes additional relationship types, including some that span the three GO hierarchies, such as *has_part* and *occurs_in*, connecting the otherwise disjoint hierarchies found in **go-basic**. This version of the GO ontology is available in two formats, OBO and OWL-RDF/XML.

go-plus: This is the most expressive edition of the GO; it includes more relationships than **go** and connections to external ontologies, including the Chemical Entities of Biological Interest ontology (ChEBI; [9]), the Uberon anatomy (or stage) ontology [10], and the Plant Ontology for plant structure/stage (PO; [11]). It also includes import modules that are minimal subsets of those ontologies. This allows for cross-ontology queries, such as "*find all genes that perform functions related to the brain*" (e.g., in AmiGO: http://amigo.geneontology.org/amigo/term/UBERON:0000955#display-associations-tab). **go-plus** [12] also includes rules encoding biological constraints, such as the spatial exclusivity between a nucleus and a cytosol. These constraints are used for validation of the ontology and annotations [13]. This version of the GO ontology is available in OWL-RDF/XML.

When working with the ontologies, the official language of the Gene Ontology is the Web Ontology Language, or **OWL**, which is a standard defined by the World Wide Web Consortium (W3C). The GO has approximately 41,000 terms covering over 4 million genes in almost 470,000 species [3]. Its organization goes beyond a simple terminology structured as a directed acyclic graph (DAG), as it consists of over 41,000 classes, but it also includes an import chain that brings in an additional 10,000 classes from additional ontologies ([10] and see "*go-plus*" above). In order to best represent the

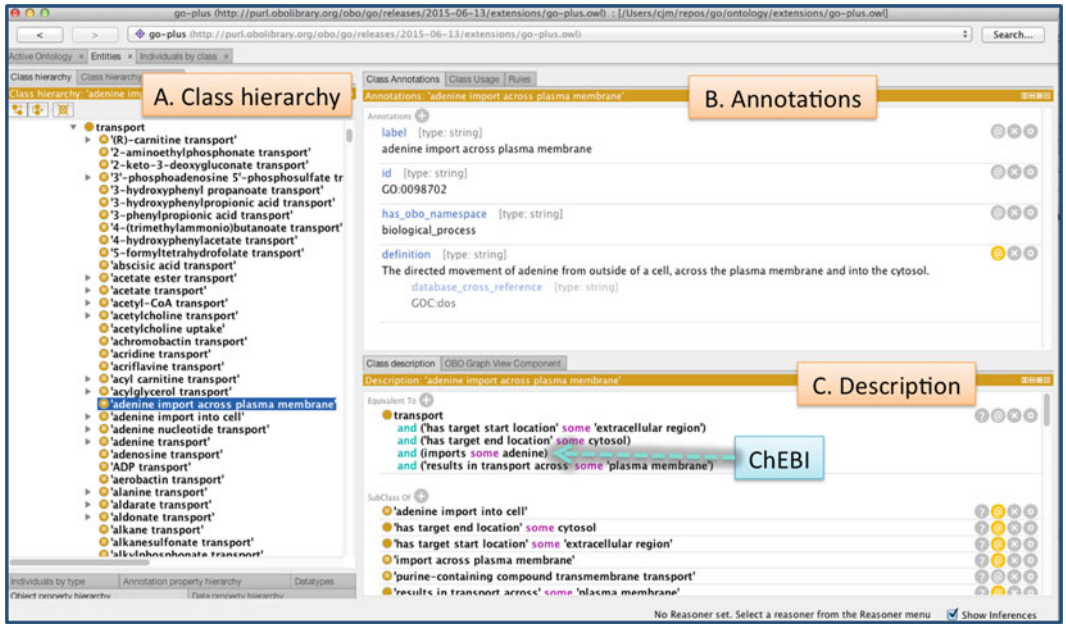

Fig. 3 Visualizing a GO term using Protégé. Protégé displays the details of the term "*adenine import across plasma membrane*" (GO:0098702). The underlying structure of the term is written in the OWL language, which adds flexibility to the expression of associations between genes and gene products and the terms in the ontology, compared to the possibilities offered in OBO. For example, in this term, inter-ontology logical definitions (OWL axioms) coming from the ChEBI ontology [9] are visible; this is not possible to see when visualizing the ontology using OBO

complexity of these classes, along with the approximately 27 million associations that connect them to each molecular entity (genes or gene products), members of the GOC software development team worked on building an axiomatic structure for GO. That is, they assigned logical definitions (known as OWL axioms or self-evidently true statements) to all the classes; the Gene Ontology has been effectively axiomatized, that is, reduced to this system of axioms in OWL, and is highly dependent on the OWL tool stack [10]. Examples of OWL stanzas for terms that are defined by a logical definition in the Gene Ontology are available from GOC—Munoz-Torres (CA), 2015 [3].

A number of tools, frameworks, and software libraries support OWL, including the ontology editor Protégé (http://protege.stanford.edu/; Fig. 3), the Java OWL API, and the OWLTools framework produced by the GO (https://github.com/owlcollab/owltools). Figure 3 shows a GO term visualized using Protégé; its underlying structure is the OWL language. We also make the ontology editions available in OBO Format, which is a simpler format used in many bioinformatics applications (note that "*go-plus*" is not available in OBO format). The two formats can be interconverted using the *Robot* tool produced by the GO Consortium, which can be found at https://github.com/ontodev/robot/.

The GO project is constantly evolving, and it welcomes feedback from all users (*see* below in Subheading 5.3). Research groups may contribute to the GO by either providing suggestions for updating the ontology (e.g., requests for new ontology terms) or by providing annotations. Requests for new synonyms or clarification of textual definitions are also welcomed.

Annotators and other data creators can search whether a term currently exists using the AmiGO browser at http://amigo.geneontology.org/, or may request new ones using either the GO issue tracker on GitHub or TermGenie. TermGenie ([14]; http://termgenie.org) is a web-based tool for requesting new Gene Ontology classes. It also allows for an ontology developer to review all generated terms before they are committed to the ontology. The system makes extensive use of OWL axioms, but can be easily used without understanding these axioms. Users not yet familiar with TermGenie, or whom do not yet have permission to use directly, may submit ontology updates and requests using the GO curator request tracker on GitHub (https://github.com/geneontology/go-ontology/issues), which allows free-text form submissions. For more information on how to best contribute to the GO, please *see* Chap. 7 [15].

3.2 Ontology Subsets

Gene Ontology subsets (also sometimes known as "*slims*") are cut-down versions of the ontologies, containing a reduced number of terms (e.g., species-specific subsets or more generic subsets with "useful" terms in various categories). They give a broad overview of the ontology content without the detail of the specific fine-grained terms. Subsets are particularly useful for giving a summary of the results of GO annotation of a genome, microarray, or cDNA collection when broad classification of gene product function is required. Further information, including Java-based tools and data downloads, is available from the GO website (http://geneontology.org/page/go-slim-and-subset-guide).

3.3 Association Files

The annotation process captures the activities and localization of a gene product using GO terms, providing a reference, and indicating the kind of available evidence in support of the assignment of each term using evidence codes. Currently, the main format for annotation information in the GO is the Gene Association File (GAF, http://geneontology.org/page/go-annotation-file-formats). This is the standardized file format that members of the Consortium use for submitting data. The annotation data is stored in tab-delimited plain text files, where each line in the file represents a single association between a gene product and a GO term, with an evidence code, the reference to support the link between them, and other information. The GAF file format has several different "flavors," with 2.1 being the most current version. Additional details about GAF files is found in Chap. 3 [16].

Recently, the GPAD/GPI files were developed, which are essentially a normalized version of GAF information. These formats are expected to have more prominence in the future, and further details about them can be found on the GO website (http://geneontology.org/page/go-annotation-file-formats).

Because they are tab-delimited text files, both the GAF and GPAD/GPI file formats are very amenable to mining with command line tools. As well, OWLTools can also be used to access this annotation information with operations such as: connecting the annotations to ontology information for exploration and reasoning, OWL translation, validation, taxon checks, and link prediction. More advanced details on this topic are further explained on the OWLTools project wiki (https://github.com/owlcollab/owltools/wiki).

Details on how to make and evaluate GO annotations are discussed in Chap. 4 [17] on "*Best Practices in Manual Annotation with the Gene Ontology*," and in Chap. 8 [18] on "*Evaluating Computational Gene Ontology Annotations.*" Information is also available in the GO Annotation Guide (http://geneontology.org/page/go-annotation-policies); more information on the meaning and use of the evidence codes in support of each annotation can be found on the GO Evidence Codes documentation (http://geneontology.org/page/guide-go-evidence-codes). The GOC is currently transitioning from using evidence codes into implementing the Evidence Ontology (ECO) to describe the evidence in support of each association between a gene product and a GO term. A detailed description of the Evidence Ontology and its use cases is included in Chap. 18 [19] on "*The Evidence and Conclusion Ontology: Supporting Conclusions & Assertions with Evidence.*"

4 Making Your Own Tools

In addition to using off-the-shelf tools provided by the GOC or other users, we also provide libraries and APIs to enable end-users to easily create their own tools for working with and analyzing GO data.

Within the Java/JVM ecosystem, the OWLTools (https://github.com/owlcollab/owltools), and OWL API (https://github.com/owlcs/owlapi) libraries are the primary tools to work with the data. Since OWL is the internal representation format used by the GOC, standard OWL reasoners and tools are all usable with the data. For slightly less general access to the data, the OWLTools(-Core) wrapper library adds numerous helper methods to access OBO-specific fields (i.e., *synonyms*, *alt_ids*), walk graphs, create closures, and other common operations.

On the JavaScript side (both client and server), AmiGO development has produced JavaScript APIs (http://wiki.geneontology.org/index.php/AmiGO_2_Manual:_JavaScript) and widgets (http://

wiki.geneontology.org/index.php/AmiGO_2_Manual:_Widgets) for better access and integration with other tools. Users interested in using the JavaScript API or widgets from AmiGO in their own site should become familiar with the manager and response interfaces, which are the core of the JavaScript interface. An introductory overview of the JavaScript API and widgets, as well as details on implementation engines, the response class, and the configuration class can be also be found on the JavaScript section of the AmiGO Manual, listed above.

As well, AmiGO provides methods for producing incoming searches to allow external sites to link to relevant information. Documentation about these methods can be found at http://wiki.geneontology.org/index.php/AmiGO_2_Manual:_Linking.

5 Additional Information

5.1 Mappings

The GO project provides mappings between GO terms and other key related systems (built for other purposes), such as Enzyme Commission numbers or Kyoto Encyclopedia of Genes and Genomes (KEGG). However, one should be aware that these mappings are neither complete nor exact and should be used with caution. A complete listing of mappings available for the resources of the GOC can be found at http://geneontology.org/page/download-mappings. Additional information about alternative and complementary resources to the GO is available on Chap. 19 [20].

5.2 Legacy Interface for GO

Currently, the AmiGO and QuickGO interfaces have moved away from SQL database derivatives of the data sets. However, to support legacy applications and queries, the GO data is regularly converted into an SQL database (MySQL). These builds can be downloaded and installed on a local machine, or queried remotely using the GO Online SQL/Solr environment (GOOSE; http://amigo.geneontology.org/goose). More information about SQL access, including various downloads and schema information, can be found in the legacy SQL section of the GO website (http://geneontology.org/page/lead-database-guide).

5.3 Help/ Troubleshooting Software and Data

In additional to other functions, the GO Helpdesk addresses user queries about the Gene Ontology and related resources. The GO Helpdesk will direct any questions or concerns with GO data, software, or analysis to the appropriate people within the consortium. You can directly contact the GO Helpdesk using the site form (http://geneontology.org/form/contact-go), which will automatically enter your query into an internal tracker to ensure responsiveness.

Funding MMT and SC were supported by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 (http://science.energy.gov/bes/), and by the U.S. National Institutes of Health, National Human Genome Research Institute grant HG002273. Open Access charges were funded by the University College London Library, the Swiss Institute of Bioinformatics, the Agassiz Foundation, and the Foundation for the University of Lausanne.

References

1. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, AmiGO Hub, Web Presence Working Group (2009) AmiGO: online access to ontology and annotation data. Bioinformatics 25(2):288–289
2. Huntley RP, Lovering RC (2016) Annotation extensions. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 17
3. Gene Ontology Consortium, Munoz-Torres MC (Corresponding Author) (2015) Gene Ontology Consortium: going forward. Nucleic Acids Res 43(Database issue):D1049–D1056. doi:10.1093/nar/gku1179
4. Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R (2009) QuickGO: a web-based tool for Gen Ontology searching. Bioinformatics 25(22):3045–3046
5. Xiang Z, Mungall C, Ruttenberg A, He Y (2011) Ontobee: a linked data server and browser for ontology terms. Proceedings of the 2nd international conference on biomedical ontologies (ICBO), Buffalo, NY, USA, 28–30 July 2011, pp 279–281
6. Mi H, Muruganujan A, Casagrande JT, Thomas PD (2013) Large-scale gene function analysis with the PANTHER classification system. Nat Protoc 8(8):1551–1566
7. Bauer S (2016) Gene-category analysis. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 13
8. Hastings J (2016) Primer on ontologies. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 1
9. Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, Muthukrishnan V, Owen G, Turner S, Williams M et al (2013) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Res 41:D456–D463
10. Mungall C, Torniai C, Gkoutos G, Lewis S, Haendel M (2012) Uberon, an integrative multi-species anatomy ontology. Genome Biol 13:R5
11. Cooper L, Walls RL, Elser J, Gandolfo MA, Stevenson DW, Smith B, Preece J, Athreya B, Mungall CJ, Rensing S et al (2013) The Plant Ontology as a tool for comparative plant anatomy and genomic analyses. Plant Cell Physiol 54:e1
12. Berardini TZ, Khodiyar VK, Lovering RC, Talmud P (2010) The Gene Ontology in 2010:

extensions and refinements. Nucleic Acids Res 38(Database Issue):D331–D335

13. Mungall CJ, Dietze H, Osumi-Sutherland D (2014) Use of OWL within the gene ontology. In Keet M, Tamma V (eds) Proceedings of the 11th international workshop on owl: experiences and directions (OWLED 2014), Riva del Garda, Italy, 17–18 October 2014, pp 25–36. doi:10.1101/010090
14. Dietze H, Berardini TZ, Foulger RE, Hill DP, Lomax J, Osumi-Sutherland D, Roncaglia P, Mungall CJ (2014) TermGenie – a web-application for pattern-based ontology class generation. J Biomed Semantics 5:48. doi:10.1186/2041-1480-5-48
15. Lovering RC (2016) How does the scientific community contribute to gene ontology? In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 7
16. Gaudet P, Škunca N, Hu JC, Dessimoz C (2016) Primer on the gene ontology. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 3
17. Poux S, Gaudet P (2016) Best practices in manual annotation with the gene ontology. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 4
18. Škunca N, Roberts RJ, Steffen M (2016) Evaluating computational gene ontology annotations. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 8
19. Chibucos MC, Siegele DA, Hu JC, Giglio M (2016) The evidence and conclusion ontology (ECO): supporting GO annotations. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 18
20. Furnham N (2016) Complementary sources of protein functional information: the far side of GO. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 19