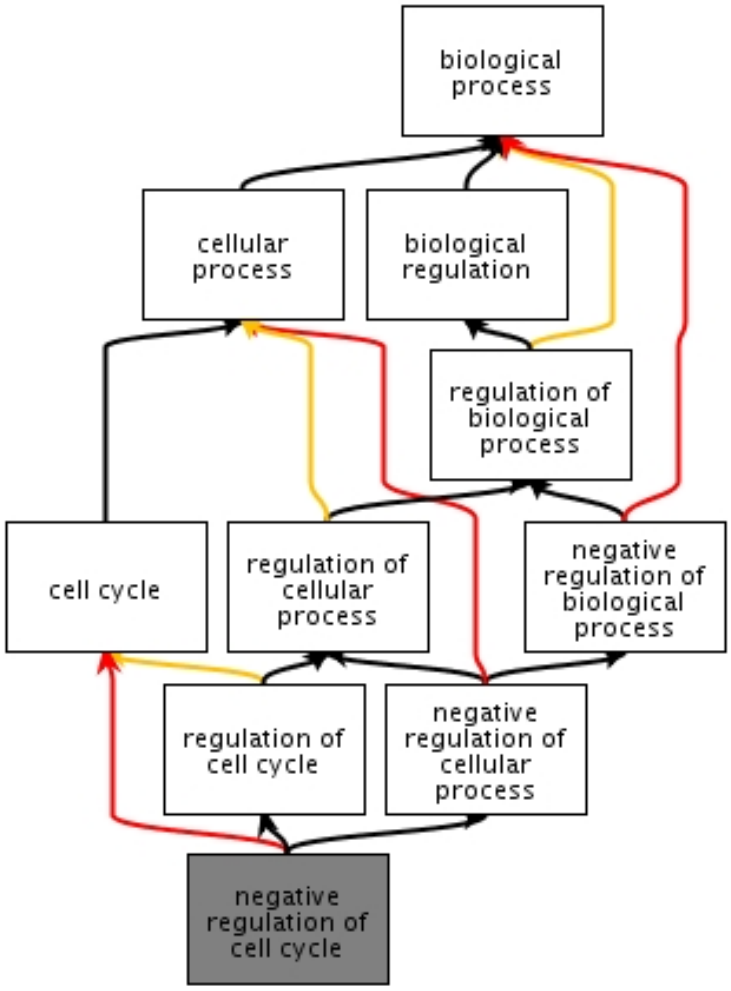
Gene Ontology Annotation at the EBI



EMBL-EBI

www.ebi.ac.uk

v8, April 2013

Table of Contents

Information on this tutorial

Pre-requisites	Ability to use a web browser
Subject area	Gene product functional annotation
Target audience	Biologists: Masters/Ph.D./Post-doc Computer biologists
Resources required	Internet browser, preferably Firefox
Approximate time needed	2.5 hours

Information

Words written in **bold** are explained in the glossary.

Course learning objectives

The aim of this course is to familiarise users with the Gene Ontology (GO) and associations of **GO terms** with gene products (**GO annotations**). The course will cover how to retrieve **GO annotations** from various sources and how to use **GO annotations** to provide a link to biological knowledge in large sequence datasets.

You will learn how to:

- Use the **QuickGO** tool to view GO data and annotations, filter annotations to create a tailored set and use **GO slims** to summarise the attributes of a gene product
- Use InterProScan to rapidly populate novel sequences with electronic GO annotation predictions
- Retrieve complete sets of **GO annotations**
- How to feedback or contribute annotations to UniProt-GOA

An introduction to the Gene Ontology

The Gene Ontology project is a major bioinformatics initiative provided by the Gene Ontology Consortium (http://www.geneontology.org/) with the aim of standardizing the representation of gene and gene product attributes across species and databases. The project provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data from GO Consortium members, as well as tools to access and process this data.

The Gene Ontology covers three domains: *cellular component*, the parts of a cell or its extracellular environment; *molecular function*, the elemental activities of a gene product at the molecular level, such as binding or catalysis; and *biological process*, operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms.

Each of the three ontologies is built from **GO terms** which describe the biological concepts within each domain. The **GO terms** are linked to each other using relationships of which there are seven currently being used within the GO Consortium; *is_a, part_of, has_part, occurs_in, regulates, positively_regulates* and *negatively_regulates*. For more information on the relationships see the documentation on the GO Consortium website: http://www.geneontology.org/GO.ontology-ext.relations.shtml.

GO annotations

Associations between gene products and these terms, termed **GO annotations**, are assigned by many biological databases providing detailed functional descriptions. A single gene product may be annotated to multiple **GO terms**, detailing a range of functional attributes, using both manual and electronic annotation methods.

A **GO annotation** is the assignment of a GO identifier, e.g. GO:0005737, with a particular sequence (either a gene or protein database identifier), e.g. UniProtKB accession Q4VCS5. There are two types of annotation; *manual annotations* are created by a curator having directly looked for functional information (either within published literature or by examining the sequence directly), whereas *electronic annotations* are produced by a number of different types of automated methods that produce high-quality, conservative predictions of GO assignments.

All annotations must provide both a reference to a source that provides information either directly for the **GO term**-gene product assignment or the method used to create the assignment, and also an **evidence code** which is a three-letter acronym indicating the type of evidence that supports the assignment of the **GO term** to the gene product.

Manual GO annotation at the EBI

High-quality manual annotations are made directly to the UniProt-GOA database, using a UniProt developed annotation tool, by over 60 highly trained biological curators within the UniProt Consortium who are located both internally and externally to the EBI. In addition, UniProt-GOA integrates manual annotations from all GO Consortium annotation groups, as well as a number of external special interest databases (e.g. LIFEdb, Human Protein Atlas, Reactome and the IntAct protein-protein interaction database). A description of how manual annotations are made by UniProt-GOA is available on our website; http://www.ebi.ac.uk/GOA/ManualAnnotationEfforts

Electronic GO annotation at the EBI

UniProt-GOA also provides large-scale assignments of **GO terms** to UniProtKB entries through a number of electronic techniques. We use existing information within database entries, including UniProt keywords (UPKW2GO), UniProt Subcellular Location (UPSL2GO), UniProt Pathway2GO (Pathway2GO), Enzyme Commission numbers (EC2GO) and cross-references to InterPro (InterPro2GO) and HAMAP (HAMAP2GO), which are manually mapped to **GO terms**. Electronically combining these mappings with a table of matching UniProtKB entries generates a set of **GO annotations**.

Finally, in collaboration with the Ensembl and EnsemblGenomes teams, manually annotated **GO annotations** are projected automatically to a number of predicted orthologs using the Compara database.

Each of these methods is described in more detail on our website; http://www.ebi.ac.uk/GOA/ElectronicAnnotationMethods

Practical use of GO and GO annotations

GO is useful in many ways. It can provide a broad overview of the functions of a proteome or specific set of proteins. It can help to generate hypotheses about underlying cellular mechanisms in disease states by providing a way to cluster

subsets of over- or under-expressed proteins according to the **GO annotations** they have in common, potentially highlighting particular pathways that may be affected by the disease. GO can also provide functional data to protein interactome sets, linking the protein binding network with the activities and locations of the interacting proteins. Additionally GO can assist researchers who are refining a particular organelle isolation technique by indicating the subcellular locations of the proteins that have been isolated.

Chapter 1 – How to view GO data using the QuickGO browser

GO annotation data is available from a number of different sources and in a variety of formats. To browse the GO hierarchy or to view annotations for individual gene products, a number of online tools are available, such as **QuickGO**, which has been developed at the EBI to display annotations assigned to UniProtKB entries.

QuickGO is highly flexible and has a number of unique features, including the ability to tailor annotation sets using multiple filtering options as well as to construct subsets of the GO (GO slims) to map-up annotations allowing a general overview of the attributes of a set of proteins.

QuickGO is available from:

http://www.ebi.ac.uk/QuickGO

For other GO browser tools, please see the 'Further Reading' section located at the end of this tutorial [1].

Chapter 1 - learning objectives

You will learn:

- How to search the Gene Ontology for individual **GO terms**
- How to search the UniProt-GOA database for **GO annotation** data

Searching for GO terms in the Gene Ontology

The **QuickGO** home page (Fig. 1.1) provides a text box to start searching for GO information.

Information

There is a search box on every page of **QuickGO**.

You may search for any aspect of a **GO annotation** including;

- **GO term** names and synonyms
- GO IDs
- UniProtKB accessions
- InterPro Ids
- Enzyme Commission numbers
- UniProt keywords.

As **QuickGO** integrates a large number of symbols and identifier types you can also query for these, for example; NCBI Gene IDs, RefSeq accessions and Ensembl IDs.

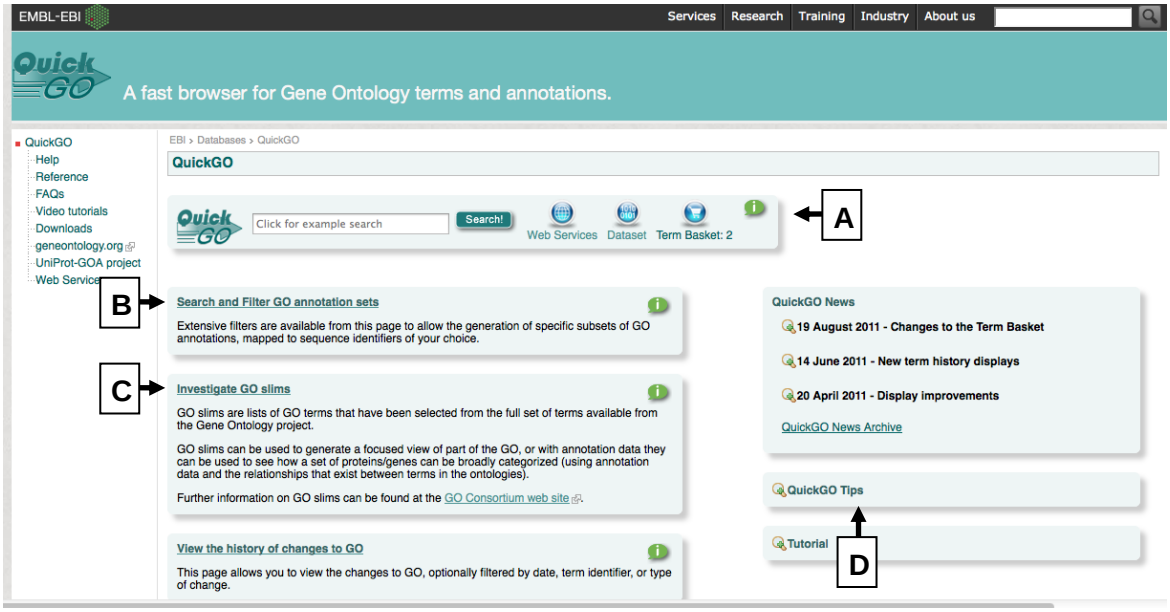


Fig 1.1 QuickGO query interface (http://www.ebi.ac.uk/QuickGO)

Figure 1.1

Notes

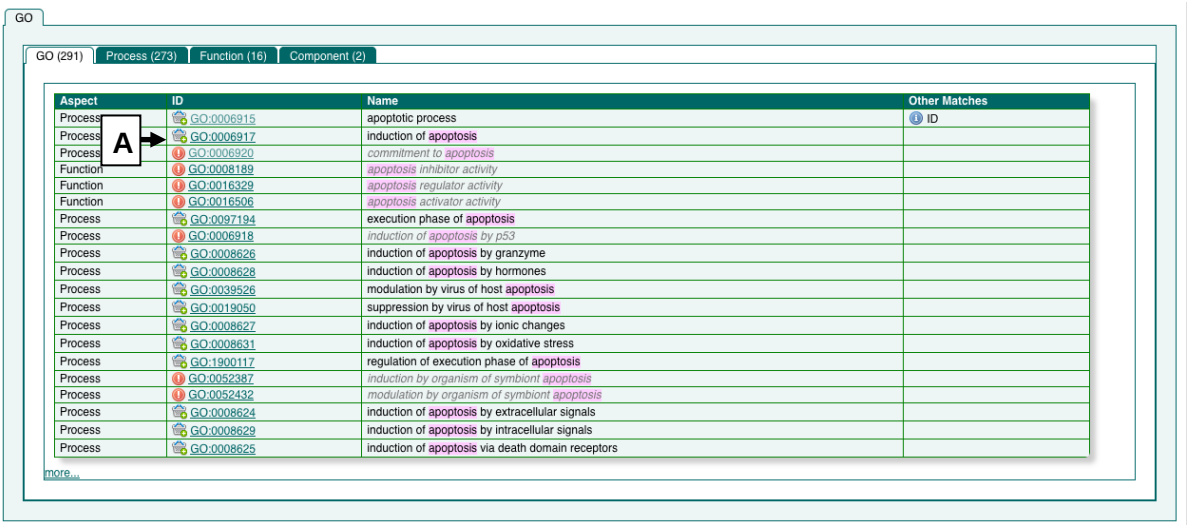
[A] The 'Global Toolbar' is visible on all pages within QuickGO. From here users can search QuickGO, access QuickGO Web Services, view the dataset that QuickGO is currently using and view 'Term Basket' selection (see Chapter 3 for more information on 'Term Basket'.

[B] The entry point for viewing, filtering and downloading annotations from the UniProt-GOA database.

[C] The entry point for GO slims. See Chapter 3 for more information.

[D] Useful tips for using GO and QuickGO

When searching for a **GO term**, **QuickGO** will return any terms relevant to your input text. The first 20 **GO terms** are shown by default, to see further terms, click on 'more' at the bottom of the list (Fig. 1.2a). Tabbed sections on this page allow you to view terms from a particular aspect of the GO, i.e. Molecular Function, Biological Process or Cellular Component. Note that some terms are retrieved due to information in their synonym, definition or cross-reference (ID) fields and this is noted to the right of the term name (Figs 1.2a & b). Obsolete terms are also retrieved and this is also indicated to the right of the term name (Fig. 1.2a & b).



Aspect	ID	Name	Other Matches
Process	GO:0006915	apoptotic process	ID
Process	GO:0006917	induction of apoptosis	
Process	GO:0006920	*commitment to apoptosis*	
Function	GO:0008189	*apoptosis inhibitor activity*	
Function	GO:0016329	*apoptosis regulator activity*	
Function	GO:0016506	*apoptosis activator activity*	
Process	GO:0097194	execution phase of apoptosis	
Process	GO:0006918	*induction of apoptosis by p53*	
Process	GO:0008626	induction of apoptosis by granzyme	
Process	GO:0008628	induction of apoptosis by hormones	
Process	GO:0039526	modulation by virus of host apoptosis	
Process	GO:0019050	suppression by virus of host apoptosis	
Process	GO:0008627	induction of apoptosis by ionic changes	
Process	GO:0008631	induction of apoptosis by oxidative stress	
Process	GO:1900117	regulation of execution phase of apoptosis	
Process	GO:0052387	*induction by organism of symbiont apoptosis*	
Process	GO:0052432	*modulation by organism of symbiont apoptosis*	
Process	GO:0008624	induction of apoptosis by extracellular signals	
Process	GO:0008629	induction of apoptosis by intracellular signals	
Process	GO:0008625	induction of apoptosis via death domain receptors	

1.2a

Function	GO:0005035	death receptor activity	Synonym
Process	GO:0001783	B cell apoptotic process	Synonym
Process	GO:0051402	neuron apoptotic process	Synonym
Process	GO:0070231	T cell apoptotic process	Synonym
Process	GO:0006309	apoptotic DNA fragmentation	Synonym
Process	GO:0033024	mast cell apoptotic process	Synonym
Process	GO:0070242	thymocyte apoptotic process	Synonym
Process	GO:0071887	leukocyte apoptotic process	Synonym
Process	GO:0001781	neutrophil apoptotic process	Synonym
Process	GO:0001913	T cell mediated cytotoxicity	Synonym
Process	GO:0034349	glial cell apoptotic process	Synonym
Process	GO:0044346	fibroblast apoptotic process	Synonym
Process	GO:0045476	nurse cell apoptotic process	Synonym
Process	GO:0070227	lymphocyte apoptotic process	Synonym
Process	GO:0071888	macrophage apoptotic process	Synonym
Process	GO:0097284	hepatocyte apoptotic process	Synonym
Process	GO:0010657	muscle cell apoptotic process	Synonym

1.2b

Fig. 1.2 A search for 'apoptosis' retrieves a choice of GO terms, tabbed sections allow for a more focused search in a particular ontology. Terms are retrieved if the word 'apoptosis' is present in their term name, synonyms, definition or cross-references.

Figure 1.2

Notes

[A] The green plus icons next to GO IDs allow you to add that term to the 'Term Basket' which you can use for comparing multiple terms in an ontology chart or for creating GO slims (see section 3).

Clicking on the GO ID for a term will take you to a page called the 'Term Information page' (Fig. 1.3a), providing full details of the selected term. Tabbed pages provide further information about the term such as Ancestor terms (Fig 1.3b), Child terms (Fig. 1.3c) and Protein Annotation to the term.

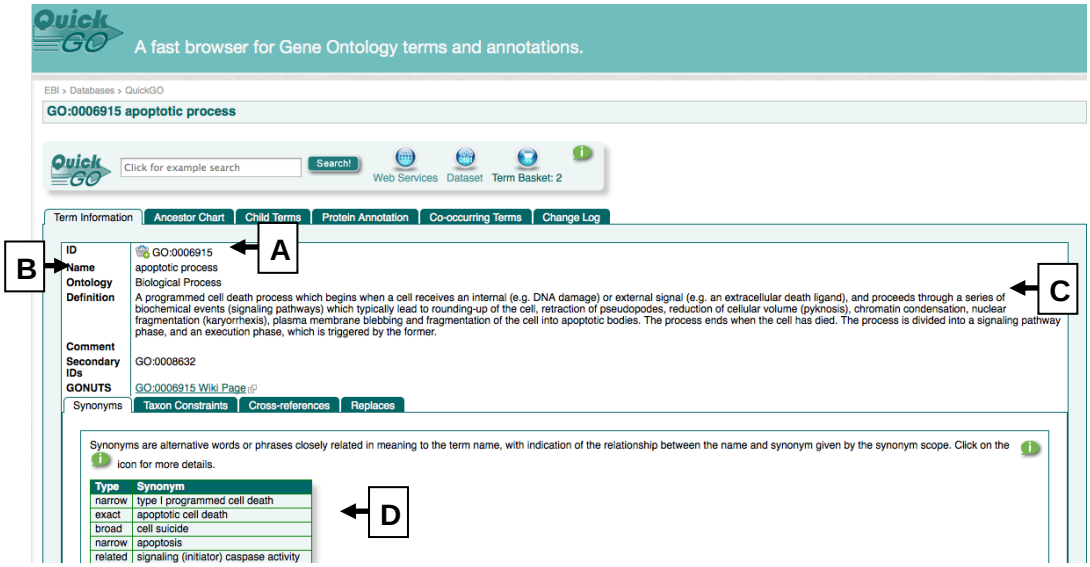


Fig. 1.3a GO Term Information page view

Figure 1.3a

Notes

[A] A unique, stable identifier for the GO term

[B] The primary GO term name

[C] The term definition, a full description indicating to what concept the term refers

[D] Term synonyms

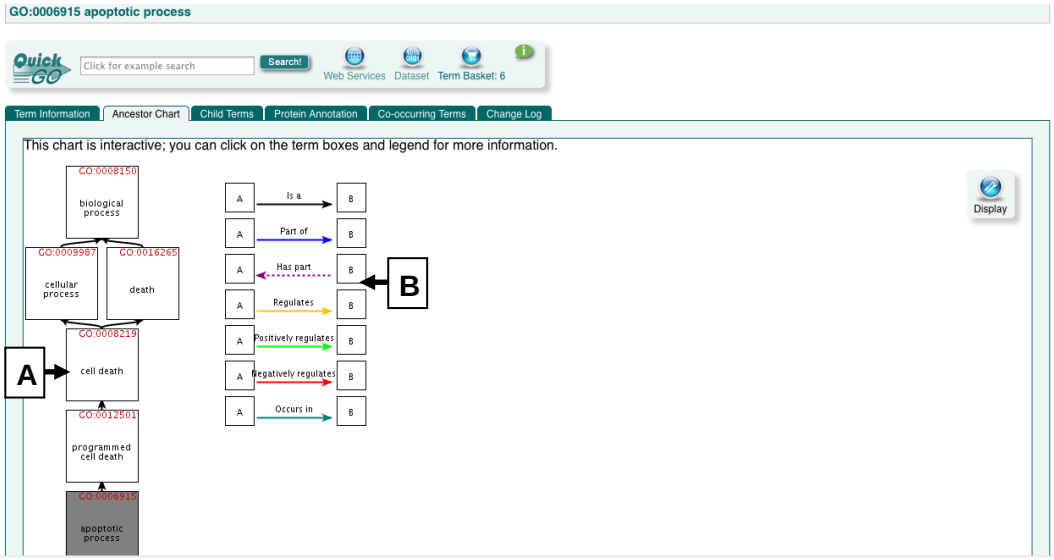


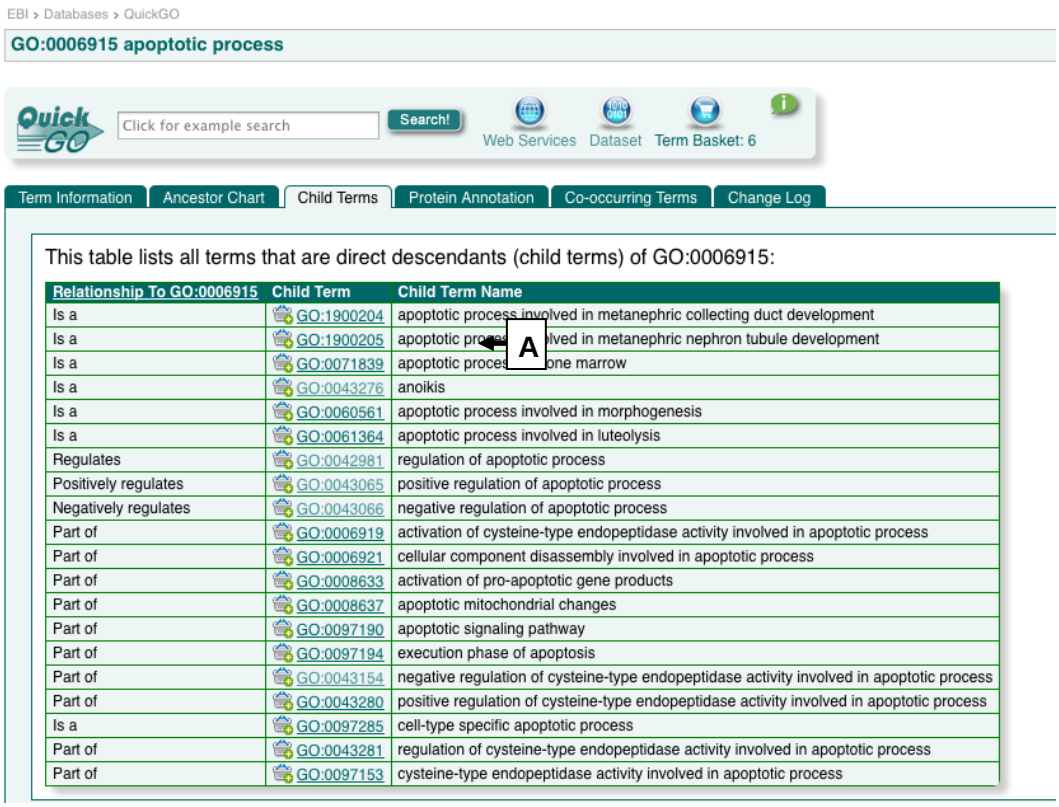
Fig. 1.3b Ancestor chart view of the GO term page.

Figure 1.3b

Notes

[A] A graphical display of the part of the Gene Ontology containing ancestor terms to the selected term.

[B] A colour-coded key to the relationships between each of the terms in the chart.



Relationship To GO:0006915	Child Term	Child Term Name
Is a	GO:1900204	apoptotic process involved in metanephric collecting duct development
Is a	GO:1900205	apoptotic process involved in metanephric nephron tubule development
Is a	GO:0071839	apoptotic process ... one marrow
Is a	GO:0043276	anoikis
Is a	GO:0060561	apoptotic process involved in morphogenesis
Is a	GO:0061364	apoptotic process involved in luteolysis
Regulates	GO:0042981	regulation of apoptotic process
Positively regulates	GO:0043065	positive regulation of apoptotic process
Negatively regulates	GO:0043066	negative regulation of apoptotic process
Part of	GO:0006919	activation of cysteine-type endopeptidase activity involved in apoptotic process
Part of	GO:0006921	cellular component disassembly involved in apoptotic process
Part of	GO:0008633	activation of pro-apoptotic gene products
Part of	GO:0008637	apoptotic mitochondrial changes
Part of	GO:0097190	apoptotic signaling pathway
Part of	GO:0097194	execution phase of apoptosis
Part of	GO:0043154	negative regulation of cysteine-type endopeptidase activity involved in apoptotic process
Part of	GO:0043280	positive regulation of cysteine-type endopeptidase activity involved in apoptotic process
Is a	GO:0097285	cell-type specific apoptotic process
Part of	GO:0043281	regulation of cysteine-type endopeptidase activity involved in apoptotic process
Part of	GO:0097153	cysteine-type endopeptidase activity involved in apoptotic process

Fig. 1.3c Child term view of the GO term page.

Figure 1.3b

Notes

[A] A list of direct children of the selected term, which can be used to browse 'down' the GO hierarchy to find more descriptive child terms.

Searching the UniProt-GOA database for GO annotations

The UniProt-GOA database contains annotations made not only by curators within the UniProt Consortium, but also by specialist groups and by members of the GO Consortium. Therefore, the UniProt-GOA database provides the most comprehensive set of **GO annotations** for over 390,000 species within UniProtKB.

QuickGO is linked from the UniProt website, which provides a condensed, representative set of **GO annotations** within UniProt records.

QuickGO can be used to search for annotations to single proteins or a group of proteins, for example a list of proteins obtained from a proteomics experiment. **GO annotations** for a single protein can be viewed in **QuickGO** by searching for various identifiers, e.g. UniProtKB accession number, NCBI Gene ID, RefSeq

accession, Ensembl ID, or a protein name (e.g. Exportin-1) in the search box of **QuickGO**. A summary of the matching accessions/protein names will be returned, by clicking on the UniProt accession of the protein you would like to view the protein annotation page of the chosen protein will be displayed, as shown in Fig. 1.4. Clickable links to more detailed information on **GO terms** and references etc. are provided within the data displayed on this page.

A summary table of **evidence codes** is provided in the Glossary (see Table 1). In addition, further information on **evidence codes** and qualifiers can be located in the GO Consortium documentation pages:

Evidence codes: http://www.geneontology.org/GO.evidence.shtml

Qualifier usage: http://www.geneontology.org/GO.annotation.shtml#qual

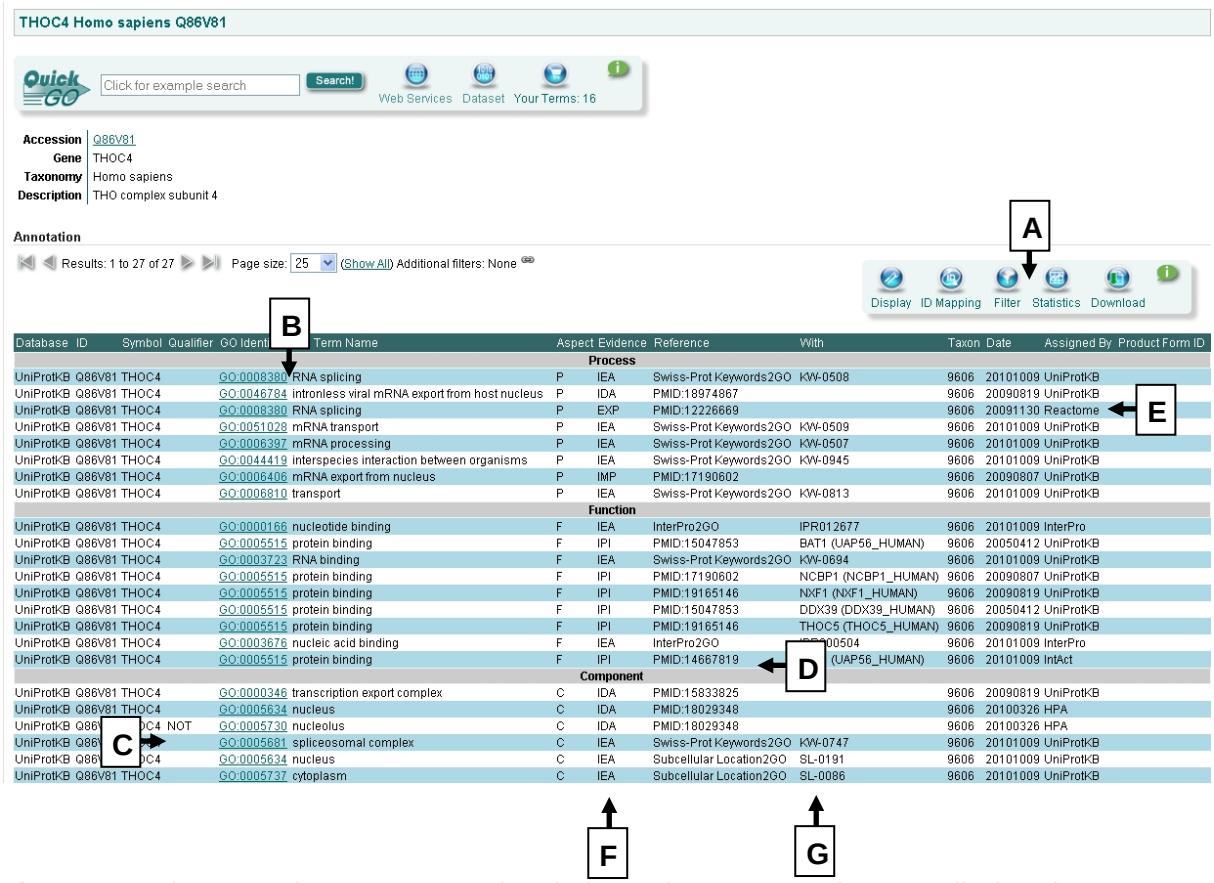


Fig. 1.4 Protein Annotation page: manual and electronic GO annotations are displayed for a queried protein together with supporting evidence codes, literature or electronic source.

Figure 1.4

Notes

[A] The Annotation Toolbar: buttons on this toolbar allow you to (i) customise your display of the annotation table, (ii) map between gene product identifiers, (iii) filter the annotation set, (iv) view the statistics associated with the annotation set, and (v) download the annotation set.

[B] Names and identifiers of GO terms that have been associated with the protein.

[C] Qualifier statements, which can alter the interpretation of the GO annotation.

[D] The reference cited as evidence to support the GO annotation. May be a literature reference (e.g. PubMed ID) or a database record (e.g. InterPro).

[E] Name of the database providing the annotation.

[F] Acronyms of GO evidence codes used to broadly categorise the types of evidence that have been found to support the association of the protein with the GO term. 'IEA' is the only electronic evidence code, all other codes are manually assigned (see Table 1 in the Glossary).

[G] 'With' data. Added to certain types of annotations to provide further information (e.g. for an InterPro2GO electronic annotation the InterPro domain that was mapped to GO is cited here).

Chapter 1: How to view GO data using the QuickGO browser - exercises

These exercises will familiarise you with the searching functionality of QuickGO.

Exercise 1 – searching for GO terms in QuickGO

1. *Open QuickGO at http://www.ebi.ac.uk/QuickGO (see Fig. 1.1).*
2. *To begin, try searching QuickGO by entering into the text box a biological process name, such as 'apoptosis'. Click 'Search'.*
3. *Click on one of the GO IDs listed.*
4. *Click through the green tabs to see what information each of them contains.*

Question 1: In the term page for 'apoptotic process' how many databases have cross-references for this term?

Question 2: How many 'part_of' child terms does 'apoptotic process' have?

Exercise 2 – searching for protein annotation in QuickGO

1. *Open QuickGO at http://www.ebi.ac.uk/QuickGO (see Fig. 1.1).*
2. *Enter into the search box the UniProt accession* Q86V81.
3. *You should see a page listing the matching protein, click on the UniProt accession to open the protein annotation page.*

Question 1: How many annotations in total does human THOC4 have? *Clue: Look for the 'Results' display.*

Question 2: What is the parent term of 'RNA splicing'? *Clue: Click on the GO ID accompanying this term.*

Question 3: What is the name of the InterPro domain that is the reference for the annotation to 'nucleotide binding'? *Clue: Follow the links.*

Chapter 2 – Using QuickGO to create a tailored set of annotations

It is possible within **QuickGO** to custom generate a set of annotations tailored to your specific requirements using extensive filtering options. Several aspects of **GO annotation** can be filtered such as taxonomic group, **evidence code**, GO ID and protein identifier, this makes **QuickGO** a uniquely powerful tool for biologists wishing to analyse specific sets of targets.

Chapter 2 - learning objectives

You will learn:

- How to filter annotations in the UniProt-GOA database to create a custom set of annotations
- How to view the statistics associated with a set of annotations

Filtering sets of annotations in QuickGO

The starting point for creating a subset of annotation is from **QuickGO**'s home page www.ebi.ac.uk/QuickGO (Fig. 1.1). Click on the link 'Search and Filter GO annotation sets', this takes you to the Annotation Download page containing all available **GO annotations** in the UniProt-GOA database. The table displays only the first 25 annotations by default, you can either page through the results using the arrows at the top of the table or increase the sample size using the box also located at the top of the table. All filtering options are located in the 'Filter' button on the Annotation Toolbar (Note [A] Fig. 2.1).

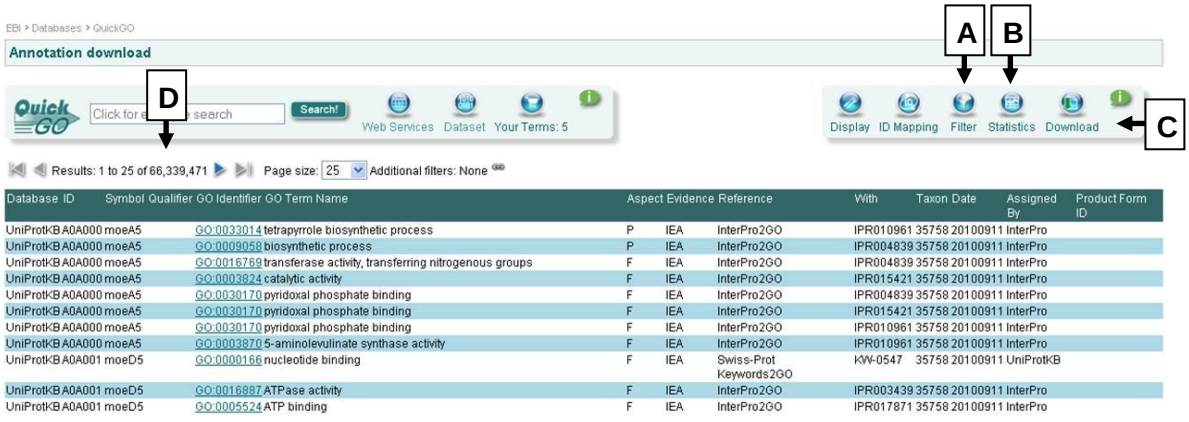


Fig. 2.1 Annotation Download page in QuickGO, the starting point for creating custom sets of GO annotations. Filtering options [A] and statistics [B] for the annotation set can be accessed via the Annotation Toolbar [C].

Figure 2.1

Notes

[A] Annotation sets can be filtered by clicking on the ‘Filter’ button. Filters include taxon, evidence code, GO ID and protein identifier.

[B] Statistics for the annotation set can be viewed by clicking the ‘Statistics’ button. Statistics are provided for counts of annotations and proteins for individual GO IDs, evidence codes, taxon IDs and sources of annotation as well as the number of unique protein accessions.

[C] The Annotation Toolbar.

[D] The total number of annotations in the set.

Clicking on the Filter button opens up a lightbox with the filtering options arranged as tabs in the window. Figure 2.2 shows the filter tab for Taxon, to retrieve annotations to a particular taxon you can either select one of the common taxon IDs from the list or enter an NCBI taxonomic identifier in the text box. Figure 2.3 shows the filter tab for Evidence where you can choose to see annotations made using certain evidence codes. For instance, it is quite common for users to remove annotations created using electronic methods, in which case you would select either ‘Manual Experimental’ or ‘Manual All’ from this filter tab. Further filtering can be done by selecting a relevant tab and entering your requirements. When you have chosen all your required filtering options, click on ‘Submit’ at the bottom of the window and the annotations will be retrieved.

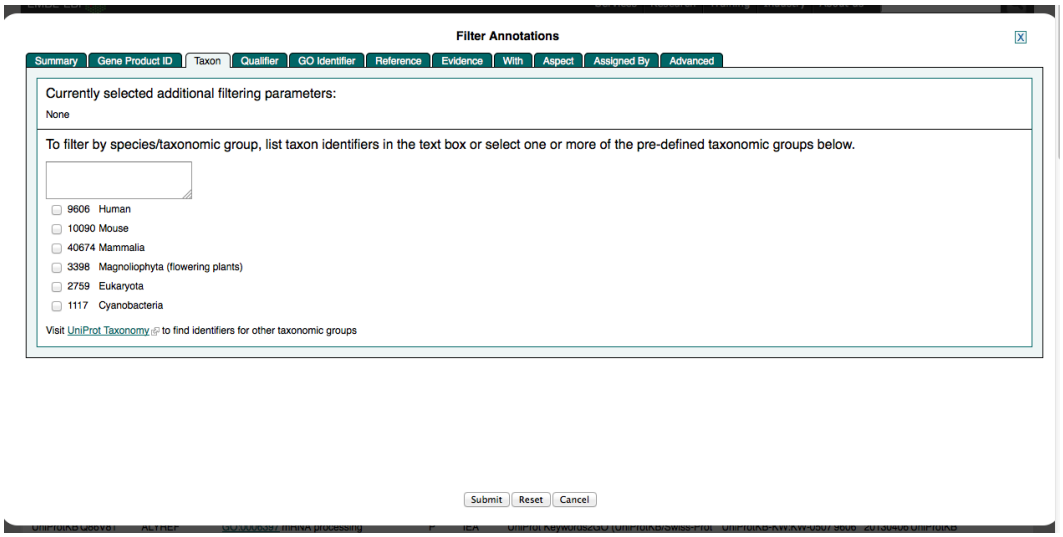


Fig. 2.2 Filter by ‘Taxon’ tab. Users can specify which taxon(s) they would like to see annotations for either by entering a list in the text box, or by selecting from the list.

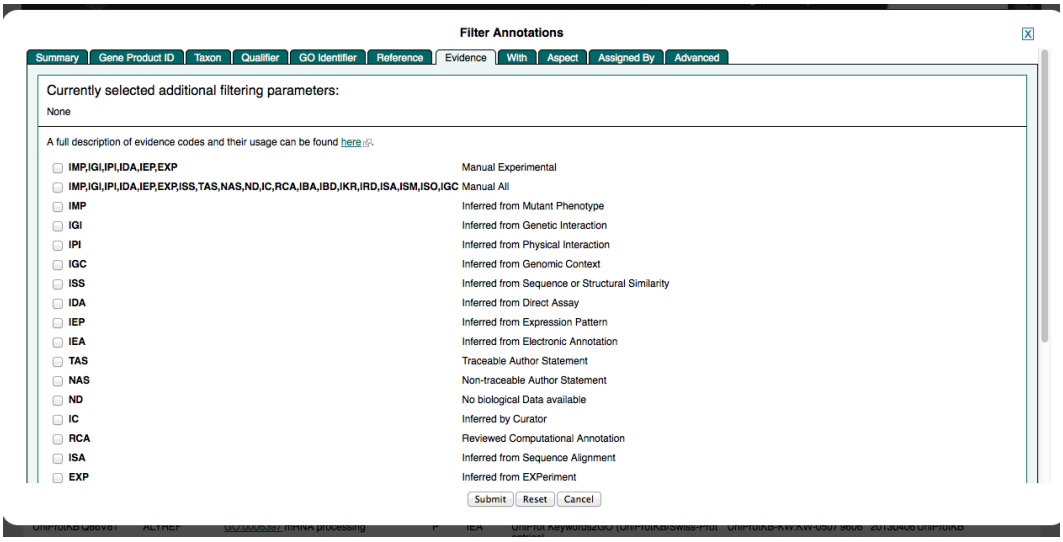


Fig. 2.3 Filter by 'Evidence' tab. Users can choose to see annotations that use only certain evidence codes by for either by selecting one or more from the list.

Viewing the statistics for a set of annotations in QuickGO

QuickGO calculates statistics for annotation sets 'on-the-fly' so are recalculated to reflect any filtering performed on the annotation set. Statistics are accessed from the 'Statistics' button on the Annotation Toolbar (Note [B] Fig. 2.1). Statistics can be obtained for counts of annotations and proteins for individual GO IDs, evidence codes, taxon IDs and sources of annotation, as well as the number of unique protein accessions, by clicking through the green tabs. N.B. The total number of annotations in the set is shown in the 'Summary' tab of the statistics section or, alternatively, at the top-left of the annotation table (Note [D] Fig. 2.1).

Clicking on the 'Statistics' button opens up a lightbox with the statistics options arranged as tabs in the window. Figure 2.4 shows the statistics for GO ID; on the left is the count of annotations per GO ID and on the right is the count of proteins per GO ID. The GO IDs are arranged in order of most used in the annotation set, e.g. in Fig. 2.4 the GO ID associated with the highest number of proteins in this set is GO:0016020 'membrane', totalling 18.25% of the proteins represented in this set. Figure 2.5 shows the statistics for evidence codes used in the set of annotations. In general, for most sets of proteins, the most common evidence code is Inferred from Electronic Annotation (IEA) simply because there are so many more electronic compared with manual annotations (139 million electronic compared with 1.25 million manual; April 2013).

The statistics are downloadable as an Excel file by clicking on the 'Download' button in the Statistics view. A bar chart is a common way of displaying the number of proteins associated with the GO IDs in an annotation set. This can be

done by using the statistics for percentage of proteins per GO ID to make a bar chart.

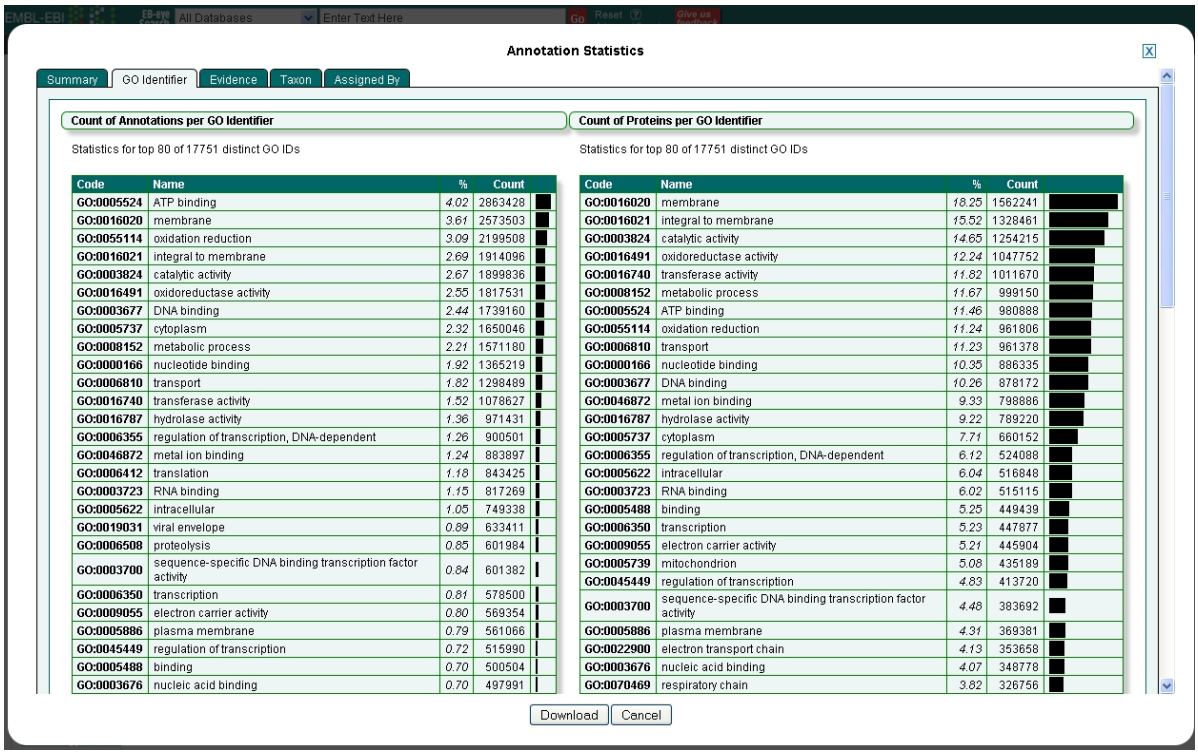


Fig. 2.4 GO ID statistics tab. Only the first 80 of the most common GO IDs are shown.

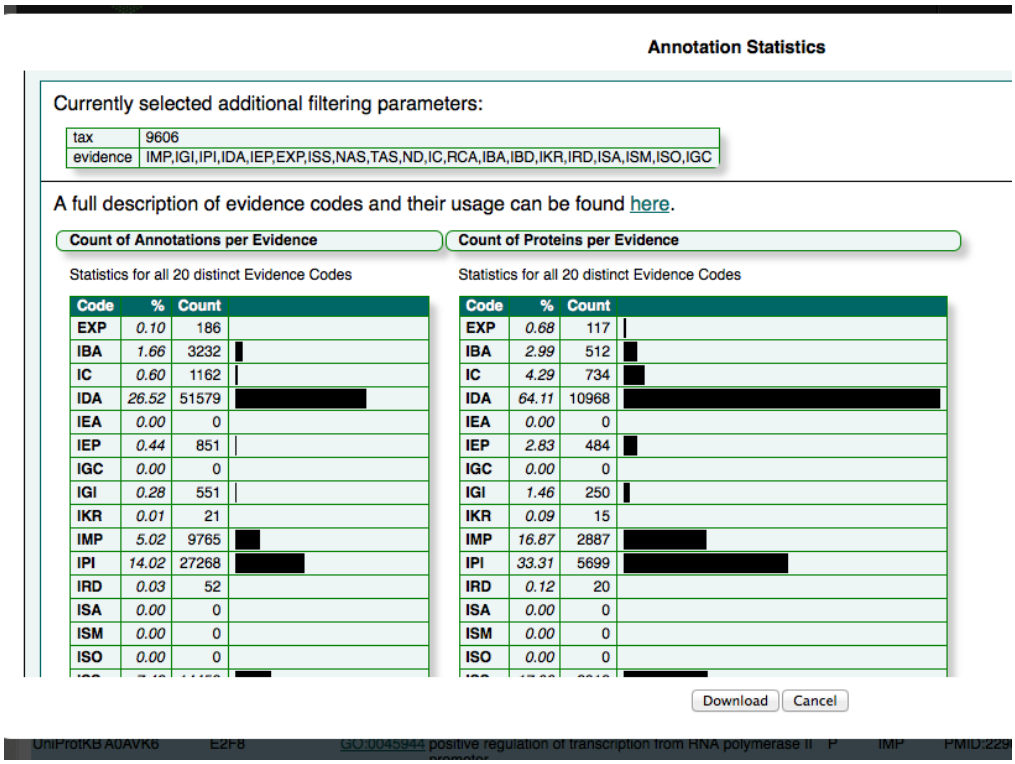


Fig. 2.5 Evidence code statistics tab. Displays the percentage and count of annotations and proteins for each evidence code used in the current set of annotations.

Chapter 2: Using QuickGO to create a tailored set of annotations - exercises

These exercises will demonstrate how to find GO annotations for a list of protein accessions, for example those obtained from a proteomics or microarray experiment, and to view the statistics of the final set of annotations.

Exercise 1 – Finding annotations for a list of protein accessions

The list of proteins used for this exercise constitute a subproteome of a Jurkat (T-cell leukaemia) cell line, originally published by Bantscheff *et al*. [2].

1. *Go to the Annotation Download page in QuickGO* (http://www.ebi.ac.uk/QuickGO/GAnnotation).
2. *Paste the 'quickgo_query.txt' list of UniProt accession numbers into the 'Gene Product ID' filter box. (This list can be found at the following URL:* ftp://ftp.ebi.ac.uk/pub/contrib/goa/Tutorial_Data*)* N.B. Do not tick any of the boxes below the text box.
3. *Click on 'Submit' to view the annotations to this list of proteins*. N.B. You may have to wait a few seconds for it to load.

Question 1: For this list of proteins, how many annotations are there using both manual and electronic evidence codes together? *Clue: See figure 2.1 [D].*

4. *Now filter these annotations to view only those made with a manual experimental evidence code using the 'Evidence' (evidence code) filter box.*

Question 2: For this list of proteins, how many annotations are there using only manual experimental evidence codes?

Exercise 2 – Viewing the annotation statistics for a list of protein accessions

1. *Use the set of annotations, filtered for manual experimental evidence codes, generated in Chapter 2, Exercise 1 to view the annotation statistics.*

Question 1: What is the GO term associated with the most proteins?

Question 2: What is the top three evidence codes used in the annotations?

Question 3: Which two annotation groups have made the most annotations for this set?

Question 4: How many proteins have electronic annotations only? *Clue: Compare total number of proteins for this set with that for the set without the manual experimental evidence filter selected.*

Chapter 3 – Using GO slims in QuickGO

GO slims are cut-down versions of the GO ontologies containing a subset of the terms in the whole GO. They give a broad overview of the ontology content without the detail of the specific fine-grained terms.

GO slims are particularly useful for giving a summary of the results of **GO annotation** of a proteome, list of genes from a microarray, or cDNA collection when broad classification of gene product function is required.

GO slims can be created by users according to their needs, and may be specific to species or to particular areas of the ontologies. There are several pre-defined GO slims available, for example a generic GO slim provided by the GO Consortium, a plant-specific slim provided by TAIR and a yeast-specific slim provided by SGD. These are available from the GO Consortium website (http://www.geneontology.org/GO.slims.shtml) or from within **QuickGO**. Alternatively, users can create their own GO slims.

Chapter 3 - learning objectives

You will learn:

- How to slim-up annotations to a subset of GO terms using QuickGO
- How to view the statistics associated with a slimmed set of annotations
- How to compare GO terms in an ontology view

Obtaining annotations to a GO slim

QuickGO allows users both to create their own GO slim by entering a list of GO IDs or by collecting **GO terms** whilst browsing the GO, and to use/modify the pre-defined GO slims.

The starting point for using or creating GO slims is from **QuickGO**'s home page www.ebi.ac.uk/QuickGO (Fig. 1.1), by clicking on the link 'Investigate GO slims'.

From the GO slim page (Fig. 3.1) you can either select a particular pre-defined GO slim or enter a list of GO IDs to create a custom GO slim.

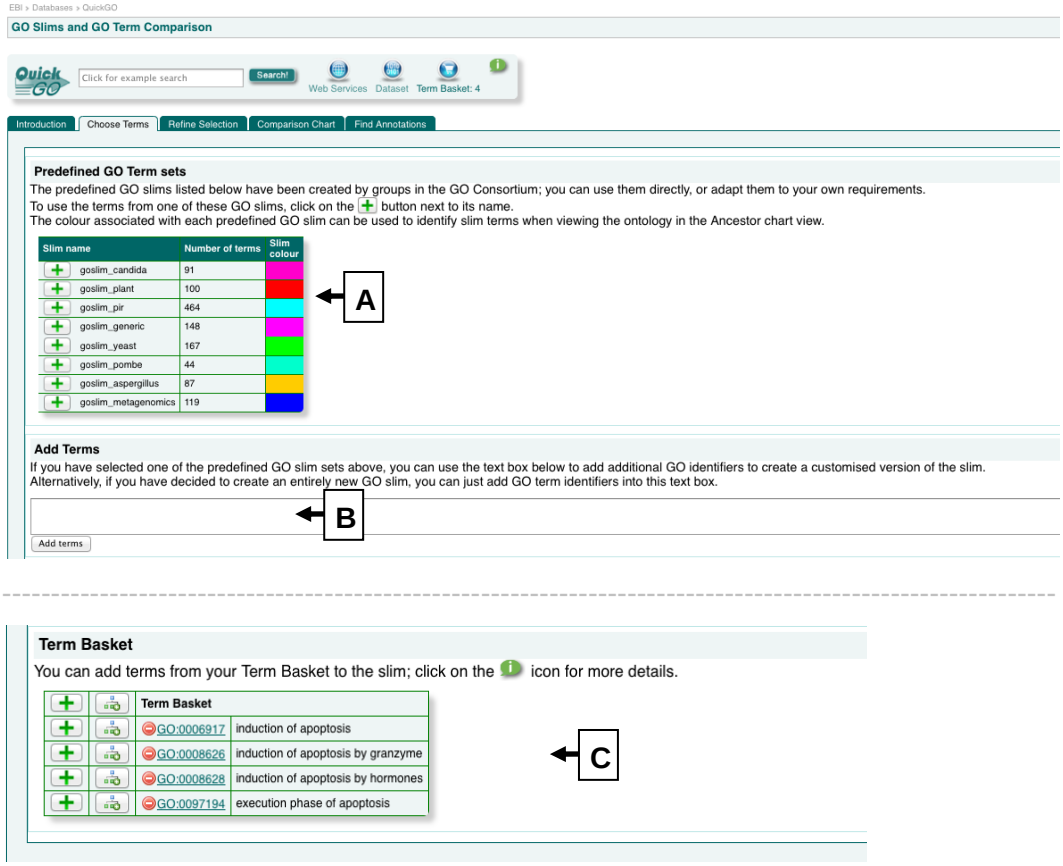


Fig. 3.1 GO Slims and GO Term Comparison page in QuickGO. From here you can create a custom set of GO terms or use a pre-defined set.

Figure 3.1

Notes

[A] To use a pre-defined set of GO terms, click on a green plus.

[B] Alternatively, GO IDs can be typed/pasted into the text box and added by clicking on 'Add terms'.

[C] Terms can also be added from the Term Basket by clicking on the green plus next to the term in the 'Choose Terms' tab of the GO Slims and GO Term Comparison page.

Another way of producing a set of **GO terms** that you would like to use as a GO slim is to collect terms as you browse **QuickGO**. Wherever you see the basket icon next to a term (see Note [A] Fig. 1.2a) you can use it to add the term to the Term Basket (Fig. 3.2), which is accessed from the Global Toolbar (Note [A] Fig. 1.1). Your collection of terms can then be used to either create a GO slim, to view/download annotations to those terms or to compare the terms in an ontology chart.

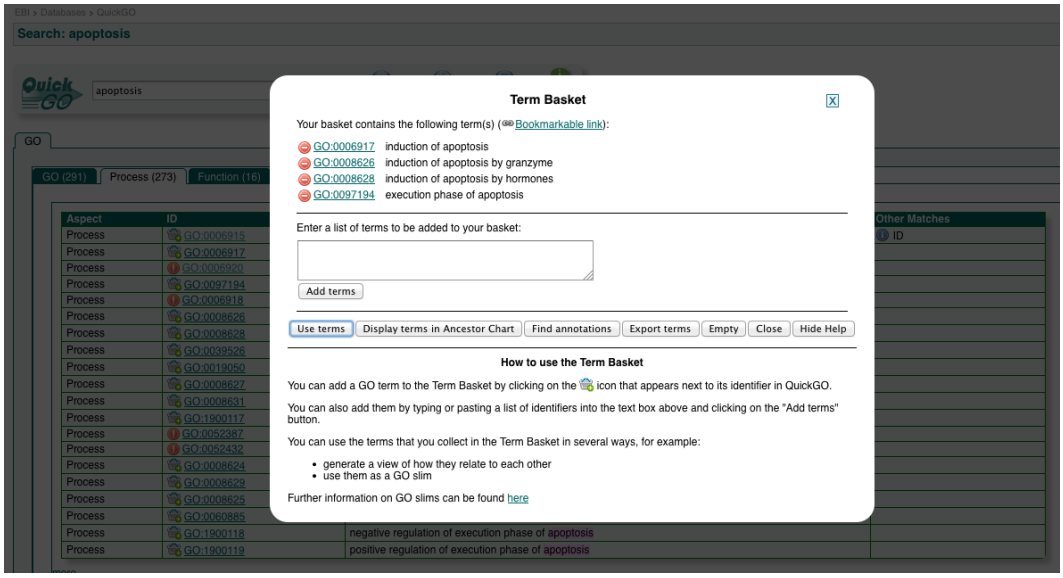


Fig. 3.2 The Term Basket contains the GO terms you have collected whilst browsing QuickGO. These can be used to create a GO slim, download annotations to these terms or compare the terms in an ontology chart.

Whether you select a predefined term set, add your own terms to the slim page or use your collection of terms, you will be directed to the ‘Refine Selection’ tab of the GO Slims and GO Term Comparison page (Fig. 3.3). Within this tab you can view the list of terms and add or remove terms as necessary. Once you have a finalised list of terms you can slim-up annotations to these terms by moving to the ‘Find annotations’ tab. The resulting table will contain all annotations mapped-up to the list of **GO terms** (Fig. 3.4). The usual procedure now would be to filter these annotations using a list of protein accessions that you are interested in. This procedure is detailed in Exercise 3.1. The GO ID statistics for protein count are useful for creating graphs or bar charts to represent the data as these tell you how many unique proteins in your list have annotations to the individual GO IDs represented in the annotation set.

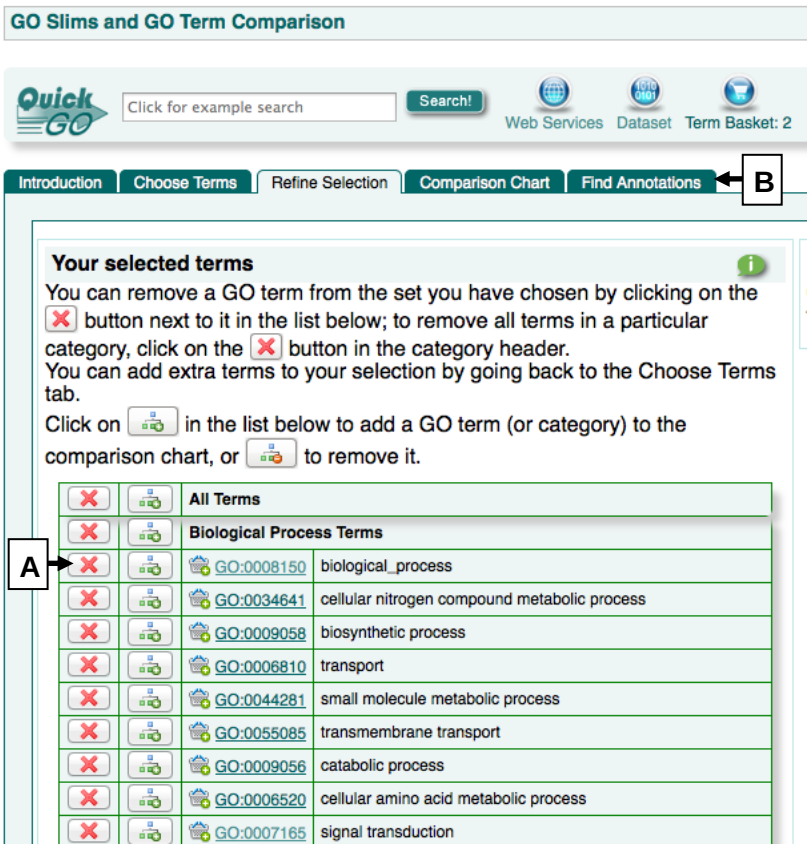


Fig. 3.3 GO Slims and GO Term Comparison page. GO terms can be either added to this list from 'Term Basket' selection or removed from the list, which can then be used in various ways such as slimming-up annotations (GO slim), finding proteins annotated to these terms or their descendents or, alternatively, comparing multiple terms as an ontology chart (see Fig. 3.7).

Figure 3.3

Notes

[A] To remove terms from the list, click on the red cross.

[B] To slim-up annotations to the terms in your list, click on the 'Find annotations' tab. This will result in a table of all the annotations in the UniProt-GOA database slimmed up to the selected terms (Fig. 3.4).

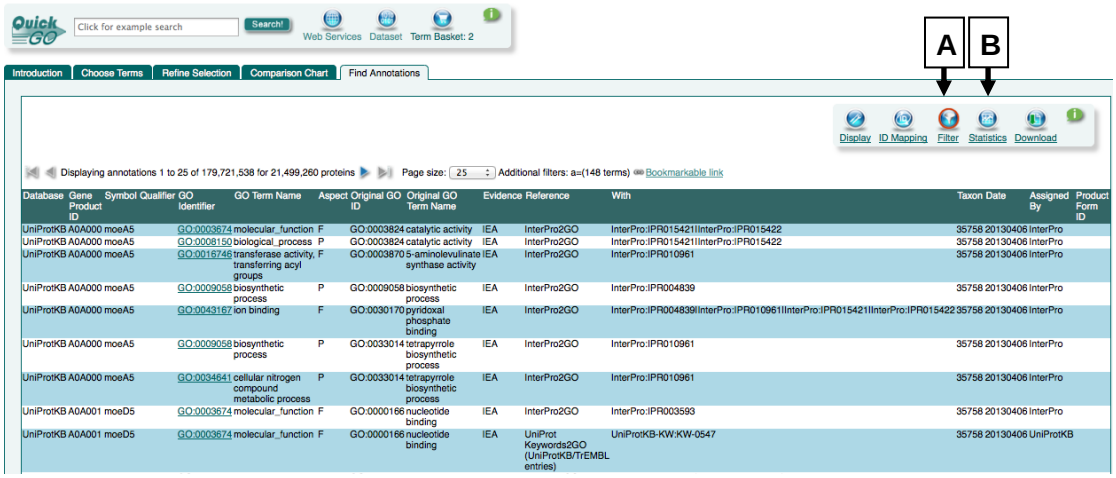


Fig. 3.4 Annotation table containing annotations mapped-up to the terms in the GO slim list.

Figure 3.4

Notes

[A] The 'Filter' button is highlighted with a red circle indicating a filter has been activated. If you click on the filter button and select the 'GO Identifier' tab (Fig. 3.5) you will see that the option to 'Use these terms as a GO slim' is selected by default.

[B] The 'Statistics' button. This provides statistics on the current annotation set and is updated with each filtering action. Statistics include counts of annotations and proteins for individual GO IDs, evidence codes, taxon IDs (Fig. 3.6) and sources of annotation, as well as the number of unique protein accessions.

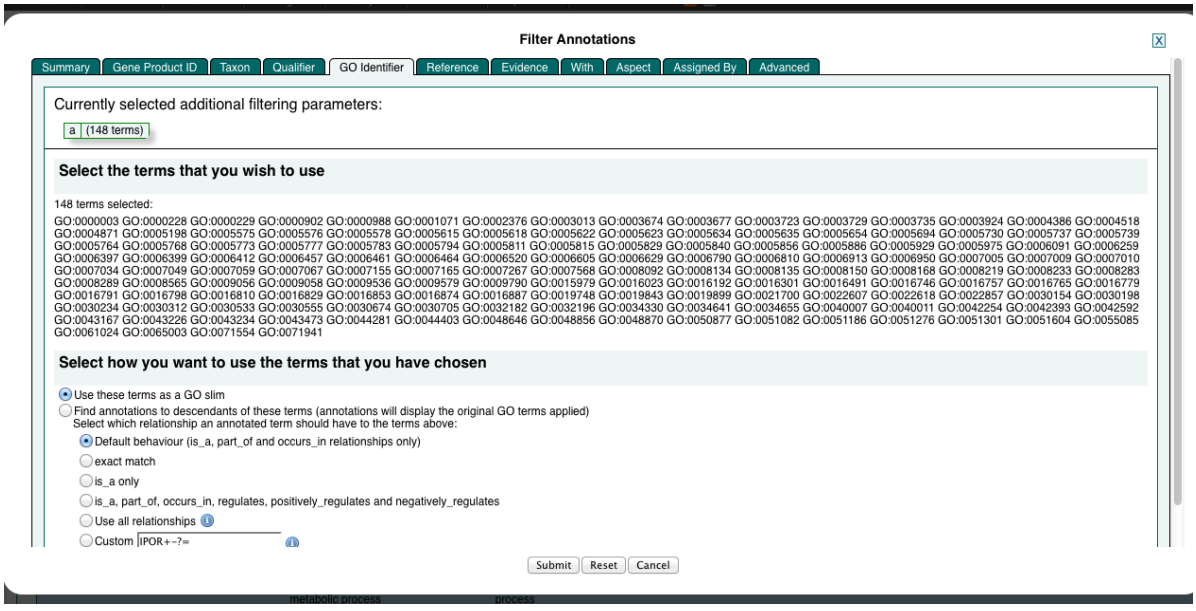


Fig. 3.5 The GO Identifier filter box. The two main options in this filter are to use the selected GO terms as a GO slim or to find annotations to the selected terms or their descendants. More advanced options for choosing which types of ontology relationships to use are also available.

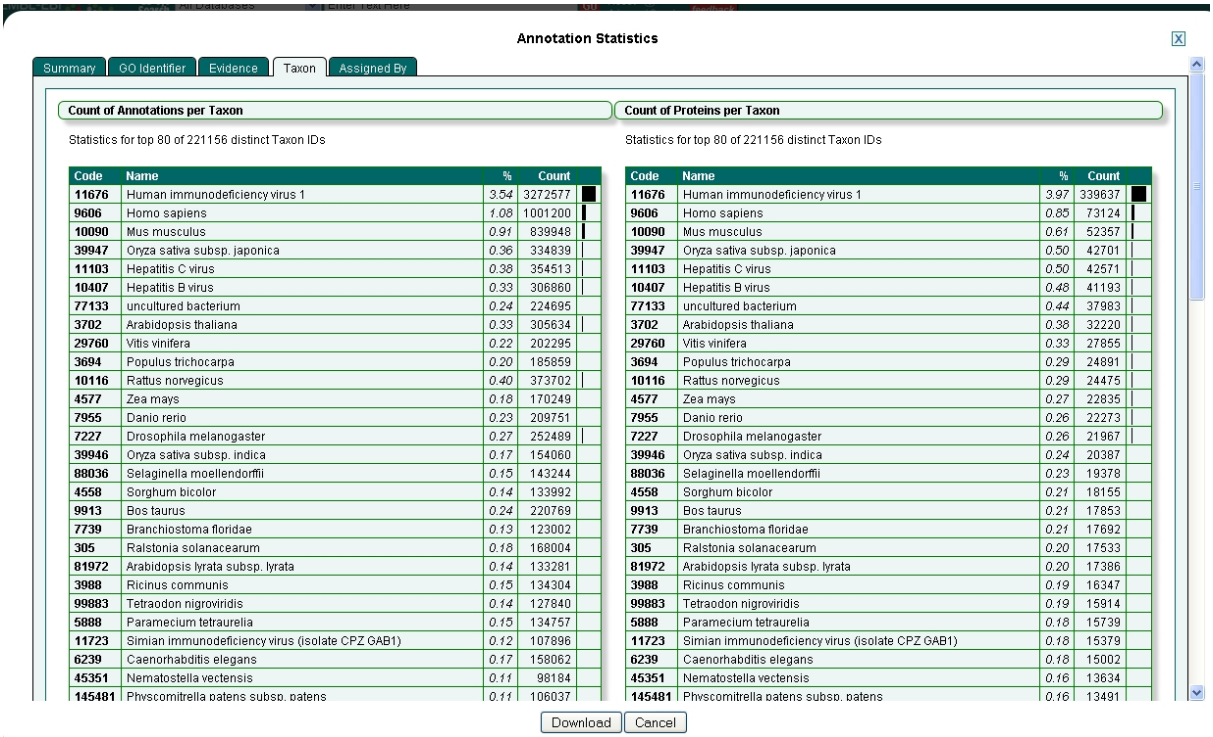


Fig. 3.6 The statistics for the taxon column. The count and percentage for both annotations and proteins per taxon is displayed. All of the statistics for the annotation set are downloadable in a text file format using the 'Download' button.

Comparing multiple terms in an ontology view

Another feature of the GO Slims and GO Term Comparison page is the ability to compare terms within the context of the GO hierarchy. Any GO term that is displayed on this page can be view in ontological context with any of the other terms on the page. To compare terms, you can either add the GO IDs to the text box in the 'Choose Terms' tab (see Fig. 3.1 Note [B]) or select terms you have collected in the Term Basket, which are also displayed in this tab (Fig. 3.1 Note [C]).

In the 'Refine Selection' tab of the GO Slims and GO Term Comparison page (Fig. 3.7), click on the chart icon next to each term (Fig. 3.7 Note [A]) to add the terms to the chart view, a larger version of which is available in the 'Comparison Chart' tab (Fig. 3.7 Note [C]). To remove terms from the chart, click again on the chart icon next to the term (Fig. 3.7 Note [B]).

When collecting terms in the Term Basket, click on the 'Display terms in Ancestor Chart' button that appears in the Term Basket lightbox (see Fig. 3.2) and you will be taken directly to the Ancestor Chart view with the chosen terms highlighted.

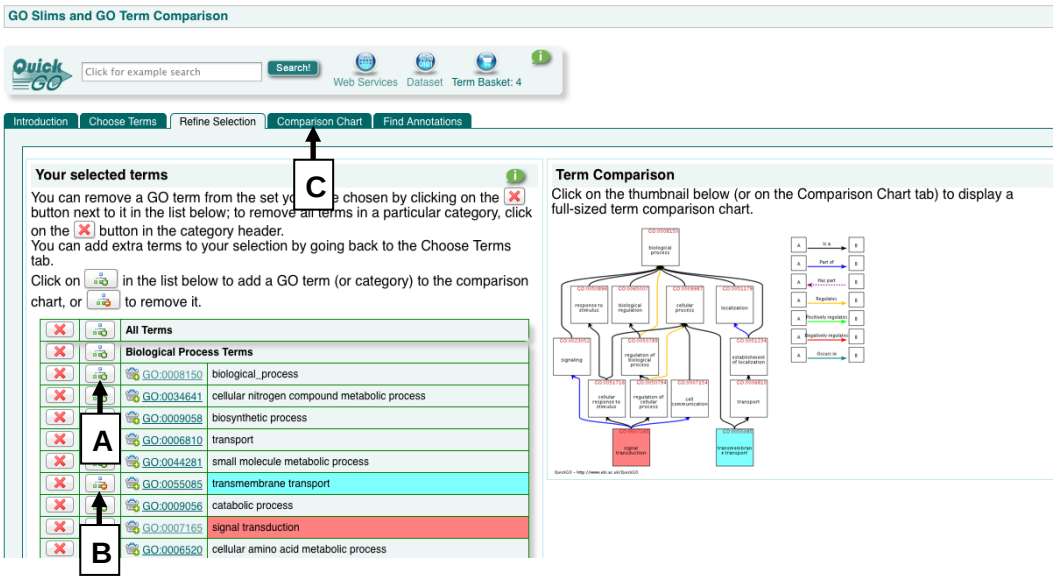


Fig. 3.7 Edit Terms tab of the GO Slims and GO Term Comparison page showing your selection of terms that may be viewed together in ontological context.

Figure 3.7

Notes

[A] To compare terms with each other as an ontology chart, click on the chart icon next to each GO ID you would like to compare. The chart containing the chosen terms will appear to the right of the page (a larger version can be seen in the 'Comparison chart' tab [C]) and the terms selected will be highlighted.

[B] To remove terms from the chart, click on the chart icon again.

[C] Comparison chart tab. A larger ontology view of the selected terms from [A].

Chapter 3: Using GO slims in QuickGO - exercises

Exercise 1 – Using QuickGO to slim-up annotations to a list of protein accessions

1. *Go to the GO slim entry page in QuickGO by clicking on the 'Investigate GO slims' link on the front page.*
2. *In the 'Choose Terms' tab, select the pre-defined 'goslim_generic' by clicking on the green plus next to its name*
3. *Click on the 'Find annotations' tab and wait for the results to load.*
4. *The resulting table contains all the annotations in the GOA database slimmed-up to the 148 terms in the GO slim. Go to the 'Filter' box in the*

Annotation Toolbar and ensure the 'Gene Product ID' tab is selected. Paste the 'quickgo_query.txt' protein list into the text box and click on 'Submit' to view the results. (*This list can be found at the following URL;* ftp://ftp.ebi.ac.uk/pub/contrib/goa/Tutorial_Data)

5. *The table now shows the annotations to the list of proteins slimmed-up to the 148 slim terms. Use the statistics calculated for this set to answer the following questions.*

Question 1: What are the GO terms associated with the most proteins in this set?

Question 2: Which evidence codes are the majority of annotations in this set made with?

Question 3: Which annotation groups have made the majority of annotations in this set?

Chapter 4 – How to retrieve bulk sets of GO annotations

While online browsers like **QuickGO** do provide detailed views for **GO annotations** to single or groups of proteins, many people like to computationally use the source data to provide a biological interpretation of large sets of genes or proteins to determine whether there is a common theme to a group of sequences, which will help in interpretation of an experiment.

Chapter 4 - learning objectives

You will learn:

- Where to retrieve complete GO annotation files (Gene Association Files) from all groups that produce them
- The format of the Gene Association File
- Where to find additional resources and files for computational analysis of GO annotations, e.g. identifier mapping files.

Each GO Consortium annotation group provides a file of annotations (called '**Gene Association Files'**) which are available to download from the GO Consortium site: http://www.geneontology.org/GO.current.annotations.shtml

This is the primary source of **GO annotation** and most annotation groups update their files monthly. There are two types of **gene association file** available on the GO Consortium website; filtered and unfiltered. The major difference between these is that filtered files are taxon-specific and provide comprehensive, non-redundant annotation files for a single organism. The unfiltered files can contain annotations for more than one taxon. There is further explanation of the filtering on the Gene Ontology Current Annotations website.

Gene Association Files can also be obtained from the websites of the groups that create them.

Web links to the current members of the GO Consortium can be found here; http://www.geneontology.org/GO.consortiumlist.shtml

UniProt-GOA provides species-specific, non-redundant **gene association files** to various species including human, mouse, rat, *Arabidopsis*, chicken, cow and zebrafish (http://www.ebi.ac.uk/GOA/downloads) as well as further annotation files for over 2000 proteomes (http://www.ebi.ac.uk/GOA/proteomes).

In addition to these files, we also provide a UniProt **gene association file,** which contains **GO annotation** to proteins from all species present in the UniProtKB.

In total, UniProt-GOA currently provides over 140 million annotations to over 390,000 species (April 2013 release).

Gene Association File Format

Gene Association Files have a simple, common format of 17 columns of annotation data that are tab-delimited (Fig. 4.1). The contents of the columns are described in Table 4.1.

1	2	3	4	5	6	7	8	9	10
Database	Object ID	Object Symbol	Qualifier	GO ID	Reference	Evidence	With' Column	Aspect	Object Name
UniProtKB	O35245	Pkd2		GO:0006874	PMID:12874107	IMP		P	Polycystin-2
UniProtKB	P98194	ATP2C1	NOT	GO:0005388	PMID:16192278	IDA		F	Calcium-transpo
UniProtKB	A0A0Y5	A0A0Y5		GO:0004634	GO_REF:0000002	IEA	InterPro:IPR000941	F	Enolase A0A0Y5
UniProtKB	Q29460	PAFAH1B3		GO:0005515	PMID:11522926	IPI	UniProtKB:P68401	F	Platelet-activati
UniProtKB	Q15459	SF3A1		GO:0000398	PMID:9731529	IC	GO:0005681	P	Splicing factor 3A
UniProtKB	Q8NFP9	NBEA		GO:0012505	GO_REF:0000024	ISS	UniProtKB:Q9EPN1	C	Neurobeachin
UniProtKB	Q8BTE6	1100001G20Rik		GO:0050872	PMID:18492766	IDA		P	Novel protein

11	12	13	14	15	16	17
Object Synonym	Object Type	Taxon ID	Date	Source DB	Annotation Extension	Gene Product Form ID
Pkd2|IPI00314	protein	taxon:10090	20031014	MGI		
ATP2C1|HUSSY-2	protein	taxon:9606	20061124	UniProtKB		UniProtKB:P98194-2
	protein	taxon:405230	20101016	InterPro		
PAFAH1B3|PAFAH	protein	taxon:9913	20101018	IntAct		
SF3A1|SAP114|IF	protein	taxon:9606	20060220	HGNC		
NBEA|BCL8B|KIA	protein	taxon:9606	20041006	UniProtKB		
RP23-430I21.5-0	protein	taxon:10090	20090316	MGI	occurs_in(CL:0000001)	

Fig. 4.1 Gene Association File format. Columns in blue are required information.

Table 4.1 Gene Association File format. Description of columns.

1. Database	The database from which the Object ID is drawn.
2. Object ID	A unique identifier in the database for the item being annotated.
3. Object Symbol	A unique symbol to which the Object ID is matched.
4. Qualifier	Flags that modify the interpretation of an annotation, e.g. NOT, contributes_to, colocalizes_with
5. GO ID	The GO identifier for the term attributed to the Object ID.
6. Reference	The source cited as an authority for the attribution of the GO ID to the Object ID, e.g. PubMed ID or a database record.
7. Evidence	The evidence code indicates the type of evidence that supports the GO annotation.
8. With or From	Required only for some evidence codes, can be, e.g. a database gene ID, sequence ID or GO ID.
9. Aspect	One of P (Biological Process), F (Molecular Function) or C (Cellular Component).
10. Object Name	Name of gene or gene product.
11. Object Synonym	Any synonym of a gene or gene product.
12. Object Type	The entity that is being annotated, e.g. protein, gene.
13. Taxon ID	The ID of the species encoding the gene product. In certain cases, such as interaction between organisms, two taxon IDs can be piped.
14. Date	The date on which the annotation was made.
15. Source DB	The database which made the annotation.
16. Annotation Extension	Contains cross-references to other ontologies that can be used to qualify or enhance the annotation. The cross-reference is prefaced by an appropriate GO relationship.
17. Gene Product Form ID	Identifier of the specific variant of the DB Object ID in Column 2, e.g. a protein isoform.

Additional resources to help users

1. A public GO MySQL Mirror now offers a remote connection to a regularly updated mirror of the GO schema, so via the command line it will be:

   $ mysql –hmysql.ebi.ac.uk –ugo_select –pamigo –P4085 latest_go

   To see when it was last built:

   SELECT * FROM instance_data

   There are example SQL queries on the GO Consortium website:

   http://wiki.geneontology.org/index.php/Example_Queries

2. UniProt provides a number of different services to help users map between different database identifiers.

   A number of different identifier mapping tools are available:

- PICR - Protein Identifier Cross-Reference Service http://www.ebi.ac.uk/Tools/picr/
- The UniProt website now also offers database identifier mapping, go to: http://uniprot.org/ and click on the 'ID Mapping' tab.

Database identifier mapping files are also available:

- Between UniProtKB and NBCI's UniGene identifiers:

  ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/gp2protein/gp2protein.unigene.gz

- Between UniProtKB and NCBI's EntrezGene identifiers:

  ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/gp2protein/gp2protein.geneid.gz

- Between UniProtKB and NCBI's RefSeq identifiers:

  ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/gp2protein/gp2protein.refseq.gz

If you require any more information on these files you can mail goa@ebi.ac.uk. The readme for these files can be found at:
http://www.ebi.ac.uk/GOA/goaHelp

3. File giving details on the annotation methods that have **not** been referenced in the annotations with public paper identifier such as PubMed identifiers (GO References):

http://www.geneontology.org/cgi-bin/references.cgi

4. File to map between GO identifiers and **GO term** names:

http://www.geneontology.org/doc/GO.terms_and_ids

Chapter 5 – Using InterProScan to rapidly populate novel sequences with electronic GO annotation predictions

http://www.ebi.ac.uk/Tools/pfa/iprscan/

Chapter 5 - learning objectives

You will learn:

- How to use the InterProScan service from InterPro to find GO annotations to novel sequences

The InterPro2GO mapping provides the highest electronic annotation coverage, supplying over 60% of UniProtKB proteins with electronic GO annotation [3].

InterPro curators create GO mappings by manually assigning **GO terms** to those InterPro identifiers that can correctly describe the function of all proteins in the UniProt/Swiss-Prot database that possess the same InterPro domain. **GO annotations** are then automatically applied to all UniProtKB proteins possessing the same InterPro identifier. The InterPro2GO mapping file is available at:

http://www.geneontology.org/external2go/interpro2go

The InterProScan service (Fig. 5.1) applies the protein signature recognition methods from InterPro member databases to user-provided genomic or protein sequences. This service also provides a quick way of obtaining electronic GO annotations to novel sequences, from the integrated InterPro2GO mapping data. This is useful for those species that do not have a dedicated Model Organism Database or group to provide manual annotations to these gene products. This

service is free to all academic and commercial users and offers interactive or e-mail job submissions.

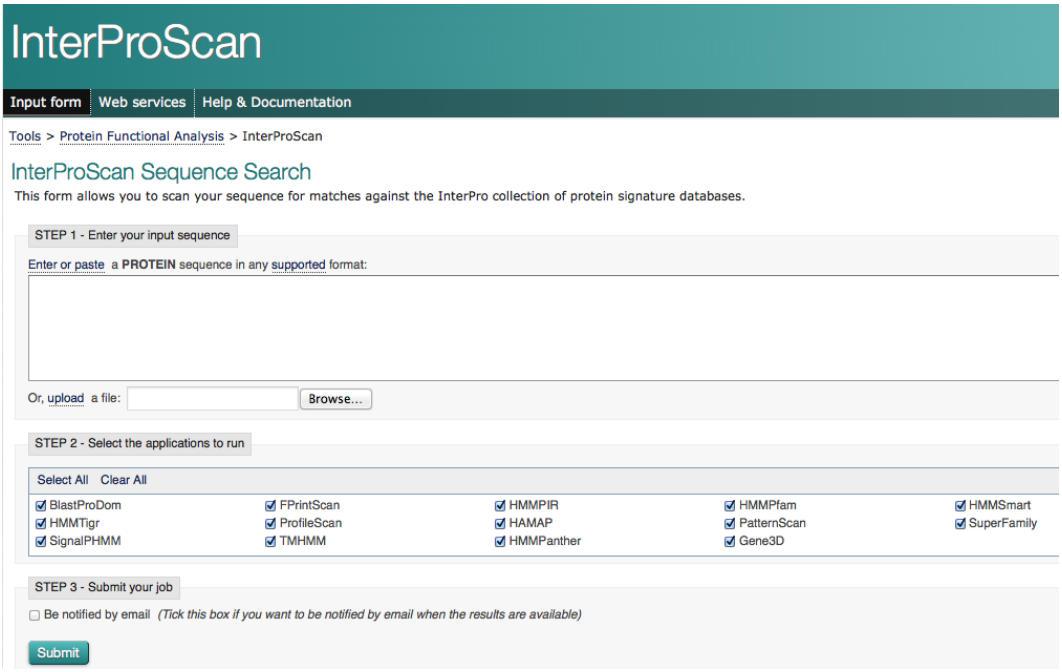


Fig. 5.1 InterProScan home page (http://www.ebi.ac.uk/Tools/pfa/iprscan/)

Chapter 5: Using InterProScan to rapidly populate novel sequences with electronic GO annotation predictions - exercises

Exercise 1 – Using InterProScan to find GO annotations for a novel sequence

1. *Copy and paste the protein sequence provided in the file 'interpro_query.txt' (also available from http://www.uniprot.org/uniprot/B2JN45.fasta) into the InterProScan input box and click 'Submit'.*
2. *On the results page, click on the 'Summary Table' button to show the GO annotation suggested by each match.*

Question 1: What functional annotation does InterProScan predict for this protein sequence?

Chapter 6 – Using GO annotation data to link biological knowledge to a set of proteins

Chapter 6 - learning objectives

You will learn:

- What to consider when choosing a suitable GO analysis tool
- Basic GO term enrichment analysis using g:Profiler

The use of high-throughput technologies, such as gene expression and systems biology as investigative tools is gaining momentum, and many users of GO are interested in evaluating a list of genes to test for the statistically significant over- or under-representation of particular pathways and functions. Such enrichment analysis often relies on the availability of gene function, process and subcellular location annotations produced by the Gene Ontology Consortium. See example case study: http://www.geneontology.org/GO.immunology.casestudy.shtml

Groups of sequences may show a correlation between their expression profiles and the GO category they are annotated to for several reasons. They may represent close family members with similar functions, genes in the same pathway or genes in alternative pathways that perform the same type of biological function.

A wide range of tools are available to analyse lists of sequence identifiers. The majority of GO tools have been developed by third parties.

A good GO tool should, at the very least:

- Be actively maintained/developed
- Provide reproducible results
- Take into account the GO Directed Acyclic Graph (DAG)
- Consider evidence codes
- Consider the importance of the 'NOT' qualifier in GO annotations
- Provide details on the version of the GO used (and carry out frequent data updates).
- Provide details on the source of the GO annotation set used (and carry out frequent data updates).
- Provide good documentation

The GO Consortium maintains a list of analysis tools in association with Neurolex (http://neurolex.org/wiki/Category:Resource:Gene_Ontology_Tools) that satisfy

most of these requirements and a recent review of tools is also available to help users [4].

Analysis with g:Profiler

http://biit.cs.ut.ee/gprofiler/

g:Profiler [5] is a public web server for characterising and manipulating lists of sequence identifiers. The tool is developed and maintained by researchers at the Institute of Computer Science, University of Tartu, Estonia.

If you have any questions about the tool, please contact the developers through their website: http://biit.cs.ut.ee/gprofiler/welcome.cgi?t=contact

g:Profiler consists of four interactive modules, however this tutorial will only use the g:GOSt module for functional profiling of a list of UniProtKB accessions with terms from the GO Molecular Function, Biological Process and Cellular Component ontologies, KEGG and Reactome pathways. The g:GOSt tool was chosen for this tutorial as it fulfils the GO tool requirements listed above, and can very quickly provide a highly informative visual presentation of the profiling results. Numerous GO tools exist which are freely available that provide users with different analyses, input and output options (see the Neurolex tools page http://neurolex.org/wiki/Category:Resource:Gene_Ontology_Tools and 4 in the 'Further Reading' section).

Lists of gene, protein or probe identifiers can be entered into the Query box on the front page of g:Profiler (Fig. 6.1).

Chapter 6: Using GO annotation data to link biological knowledge to a set of proteins - exercises

Exercise 1 – Using g:Profiler to perform GO term enrichment analysis on a list of protein accessions

1. *Copy and paste the list of human protein accessions in 'gprofiler_query.txt' file provided into the Query box.*
   (*This list can be found at the following URL;* ftp://ftp.ebi.ac.uk/pub/contrib/goa/Tutorial_Data)

2. *As all of the accessions provided are human select 'Homo sapiens' in the organism box.*

3. *To see the most significant terms in order of P-value, de-select the 'Hierarchical sorting' button, if this button is selected the results will be shown sorted by GO domain.*
4. *Leave all other options as the provided default.*
5. *Click 'g:Profile!'*

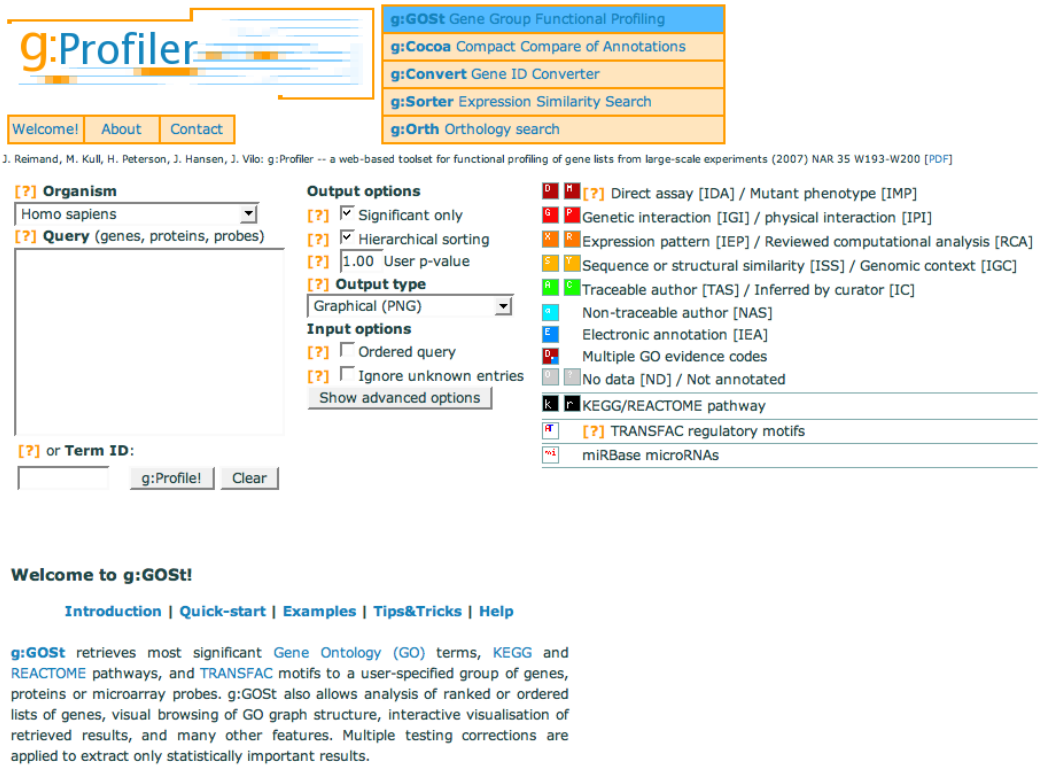


Fig. 6.1 The g:GOSt query interface from g:Profiler.

Question 1: What term(s) appear to be over-represented within the queried protein set? *Clue: See Fig. 6.2 for the output from this query.*

Fig. 6.2 Output of the g:GOSt analysis, showing enriched functional terms from GO and other relevant biological databases for the queried proteins. GO terms are shown in a tree-like top-down group order, grouped either by domain or ranked by statistical significance. Each term is accompanied by the size of the query and term gene lists, their overlap and the statistical significance (p-value) of such enrichment. The column numbers are explained in the section entitled ‘g:GOSt output explained’.

g:GOSt output explained

1. The rows of boxes indicate what annotation data is available for each queried protein. If a box is coloured, this indicates that annotation(s) have been found to link the protein and the GO term in the corresponding row. Boxes are coloured differently depending on the evidence category that the supplied annotation was found to have. Where multiple annotations were found to support the association between GO term and protein, 3/4 of the square is filled with the colour representing the highest quality evidence code, and 1/4 with second-best evidence code 'colour'.
2. P-value**.** The statistical significance of a GO term being associated with the set of identifiers queried. The accompanying red horizontal bars represent the p-value strength. The darker red the lines, the stronger the p-value evidence.
3. Term size: the total number of sequences that have been annotated to the corresponding GO term displayed to the right of the screen. This number is used as the background count for p-value calculation. *In this worked example, these values will be the total number of human sequences annotated to the represented GO terms, as obtained by the tool's analysis of the human gene association file.*
4. Query size: the number of sequence identifiers being analysed by the user (with an ordered query, these Q values represent the number of queried proteins in a group which provide the best p-value).
5. The number of genes from the query that have been annotated to the corresponding term.
6. The proportion of the query annotated with a given term. This corresponds to the value in column 5 divided by column 4. This value is called the precision, or positive prediction value.
7. Proportion of all genes annotated to a given term (sensitivity).
8. GO stable term identifiers.
9. Domain of a term group, either MF (molecular function), BP (biological process), CC (cellular component) for GO.
10. Name of term and a number displaying the term's depth in local hierarchy. In case of hierarchical sorting, terms are preceded with spaces according to their relative depth in the hierarchy. The displayed section of the GO hierarchy is always relative to p-value threshold, and terms with p-values above the threshold are not shown.

Chapter 7 – Case study for the course

Case study 1 – How to find GO annotation in a particular area of biology using QuickGO

Many users of the GO are research scientists interested in a certain area of biology. One common question is “How do I find the gene products that are involved in a particular process?”. The following case study is taken from a user query that was sent to the UniProtKB-GOA group.

Topics covered include:

- Customizing a set of annotations by filtering on **GO term** and taxonomic identifier.
- Mapping annotations to a different sequence identifier.
- Reviewing a filtered annotation set using **QuickGO**'s annotation statistics.
- Downloading a filtered protein list.

Scenario:

“I'm currently working on zebrafish, and I would like to get a list of all genes implied in the development. What is the easiest way to get that list? The best for me would be to get Ensembl IDs of the genes, but other IDs would be ok.”

The user may find that he has to go to several different resources to find the information he requires; however, this question can be answered completely and very easily with **QuickGO** alone by utilizing its filtering and identifier mapping capabilities. Here, users can choose to see annotations with identifiers such as UniProtKB, Ensembl, RefSeq, FlyBase, etc.

Here is how to obtain these results with **QuickGO**:

(i) The starting point for obtaining a custom set of annotations is the Annotation Download (http://www.ebi.ac.uk/QuickGO/GAnnotation) page. This page displays a table of an initial set of all **GO annotations** available from UniProt-GOA (Fig. 2.1). This page is also accessed from the ‘Search and Filter GO annotation sets’ link on the front page of **QuickGO**. The annotations in this table can be filtered according to your needs by using the ‘Filter’ button on the Annotation Toolbar (see Fig. 2.1 Note [A]).

(ii) To obtain the set of annotations the user requires, the annotations must be filtered by database identifier, taxon and **GO term**. First, click on the 'Filter' button on the Annotation Toolbar, select the 'Taxon' tab and type '7955', the taxon identifier for *Danio rerio*, into the text box. Then select the 'GO Identifier' tab and type 'GO:0032502' the GO identifier for 'developmental process' into the free text box, ensuring the option to 'Find annotations to descendants of these terms' is selected.

(iii) Click on 'Submit' to load the set of annotations to 'developmental process' or its child terms for zebrafish.

(iv) To map the UniProt accessions, which are displayed by default, to Ensembl identifiers click on the 'ID mapping' button on the Annotation Toolbar. Select which type of Ensembl identifier you require (either gene, protein or transcript ID) and click on 'Submit' to load the annotations.

(v) Note that there are likely fewer annotations in the mapped set because not all of these UniProt accessions can be mapped to Ensembl. Initially, only the first 25 annotations will be displayed in the table, to see more annotations you can either page through or increase the number of annotations per page using the drop-down menu at the top of the table. Additionally, the whole set of annotations can be downloaded in various formats by using the 'Download' button on the Annotation Toolbar (Fig. 2.1 Note [C]); when downloading, always ensure that you have increased the download limit to equal to or greater than the number of annotations in the set.

(vi) To view or download the annotation or protein statistics associated with the annotation set, use the 'Statistics' button located on the Annotation Toolbar. From here you can see the most common **GO terms** appearing in the set (Fig. 7.1), how many distinct proteins are represented in the set (Fig. 7.2) and what types of evidence codes were used for the annotations.

(vii) In the original query the user required a list of Ensembl identifiers. A list of unique identifiers present in this annotation set can be downloaded by selecting the 'proteinList' download option. To ensure the entire list of sequence identifiers is included in the download, increase the Download 'limit' to the number of distinct proteins in the set; this number is shown in the dialog of the Download box.

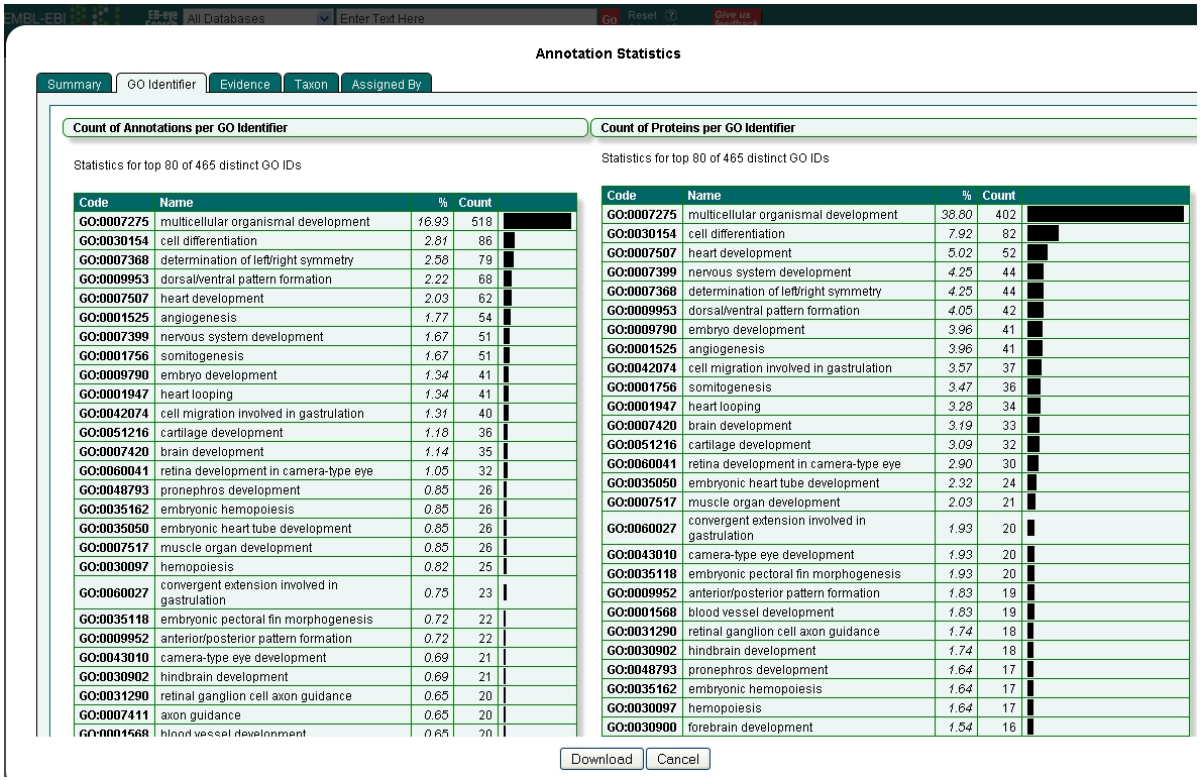


Fig. 7.1 The GO Identifier statistics associated with the set of annotations to developmental processes for zebrafish Ensembl identifiers

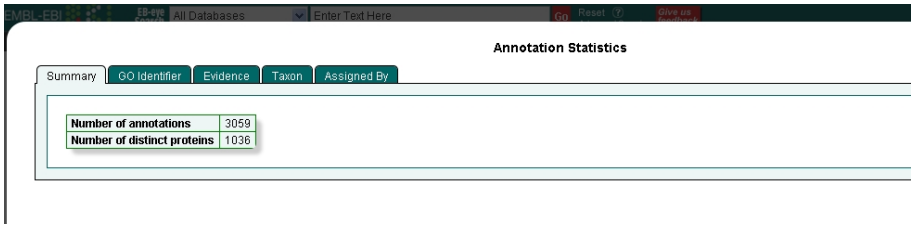


Fig. 7.2 The summary statistics for the set of annotations to developmental processes for zebrafish Ensembl identifiers, showing total number of annotations and number of distinct protein identifiers.

Therefore, in these few simple steps we have been able to retrieve a set of annotations that would have either taken several resources to obtain or more advanced computing knowledge to extract the information from a **gene association file**.

Chapter 8 – Programmatic access to GO terms and annotations

QuickGO can supply GO term information and GO annotation data via REST web services.

Documentation and example client code is available at;

http://www.ebi.ac.uk/QuickGO/WebServices.html

GTerm webservice

To access GO term information as a web service, the URI template is;

GTerm?id=*termID*&format=format

Format	Description
mini	Mini HTML, suitable for dynamically embedding in popup boxes
obo	OBO format snippet
oboxml	OBO XML format snippet

Example:

```
#!/usr/bin/perl
use LWP::UserAgent;
use XML::XPath;
use XML::XPath::XMLParser;
my $ua = LWP::UserAgent->new;
my $req = HTTP::Request->new(GET => 'http://www.ebi.ac.uk/QuickGO/GTerm?id=GO:0003824&format=oboxml');
my $res = $ua->request($req, "term.xml");
my $xpath = XML::XPath->new(filename => "term.xml");
my $nodeset = $xpath->find("/obo/term/name");
foreach my $node ($nodeset->get_nodelist) {
  print XML::XPath::XMLParser::as_string($node) . "\n";
}
exit;
```

GAnnotation webservice

To access GO annotation information as a web service, the URI template is;

GAnnotation?format=*download-format*&limit=*max-number-of-rows*{&gz}{&field=value{,value}}

The description of the field names and acceptable values are available at;

http://www.ebi.ac.uk/QuickGO/WebServices.html

Example:

```
#!/usr/bin/perl
use LWP::UserAgent;
my $ua = LWP::UserAgent->new;
my $req = HTTP::Request->new(
  GET => 'http://www.ebi.ac.uk/QuickGO/GAnnotation?protein=P12345,Q4VCS5&format=gaf&limit=-1');
my $res = $ua->request($req, "gene_association.example");
open (GAF, "gene_association.example");
while (<GAF>) {
  next if /^!/;
  chomp;
  my (
    $db, $db_object_id, $db_object_symbol, $qualifier, $go_id, $reference, $evidence, $with_string,
    $aspect, $db_object_name, $synonym, $db_object_type, $taxon_id, $annotation_date, $assigned_by,
    $annotation_extension, $gene_product_form_id
  ) = split(/\t/);
  print "$db_object_id => $go_id $evidence $reference $with_string $assigned_by\n";
}
close GAF;
exit;
```

Chapter 8: Programmatic access to GO terms and annotations - exercises

Exercise 1 – Using QuickGO web service to find specific annotations

1. Using the QuickGO web services, retrieve the following;

   Annotations to the GO term ‘kidney development’ (GO:0001822) for human proteins (taxon = 9606) that use the evidence codes ‘IDA’ and ‘IMP’.

Question 1: How many annotations were retrieved?

Whole course summary

After completing this course you should be familiar with the content and structure of the Gene Ontology and how **GO terms** are associated with gene products. You should understand the composition of a **GO annotation** and the different methods by which they are created. You should be able to retrieve **GO annotations** from several sources and have a basic understanding of the use of GO in a biological context.

We have shown how to search for both **GO terms** and **GO annotations** using the EBI's GO tool **QuickGO**. In addition, you should be able to use **QuickGO** both to make a custom set of annotations using the extensive filtering options built into **QuickGO**, and to slim-up a set of **GO annotations** for a particular set of protein identifiers.

We have hopefully shown that **QuickGO** is a simple, yet powerful tool for viewing and querying **GO annotations** associated with the hundreds of thousands of species in the UniProt Knowledgebase.

We have shown how you can use InterProScan to populate novel sequences with **GO annotation**, which is useful for species which do not have a dedicated manual GO annotation effort, and have introduced you to one of the many GO analysis tools that are available to researchers who would like to characterise discrete sets of genes or proteins.

Finally, we have demonstrated how to programmatically access GO term and annotation data via QuickGO web services.

Glossary

GO annotation

A GO annotation is the assignment of a GO identifier with a particular sequence (either a gene or protein database identifier). Manual annotations are created by a curator having directly looked for functional information (either within published literature or by examining the sequence directly), whereas electronic annotations are produced by a number of different types of automated methods that produce high-quality, conservative predictions of GO assignments.

All annotations must provide both a reference to a source that provides information either directly for the GO term-gene product assignment or the method used to create the assignment, and also an evidence code (see below).

GO term

The Gene Ontology is a controlled vocabulary of GO terms which describe a particular attribute of a gene product within three categories; Molecular Function, Biological Process and Cellular Component (Subcellular Location). Each GO term has a unique, computer-readable ID and has a definition. GO terms may also have cross-references to external databases that describe an identical or similar concept, e.g. the Enzyme Commission.

QuickGO

The Gene Ontology browser developed by the UniProt-GOA group at the EBI. Within QuickGO the user is able to view GO terms and all associated term information and protein annotation, view GO annotations for single or lists of proteins, customise sets of GO annotation using extensive filtering options and use pre-existing or create new GO slims for use in summarising the functional information for a list of proteins. Sets of annotations and their associated statistics are available for download.

Evidence codes

Every GO annotation must indicate the type of evidence that supports it; these evidence codes correspond to broad categories of experimental or other support. The codes are listed in Table 1 together with some examples (not exhaustive lists) of the kinds of experiments that each code covers.

Full information on evidence codes can be found on the GO Consortium website; http://www.geneontology.org/GO.evidence.shtml

Evidence code	Full Name	Examples of Usage in Annotations
IDA	Inferred from Direct Assay	Enzyme assays In vitro reconstitution (transcription) Immunofluorescence or cell fractionation (for cellular component) Physical interaction/binding assay
IEP	Inferred from Expression Pattern	Transcript levels (e.g. Northern blotting, microarray data) Protein levels (e.g. Western blotting)
IGI	Inferred from Genetic Interaction	'Traditional' genetic interactions such as suppressors, synthetic lethals Functional complementation Rescue Experiments Inference from the phenotype of a mutation in a different gene.
IMP	Inferred from Mutant Phenotype	Any gene knockout/mutation Overexpression/ectopic expression of a wild-type or mutant gene Anti-sense experiments Specific protein inhibitors
IPI	Inferred from Physical Interaction	Immunoprecipitation Binding assay (ion/protein) Two-hybrid interactions Co-purification Co-immunoprecipitation
ISS	Inferred from Sequence Similarity	Sequence or structural similarity Recognised domains
RCA	Inferred from Reviewed Computational Analysis	Predictions from large-scale experiments (e.g. genome-wide two-hybrids, genome-wide synthetic interactions) or text-based computation (e.g. text mining)
IGC	Inferred from Genomic Context	Operon structure Syntenic regions Pathway analysis Genome-scale analysis of processes
TAS	Traceable Author Statement	Referenced statement on experimental findings which is traceable to a primary paper (e.g. a review paper)
NAS	Nontraceable Author Statement	Author statement that is not supported by any references, such as a hypothesis by the author or information from a paper's abstract.
ND	No Biological Data available	Where no biological data found (this is only used with 'unknown' GO terms).
IEA	Inferred from Electronic Annotation	Computational techniques (e.g. GO mappings, function prediction based on sequence similarity or pattern matching).

Table 1. GO evidence codes. Descriptions of a selection of GOC-agreed evidence codes, which describe the broad categories of evidence found to support a GO term-gene product association. Apart from the 'IEA' code, all others are used in manual annotation.

Gene association file

A file representing annotation data using a tab-delimited format, where each line represents a single association between a gene product (protein, gene, transcript, etc.) and a GO term with a certain evidence code and the reference to support the association. A guide to the format of gene association files can be viewed on the GO Consortium website:
http://www.geneontology.org/GO.format.annotation.shtml

Further reading

1 Alternative, major GO browsers:

AmiGO (official GO Consortium browser):
http://amigo.geneontology.org/cgi-bin/amigo/go.cgi

Ontology Lookup Service:
http://www.ebi.ac.uk/ontology-lookup/

OBO-Edit:
https://sourceforge.net/project/showfiles.php?group_id=36855&package_id=192411

2 Bantscheff M, Eberhard D, Abraham Y, Bastuck S *et al*. (2007) **Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors.** *Nat. Biotechnol.* **25**, 1035-1044.

3 Quevillon, E, Silventoinen, V, Pillai, S, Harte, N, Mulder, N, Apweiler, R and Lopez, R (2005) **InterProScan: protein domains identifier.** *Nucleic Acids Res.* **33,** W116-W120.

4 Khatri, P. and Draghici, S. (2005) **Ontological analysis of gene expression data: current tools, limitations, and open problems**. *Bioinformatics* **21,** 3587-3595.

5 Reimand, J., Kull, M., Peterson, H., Hansen, J. and Viol, J. (2007) **g:Profiler – a web-based toolset for functional profiling of gene lists from large-scale experiments.** *Nucleic Acids Res.* W193-W200.

6 Lomax J, The Gene Ontology Consortium (2005) **Get ready to GO! A biologist's guide to the Gene Ontology.** *Brief Bioinform*. ***6***: 298-304.

7 Barrell D, Dimmer E, Huntley RP, Binns D, O'Donovan C, Apweiler R (2008) *The GOA database in 2009--an integrated Gene Ontology Annotation resource. Nucleic Acids Res.* ***37***, D396-403.

8 Dimmer EC, Huntley RP, Barrell DG, Binns D, Draghici S, Camon EB, Hubank M, Talmud PJ, Apweiler R, Lovering RC (2008) **The Gene Ontology - Providing a Functional Role in Proteomic Studies.** *Proteomics*. Jul 17 [Epub ahead of print].

9 Huntley RP, Binns D, Dimmer E, Barrell D, O'Donovan C, Apweiler R (2009) **QuickGO: a user tutorial for the web-based Gene Ontology browser.** *Database*. Sep 29. doi: 10.1093/database/bap010

Where to find out more

UniProt-GOA: http://www.ebi.ac.uk/GOA/

GO Consortium: http://www.geneontology.org/

UniProt: http://www.uniprot.org/

InterPro: http://www.ebi.ac.uk/interpro/

How to feedback or contribute annotations

If you find that the Gene Ontology is missing terms that fall within the ontology's scope, you can contribute GO content by entering suggestions into the GO ontology tracker on the SourceForge site:

http://sourceforge.net/tracker/?group_id=36855&atid=440764

Any questions concerning the GO should be e-mailed to: go-helpdesk@ebi.ac.uk

Regarding GO annotations:

1. if you have found that your protein set has not been fully annotated
2. you would like to contribute annotation data
3. you would like to be added to UniProt-GOA's expert panel for reviewing final sets of annotations,

then you can let us know by either emailing us at goa@ebi.ac.uk or alternatively fill in the GOA web form, at: http://www.ebi.ac.uk/GOA/contactus

A manual GO annotation project is underway in the area of cardiovascular processes. To find out more, please see the wiki; http://wiki.geneontology.org/index.php/Cardiovascular

‘Gene Ontology Annotation at the EBI’ course exercise answers

Chapter 1: How to view GO data using the QuickGO browser - answers

Exercise 1 answers: searching for GO terms in QuickGO

(correct at time of writing: April 2013)

Question 1: In the term page for ‘apoptotic process’ how many databases have cross-references for this term?

Answer 1: Three; InterPro, UniProtKB and Wikipedia

Question 2: How many ‘part_of’ child terms does ‘apoptotic process’ have?

Answer 2: Eight

Exercise 2 answers: searching for protein annotation in QuickGO

(correct at time of writing: April 2013)

Question 1: How many annotations in total does human THOC4 have? *Clue: Look for the ‘Results’ display.*

Answer 1: 64 annotations

Question 2: What is the parent term of ‘RNA splicing’? *Clue: Click on the GO ID accompanying this term.*

Answer 2: ‘RNA processing’.

Question 3: What is the name of the InterPro domain that is the reference for the annotation to ‘nucleotide binding’? *Clue: Follow the links.*

Answer 3: IPR012677 Nucleotide-binding, alpha-beta plait.

Chapter 2: Using QuickGO to create a tailored set of annotations - answers

Exercise 1 answers: finding annotations for a list of protein accessions

(correct at time of writing: April 2013)

Question 1: For this list of proteins, how many annotations are there using both manual and electronic evidence codes together?

Answer 1: 16,354 annotations

Question 2: For this list of proteins, how many annotations are there using only manual experimental evidence codes?

Answer 1: 4,140 annotations

Information

From these answers you can see that the majority of annotations for this set of proteins (and generally for most sets of proteins) are produced through electronic methods. Therefore this is a very powerful type of method of creating high-quality GO annotations in a short amount of time.

Exercise 2 answers: Viewing the annotation statistics for a list of protein accessions

(correct at time of writing: April 2013)

Question 1: What is the GO term associated with the most proteins?

Answer 1: GO:0005515 protein binding

Question 2: What are the top three evidence codes used in the annotations?

Answer 2: IDA, IPI and IMP

Question 3: Which two annotation groups have made the most annotations for this set?

Answer 3: IntAct and UniProt

Question 4: How many proteins have electronic annotations only? *Clue: Compare total number of proteins for this set with that for the set without the manual experimental evidence filter selected.*

Answer 4: 29 proteins

Chapter 3: Using GO slims in QuickGO - answers

Exercise 1 answers: Using QuickGO to slim-up annotations to a list of protein accessions

(correct at time of writing: April 2013)

Question 1: What are the GO terms associated with the most proteins in this set?

Answer 1: Biological process, molecular function and cellular modification process

Question 2: Which evidence code(s) are the majority of annotations in this set made with?

Answer 2: IEA: Inferred from Electronic Annotation

Question 3: Which annotation groups have made the majority of annotations in this set?

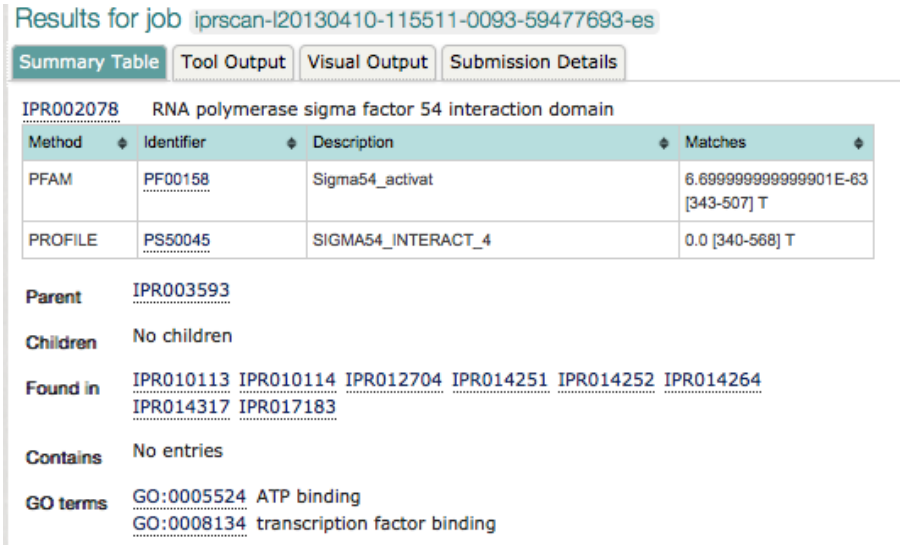
Answer 3: UniProt and Reactome

Chapter 5: Using InterProScan to rapidly populate novel sequences with electronic GO annotation predictions - answers

Exercise 1 answers: Using InterProScan to find GO annotations for a novel sequence

Question 1: What functional annotation does InterProScan predict for this protein sequence?

Answer 1: ATP binding, transcription factor binding and sequence-specific DNA binding transcription factor activity, nucleotide binding and nucleoside-triphosphate activity (snippet shown below)



Chapter 6: Using GO annotation data to link biological knowledge to a set of proteins - answers

Exercise 1 answers: Using g:Profiler to perform GO term enrichment analysis on a list of protein accessions

 (correct at time of writing: April 2013)

Question 1: What term(s) appear to be over-represented within the queried protein set?

Answer 1:

Molecular Function: GO:0004672 protein kinase activity;
Cellular Component: GO:0005829 cytosol;
Biological Process: GO:0044267 cellular protein metabolic process

Chapter 8: Programmatic access to GO terms and annotations - answers

Exercise 1 answers – Using QuickGO web service to find specific annotations

 (correct at time of writing: September 2012)

Question 1: How many annotations were retrieved?

Answer 1: 38 annotations

Contributors

Rachael Huntley, UniProt-GOA group