GO annotation

EMBL-EBI

GOA Group

goa@ebi.ac.uk

Emily Dimmer
(GOA coordinator)

Rachael Huntley
(GOA curator)

Yasmin Alam-Faruque
(KRUK Renal GO annotation project)

Tony Sawford
(QuickGO, P2GO and database)

In addition to:

Manual and electronic annotation and release pipeline contributions from:

UniProt, InterPro, IntAct, InterPro, Integr8, Ensembl.

GO editors at the EBI and other GO consortium groups

EMBL-EBI

Gene Ontology Annotation (GOA) Database

- Member of the GO Consortium since 2001
- Provides over 72 million GO annotations for over 283,000 taxonomic groups in the UniProt KnowledgeBase
- Sole provider of electronic annotations to many species
- Integrates manual annotations from GO Consortium groups
- Manual annotation priority is the human proteome
- Providers of the QuickGO and Protein2GO tools.

Core information needed for a GO annotation

1. Gene or gene product identifier

   e.g. Q9ARH1

2. GO term ID

   e.g. GO:0004674 (protein serine threonine kinase)

3. Reference ID

   e.g. PubMed ID: 12374299

   GO_REF:0000001

4. Evidence code

   e.g. IDA

..and also in some cases:

- Qualifiers available to modify interpretation of annotation

NOT

contributes_to

colocalizes_with

- with column (8)

- annotation extension column (16)

Isoform annotation

- The Protein2GO tool allows both UniProt accessions and isoform identifiers to be annotated with GO terms.

“The thapsigargin-insensitive ability of each of the transiently overexpressed SPCA1 isoforms to actively transport Ca2+ into a membrane-delineated Ca2+ store was assessed following expression in COS-1 cells as previously described… the level of 45Ca2+ accumulated in the presence of oxalate by SPCA1a, SPCA1b, and SPCA1d, respectively, was 2.8-, 2.9-, and 4.0-fold increased relative to that of control cells….” PMID:16192278

SPCA1a	**calcium-transporting ATPase activity**	**IDA**
SPCA1b	**calcium-transporting ATPase activity**	**IDA**
SPCA1d	**calcium-transporting ATPase activity**	**IDA**

References

- Every electronic annotation cites a GO reference, which describes the type of method applied to generate a particular annotation (a GO_REF);

Example:

Protein	**GO term identifier**	**Reference**	**Evid.**	**with**
A0A000	GO:0030170 pyridoxal phosphate binding	GO_REF:0000002	IEA	IPR010961

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

References

- Manual annotations tend to use PubMed identifiers to provide support for an annotation.

Protein	GO term identifier	Reference	Evid.	with
A0A181	GO:0007165 signal transduction	PMID:17283332	IDA	

- Although there are occasions where a certain type of manual annotation will require a GO Reference (for instance for ND or ISS-evidenced annotations)

... however these alternative identifiers will be added for you by the Protein2GO tool

Evidence Codes

IEA	Inferred from Electronic Annotation
IDA	Inferred from Direct Assay
IMP	Inferred from Mutant Phenotype
IPI	Inferred from Protein Interaction
IEP	Inferred from Expression Pattern
IGI	Inferred from Genetic Interaction
ISS	Inferred from Sequence or Structural Similarity
IGC	Inferred from Genomic Context
RCA	Reviewed Computational Analysis
TAS	Traceable Author Statement
NAS	Non-traceable Author Statement
IC	Inferred from Curator judgement
ND	No Data available

IDA:

- **Enzyme assays**
- ***In vitro* reconstitution (transcription)**
- **Immunofluorescence**
- **Cell fractionation**

TAS:

- **In the literature source the original experiments are referenced.**

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

Evidence codes (cont'd)

IGC	Inferred from Genomic Context
RCA	Reviewed Computational Analysis
TAS	Traceable Author Statement
NAS	Non-traceable Author Statement
IC	Inferred from Curator judgement
ISS	Inferred from Sequence or Structural Similarity
ND	No Data available

Inferred from Genomic Context (IGC)

- operon structure
- syntenic regions
- pathway analysis
- genome scale analysis of processes

Genomic context includes: the identity of the genes neighboring the gene product in question (i.e. synteny), operon structure, and phylogenetic or other whole genome analysis.

IGC may be used in situations where part of the evidence for the function of a protein is that it is present in a putative operon for which the other members of the operon have strong sequence or literature based evidence for function.

It is encouraged that when using this code with operon structure that the id numbers for the genes in the operon be put in the with/from field.

The IGC evidence code can also be used to annotate gene products encoded by genes within a region of conserved synteny.

http://www.geneontology.org/GO.evidence.shtml?all#igc

Inferred from Reviewed Computational Analysis (RCA)

Used for annotations made from predictions based on computational analyses of large-scale experimental data sets, or on computational analyses that integrate multiple types of data into the analysis.

Acceptable experimental data types include:

- protein-protein interaction data
- synthetic genetic interactions
- sequence-based structural predictions

http://www.geneontology.org/GO.evidence.shtml?all#rca

RCA example:

The mouse kinome: discovery and comparative genomics of all mouse protein kinases PMID:15289607

‘Our use of multiple sequence sources, multiple prediction methods, homology to the human kinome, and manual curation enabled the discovery of previously unreported mouse kinase genes and the extension or correction of >150 known kinase sequences….**Catalytically Inactive Kinases.** Several kinases are known to lack catalytic function and instead serve as scaffolds or kinase substrates. .. The mouse kinome shows an almost identical set of predicted inactive kinases (Table 6)’

MGI:2445052 NOT GO:protein kinase activity RCA

Inferred by Curator (IC)

The IC evidence code is to be used for those cases where an annotation is not supported by any direct evidence, but can be reasonably inferred by a curator from other GO annotations, for which evidence is available.

Note that the with/from field must *always* be filled in with a GO ID when using this evidence code.

http://www.geneontology.org/GO.evidence.shtml?all#ic

Inferred from Sequence Similarity (ISS)

Used when a sequence-based analysis forms the basis for an annotation and *review of the evidence and annotation has been done manually*.

If the annotation has not been reviewed manually, the correct evidence code is IEA

GOA is very restrictive as to the use of ISS annotations. Has not yet enabled the use of the ISS child codes (ISA, ISO or ISM) in Protein2GO.

http://www.geneontology.org/GO.evidence.shtml?all#iss

No Data (ND)

- Can only be used with 3 GO terms:

molecular_function GO:0003674

biological_process GO:0008150

cellular_component GO:0005575

- ND should be used when you have exhausted the literature search and can find no annotation. No need to cite a reference.

- If an author states that a protein has unknown function and the paper is recent (after 2004) then you can assign NAS code.

e.g. **'***SH3P17 has unknown function but contains four SH3 domains'.*

How does GOA annotate to the GO ?

💻 Electronic Annotation

📖 Manual Annotation

- Both these methods have their advantages
- They can be easily distinguished by the evidence code used.

GOA Electronic Annotation methods

1. Mapping of external concepts to GO terms

- InterPro2GO (protein domains)
- SPKW2GO (UniProt/Swiss-Prot keywords)
- HAMAP2GO (Microbial protein annotation)
- EC2GO (Enzyme Commission numbers)
- SPSL2GO (Swiss-Prot subcellular locations)

2. Automatic transfer of annotations to orthologs

- Ensembl Compara projections between orthologs

Manual annotations

Are both internally created...

UniProt, IntAct, InterPro

AgBase SIB

BHF-UCL DFLAT (Tuft's)

HGNC

PINC

Roslin Institute

All use the Protein2GO curation tool and are therefore directly editable

...and integrated from external files:

DictyBase, FlyBase, GDB, GeneDB(S.pombe), Gramene, MGI, Reactome, RGD, SGD, TAIR, TIGR, WormBase, ZFIN, IntAct, LIFEdb and Human Protein Atlas datasets.

Annotation exchange between GO Consortium groups

- Other GO Consortium groups are obliged to integrate manual GO annotations from GOA, for their species

- Groups may decide whether to take both electronic and manual or just manual annotations

- If there are any annotation issues, curators contact the group which generated the annotation to make changes to their files, by;

  - direct email
  - via a GO SourceForge tracker

Dual Taxon Annotations - Annotating gene products that interact with other organisms

- Used when characterizing gene products encoded by one organism that act on or in other organisms
  e.g. from obligate parasitic species

(interactions may be between organisms of the same or different species)

- There is a special set of biological process terms in the GO to describe such activities (child terms of 'multi-organism process' GO: 0051704)

- The second species in the interaction is recorded using an additional Taxon identifier column.

Dual taxon annotation examples:

1. **Bacteria living as endosymbiont in plant cell; secretes protein esp1 into host cytoplasm (where the** Host taxon: 123)

- **Annotation of esp1:**

esp1	GO:host cell cytoplasm	IDA	dual taxon:123

2. **Bacteria secretes protein bad1 which kills the host cell**

- **Annotation of bad1:**

bad1	GO: killing of host cells	IDA	dual taxon:123

3. **Bacterial protein lig1 (taxon: 666) interacts with rec5 from bacteria of taxon 999, enabling them to form a biofilm**

- **Annotation of lig1 and rec1:**

lig1	GO:multi-species biofilm formation	IPI	'with'	rec1	dual taxon:999
rec1	GO: multi-species biofilm formation	IPI	'with'	lig1	dual taxon:666

The ‘Qualifier’ Column

The Qualifier column is used to modify the interpretation of an annotation.

Allowable values are: **NOT**

colocalizes_with

contributes_to

http://www.geneontology.org/GO.annotation.conventions.shtml

The ‘NOT’ qualifier

- 'NOT' **is used to make an explicit note that the gene product is not associated with the GO term.**

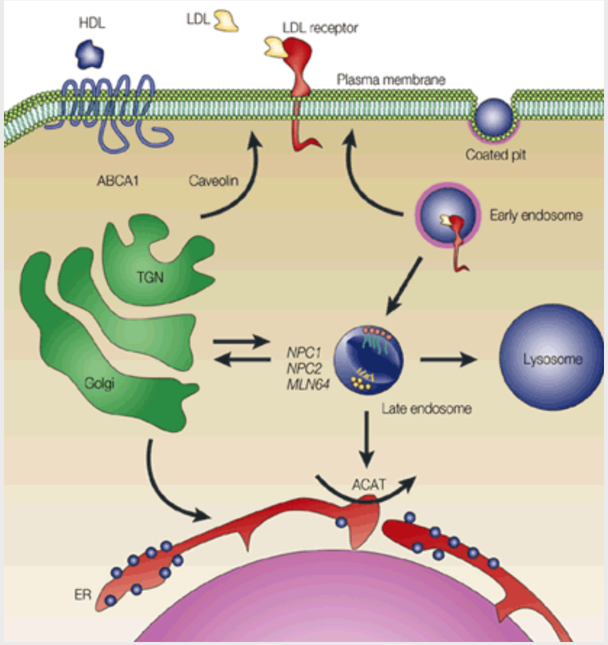
… particularly important when associating a GO term with a gene product should be **avoided** (but might otherwise be made, especially by an automated method).

e.g. This protein does not have ‘kinase activity’ because it has been found that this protein has a disrupted/missing an ‘ATP binding’ domain.

Also used to document conflicting claims in the literature.

NOT can be used with ALL three GO Ontologies.

The ‘colocalizes_with’ qualifier



- Gene products that are **transiently** or **peripherally** associated with an organelle or complex may be annotated to the relevant cellular component term, using the 'colocalizes_with' qualifier.

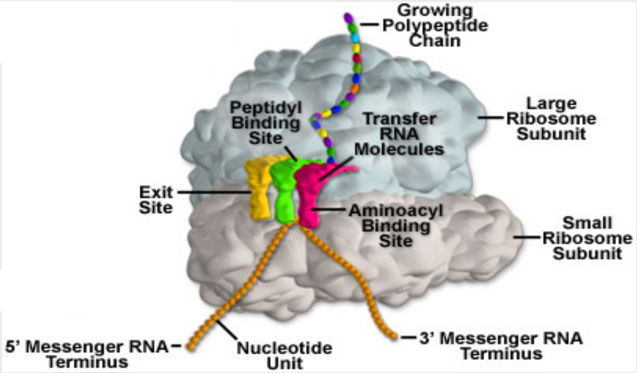
<u>Only</u> used with GO Component Ontology

Colocalizes_with example:

“Interestingly, in quiescent cells, centrosomes are not stained by topoisomerase IIα specific antibodies, indicating that the localization of topoisomerase IIα to the centrioles is restricted to cycling cells.”

TOP2	colocalizes_with	GO:centrioles	IDA

The ‘contributes_to’ qualifier



Individual gene products that are part of a complex can be annotated to terms that describe the action (function or process) of the whole complex.

i.e. annotating 'to the potential of the complex‘

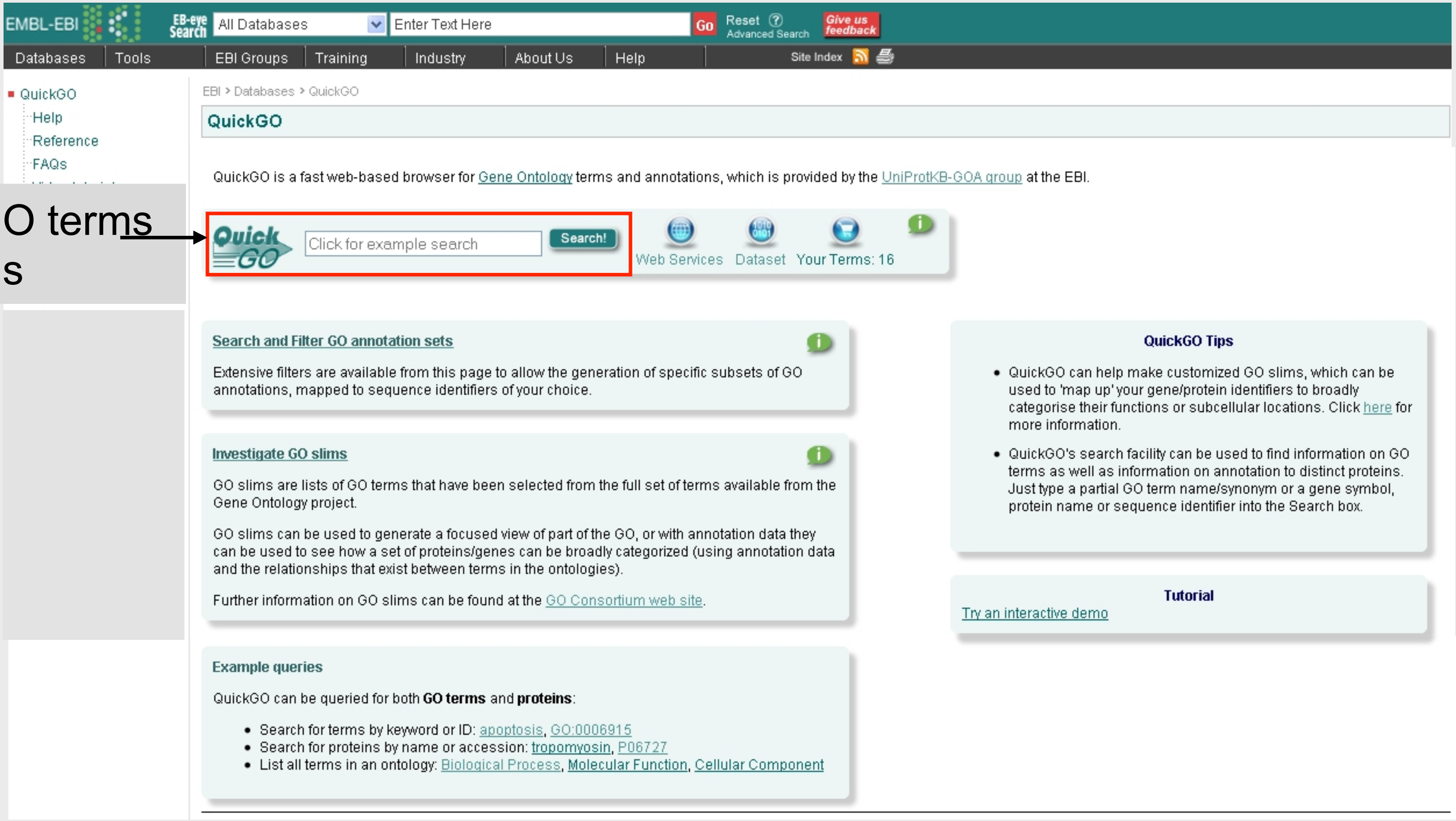
- distinguishes an individual subunit from complex functions

All gene products annotated using 'contributes_to' must also be annotated to a cellular component term representing the complex that possesses the activity.

<u>Only</u> used with GO Function Ontology

The EBI's QuickGO browser

Search GO terms or proteins



QuickGO is a fast web-based browser for Gene Ontology terms and annotations, which is provided by the UniProtKB-GOA group at the EBI.

Search and Filter GO annotation sets

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

Investigate GO slims

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.

Example queries

QuickGO can be queried for both **GO terms** and **proteins**:

- Search for terms by keyword or ID: apoptosis, GO:0006915
- Search for proteins by name or accession: tropomyosin, P06727
- List all terms in an ontology: Biological Process, Molecular Function, Cellular Component

QuickGO Tips

- QuickGO can help make customized GO slims, which can be used to 'map up' your gene/protein identifiers to broadly categorise their functions or subcellular locations. Click here for more information.
- QuickGO's search facility can be used to find information on GO terms as well as information on annotation to distinct proteins. Just type a partial GO term name/synonym or a gene symbol, protein name or sequence identifier into the Search box.

Tutorial

Try an interactive demo

www.ebi.ac.uk/QuickGO

Help for new curators

See the Confluence page;

http://www.ebi.ac.uk/seqdb/confluence/display/GOA/Aids+for+New+GOA+Curators