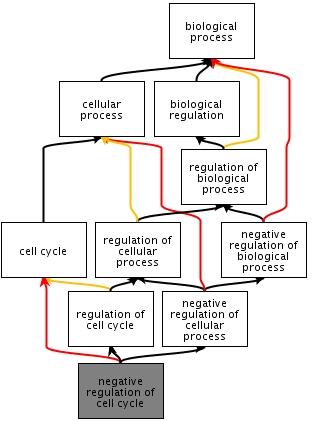
An Introduction to the Gene Ontology (GO) and Gene Ontology Annotations



The Bioinformatics Roadshow

Düsseldorf, Germany

March 15th 2011

Description: Information Information

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

# Learning Objectives

The aim of this tutorial is to familiarise users with the Gene Ontology (GO) and associations of **GO terms** with gene products (**GO annotations**).

You will learn how to:

* Use the **QuickGO** tool to view GO terms.
* Use the **QuickGO** tool to view GO annotations, filter annotations to create a tailored set, and use **GO slims** to summarise the attributes of a gene product.
* Retrieve complete sets of **GO annotations.**

# 1 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** that describe the biological concepts. The **GO terms** are linked to each other using six existing relationships; *is\_a, part\_of, has\_part, regulates, positively\_regulates* and *negatively\_regulates*. For more information on the relationships, see the documentation on the GO Consortium website:

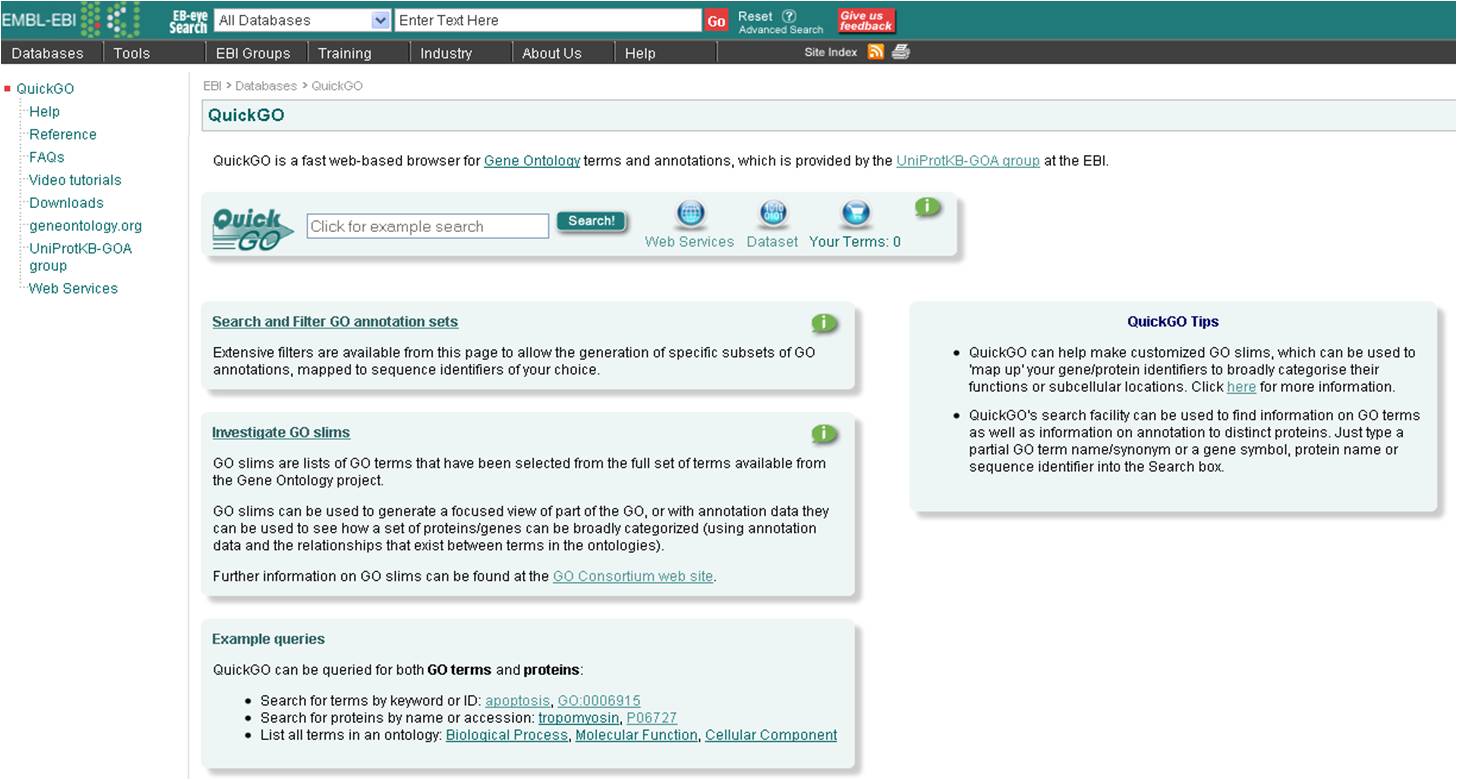
<http://www.geneontology.org/GO.ontology-ext.relations.shtml>.

# **1.1 Viewing GO terms using the QuickGO browser**

To browse the GO hierarchy or to view annotations for individual gene products, a number of online tools are available, such as: **QuickGO** (<http://www.ebi.ac.uk/QuickGO>), which has been developed at the EBI, and AmiGO (<http://amigo.geneontology.org>) which is developed by the GO Consortium. For other GO browser tools, please see the ‘Further Reading’ section located at the end of this tutorial.

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

The **QuickGO** home page (Fig. 1.1) provides a text box [A] to start searching for GO information. You may search for any aspect of a **GO annotation** including **GO term** names and synonyms, GO IDs, UniProtKB accessions, InterPro IDs, Enzyme Commission (EC) numbers, and UniProtKB 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.



**A**

**B**

**C**

**D**

**E**

**Figure 1.1 QuickGO query interface (**[**http://www.ebi.ac.uk/QuickGO**](http://www.ebi.ac.uk/QuickGO)**)**

**[A]** The ‘Global Toolbar’ is visible on all pages within QuickGO. From here you can search QuickGO, access QuickGO Web Services, view the dataset that QuickGO is currently using and view ‘Your Terms’ selection. See Section 1.2 and Section 3 for more information.

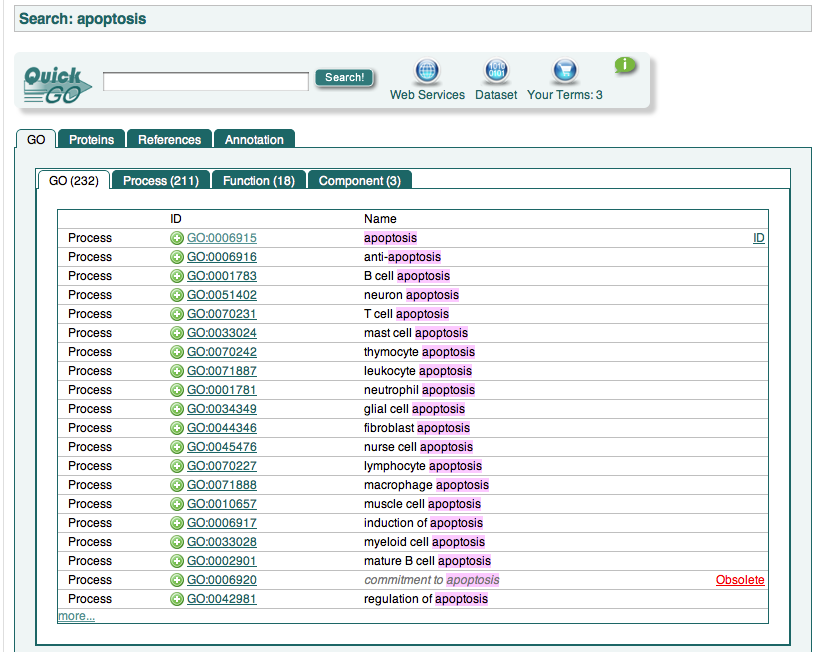
**[B]** The entry point for viewing, filtering and downloading annotations from the GOA database.

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

**[D]** Examples of simple queries that can be performed in QuickGO.

**[E]** Useful tips for using GO and QuickGO.

A search for ‘apoptosis’ retrieves terms if the word ‘apoptosis’ is present in the term name, synonyms, definition or cross-references (Fig. 1.2).

****

**A**

**B**

**C**

**D**

**Figure 1.2. A search for ‘apoptosis’.**

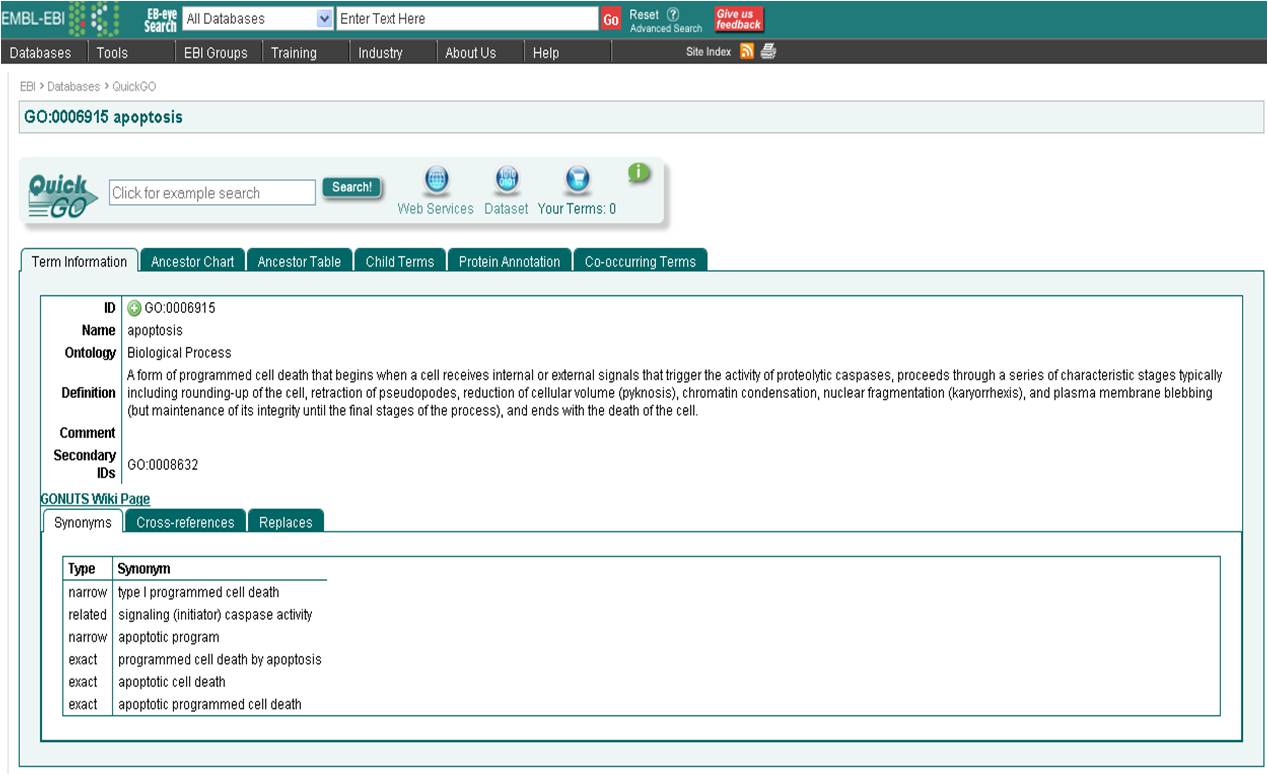
**[A]** The green plus icons next to GO IDs allow you to add that term to ‘Your terms’ selection which you can use for comparing multiple terms in an ontology chart (see section 1.2) or for creating GO slims (see section 3).

**[B]** The first 20 GO terms are shown by default to see further terms, click on ‘more’ at the bottom of the list.

**[C]** Tabbed sections allow for a more focused search in a particular ontology

**[D]** Obsolete terms are also retrieved, and this is indicated to the right of the term name.

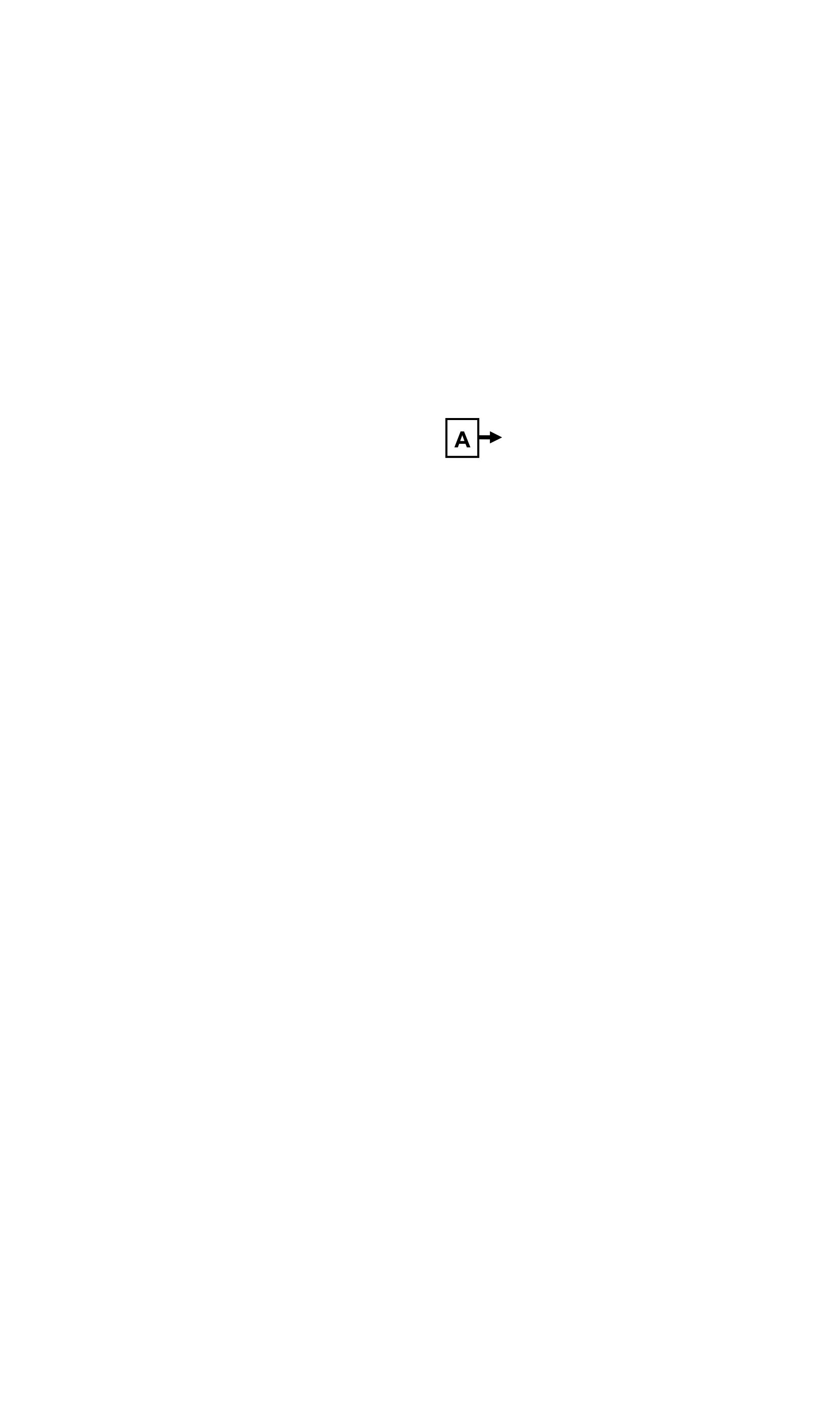
Clicking on the GO ID for a term will take you to a page called the ‘Term Information Page’ (Fig. 1.3), providing full details of the selected term.



**B**

**A**

**D**

**C**

**Figure 1.3. GO Term Information Page View**

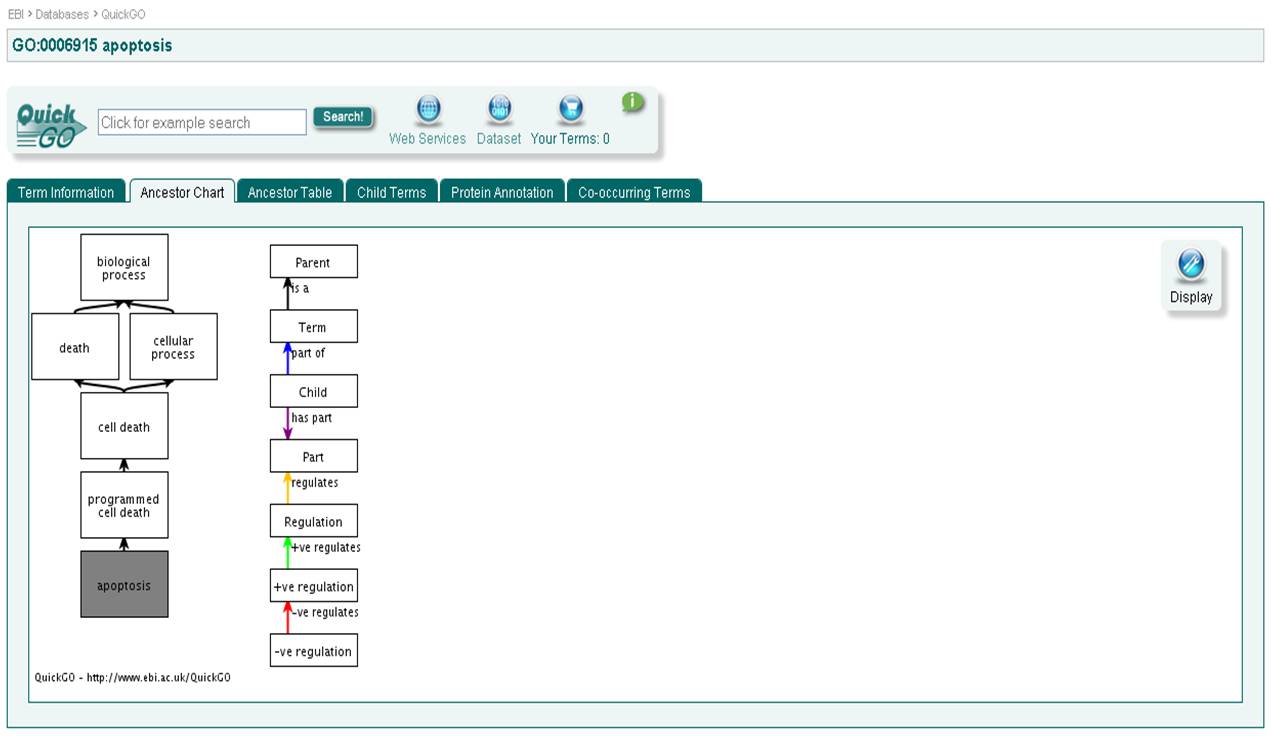
**[A]** A unique, stable identifier for the GO term

**[B]** The primary GO term name

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

**[D]** Term synonyms

Tabbed pages provide further information about the term such as Ancestor terms (Fig. 1.4), Child terms and Protein Annotation to the term.



**A**

**B**

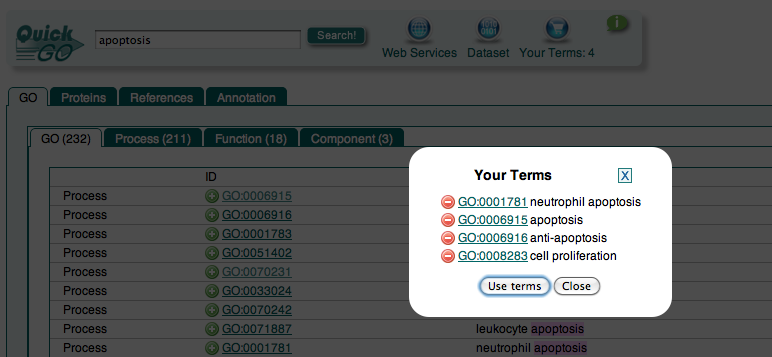
**Figure 1.4. Ancestor chart view of the GO term page.**

**[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.

## **1.2 Comparing multiple GO terms in an ontology view**

You can compare **GO terms** within the ontology. To compare a list of **GO** **terms**, add them one by one to ‘Your Terms’ list by searching for them in **QuickGO** and then clicking on the green plus icon that appears next to each GO ID (Fig. 1.2, note [A]). A ‘Your Terms’ lightbox will appear, listing your selection of terms (Fig. 1.5).



**A**

**C**

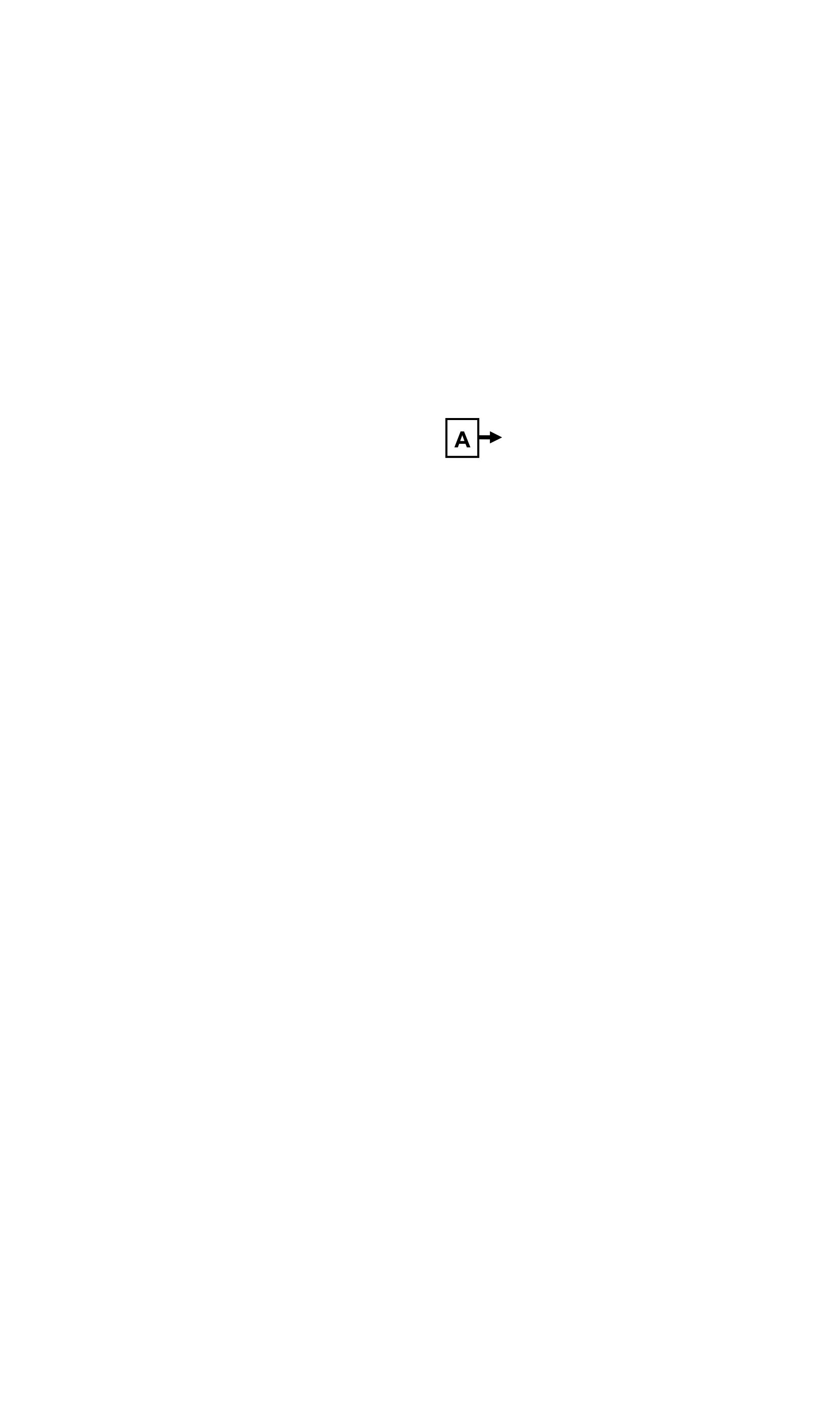
**Figure. 1.5** ‘Your terms’ basket contains the GO terms you have collected whilst browsing QuickGO. These can be used to compare the terms in an ontology chart.

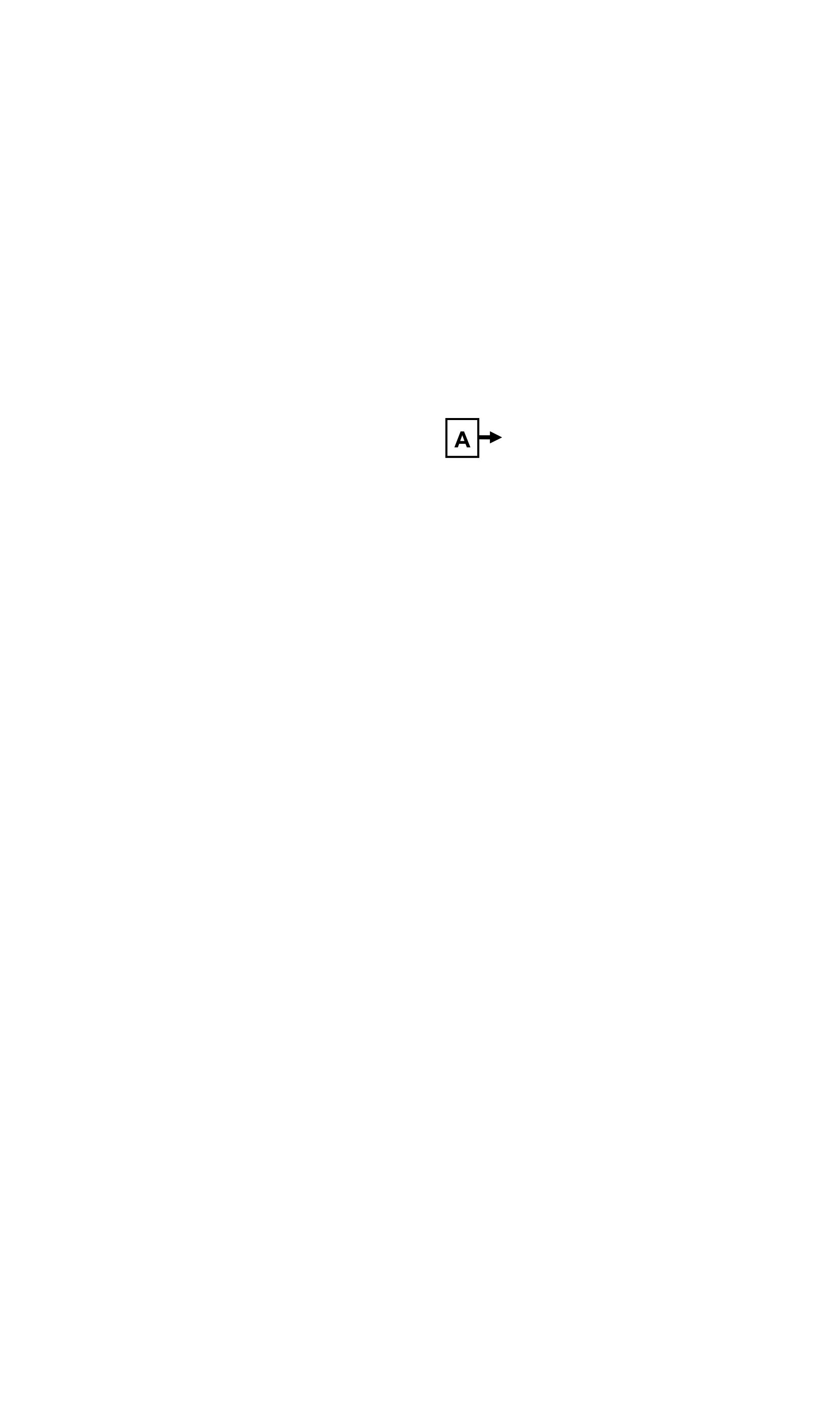
**[A]** Click on ‘Use Terms’ and you will be directed to the ‘Edit Terms’ tab (Fig. 1.6).

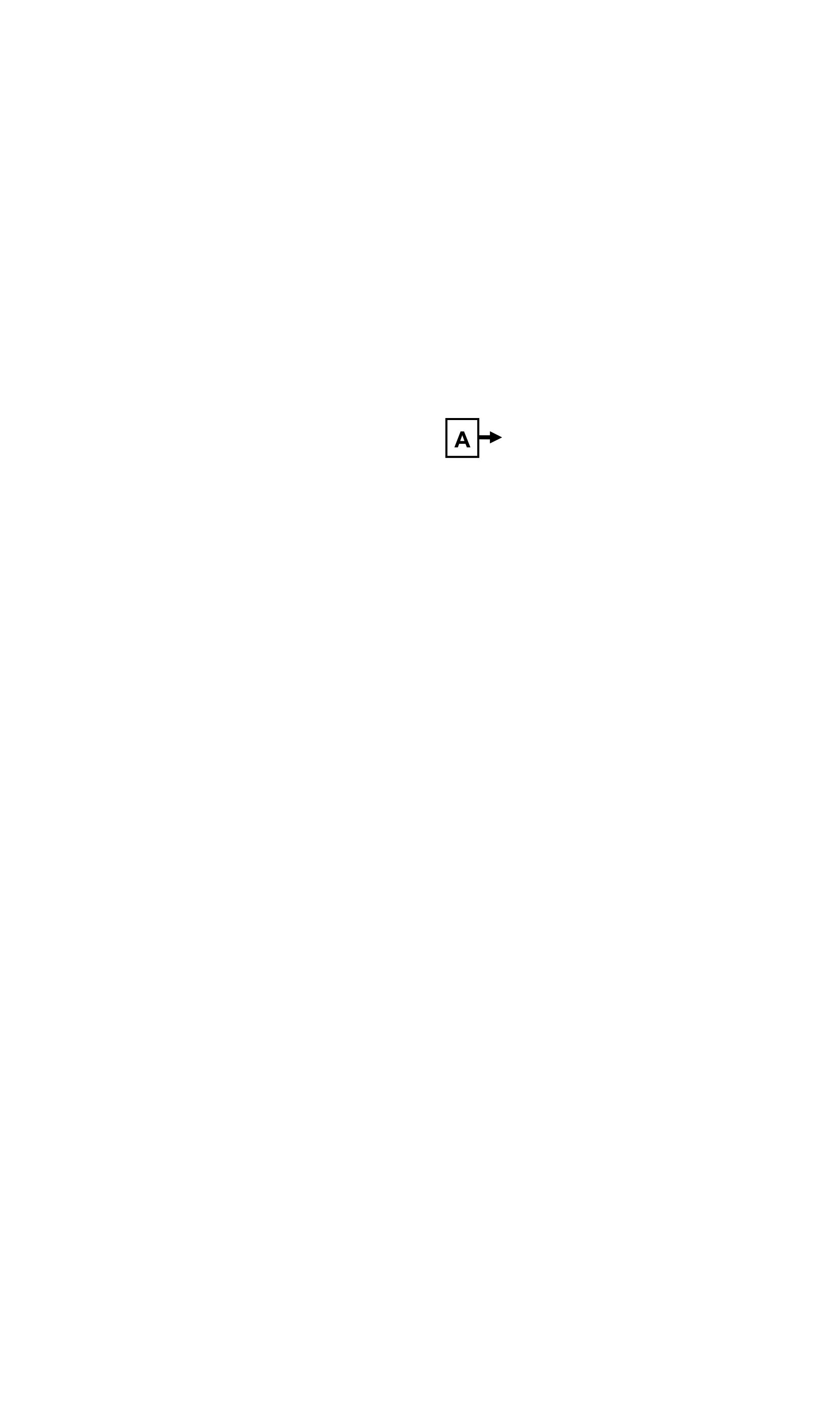
## Macintosh HD:Users:rfoulger:Desktop:Screen shot 2011-02-02 at 21.26.06.png

**B**

**A**

**C**

**B**

**A**

**Figure 1.6. The ‘Edit Terms’ tab** of the GO Term Comparison page showing your selection of terms that may be viewed together in ontological context.

**[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 and the terms selected will be highlighted.

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

**[C]** Comparison chart tab.

## **Exercise 1 **Searching for GO terms using QuickGO****

This exercise will familiarise you with the functionality of QuickGO for browsing GO.

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.*
   * + - *Look at the ancestor chart for a GO term.*
       - *Look at the child terms and their relationships.*
5. *Use the green ‘plus’ buttons to select two or more GO terms and see how they are related within the ontology.*

Question

**Question 1:** Which cellular component terms are retrieved in a search for ‘apoptosis’?

**Question 2:** What is the GO ID for ‘anoikis’?

**Question 3:** What is the secondary GO ID for ‘apoptosis’?

**Question 4:** What are the synonyms for ‘nurse cell apoptosis’?

**Question 5:** What is the parent of ‘apoptosis’, and what relationship connects the terms?

**Question 6:** In the term page for ‘apoptosis’ how many databases have cross-references for this term?

**Question 7:** How many ‘part\_of’ child terms does ‘apoptosis’ have?

## 

## **1.3 Browsing GO annotations**

Associations between gene products and GO terms are termed **GO annotations** and are assigned by many biological databases. A single gene product may be annotated to multiple **GO terms** using both manual and electronic annotation methods.

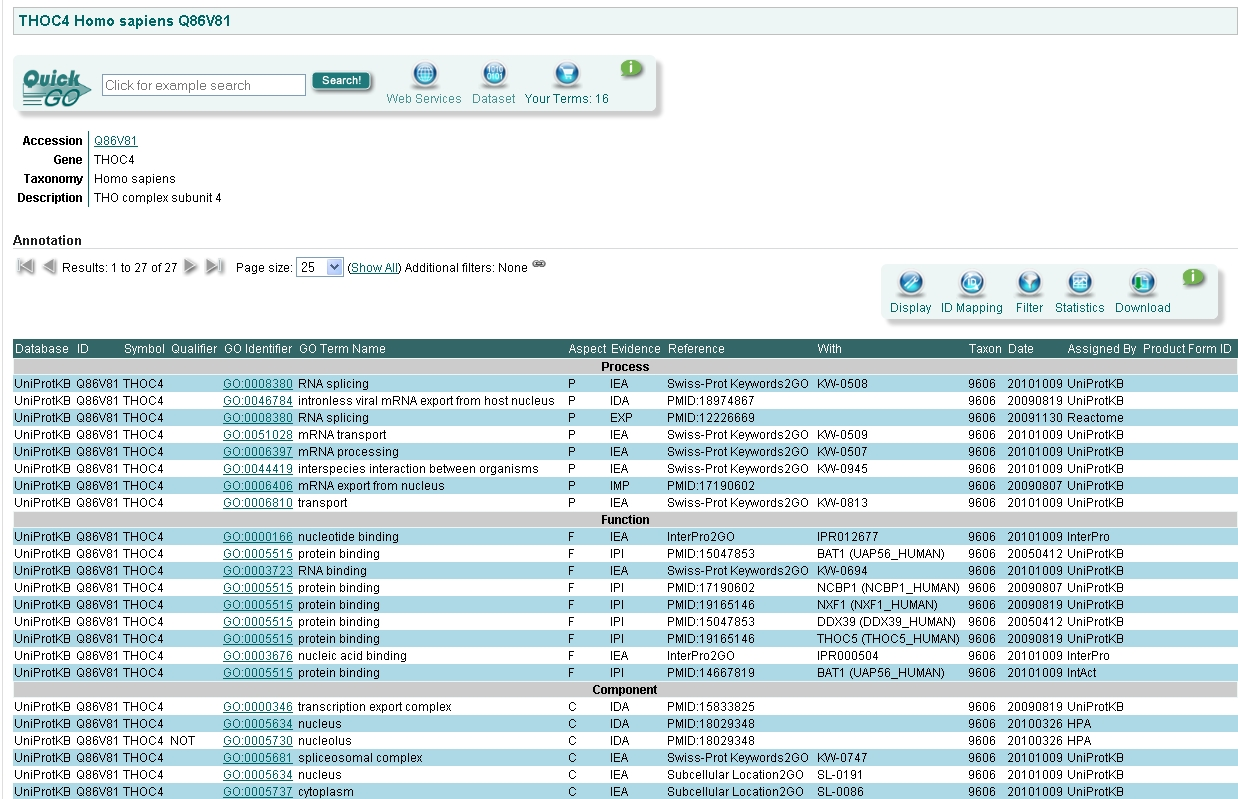
A **GO annotation** is the assignment of a GO identifier, e.g. GO:0005737, with a particular sequence (either a gene or protein database identifier). *Manual annotations* are created by a curator having directly looked for functional information (either within published literature or by examining the sequence directly). *Electronic annotations* are produced by automated methods that produce high-quality, conservative predictions.

All annotations must provide both a reference to a source that provides information (For example the PubMed identifier for a primary paper) 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. In addition, qualifiers can alter the interpretation of a **GO annotation**. For more information on evidence codes and qualifier usage, see:

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

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

### 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, a protein name (e.g. Exportin-1), or a GO name or GO ID in the search box of QuickGO. Fig. 1.7 shows the GO annotation for the human gene THOC4 (UniProt accession Q86V81).



**C**

**B**

**D**

**A**

**E**

**F**

**G**

**B**

**Figure 1.7. Protein Annotation page for human THOC4**: manual and electronic GO annotations are displayed for a queried protein together with supporting evidence codes, literature or electronic source.

**[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. For a list of evidence codes, see <http://www.geneontology.org/GO.evidence.shtml>.

**[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).

### **Exercises Searching for GO Annotations**

### These exercises will familiarise you with searching for all proteins annotated to a particular GO term, and all GO annotations for a particular protein.

### **Exercise 2 Searching for GO annotations in QuickGO using a GO term**

1. *Open QuickGO at* [*http://www.ebi.ac.uk/QuickGO*](http://www.ebi.ac.uk/QuickGO) *(see Fig. 1.1).*
2. *Enter the query ‘apoptosis’ into the search box*.
3. *From the retrieved options, select the term ‘nurse cell apoptosis’.*
4. *Click on the ‘Protein Annotation’ green tab. You will see the proteins that have been annotated with ‘nurse cell apoptosis’.*

Question

**Question 1:** How many annotations are there to ‘nurse cell apoptosis’? *Clue: Look for the ‘Results’ display.*

**Question 2:** Which organism do the annotated proteins come from? *Clue: Click on the Taxon ID for one of the annotations.*

**Question 3:** Which gene products are not involved in ‘nurse cell apoptosis’?

### **Exercise 3 Searching for GO annotations in QuickGO using a protein identifier**

1. *Open QuickGO at* [*http://www.ebi.ac.uk/QuickGO*](http://www.ebi.ac.uk/QuickGO) *(see Fig. 1.1).*
2. *Enter into the search box either the UniProtKB accession* Q86V81 or the protein name THOC4.
3. *If you searched for the UniProtKB accession, you should see a page listing the matching protein, click on the UniProtKB accession to open the protein annotation page. If you searched for the protein name, you will see a list of proteins matching that name, click on the human THOC4 to see the protein annotation page.*

Question

**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.*

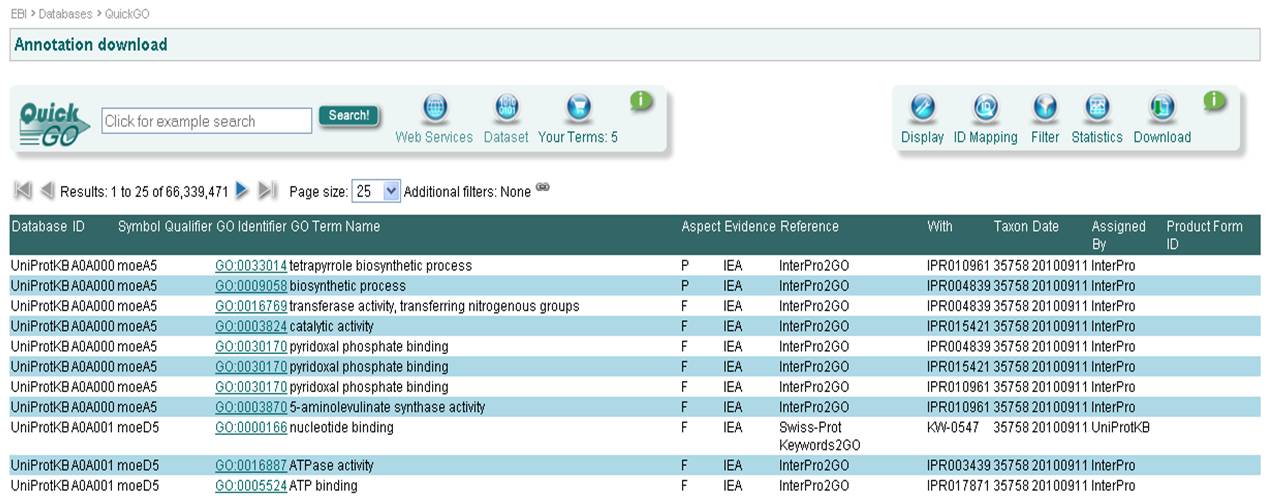
# 2 **Using QuickGO to create a tailored set of annotations**

It is possible within **QuickGO** to:

* Filter annotations by taxonomic group, **evidence code**, GO ID and protein identifier.
* View the statistics associated with a set of annotations.

## **2.1 Filtering**

On the **QuickGO** home page ([www.ebi.ac.uk/QuickGO](http://www.ebi.ac.uk/QuickGO)), click on the link ‘Search and Filter GO annotation sets’: this takes you to a page containing all available **GO annotations** in the 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 page 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:



**C**

**D**

**B**

**A**

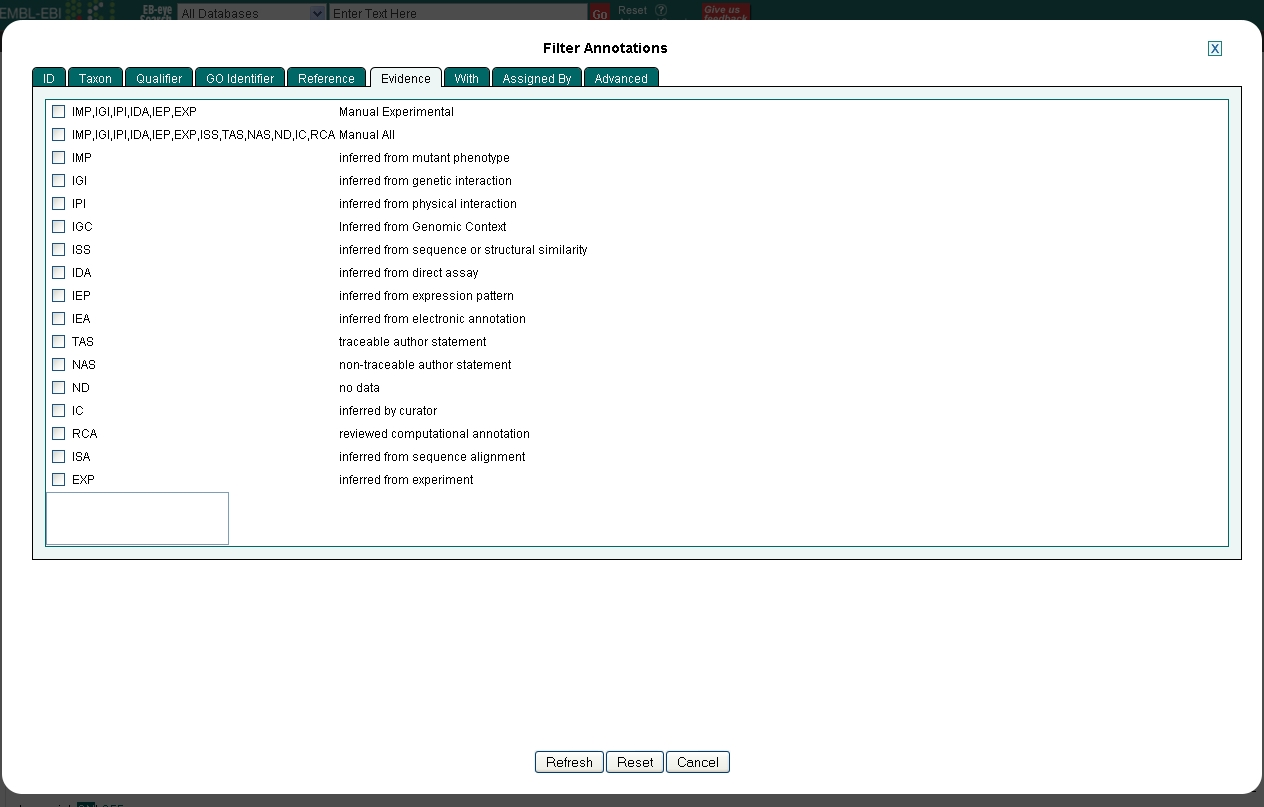
**Figure 2.1. Filtering annotations in QuickGO**

**[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. For example, you can retrieve annotations to a particular taxon or made using certain evidence codes. Fig. 2.2 shows the filter tab for evidence codes: 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. When you have chosen all your required filtering options, click on ‘Refresh’ at the bottom of the window and the annotations will be retrieved.

**Figure 2.2** **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.

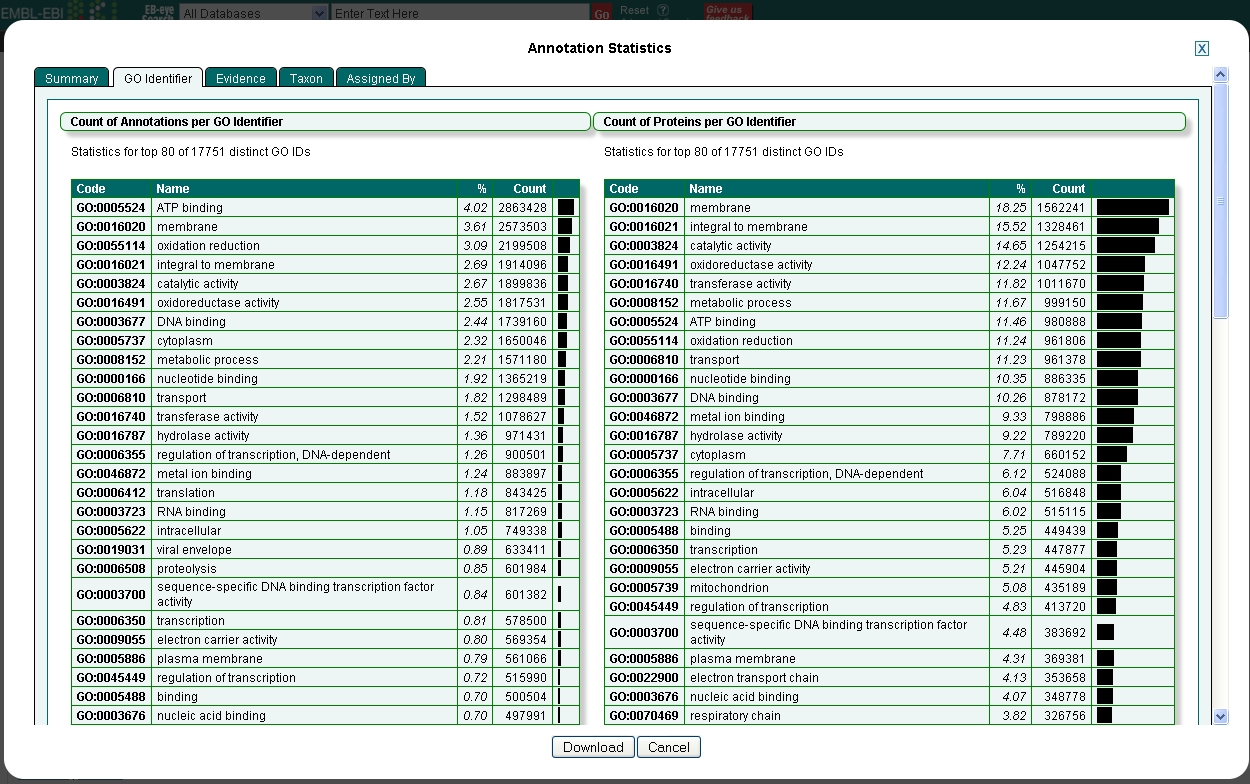
In general, for most sets of proteins, the most common evidence code is ‘Inferred from Electronic Annotation (IEA)’, simply because there are so many more electronic annotations compared with manual annotations.

## **2.2 Statistics**

**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.

Clicking on the ‘Statistics’ button opens up a lightbox with the statistics options arranged as tabs in the window. Fig. 2.3 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.

The statistics are downloadable as a text 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 copying the downloaded statistics for percentage of proteins per GO ID into a graph drawing program.



**Figure 2.3** **GO ID statistics tab.** Only the first 80 of the most common GO IDs are shown.

### **Exercises Filtering and Statistics**

## 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 how to view the statistics of the final set of annotations.

### **Exercise 4 Finding annotations for a list of proteins**

This exercise will use a pre-generated list of UniProtKB accession numbers. The list is a subproteome of a Jurkat (T-cell leukaemia) cell line, originally published by Bantscheff *et al*. [2]. The list can be found at: <ftp://ftp.ebi.ac.uk/pub/contrib/goa/Tutorial_Data/Dec_2010> in the file ‘quickgo\_query.txt’.

1. *From the QuickGO home page, click on ‘Search and Filter GO annotation sets’.*
2. *Click on the ‘Filter’ button.*
3. *Click on the ‘ID’ tab.*
4. *Paste the list of UniProtKB accession numbers from the ‘quickgo\_query.txt’ file into the ‘ID’ filter box. N.B. Do not tick any of the boxes below the text box.*
5. *Click on ‘Refresh’ to view the annotations to this list of proteins. N.B. You may have to wait a few seconds for it to load.*

Question

**Question 1:** For this list of proteins, how many annotations are there using both manual and electronic evidence codes together? *Clue: See ‘Results’.*

1. *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 5 Viewing annotation statistics**

1. *Use the set of annotations, filtered for manual experimental evidence codes, generated in Exercise 4 to view the annotation statistics (Click on the ‘Statistics’ tab).*

Question

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

**Question 2:** What are 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 manual experimental annotations?

# 3 GO Slims

GO slims are cut-down versions of GO, containing a subset of the terms in the whole GO. They give a broad overview of the ontology 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.

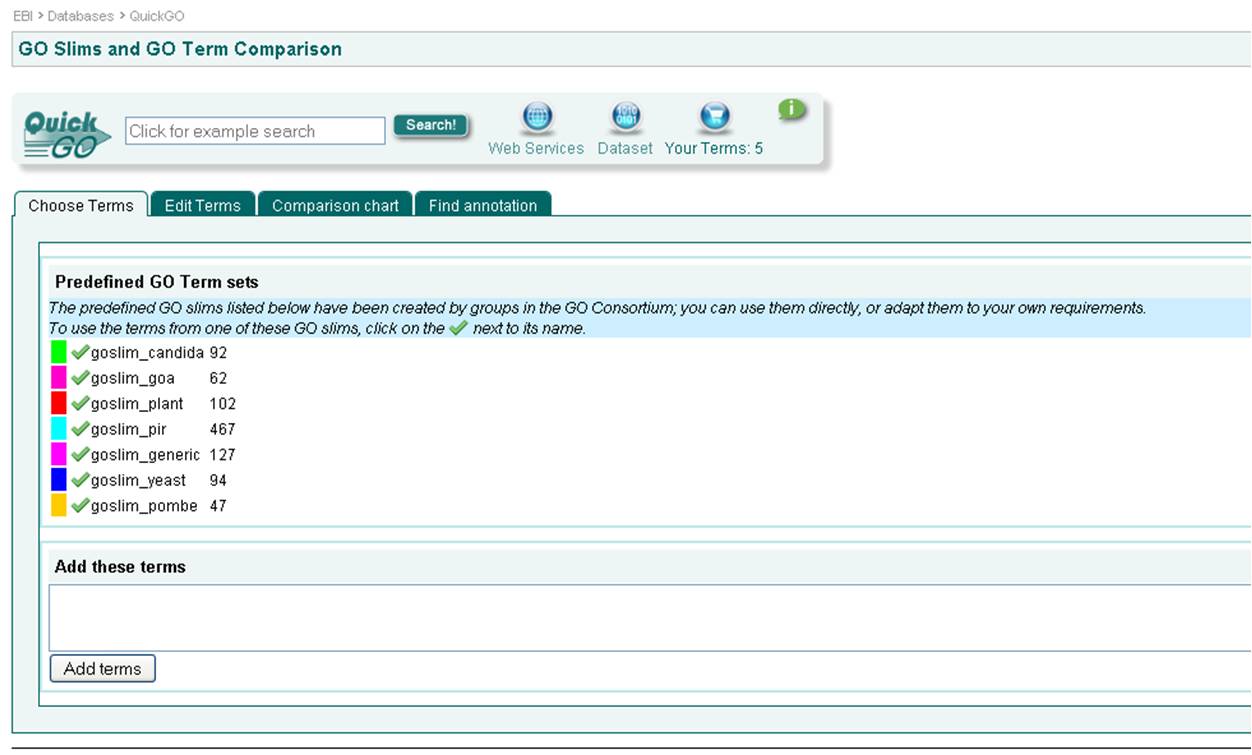
GO slims can be created by users according to their needs, and may be specific to species or to particular areas of the ontologies. There are also several ready-made GO slims, including a generic GO slim provided by the GO Consortium, a plant-specific slim provided by TAIR and a yeast-specific slim provided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

**In this section, 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.

## **3.1 Choosing your GO Slim terms**

On **QuickGO**’s home page [www.ebi.ac.uk/QuickGO](http://www.ebi.ac.uk/QuickGO), click on the link ‘Investigate GO slims’. Here, you can either select a particular pre-defined GO slim or enter a list of GO IDs to create a custom GO slim.



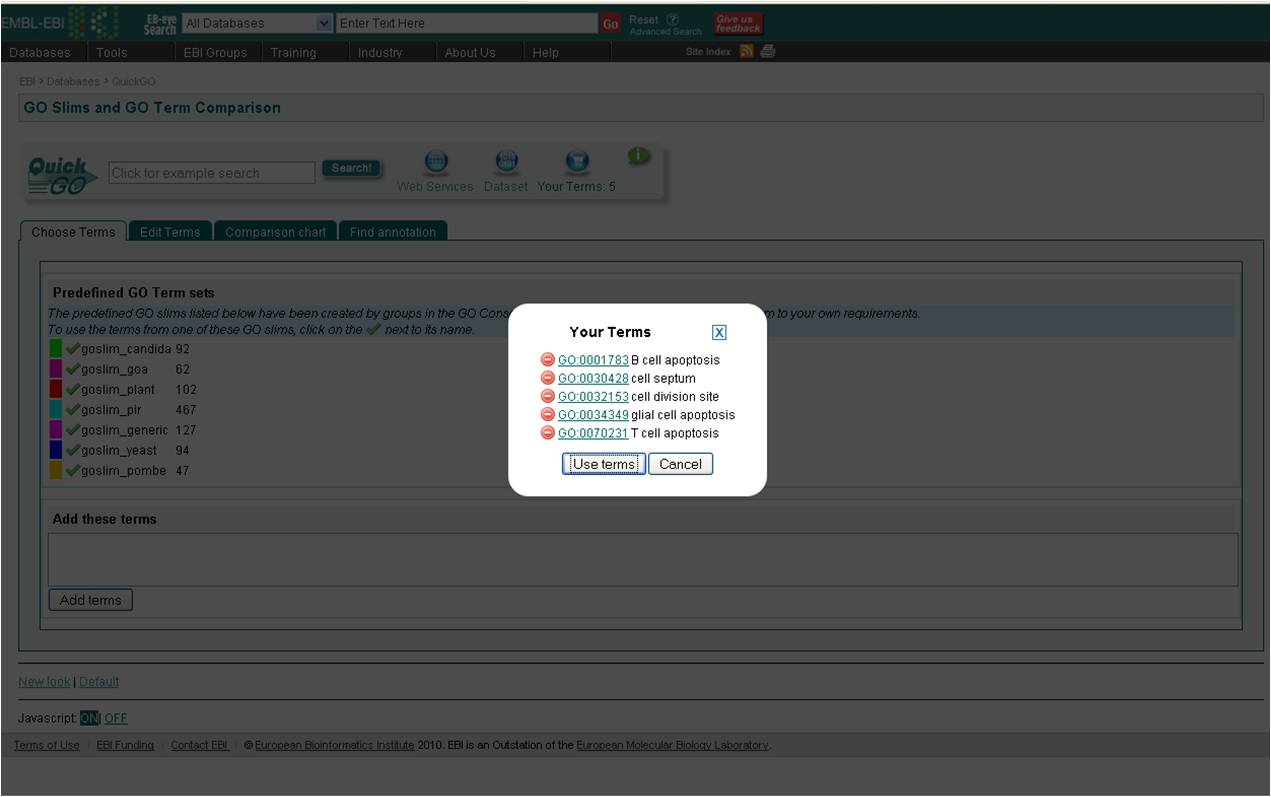
**Figure 3.1. GO Slims in QuickGO**

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

**[B]** To create your own set of terms, GO IDs can be typed/pasted into the text box and added by clicking on ‘Add terms’.

**A**

**B**

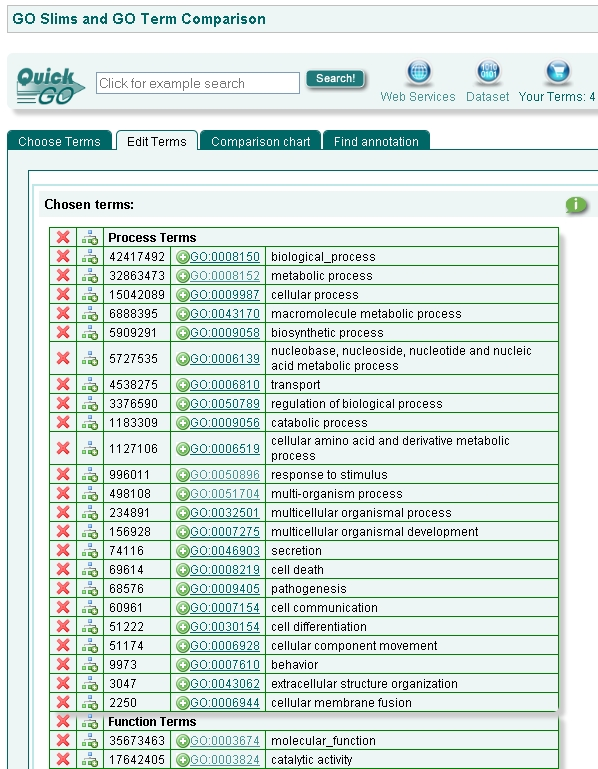
Another way of selecting your own set of **GO terms** that you would like to use as a **GO Slim** is to collect terms as you browse **QuickGO**. Wherever you see the green and white cross icon next to a term, you can use it to add the term to ‘Your terms’ basket (Fig. 3.2). Your collection of terms can then be used to create a GO slim.

**Fig. 3.2** ‘**Your Terms’ basket** contains the GO terms you have collected whilst browsing QuickGO. These can be used to create a GO slim.

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 ‘Edit terms’ 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. You should recognise this view from when you were comparing GO terms in section 1.2.

## **3.2 Slimming-up annotations to GO slim terms**

Once you have a finalised list of terms, you can slim-up annotations to these terms by moving to the ‘Find annotation’ 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 6. 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.

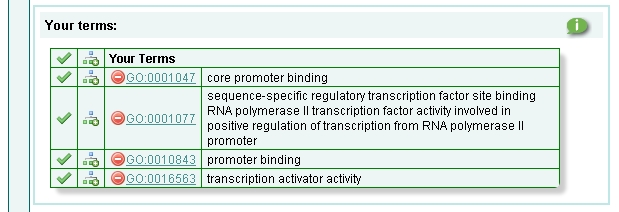


**C**

**A**

**B**

------------------------------------------------------------------------------------------------------------

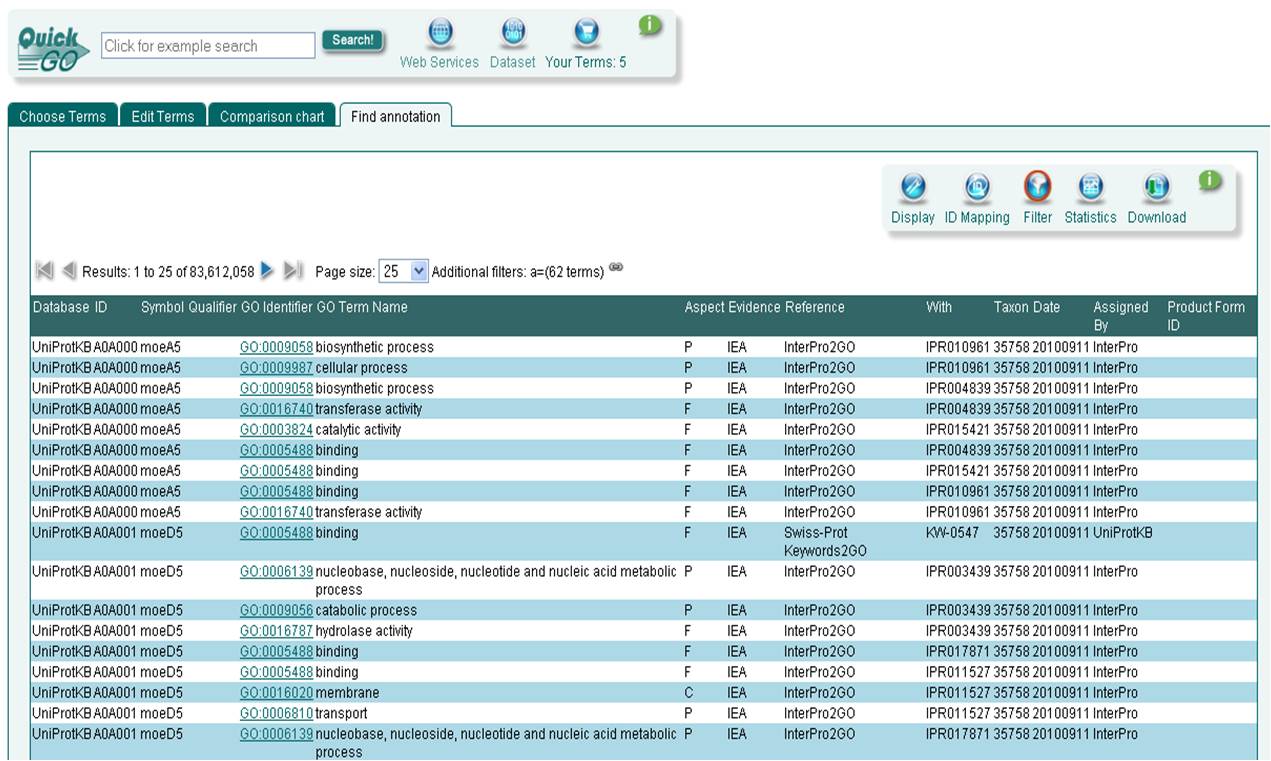


**Figure 3.3. Edit Terms Tab for GO Slims**

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

**[B]** To add terms to the list, first add them to ‘Your terms’ selection, by clicking on the green plus icon as described in Note [A] of Fig. 1.2, and then click on the green tick next to the term in the GO Slims and GO Term Comparison page.

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



**B**

**A**

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

**[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, 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 and sources of annotation, as well as the number of unique protein accessions.

## **GO Slim Exercises**

### **Exercise 6 Using QuickGO to slim-up annotations to a selected list of protein accessions**

In this exercise, we will see which GO slim terms the selected list of UniProtKB accessions slim up to.

You will use the same list of protein accessions as before. The list can be found at: <ftp://ftp.ebi.ac.uk/pub/contrib/goa/Tutorial_Data/Dec_2010> in the file ‘quickgo\_query.txt’.

1. *Go to the GO slim entry page in QuickGO by clicking on the ‘Investigate GO slims’ link on the front page.*
2. *Go to the ‘Choose Terms’ tab, and select the pre-defined ‘goslim\_generic’ by clicking on the green tick next to its name.*
3. *Click on the ‘Find annotation’ tab and wait for the results to load.*
4. *The resulting table contains all the annotations in the GOA database slimmed-up to the 105 terms in the GO slim. Go to the ‘Filter’ box in the Annotation Toolbar and ensure the ‘ID’ tab is selected. Paste the list from the ‘quickgo\_query.txt’ file* *into the text box in the ID tab, then click on ‘Refresh’ to view the results.*
5. *The table now shows the annotations to the list of proteins slimmed-up to the 105 slim terms. Use the statistics calculated for this set to answer the following questions.*

Question

**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?

# 4 Using GO annotation data to link biological knowledge to a set of proteins (GO enrichment analysis)

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 tools page (<http://www.geneontology.org/GO.tools.html>) lists a number of tools that satisfy most of these requirements and a recent review of tools is also available to help users [4].

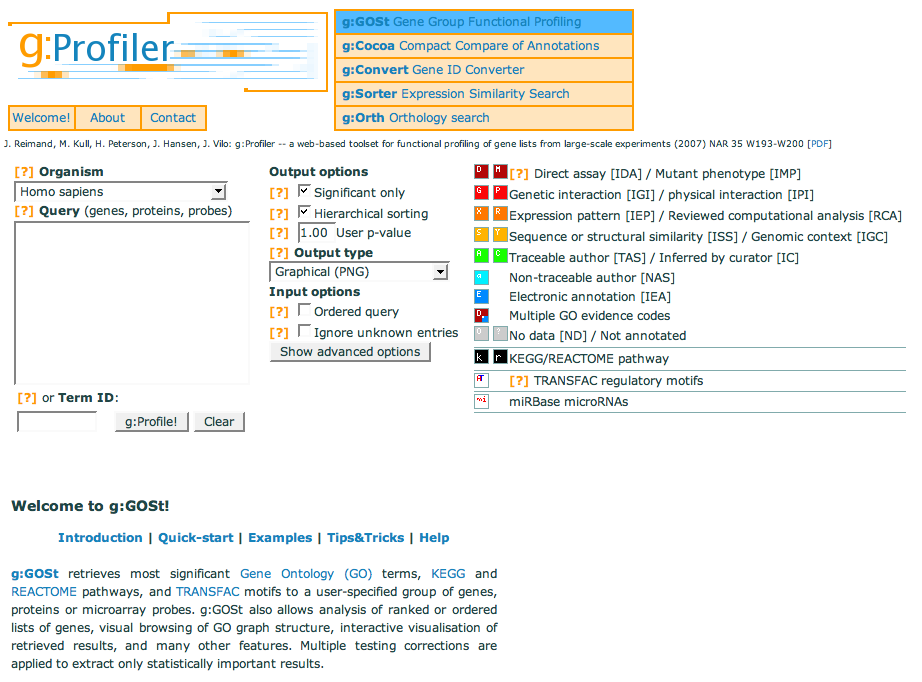
## **4.1 Analysis with g:Profiler**

<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 GO Consortium tools page <http://www.geneontology.org/GO.tools.html> 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. 4.1).

**Figure 4.1** The g:GOSt query interface from g:Profiler.

### 

### **Exercise 7 Using g:Profiler to perform GO term enrichment analysis on a list of protein accessions**

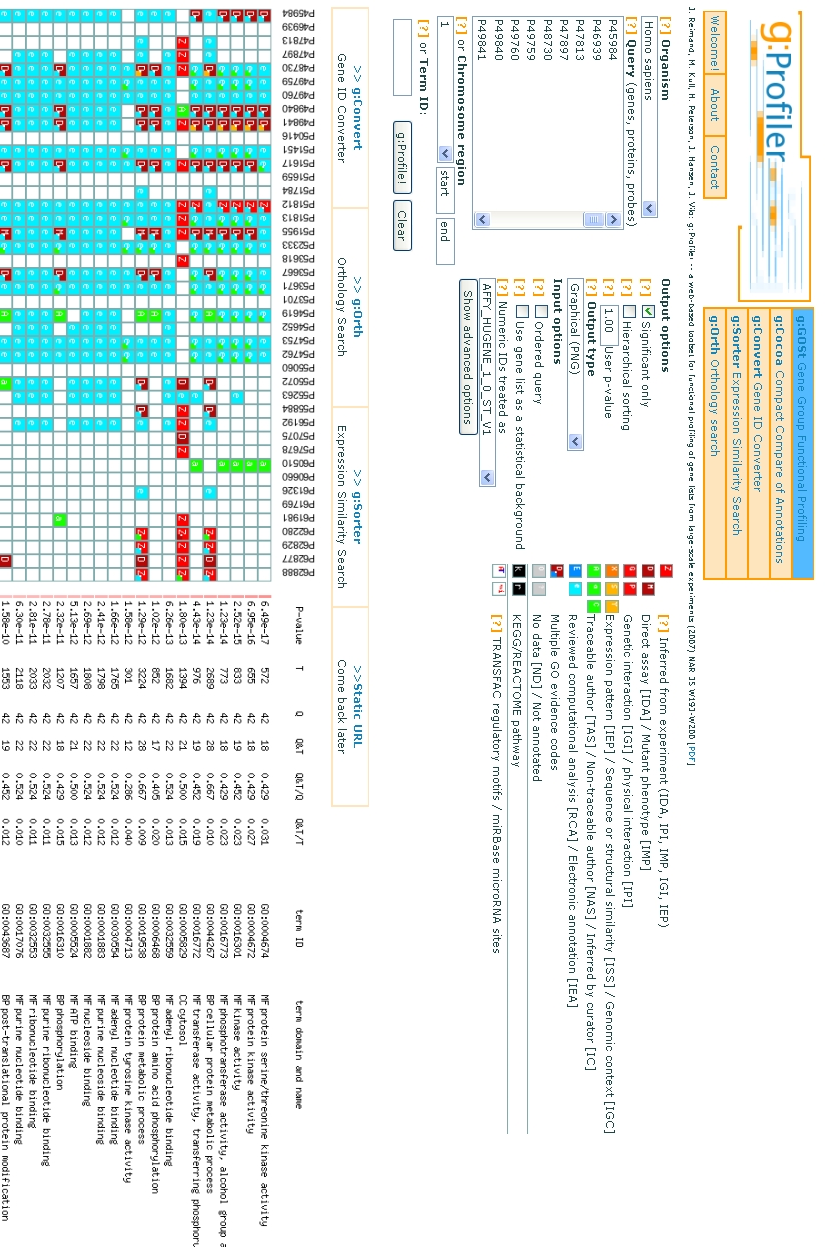
The list of protein accessions you will use in this exercise can be found in the file ‘gprofiler\_query.txt’ at: <ftp://ftp.ebi.ac.uk/pub/contrib/goa/Tutorial_Data/Dec_2010>

1. *Go to the gprofiler website:* (<http://biit.cs.ut.ee/gprofiler/>)
2. *Copy and paste the list of human protein accessions in ‘gprofiler\_query.txt’ file provided into the Query box.*
3. *As all of the accessions provided are human select ‘Homo sapiens’ in the organism box.*
4. *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.*
5. *Leave all other options as the provided default.*
6. *Click ‘g:Profile!’*

enriched GO terms

Question

**Question 1:** What term(s) appear to be over-represented within the queried protein set? *Clue: See Fig. 4.2 for the output from this query.*



1

2

3

4

5

6

7

8

9

10

**Fig. 4.2** 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.

# 

# 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.

# 

# 

# 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 UniProtKB-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. More information on evidence codes can be found on the GO Consortium website; <http://www.geneontology.org/GO.evidence.shtml>

### **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>

1. 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.
2. 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.
3. Khatri, P. and Draghici, S. (2005) **Ontological analysis of gene expression data: current tools, limitations, and open problems**. *Bioinformatics* **21,** 3587-3595.
4. 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.
5. Lomax J, The Gene Ontology Consortium (2005) **Get ready to GO! A biologist's guide to the Gene Ontology.** *Brief Bioinform*. ***6***: 298-304.
6. 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.
7. 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].
8. 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

GOA: <http://www.ebi.ac.uk/GOA/>

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

UniProtKB: <http://www.uniprot.org/>

# 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: [gohelp@genome.stanford.edu](mailto:gohelp@genome.stanford.edu)

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 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](mailto:goa@ebi.ac.uk) or alternatively fill in the GOA web form at: <http://www.ebi.ac.uk/GOA/contactus.html>

EXERCISE ANSWERS (correct at time of writing: February 2011)

## **Exercise 1 answers: Searching for GO terms using QuickGO**

Answer

**Question 1:** Which cellular component terms are retrieved in a search for ‘apoptosis’?

**Answer 1:** GO:0043293 apoptosome

GO:0008303 caspase complex

GO:0005868 cytoplasmic dynein complex

**Question 2:** What is the GO ID for ‘anoikis’?

**Answer 2:** GO:0043276

**Question 3:** What is the secondary GO ID for ‘apoptosis’?

**Answer 3:** GO:0008632 (Some GO terms have secondary IDs if two or more terms have been merged together).

**Question 4:** What are the synonyms for ‘nurse cell apoptosis’?

**Answer 4:** apoptosis of nurse cells

nurse cell programmed cell death by apoptosis

programmed cell death of nurse cells by apoptosis

**Question 5:** What is the parent term of ‘apoptosis’, and what relationship connects the terms?

**Answer 5:** GO:0012501 programmed cell death

‘apoptosis’ is\_a ‘programmed cell death

**Question 2:** In the term page for ‘apoptosis’ how many databases have cross-references for this term?

**Answer 2:** Two; InterPro and Wikipedia

**Question 2:** How many ‘part\_of’ child terms does ‘apoptosis’ have?

**Answer 2:** Five;

GO:0070782 phosphatidylserine exposure on apoptotic cell surface

GO:0006919 activation of caspase activity

GO:0006921 cellular component disassembly involved in apoptosis

GO:0008633 activation of pro-apoptotic gene products

GO:0008637 apoptotic mitochondrial changes.

## **Exercise 2 answers: Searching for GO annotations in QuickGO using a GO term.**

Answer

**Question 1:** How many annotations are there to ‘nurse cell apoptosis’?

**Answer 1:** 24 annotations

**Question 2:** Which organism do the annotated proteins come from?

**Answer 2:** *Drosophila melanogaster* (Fruit fly)

**Question 3:** Which gene products are not involved in ‘nurse cell apoptosis’?

**Answer 3:** W (Wrinkled), rpr (Reaper), grim (Grim)

## **Exercise 3 answers: Searching for GO annotations in QuickGO using a protein identifier.**

Answer

**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.

## **Exercise 4 answers: Finding annotations for a list of protein accessions**

Answer

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

**Answer 1:** 14,246 annotations

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

**Answer 1:** 4,041 annotations

Description: Information 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 5 answers: Viewing annotation statistics for a list of protein accessions**

Answer

**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:** IPI, IDA and EXP

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

**Answer 3:** IntAct and UniProtKB

**Question 4:** How many proteins have manual experimental annotations?

**Answer 4:** 292 (Look in the ‘Annotation Statistics Summary: Number of distinct proteins).

### **Exercise 6 answers: Using QuickGO to slim-up annotations to a list of protein accessions**

Answer

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

**Answer 1:** cell ; GO:0005623

transport ; GO:0006810

biosynthetic process ; GO:0009058

**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:** InterPro and GOC (GOC is where GO terms have been inferred from existing annotations).

### 

### **Exercise 7 answers: Using g:Profiler to perform GO term enrichment analysis on a list of protein accessions**

Answer (correct at time of writing: October 2010)

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

**Answer 1:**

MolecularFunction:GO:0004674 protein serine/threonine kinase activity;

Cellular Component: GO:0005829 cytosol;

Biological Process: GO:0006468 protein amino acid phosphorylation

# Contributors

Emily Dimmer, Rachael Huntley, Rebecca Foulger, UniProtKB-GOA group.