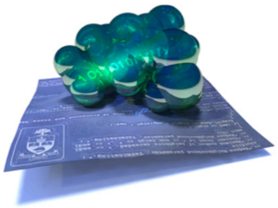
A
Bioinformatics
Course

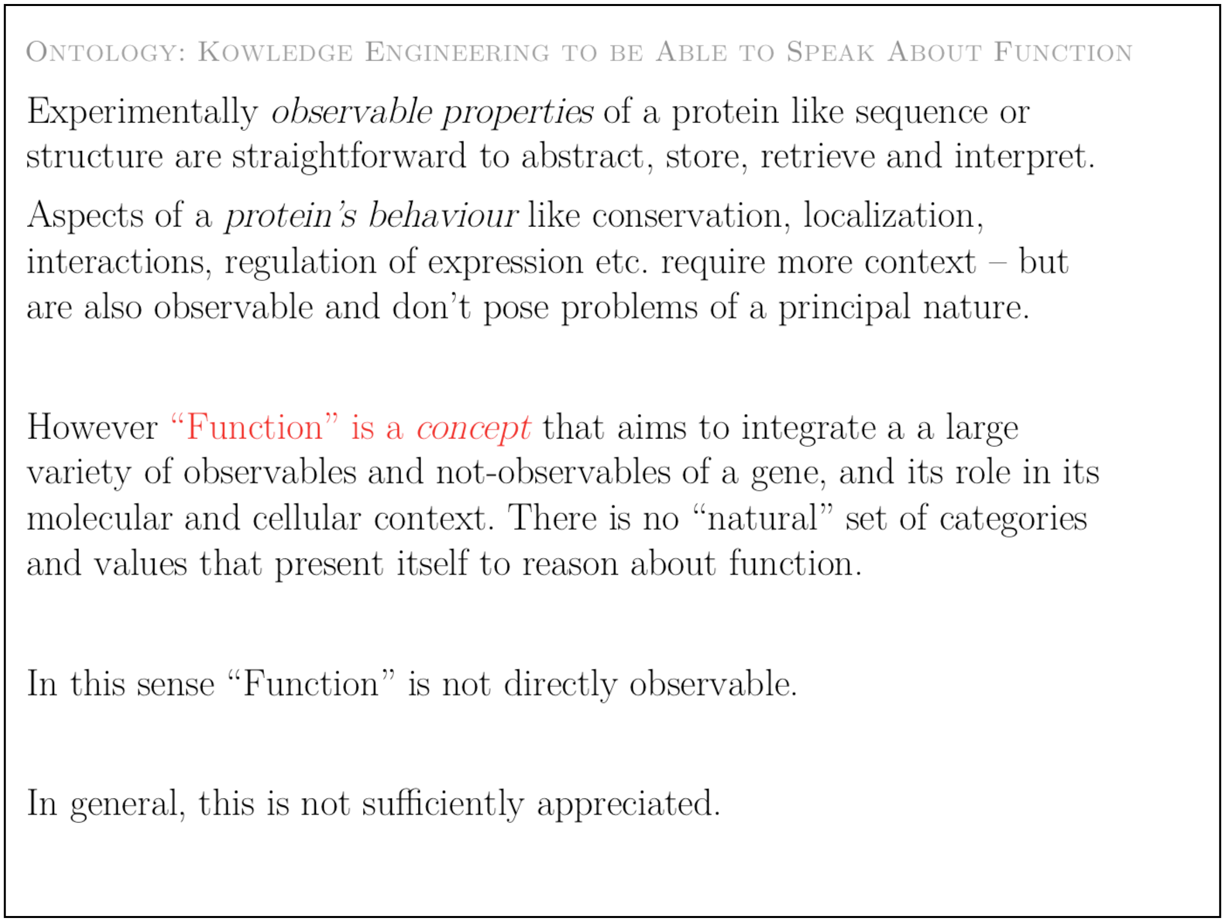
G O

(Gene Ontology)



Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto



Ontology: Kowledge Engineering to be Able to Speak About Function

Experimentally *observable properties* of a protein like sequence or structure are straightforward to abstract, store, retrieve and interpret.

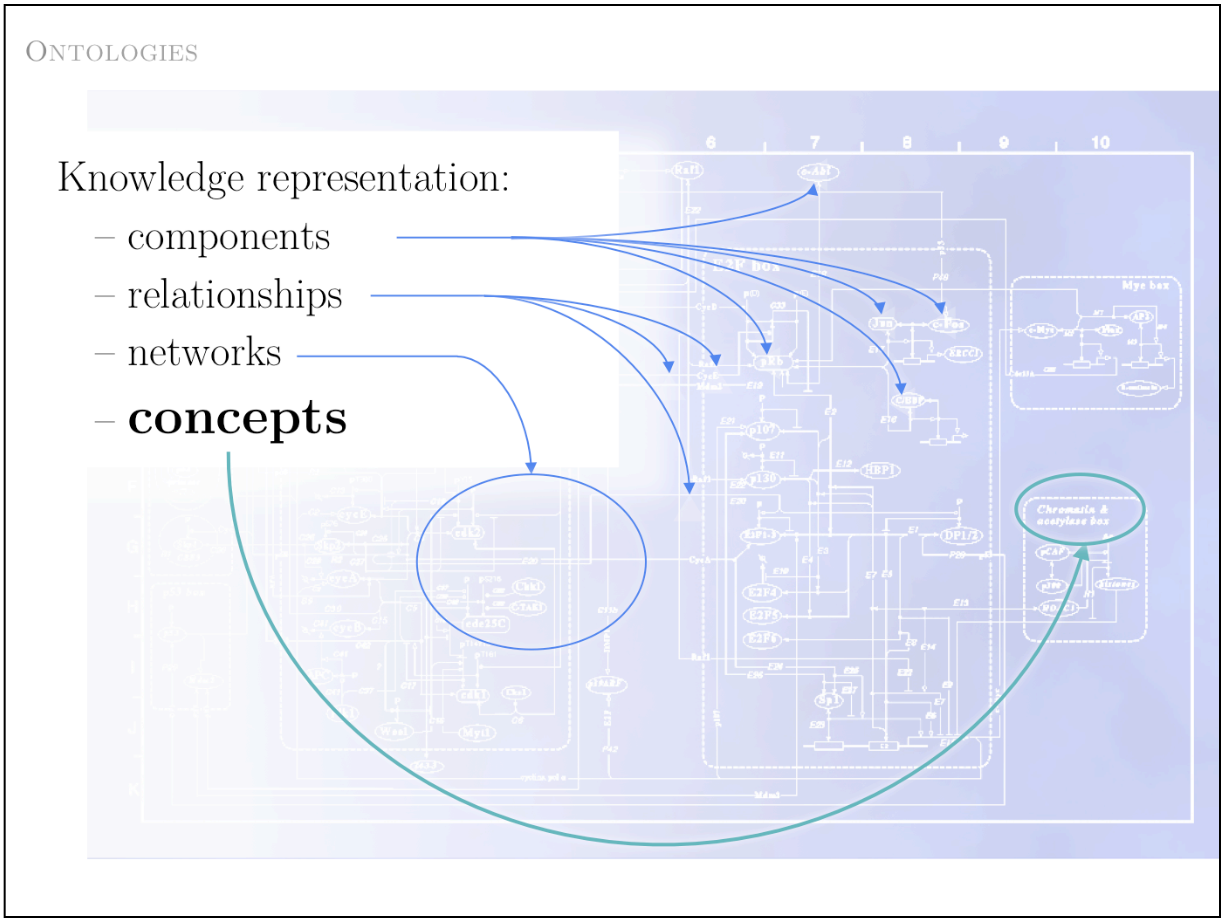
Aspects of a *protein's behaviour* like conservation, localization, interactions, regulation of expression etc. require more context – but are also observable and don't pose problems of a principal nature.

However "Function" is a *concept* that aims to integrate a a large variety of observables and not-observables of a gene, and its role in its molecular and cellular context. There is no "natural" set of categories and values that present itself to reason about function.

In this sense "Function" is not directly observable.

In general, this is not sufficiently appreciated.

That aspect of biology that realizes the ultimate goal of our endeavours is the most elusive: we can't even properly talk about function. The first task of function analysis therefore is knowledge engineering: to define a "language" in which we may categorize, collect and compare functions.



ONTOLOGIES

Knowledge representation:

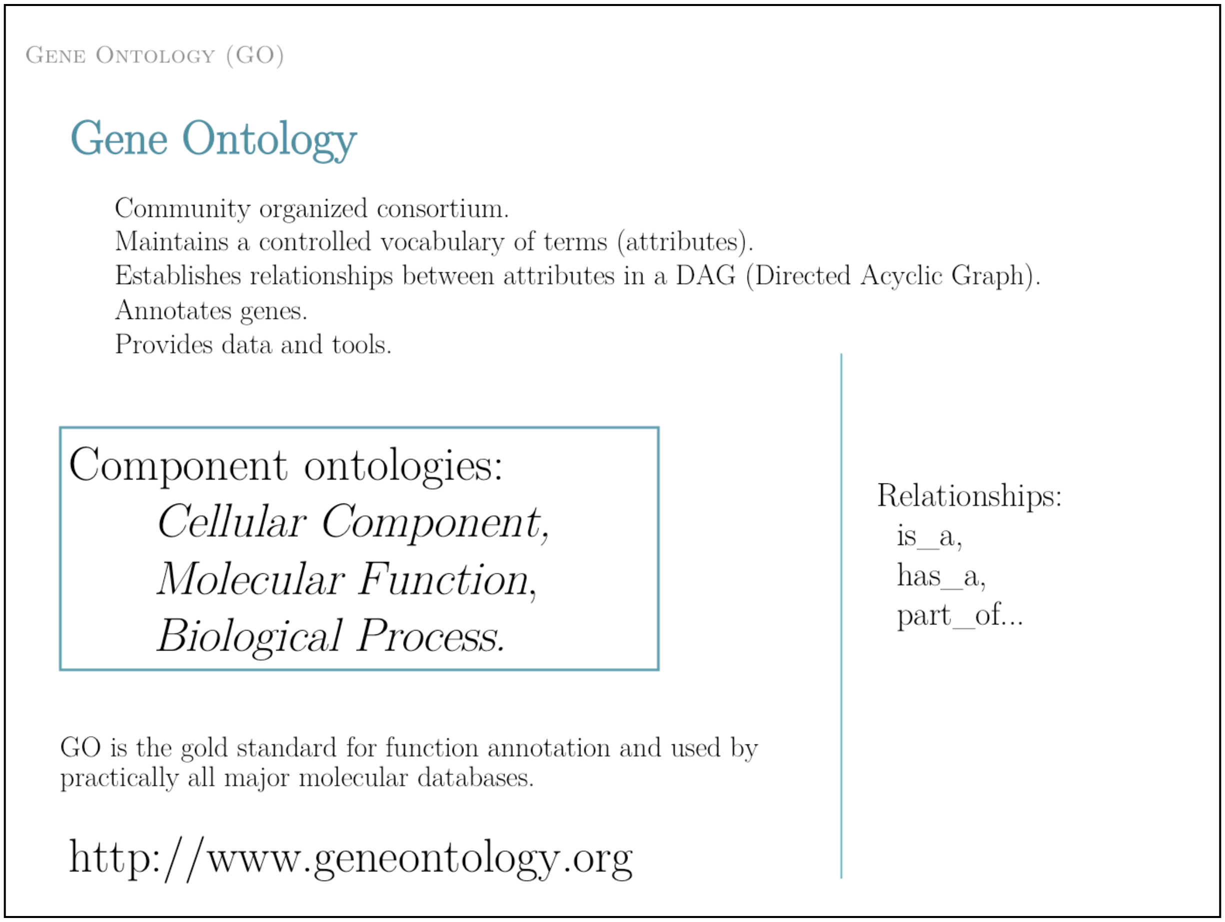
- components
- relationships
- networks
- **concepts**

If we believe everything of interest about biology can be expressed as entities*, their attributes, and the relationships between them, we are missing the importance of “concepts” that represent our ideas about *why* systems are composed the way they are.

Considering biomolecular sequences, and the structure of activities of biomolecules, it is obvious that it is possible to order and arrange those to pathways and networks of activities (albeit that is not trivial in practice).

However there is no automatic way to structure and categorize sets of such relationships into a higher level “understanding”, into insight *why* the observed entities and relationships give rise to their behaviour, and, *what the “purpose” of the biomolecular systems is*, that {purpose / meaning / objective / fitness function} that has been selected by evolution.

* an “entity” can be a gene, protein or metabolite, a compartment, an activity ...



GENE ONTOLOGY (GO)

Gene Ontology

Community organized consortium.
Maintains a controlled vocabulary of terms (attributes).
Establishes relationships between attributes in a DAG (Directed Acyclic Graph).
Annotates genes.
Provides data and tools.

Component ontologies:
Cellular Component,
Molecular Function,
Biological Process.

Relationships:
is_a,
has_a,
part_of...

GO is the gold standard for function annotation and used by practically all major molecular databases.

http://www.geneontology.org

In order to speak about function, we need a common language. A common language means that we use the same terms to describe the same facts. To establish a common language about function for molecular biology is the goal of the Gene Ontology (GO) consortium.

The term “ontology” comes from the domain of *knowledge engineering*. An ontology collects terms that describe facts, their definitions, and their relationships.

GO has grown over many years and it is a well funded, large and mature project.

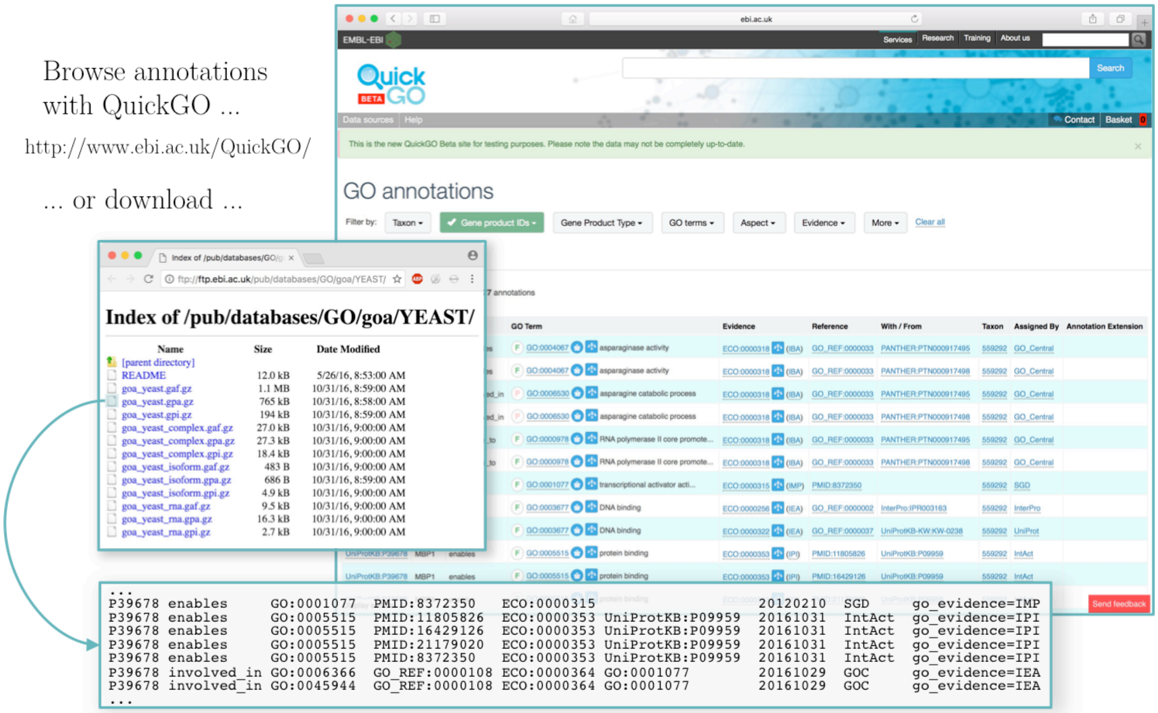
The graph of GO definitions contains three distinct component ontologies. Their root nodes are Cellular Componen, Molecular Function, and Biological Process.

Gene Ontology Annotations

Browse annotations with QuickGO ...

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

... or download ...



Index of /pub/databases/GO/goa/YEAST/

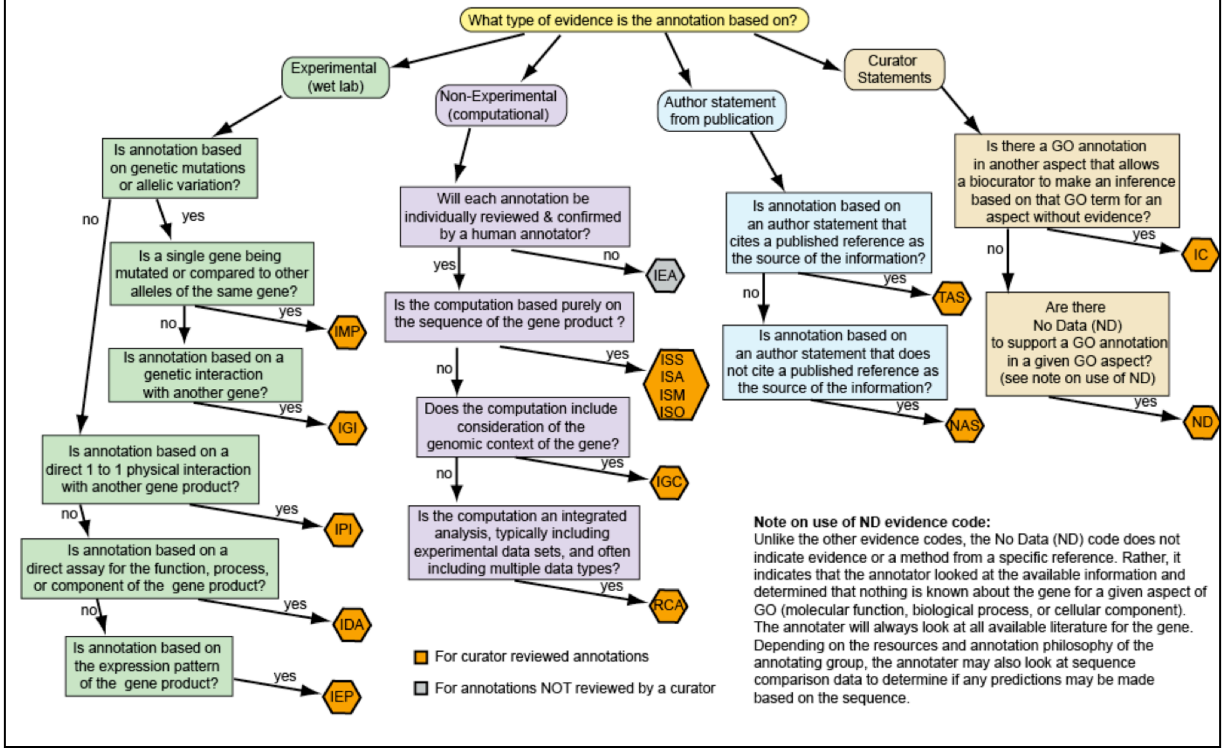
ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/YEAST/

Name	Size	Date Modified
[parent directory]		
README	12.0 kB	5/26/16, 8:53:00 AM
goa_yeast.gaf.gz	1.1 MB	10/31/16, 8:59:00 AM
goa_yeast.gpa.gz	765 kB	10/31/16, 8:58:00 AM
goa_yeast.gpi.gz	194 kB	10/31/16, 8:59:00 AM
goa_yeast_complex.gaf.gz	27.0 kB	10/31/16, 9:00:00 AM
goa_yeast_complex.gpa.gz	27.3 kB	10/31/16, 9:00:00 AM
goa_yeast_complex.gpi.gz	18.4 kB	10/31/16, 9:00:00 AM
goa_yeast_isoform.gaf.gz	483 B	10/31/16, 9:00:00 AM
goa_yeast_isoform.gpa.gz	686 B	10/31/16, 8:59:00 AM
goa_yeast_isoform.gpi.gz	4.9 kB	10/31/16, 9:00:00 AM
goa_yeast_rna.gaf.gz	9.5 kB	10/31/16, 9:00:00 AM
goa_yeast_rna.gpa.gz	16.3 kB	10/31/16, 9:00:00 AM
goa_yeast_rna.gpi.gz	2.7 kB	10/31/16, 9:00:00 AM

```
...
P39678 enables     GO:0001077 PMID:8372350   ECO:0000315                 20120210  SGD      go_evidence=IMP
P39678 enables     GO:0005515 PMID:11805826  ECO:0000353 UniProtKB:P09959 20161031  IntAct   go_evidence=IPI
P39678 enables     GO:0005515 PMID:16429126  ECO:0000353 UniProtKB:P09959 20161031  IntAct   go_evidence=IPI
P39678 enables     GO:0005515 PMID:21179020  ECO:0000353 UniProtKB:P09959 20161031  IntAct   go_evidence=IPI
P39678 enables     GO:0005515 PMID:8372350   ECO:0000353 UniProtKB:P09959 20161031  IntAct   go_evidence=IPI
P39678 involved_in GO:0006366 GO_REF:0000108 ECO:0000364 GO:0001077       20161029  GOC      go_evidence=IEA
P39678 involved_in GO:0045944 GO_REF:0000108 ECO:0000364 GO:0001077       20161029  GOC      go_evidence=IEA
...
```

Gene Ontology Annotation (GOA) is part of GO's biocuration effort, aiming to associate proteins in UniProtKB with their repsective GO terms. After searching for genes of interest in the QuickGO browser, you can select GO terms that you would like to explore, find all proteins annotated to these nodes, and filter for particular taxons.

Gene Ontology Annotations (GOA) – Evidence Code Decision Tree


What type of evidence is the annotation based on?
Experimental (wet lab)
Non-Experimental (computational)
Author statement from publication
Curator Statements
Is annotation based on genetic mutations or allelic variation?
yes
no
Is a single gene being mutated or compared to other alleles of the same gene?
yes
IMP
no
Is annotation based on a genetic interaction with another gene?
yes
IGI
Is annotation based on a direct 1 to 1 physical interaction with another gene product?
yes
IPI
no
Is annotation based on a direct assay for the function, process, or component of the gene product?
yes
IDA
no
Is annotation based on the expression pattern of the gene product?
yes
IEP
Will each annotation be individually reviewed & confirmed by a human annotator?
no
IEA
yes
Is the computation based purely on the sequence of the gene product ?
yes
ISS
ISA
ISM
ISO
no
Does the computation include consideration of the genomic context of the gene?
yes
IGC
no
Is the computation an integrated analysis, typically including experimental data sets, and often including multiple data types?
yes
RCA
Is annotation based on an author statement that cites a published reference as the source of the information?
yes
TAS
no
Is annotation based on an author statement that does not cite a published reference as the source of the information?
yes
NAS
Is there a GO annotation in another aspect that allows a biocurator to make an inference based on that GO term for an aspect without evidence?
yes
IC
no
Are there No Data (ND) to support a GO annotation in a given GO aspect? (see note on use of ND)
yes
ND
🟧 For curator reviewed annotations
⬜ For annotations NOT reviewed by a curator

Note on use of ND evidence code:
Unlike the other evidence codes, the No Data (ND) code does not indicate evidence or a method from a specific reference. Rather, it indicates that the annotator looked at the available information and determined that nothing is known about the gene for a given aspect of GO (molecular function, biological process, or cellular component). The annotater will always look at all available literature for the gene. Depending on the resources and annotation philosophy of the annotating group, the annotater may also look at sequence comparison data to determine if any predictions may be made based on the sequence.

GO as based on definitions, but annotations in GOA reqiure evidence. GOA evidence codes make the evidence explicit for every annotation. They are a crucial part of GOA that you **must** be familier with to work with the data and subset it according to your application's needs.

Experimental Evidence Codes

DA: Direct Assay
IPI: Physical Interaction
IMP: Mutant Phenotype
IGI: Genetic Interaction
IEP: Expression Pattern

Computational Analysis Evidence Codes

ISS: Sequence or Structural Similarity
ISO: Sequence Orthology
ISA: Sequence Alignment
ISM: Sequence Model
IGC: Genomic Context
RCA: Reviewed Computational Analysis

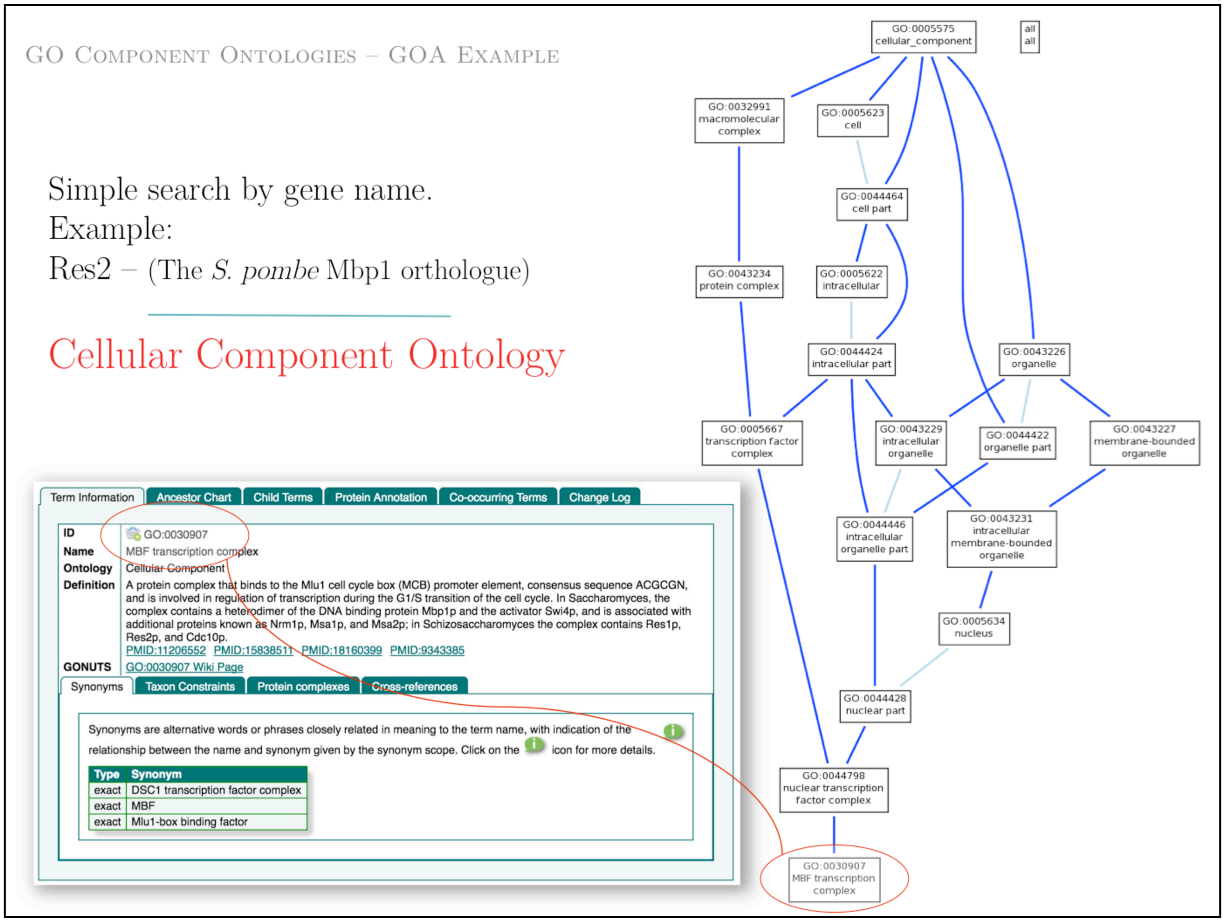
Author Statement Evidence Codes

TAS: Traceable Author Statement
NAS: Non-traceable Author Statement

Curator Statement Evidence Codes

IC: Inferred by Curator
ND: No biological Data available

http://www.geneontology.org/page/evidence-code-decision-tree



GO Component Ontologies – GOA Example

Simple search by gene name.
Example:
Res2 – (The *S. pombe* Mbp1 orthologue)

Cellular Component Ontology

ID GO:0030907
Name MBF transcription complex
Ontology Cellular Component
Definition A protein complex that binds to the Mlu1 cell cycle box (MCB) promoter element, consensus sequence ACGCGN, and is involved in regulation of transcription during the G1/S transition of the cell cycle. In Saccharomyces, the complex contains a heterodimer of the DNA binding protein Mbp1p and the activator Swi4p, and is associated with additional proteins known as Nrm1p, Msa1p, and Msa2p; in Schizosaccharomyces the complex contains Res1p, Res2p, and Cdc10p.
PMID:11206552 PMID:15838511 PMID:18160399 PMID:9343385
GONUTS GO:0030907 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 i icon for more details.

Type	Synonym
exact	DSC1 transcription factor complex
exact	MBF
exact	Mlu1-box binding factor

GO is divided into three complementary views of function: *Cellular Components* ...

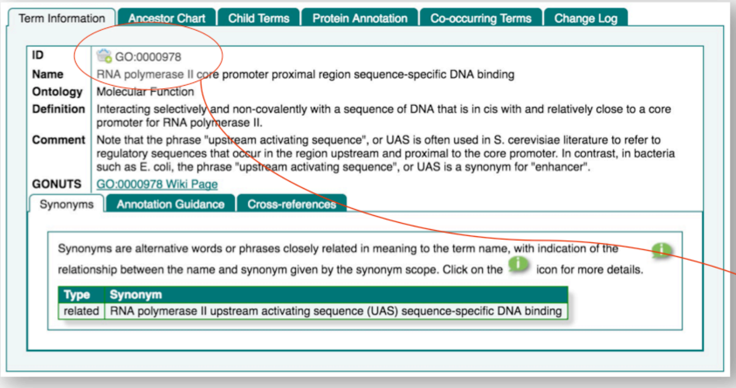
> "These terms describe a component of a cell that is part of a larger object, such as an anatomical structure (e.g. rough endoplasmic reticulum or nucleus) or a gene product group (e.g. ribosome, proteasome or a protein dimer)."

http://www.geneontology.org/page/ontology-documentation

GO Component Ontologies – GOA Example

Simple search by gene name.
Example:
Res2 – (The *S. pombe* Mbp1 orthologue)

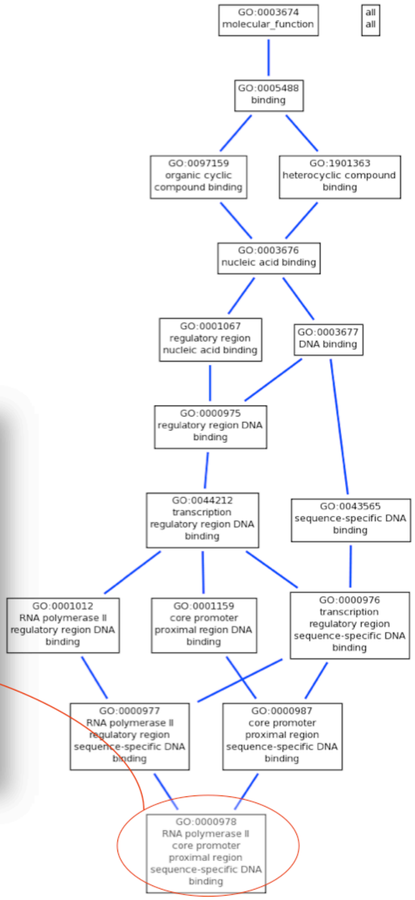
Molecular Function Ontology



ID GO:0000978
Name RNA polymerase II core promoter proximal region sequence-specific DNA binding
Ontology Molecular Function
Definition Interacting selectively and non-covalently with a sequence of DNA that is in cis with and relatively close to a core promoter for RNA polymerase II.
Comment Note that the phrase "upstream activating sequence", or UAS is often used in S. cerevisiae literature to refer to regulatory sequences that occur in the region upstream and proximal to the core promoter. In contrast, in bacteria such as E. coli, the phrase "upstream activating sequence", or UAS is a synonym for "enhancer".
GONUTS GO:0000978 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.

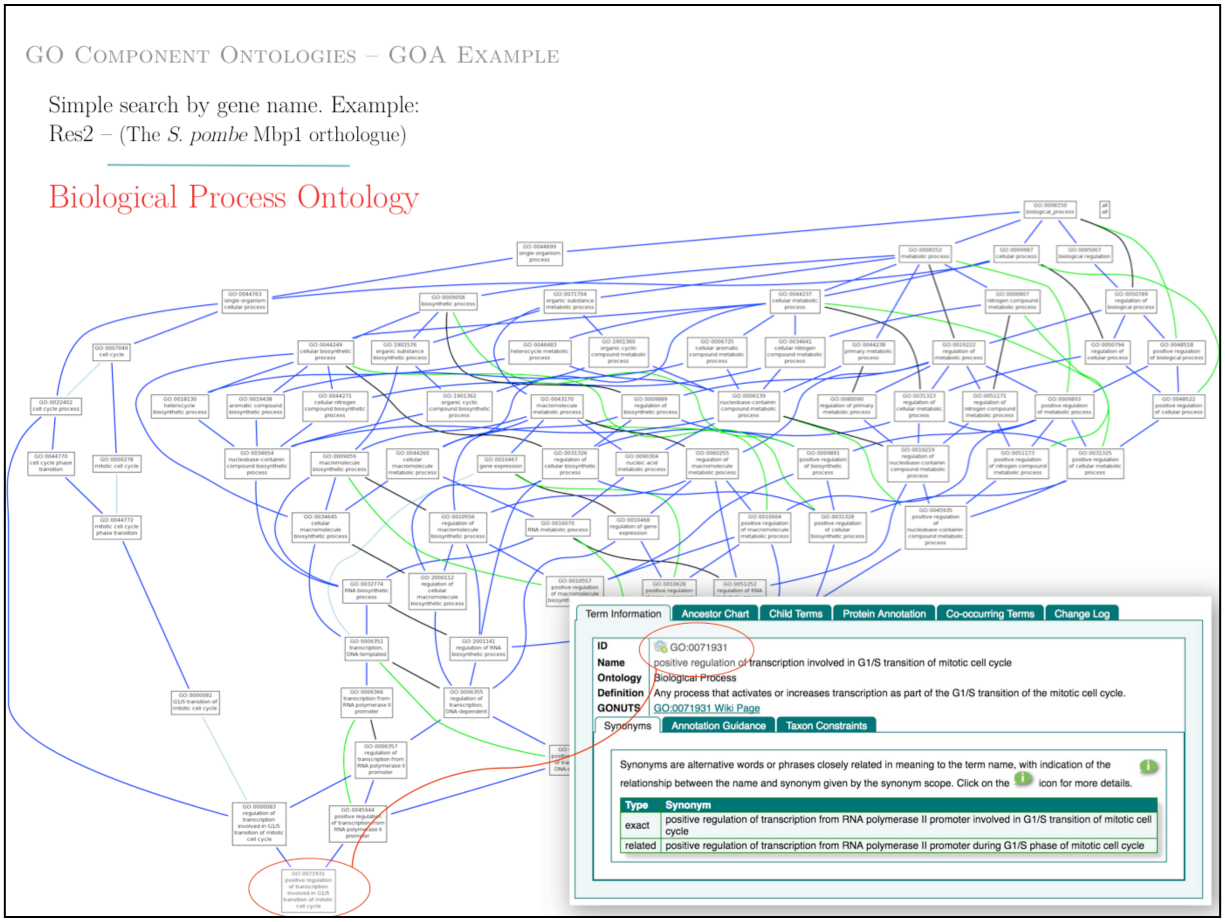
Type	Synonym
related	RNA polymerase II upstream activating sequence (UAS) sequence-specific DNA binding


GO:0003674 molecular_function
all all
GO:0005488 binding
GO:0097159 organic cyclic compound binding
GO:1901363 heterocyclic compound binding
GO:0003676 nucleic acid binding
GO:0001067 regulatory region nucleic acid binding
GO:0003677 DNA binding
GO:0000975 regulatory region DNA binding
GO:0044212 transcription regulatory region DNA binding
GO:0043565 sequence-specific DNA binding
GO:0001012 RNA polymerase II regulatory region DNA binding
GO:0001159 core promoter proximal region DNA binding
GO:0000976 transcription regulatory region sequence-specific DNA binding
GO:0000977 RNA polymerase II regulatory region sequence-specific DNA binding
GO:0000987 core promoter proximal region sequence-specific DNA binding
GO:0000978 RNA polymerase II core promoter proximal region sequence-specific DNA binding

… *Molecular Function* …

> "Molecular function terms describes activities that occur at the molecular level, such as "catalytic activity" or "binding activity". GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where, when, or in what context the action takes place. Molecular functions generally correspond to activities that can be performed by individual gene products, but some activities are performed by assembled complexes of gene products. Examples of broad functional terms are "catalytic activity" and "transporter activity"; examples of narrower functional terms are "adenylate cyclase activity" or "Toll receptor binding".
>
> It is easy to confuse a gene product name with its molecular function; for that reason GO molecular functions are often appended with the word "activity"."

http://www.geneontology.org/page/ontology-documentation



GO Component Ontologies – GOA Example

Simple search by gene name. Example:
Res2 – (The *S. pombe* Mbp1 orthologue)

Biological Process Ontology

GO:0006366
transcription from RNA polymerase II promoter
GO:0006357
regulation of transcription from RNA polymerase II promoter
GO:0000083
regulation of transcription involved in G1/S transition of mitotic cell cycle
GO:0045944
positive regulation of transcription from RNA polymerase II promoter
GO:0071931
positive regulation of transcription involved in G1/S transition of mitotic cell cycle
GO:0016070
RNA metabolic process
GO:0010468
regulation of gene expression
GO:0010604
positive regulation of macromolecule metabolic process
GO:0031328
positive regulation of cellular biosynthetic process
GO:0045935
positive regulation of nucleobase-containin compound metabolic process
GO:0010557
positive regulation of macromolecule biosynth
GO:0010628
positive regulation
GO:0051252
regulation of RNA
Term Information
Ancestor Chart
Child Terms
Protein Annotation
Co-occurring Terms
Change Log
Synonyms
Annotation Guidance
Taxon Constraints

ID	GO:0071931
Name	positive regulation of transcription involved in G1/S transition of mitotic cell cycle
Ontology	Biological Process
Definition	Any process that activates or increases transcription as part of the G1/S transition of the mitotic cell cycle.
GONUTS	GO:0071931 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
exact	positive regulation of transcription from RNA polymerase II promoter involved in G1/S transition of mitotic cell cycle
related	positive regulation of transcription from RNA polymerase II promoter during G1/S phase of mitotic cell cycle

… and *Biological Process*.

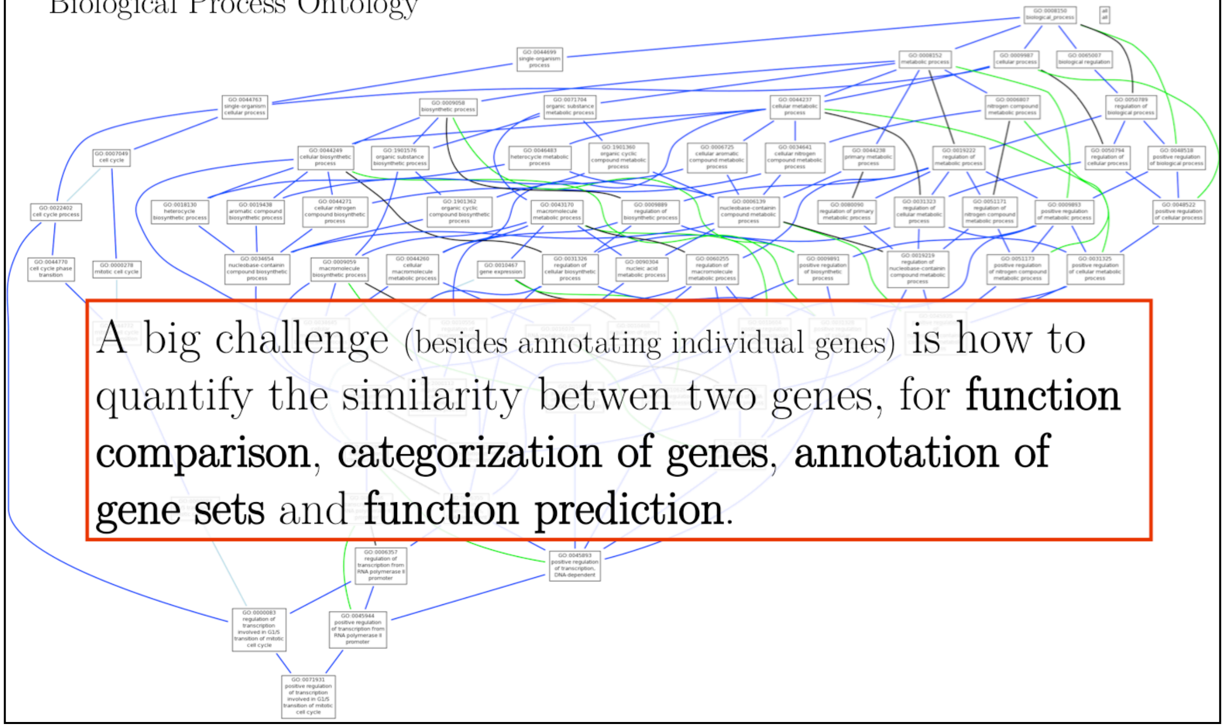
> “A biological process term describes a series of events accomplished by one or more organized assemblies of molecular functions. Examples of broad biological process terms are "cellular physiological process" or "signal transduction". Examples of more specific terms are "pyrimidine metabolic process" or "alpha-glucoside transport". The general rule to assist in distinguishing between a biological process and a molecular function is that a process must have more than one distinct steps.
>
> A biological process is not equivalent to a pathway. At present, the GO does not try to represent the dynamics or dependencies that would be required to fully describe a pathway.”

http://www.geneontology.org/page/ontology-documentation

GO Component Ontologies – GOA Example

Simple search by gene name.mExample:
Res2 – (The *S. pombe* Mbp1 orthologue)

Biological Process Ontology


GO:0008150 biological_process
all all
GO:0044699 single-organism process
GO:0008152 metabolic process
GO:0009987 cellular process
GO:0065007 biological regulation
GO:0044763 single-organism cellular process
GO:0009058 biosynthetic process
GO:0071704 organic substance metabolic process
GO:0044237 cellular metabolic process
GO:0006807 nitrogen compound metabolic process
GO:0050789 regulation of biological process
GO:0007049 cell cycle
GO:0044249 cellular biosynthetic process
GO:1901576 organic substance biosynthetic process
GO:0046483 heterocycle metabolic process
GO:1901360 organic cyclic compound metabolic process
GO:0006725 cellular aromatic compound metabolic process
GO:0034641 cellular nitrogen compound metabolic process
GO:0044238 primary metabolic process
GO:0019222 regulation of metabolic process
GO:0050794 regulation of cellular process
GO:0048518 positive regulation of biological process
GO:0022402 cell cycle process
GO:0018130 heterocycle biosynthetic process
GO:0019438 aromatic compound biosynthetic process
GO:0044271 cellular nitrogen compound biosynthetic process
GO:1901362 organic cyclic compound biosynthetic process
GO:0043170 macromolecule metabolic process
GO:0009889 regulation of biosynthetic process
GO:0006139 nucleobase-containin compound metabolic process
GO:0080090 regulation of primary metabolic process
GO:0031323 regulation of cellular metabolic process
GO:0051171 regulation of nitrogen compound metabolic process
GO:0009893 positive regulation of metabolic process
GO:0048522 positive regulation of cellular process
GO:0044770 cell cycle phase transition
GO:0000278 mitotic cell cycle
GO:0034654 nucleobase-containin compound biosynthetic process
GO:0009059 macromolecule biosynthetic process
GO:0044260 cellular macromolecule metabolic process
GO:0010467 gene expression
GO:0031326 regulation of cellular biosynthetic process
GO:0090304 nucleic acid metabolic process
GO:0060255 regulation of macromolecule metabolic process
GO:0009891 positive regulation of biosynthetic process
GO:0019219 regulation of nucleobase-containin compound metabolic process
GO:0051173 positive regulation of nitrogen compound metabolic process
GO:0031325 positive regulation of cellular metabolic process
GO:0006357 regulation of transcription from RNA polymerase II promoter
GO:0045893 positive regulation of transcription, DNA-dependent
GO:0000083 regulation of transcription involved in G1/S transition of mitotic cell cycle
GO:0045944 positive regulation of transcription from RNA polymerase II promoter
GO:0071931 positive regulation of transcription involved in G1/S transition of mitotic cell cycle

A big challenge (besides annotating individual genes) is how to quantify the similarity betwen two genes, for **function comparison**, **categorization of genes**, **annotation of gene sets** and **function prediction**.

http://steipe.biochemistry.utoronto.ca/abc

BORIS.STEIPE@UTORONTO.CA

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA