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

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.

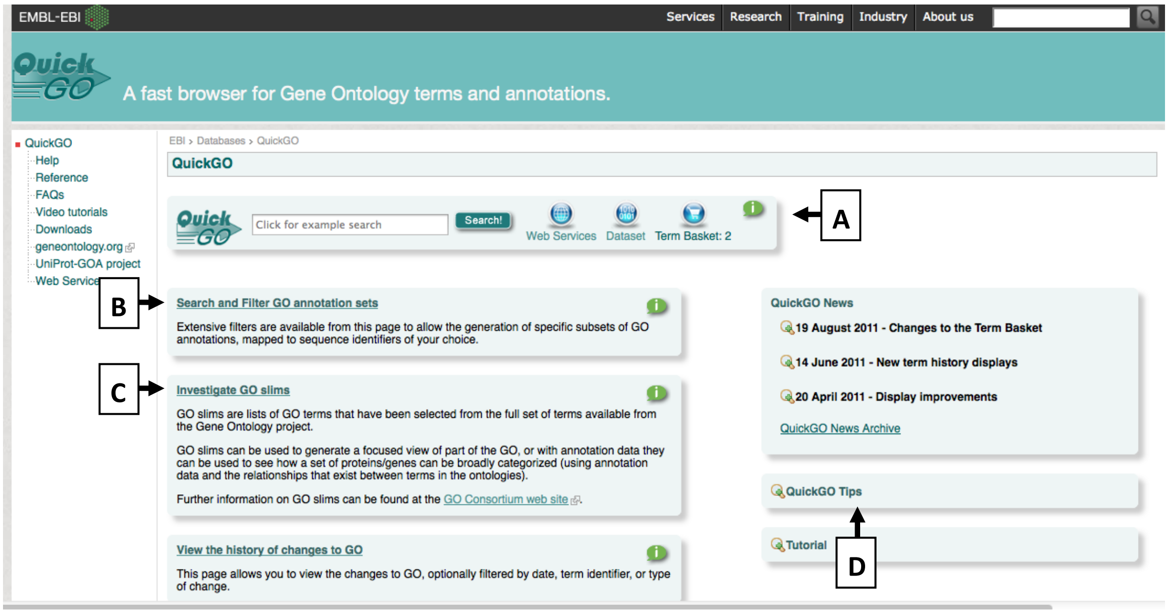
ℹ Information

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.



Extensive filters are available from this page to allow the generation of specific subsets of GO annotations, mapped to sequence identifiers of your choice.

GO slims are lists of GO terms that have been selected from the full set of terms available from the Gene Ontology project.

GO slims can be used to generate a focused view of part of the GO, or with annotation data they can be used to see how a set of proteins/genes can be broadly categorized (using annotation data and the relationships that exist between terms in the ontologies).

Further information on GO slims can be found at the GO Consortium web site.

This page allows you to view the changes to GO, optionally filtered by date, term identifier, or type of change.

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

Figure 1.1

Notes

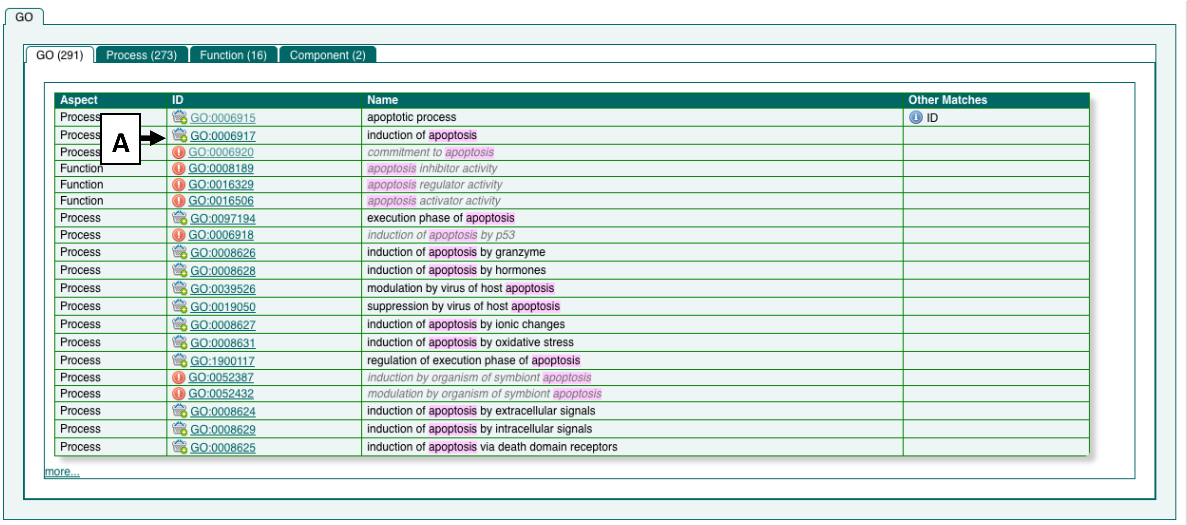
[A] The 'Global Toolbar' is visible on all pages within QuickGO. From here users can search QuickGO, access QuickGO Web Services, view the dataset that QuickGO is currently using and view 'Term Basket' selection (see Chapter 3 for more information on 'Term Basket'.

[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

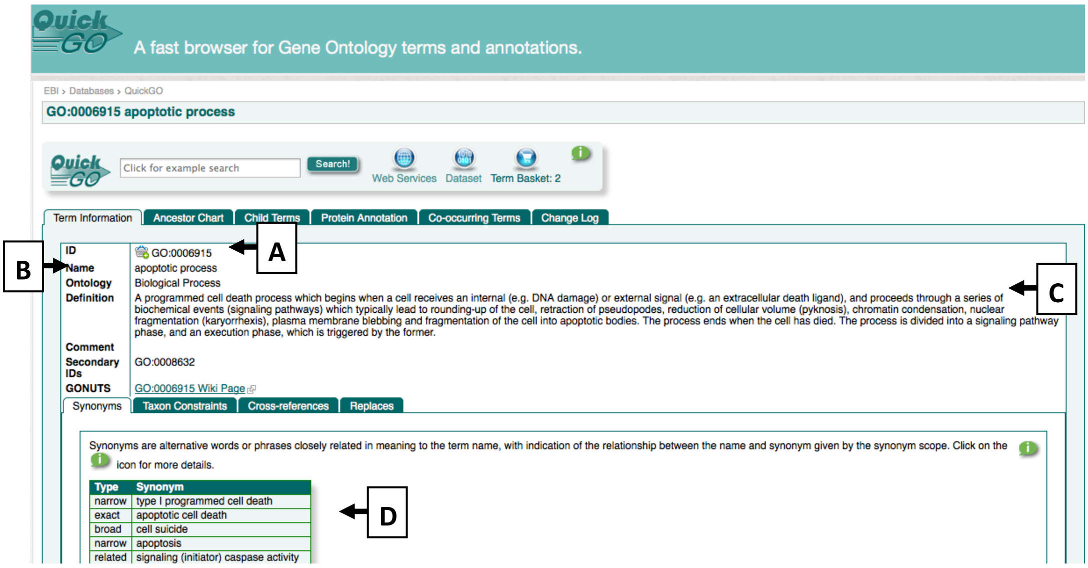
Fig. 1.2 A search for ‘apoptosis’ retrieves a choice of GO terms, tabbed sections allow for a more focused search in a particular ontology. Terms are retrieved if the word ‘apoptosis’ is present in their term name, synonyms, definition or cross-references.

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.



GO:0006915 apoptotic process

ID	GO:0006915
Name	apoptotic process
Ontology	Biological Process
Definition	A programmed cell death process which begins when a cell receives an internal (e.g. DNA damage) or external signal (e.g. an extracellular death ligand), and proceeds through a series of biochemical events (signaling pathways) which typically lead to rounding-up of the cell, retraction of pseudopodes, reduction of cellular volume (pyknosis), chromatin condensation, nuclear fragmentation (karyorrhexis), plasma membrane blebbing and fragmentation of the cell into apoptotic bodies. The process ends when the cell has died. The process is divided into a signaling pathway phase, and an execution phase, which is triggered by the former.
Comment	
Secondary IDs	GO:0008632
GONUTS	GO:0006915 Wiki Page

Synonyms are alternative words or phrases closely related in meaning to the term name, with indication of the relationship between the name and synonym given by the synonym scope. Click on the icon for more details.

Type	Synonym
narrow	type I programmed cell death
exact	apoptotic cell death
broad	cell suicide
narrow	apoptosis
related	signaling (initiator) caspase activity

Fig. 1.3a GO Term Information page view

Figure 1.3a

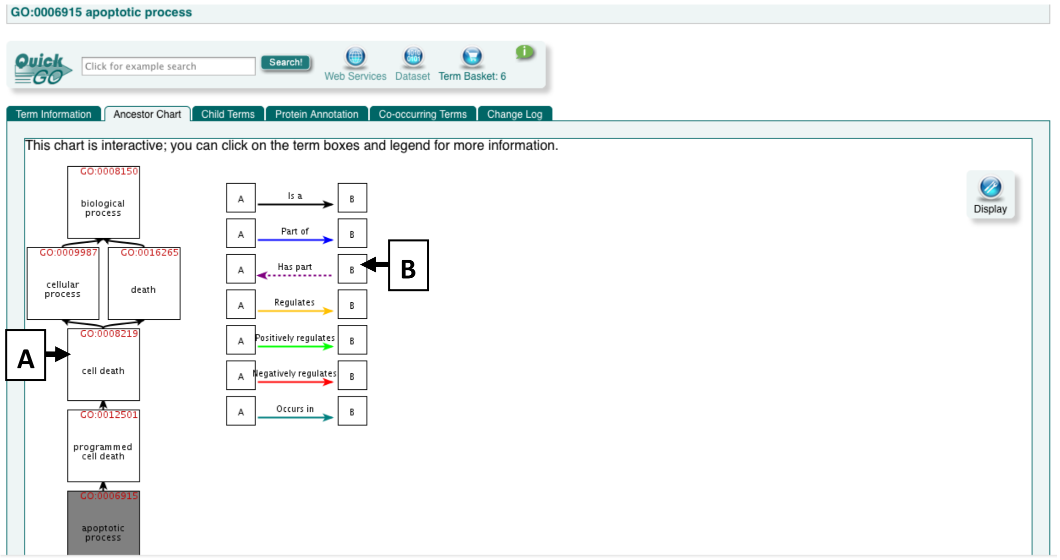
Notes

[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



GO:0006915 apoptotic process

This chart is interactive; you can click on the term boxes and legend for more information.

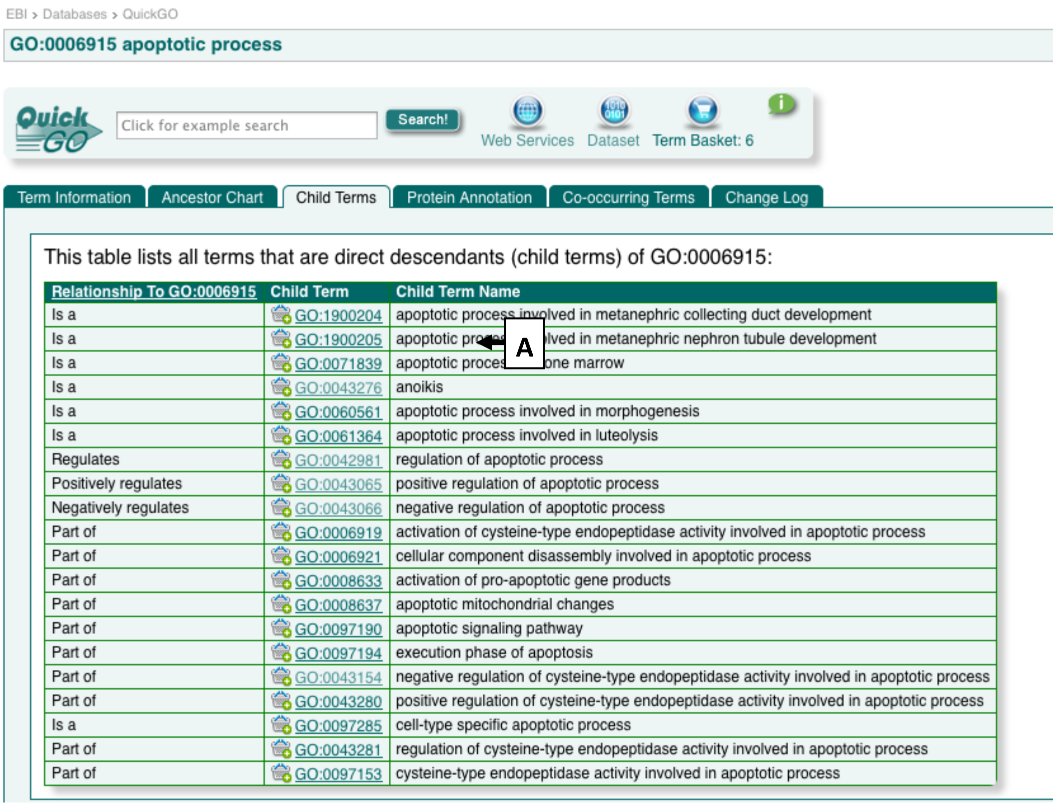
Fig. 1.3b Ancestor chart view of the GO term page.

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.



This table lists all terms that are direct descendants (child terms) of GO:0006915:

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 pr lved in metanephric nephron tubule development
Is a	GO:0071839	apoptotic proces 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

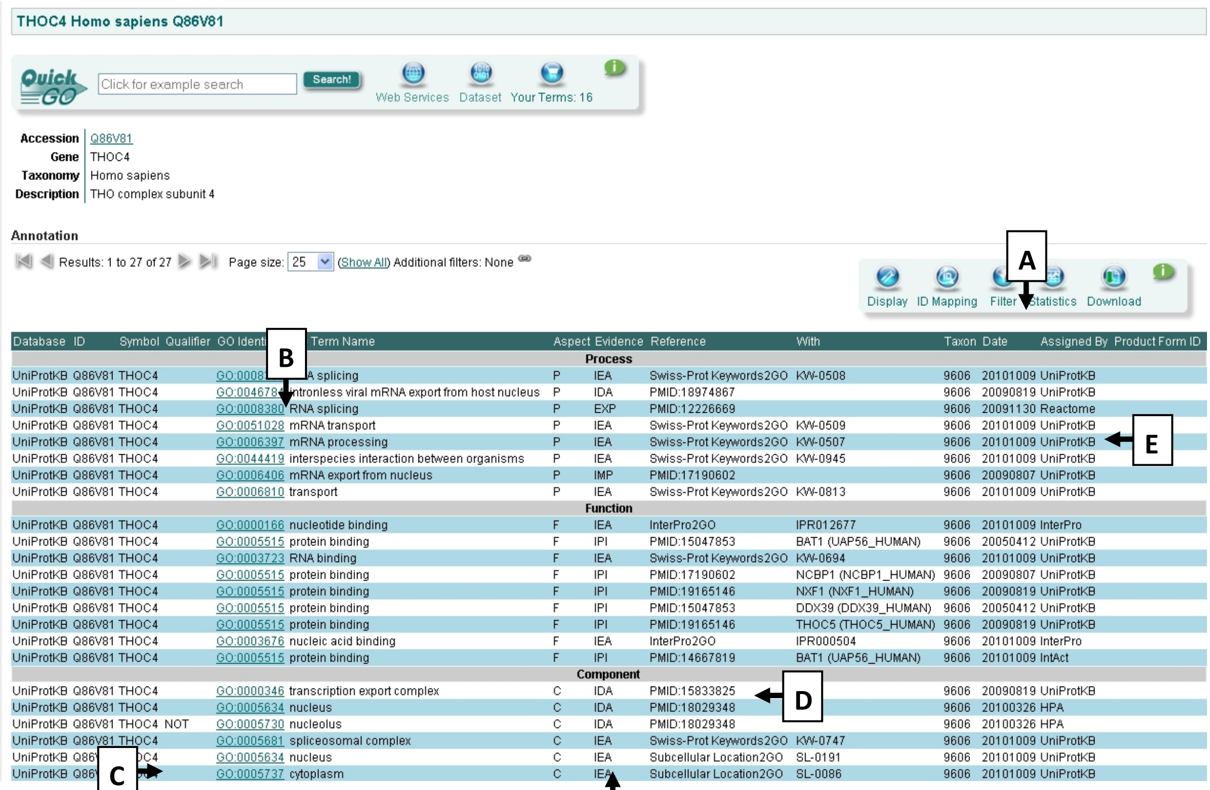
The UniProt-GOA database contains annotations made not only by curators within the UniProt Consortium, but also by specialist groups and by members of the GO Consortium. Therefore, the UniProt-GOA database provides the most comprehensive set of **GO annotations** for over 390,000 species within UniProtKB.

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.

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



THOC4 Homo sapiens Q86V81

Accession	Q86V81
Gene	THOC4
Taxonomy	Homo sapiens
Description	THO complex subunit 4

Annotation

Results: 1 to 27 of 27 Page size: 25 (Show All) Additional filters: None

Database	ID	Symbol	Qualifier	GO Identi	Term Name	Aspect	Evidence	Reference	With	Taxon	Date	Assigned By	Product Form ID
							Process						
UniProtKB	Q86V81	THOC4		GO:0008	A splicing	P	IEA	Swiss-Prot Keywords2GO	KW-0508	9606	20101009	UniProtKB	
UniProtKB	Q86V81	THOC4		GO:004678	intronless viral mRNA export from host nucleus	P	IDA	PMID:18974867		9606	20090819	UniProtKB	
UniProtKB	Q86V81	THOC4		GO:0008380	RNA splicing	P	EXP	PMID:12226669		9606	20091130	Reactome	
UniProtKB	Q86V81	THOC4		GO:0051028	mRNA transport	P	IEA	Swiss-Prot Keywords2GO	KW-0509	9606	20101009	UniProtKB	
UniProtKB	Q86V81	THOC4		GO:0006397	mRNA processing	P	IEA	Swiss-Prot Keywords2GO	KW-0507	9606	20101009	UniProtKB	
UniProtKB	Q86V81	THOC4		GO:0044419	interspecies interaction between organisms	P	IEA	Swiss-Prot Keywords2GO	KW-0945	9606	20101009	UniProtKB	
UniProtKB	Q86V81	THOC4		GO:0006406	mRNA export from nucleus	P	IMP	PMID:17190602		9606	20090807	UniProtKB	
UniProtKB	Q86V81	THOC4		GO:0006810	transport	P	IEA	Swiss-Prot Keywords2GO	KW-0813	9606	20101009	UniProtKB	
							Function						
UniProtKB	Q86V81	THOC4		GO:0000166	nucleotide binding	F	IEA	InterPro2GO	IPR012677	9606	20101009	InterPro	
UniProtKB	Q86V81	THOC4		GO:0005515	protein binding	F	IPI	PMID:15047853	BAT1 (UAP56_HUMAN)	9606	20050412	UniProtKB	
UniProtKB	Q86V81	THOC4		GO:0003723	RNA binding	F	IEA	Swiss-Prot Keywords2GO	KW-0694	9606	20101009	UniProtKB	
UniProtKB	Q86V81	THOC4		GO:0005515	protein binding	F	IPI	PMID:17190602	NCBP1 (NCBP1_HUMAN)	9606	20090807	UniProtKB	
UniProtKB	Q86V81	THOC4		GO:0005515	protein binding	F	IPI	PMID:19165146	NXF1 (NXF1_HUMAN)	9606	20090819	UniProtKB	
UniProtKB	Q86V81	THOC4		GO:0005515	protein binding	F	IPI	PMID:15047853	DDX39 (DDX39_HUMAN)	9606	20050412	UniProtKB	
UniProtKB	Q86V81	THOC4		GO:0005515	protein binding	F	IPI	PMID:19165146	THOC5 (THOC5_HUMAN)	9606	20090819	UniProtKB	
UniProtKB	Q86V81	THOC4		GO:0003676	nucleic acid binding	F	IEA	InterPro2GO	IPR000504	9606	20101009	InterPro	
UniProtKB	Q86V81	THOC4		GO:0005515	protein binding	F	IPI	PMID:14667819	BAT1 (UAP56_HUMAN)	9606	20101009	IntAct	
							Component						
UniProtKB	Q86V81	THOC4		GO:0000346	transcription export complex	C	IDA	PMID:15833825		9606	20090819	UniProtKB	
UniProtKB	Q86V81	THOC4		GO:0005634	nucleus	C	IDA	PMID:18029348		9606	20100326	HPA	
UniProtKB	Q86V81	THOC4	NOT	GO:0005730	nucleolus	C	IDA	PMID:18029348		9606	20100326	HPA	
UniProtKB	Q86V81	THOC4		GO:0005681	spliceosomal complex	C	IEA	Swiss-Prot Keywords2GO	KW-0747	9606	20101009	UniProtKB	
UniProtKB	Q86	C4		GO:0005634	nucleus	C	IEA	Subcellular Location2GO	SL-0191	9606	20101009	UniProtKB	
UniProtKB	Q86			GO:0005737	cytoplasm	C	IEA	Subcellular Location2GO	SL-0086	9606	20101009	UniProtKB	

Fig. 1.4 Protein Annotation page: manual and electroni O 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.*

Q

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.*

Q

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.

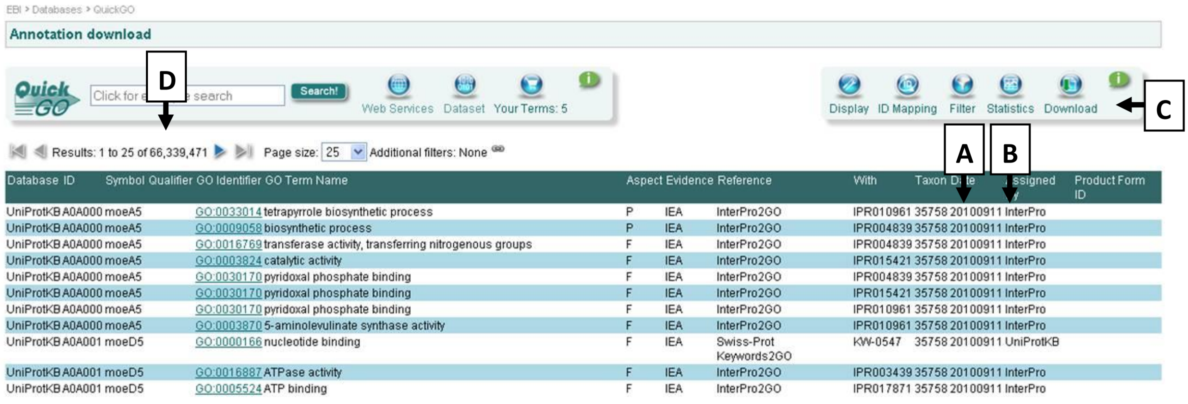
Chapter 2 - learning objectives

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).



Results: 1 to 25 of 66,339,471 Page size: 25 Additional filters: None

Database	ID	Symbol	Qualifier	GO Identifier	GO Term Name	Aspect	Evidence	Reference	With	Taxon	D te	ssigned y	Product Form ID
UniProtKB	A0A000	moeA5		GO:0033014	tetrapyrrole biosynthetic process	P	IEA	InterPro2GO	IPR010961	35758	20100911	InterPro	
UniProtKB	A0A000	moeA5		GO:0009058	biosynthetic process	P	IEA	InterPro2GO	IPR004839	35758	20100911	InterPro	
UniProtKB	A0A000	moeA5		GO:0016769	transferase activity, transferring nitrogenous groups	F	IEA	InterPro2GO	IPR004839	35758	20100911	InterPro	
UniProtKB	A0A000	moeA5		GO:0003824	catalytic activity	F	IEA	InterPro2GO	IPR015421	35758	20100911	InterPro	
UniProtKB	A0A000	moeA5		GO:0030170	pyridoxal phosphate binding	F	IEA	InterPro2GO	IPR004839	35758	20100911	InterPro	
UniProtKB	A0A000	moeA5		GO:0030170	pyridoxal phosphate binding	F	IEA	InterPro2GO	IPR015421	35758	20100911	InterPro	
UniProtKB	A0A000	moeA5		GO:0030170	pyridoxal phosphate binding	F	IEA	InterPro2GO	IPR010961	35758	20100911	InterPro	
UniProtKB	A0A000	moeA5		GO:0003870	5-aminolevulinate synthase activity	F	IEA	InterPro2GO	IPR010961	35758	20100911	InterPro	
UniProtKB	A0A001	moeD5		GO:0000166	nucleotide binding	F	IEA	Swiss-Prot Keywords2GO	KW-0547	35758	20100911	UniProtKB	
UniProtKB	A0A001	moeD5		GO:0016887	ATPase activity	F	IEA	InterPro2GO	IPR003439	35758	20100911	InterPro	
UniProtKB	A0A001	moeD5		GO:0005524	ATP binding	F	IEA	InterPro2GO	IPR017871	35758	20100911	InterPro	

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

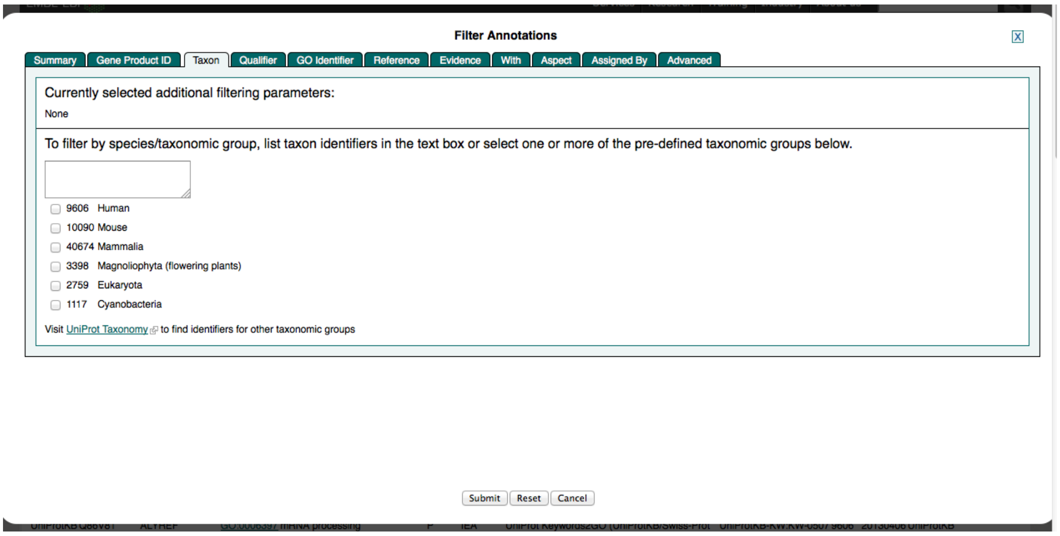
[A] Annotation sets can be filtered by clicking on the 'Filter' button. Filters include taxon, evidence code, GO ID and protein identifier.

[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.



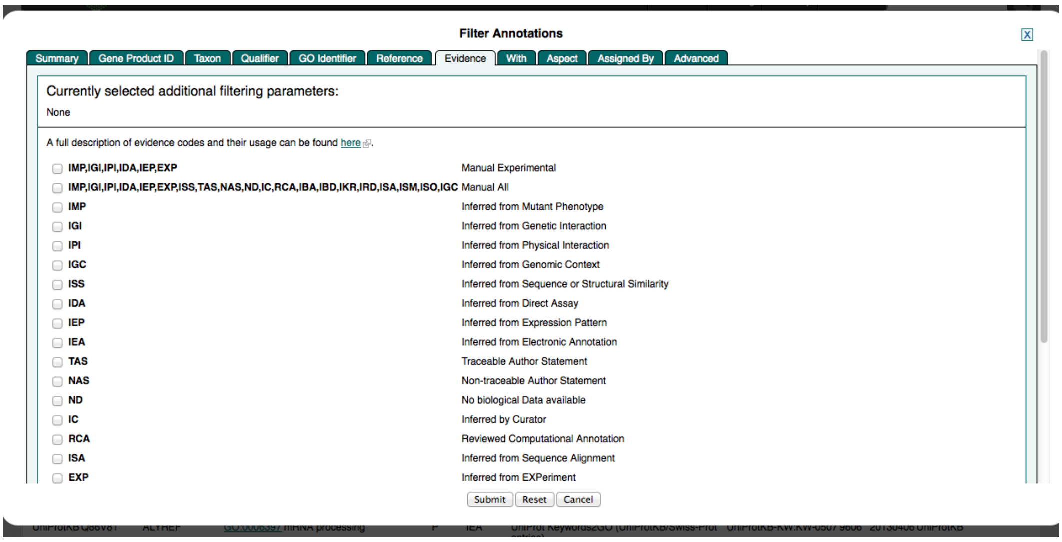
Currently selected additional filtering parameters:
None

To filter by species/taxonomic group, list taxon identifiers in the text box or select one or more of the pre-defined taxonomic groups below.

☐ 9606 Human
☐ 10090 Mouse
☐ 40674 Mammalia
☐ 3398 Magnoliophyta (flowering plants)
☐ 2759 Eukaryota
☐ 1117 Cyanobacteria

Visit UniProt Taxonomy to find identifiers for other taxonomic groups

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.



Currently selected additional filtering parameters:

None

A full description of evidence codes and their usage can be found here.

☐ **IMP,IGI,IPI,IDA,IEP,EXP**	Manual Experimental
☐ **IMP,IGI,IPI,IDA,IEP,EXP,ISS,TAS,NAS,ND,IC,RCA,IBA,IBD,IKR,IRD,ISA,ISM,ISO,IGC**	Manual All
☐ **IMP**	Inferred from Mutant Phenotype
☐ **IGI**	Inferred from Genetic Interaction
☐ **IPI**	Inferred from Physical Interaction
☐ **IGC**	Inferred from Genomic Context
☐ **ISS**	Inferred from Sequence or Structural Similarity
☐ **IDA**	Inferred from Direct Assay
☐ **IEP**	Inferred from Expression Pattern
☐ **IEA**	Inferred from Electronic Annotation
☐ **TAS**	Traceable Author Statement
☐ **NAS**	Non-traceable Author Statement
☐ **ND**	No biological Data available
☐ **IC**	Inferred by Curator
☐ **RCA**	Reviewed Computational Annotation
☐ **ISA**	Inferred from Sequence Alignment
☐ **EXP**	Inferred from EXPeriment

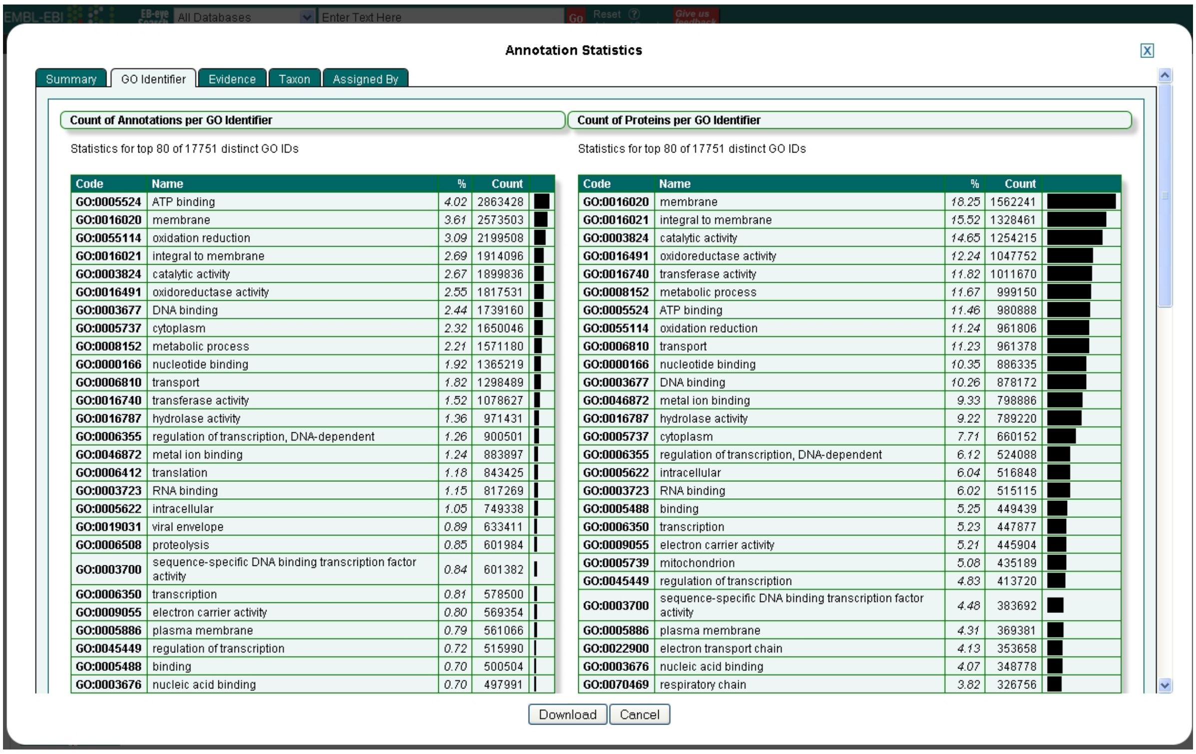
Fig. 2.3 Filter by 'Evidence' tab. Users can choose to see annotations that use only certain evidence codes by for either by selecting one or more from the list.

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.



Annotation Statistics

Count of Annotations per GO Identifier

Statistics for top 80 of 17751 distinct GO IDs

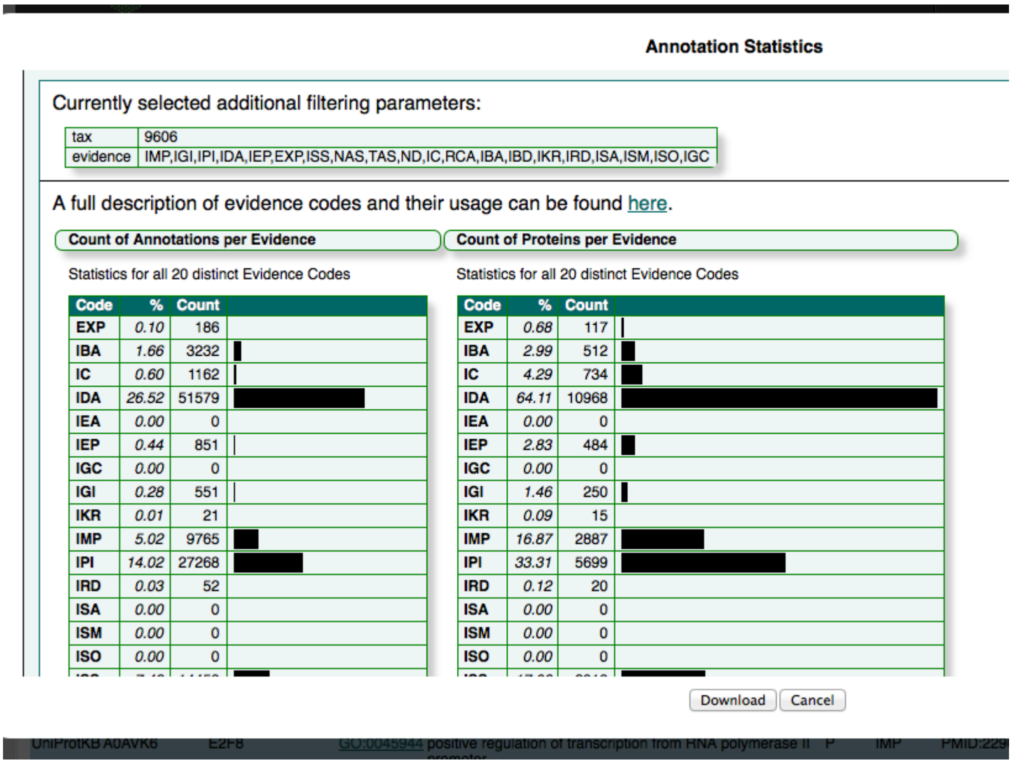
Code	Name	%	Count
GO:0005524	ATP binding	*4.02*	2863428
GO:0016020	membrane	*3.61*	2573503
GO:0055114	oxidation reduction	*3.09*	2199508
GO:0016021	integral to membrane	*2.69*	1914096
GO:0003824	catalytic activity	*2.67*	1899836
GO:0016491	oxidoreductase activity	*2.55*	1817531
GO:0003677	DNA binding	*2.44*	1739160
GO:0005737	cytoplasm	*2.32*	1650046
GO:0008152	metabolic process	*2.21*	1571180
GO:0000166	nucleotide binding	*1.92*	1365219
GO:0006810	transport	*1.82*	1298489
GO:0016740	transferase activity	*1.52*	1078627
GO:0016787	hydrolase activity	*1.36*	971431
GO:0006355	regulation of transcription, DNA-dependent	*1.26*	900501
GO:0046872	metal ion binding	*1.24*	883897
GO:0006412	translation	*1.18*	843425
GO:0003723	RNA binding	*1.15*	817269
GO:0005622	intracellular	*1.05*	749338
GO:0019031	viral envelope	*0.89*	633411
GO:0006508	proteolysis	*0.85*	601984
GO:0003700	sequence-specific DNA binding transcription factor activity	*0.84*	601382
GO:0006350	transcription	*0.81*	578500
GO:0009055	electron carrier activity	*0.80*	569354
GO:0005886	plasma membrane	*0.79*	561066
GO:0045449	regulation of transcription	*0.72*	515990
GO:0005488	binding	*0.70*	500504
GO:0003676	nucleic acid binding	*0.70*	497991

Count of Proteins per GO Identifier

Statistics for top 80 of 17751 distinct GO IDs

Code	Name	%	Count
GO:0016020	membrane	*18.25*	1562241
GO:0016021	integral to membrane	*15.52*	1328461
GO:0003824	catalytic activity	*14.65*	1254215
GO:0016491	oxidoreductase activity	*12.24*	1047752
GO:0016740	transferase activity	*11.82*	1011670
GO:0008152	metabolic process	*11.67*	999150
GO:0005524	ATP binding	*11.46*	980888
GO:0055114	oxidation reduction	*11.24*	961806
GO:0006810	transport	*11.23*	961378
GO:0000166	nucleotide binding	*10.35*	886335
GO:0003677	DNA binding	*10.26*	878172
GO:0046872	metal ion binding	*9.33*	798886
GO:0016787	hydrolase activity	*9.22*	789220
GO:0005737	cytoplasm	*7.71*	660152
GO:0006355	regulation of transcription, DNA-dependent	*6.12*	524088
GO:0005622	intracellular	*6.04*	516848
GO:0003723	RNA binding	*6.02*	515115
GO:0005488	binding	*5.25*	449439
GO:0006350	transcription	*5.23*	447877
GO:0009055	electron carrier activity	*5.21*	445904
GO:0005739	mitochondrion	*5.08*	435189
GO:0045449	regulation of transcription	*4.83*	413720
GO:0003700	sequence-specific DNA binding transcription factor activity	*4.48*	383692
GO:0005886	plasma membrane	*4.31*	369381
GO:0022900	electron transport chain	*4.13*	353658
GO:0003676	nucleic acid binding	*4.07*	348778
GO:0070469	respiratory chain	*3.82*	326756

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



Annotation Statistics

Currently selected additional filtering parameters:

tax	9606
evidence	IMP,IGI,IPI,IDA,IEP,EXP,ISS,NAS,TAS,ND,IC,RCA,IBA,IBD,IKR,IRD,ISA,ISM,ISO,IGC

A full description of evidence codes and their usage can be found here.

Count of Annotations per Evidence

Statistics for all 20 distinct Evidence Codes

Code	%	Count
EXP	*0.10*	186
IBA	*1.66*	3232
IC	*0.60*	1162
IDA	*26.52*	51579
IEA	*0.00*	0
IEP	*0.44*	851
IGC	*0.00*	0
IGI	*0.28*	551
IKR	*0.01*	21
IMP	*5.02*	9765
IPI	*14.02*	27268
IRD	*0.03*	52
ISA	*0.00*	0
ISM	*0.00*	0
ISO	*0.00*	0

Count of Proteins per Evidence

Statistics for all 20 distinct Evidence Codes

Code	%	Count
EXP	*0.68*	117
IBA	*2.99*	512
IC	*4.29*	734
IDA	*64.11*	10968
IEA	*0.00*	0
IEP	*2.83*	484
IGC	*0.00*	0
IGI	*1.46*	250
IKR	*0.09*	15
IMP	*16.87*	2887
IPI	*33.31*	5699
IRD	*0.12*	20
ISA	*0.00*	0
ISM	*0.00*	0
ISO	*0.00*	0

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/quickgo_query.txt)* 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.

Q

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.*

Q

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 InterProScan to rapidly populate novel sequences with electronic GO annotation predictions

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

Chapter 3 - learning objectives

You will learn:

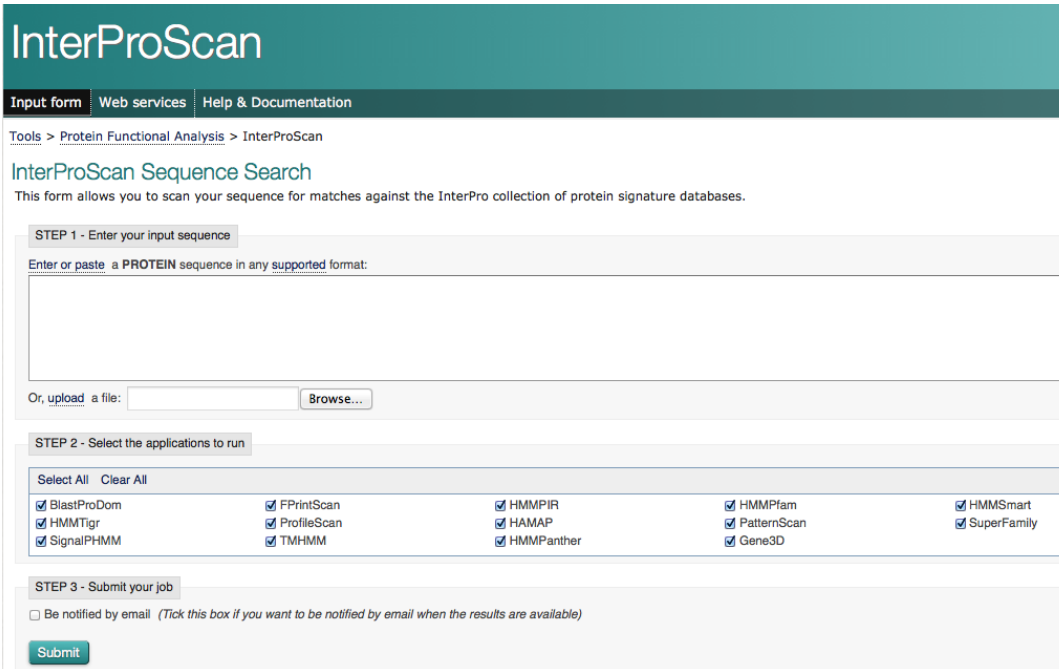
How to use the InterProScan service from InterPro to find GO annotations to novel sequences

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

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. 3.1) 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.



InterProScan Sequence Search

This form allows you to scan your sequence for matches against the InterPro collection of protein signature databases.

STEP 1 - Enter your input sequence

Enter or paste a **PROTEIN** sequence in any supported format:

Or, upload a file:

STEP 2 - Select the applications to run

☑ BlastProDom
☑ HMMTigr
☑ SignalPHMM
☑ FPrintScan
☑ ProfileScan
☑ TMHMM
☑ HMMPIR
☑ HAMAP
☑ HMMPanther
☑ HMMPfam
☑ PatternScan
☑ Gene3D
☑ HMMSmart
☑ SuperFamily

STEP 3 - Submit your job

☐ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Fig. 3.1 InterProScan home page (http://www.ebi.ac.uk/Tools/pfa/iprscan/)

Chapter 3: 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.*

Q

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