# Bioinformatics Algorithms
# COS-BIOL-530/630
# Lecture07

| Days & Times | Room | Meeting Dates |
|---|---|---|
| Tu 2:00PM - 3:50PM | Thomas Gosnell Hall (GOS)-2178 | 01/13/2025 - 04/28/2025 |
| Th 2:00PM - 3:50PM | Thomas Gosnell Hall (GOS)-2178 | 01/13/2025 - 04/28/2025 |

Instructor:
Fernando Rodriguez
email: frvsbi@rit.edu
Office: Orange Hall 1311

# Function of Genes: Knowledgebases - Lecture07-

**Announcements**

**Week 7**
Lecture07

**Exam 1: Lecture01-06/Lab01-06**
- Thursday, February 27th 2pm.

**Week8**
Lab07 on Tuesday

Lecture08/Lab08 on Thursday

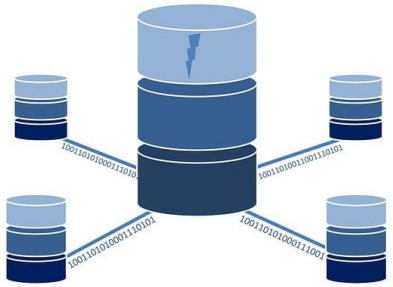
# Function of Genes: Knowledgebases - Lecture07-

**Topics:**

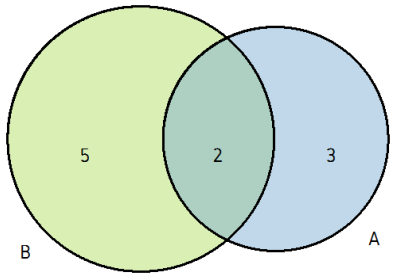
- Gene Ontology (GO)
- GO Terms
- Gene Annotations
- Gene Enrichment Analysis

# Why do we need Vocabularies/Ontologies?

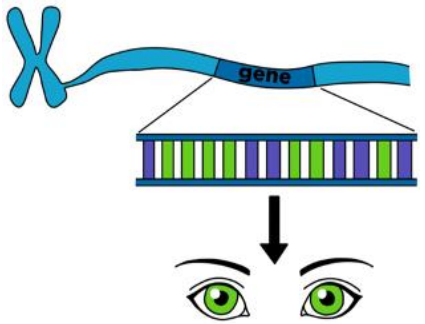
- To search for information across databases.
  - Looking for one term in multiple databases



- To map from one database to another.
  - Looking for shared vocabulary between two databases

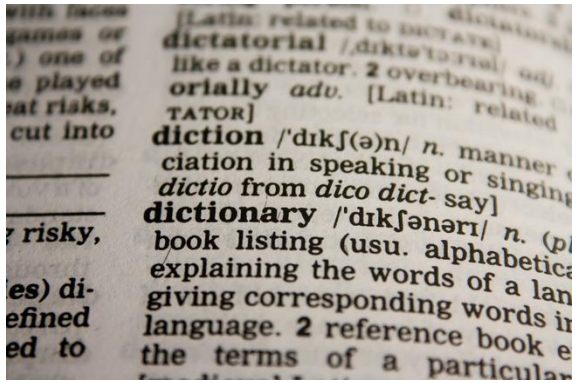



- To Annotate biological data
  - One Gene -> One Function (is that right?)




✓ Microarrays
✓ Genomic sequence
✓ Transcriptomes

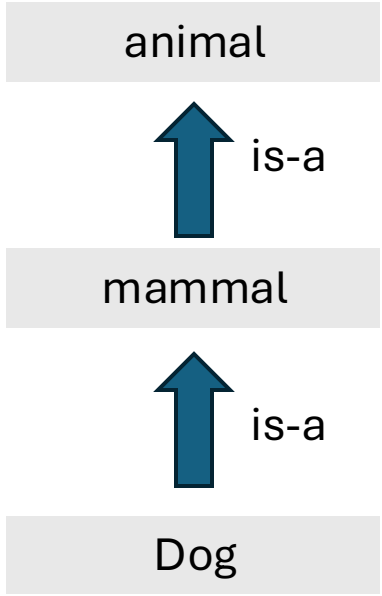
# What is a vocabulary?




- A standardized list of terms and their meaning
- Relationships that place the terms into context

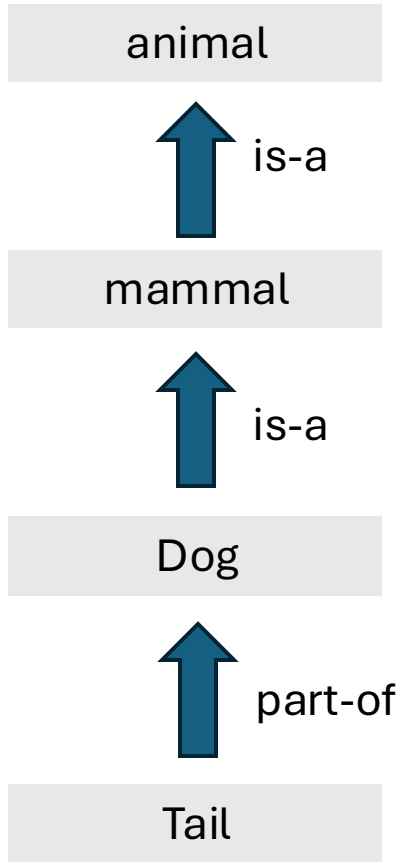
# What is a vocabulary?

- Relationships that place the terms into context:
  - Hierarchical relationship
    - “is-a”



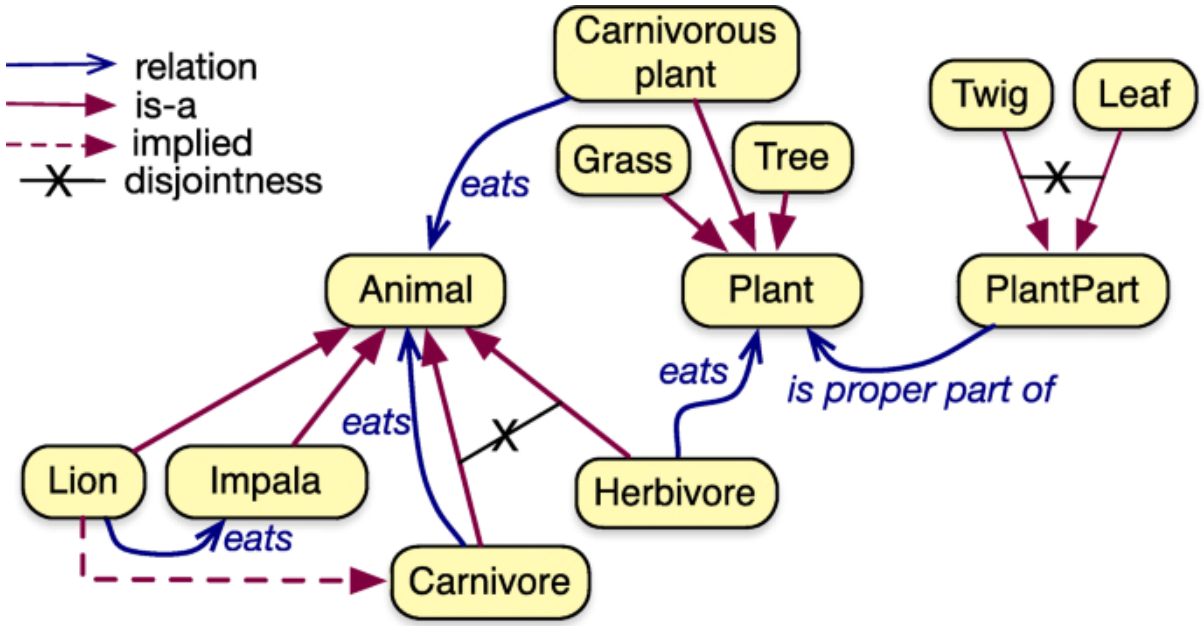
# What is a vocabulary?

- Relationships that place the terms into context:
  - Hierarchical relationship
    - “is-a”
  - Non-hierarchical relationships
    - “part-of”
    - ”has-a”



# What is a vocabulary?

- Relationships that place the terms into context:
  - Hierarchical relationship
    - “is-a”
  - Non-hierarchical relationships
    - “part-of”
    - ”has-a”
    - Other relationships



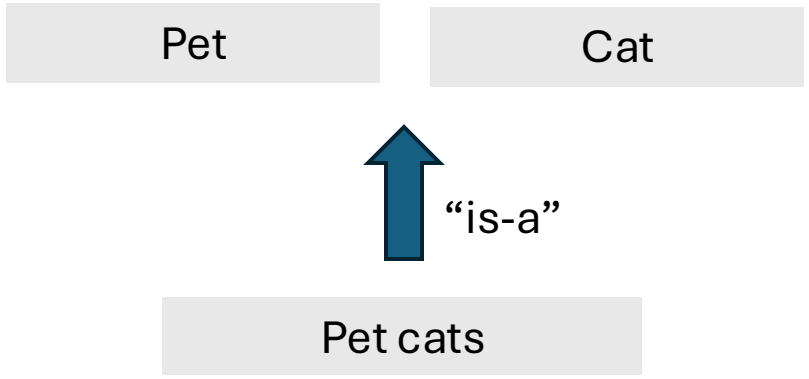

Source: Keet, 2020

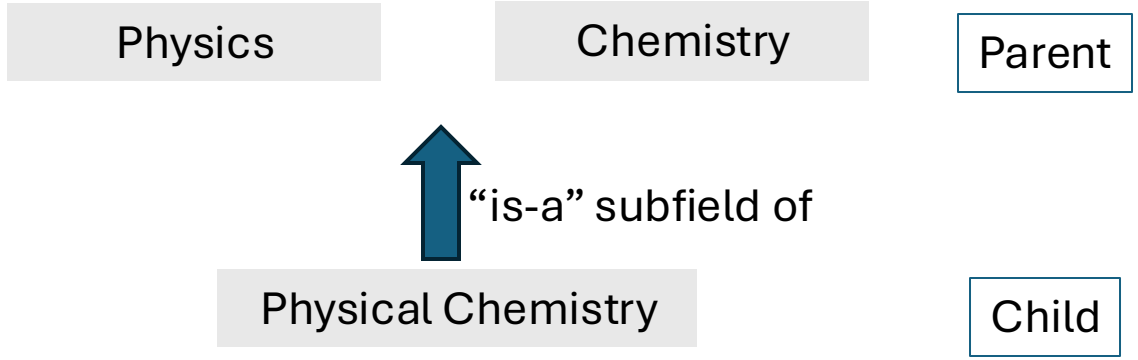
# What should be in an ontology/Knowledgebase?

- A standardized list of terms and their meaning (vocabulary).

- Relationships that place the terms into context.
  - Hierarchical relationship
  - Non-Hierarchical relationship

- Can be a polyhierarchy: any hierarchical structure that allows a term to have multiple parents.



# Ontologies / Knowledgebases

- Gene Ontology (GO) – gene product properties
  - https://geneontology.org/

- Metabolomics:
  - KEGG: Kyoto Encyclopedia of Genes and Genomes
    - https://www.genome.jp/kegg/
  - BioCyc: Pathway/Genome Database Collection
    - https://biocyc.org/
  - SMPDB: Small Molecule Pathway Database
    - http://www.smpdb.ca/

- Chemistry
  - PubChem
    - https://pubchem.ncbi.nlm.nih.gov/

# What is Gene Ontology (GO)?

- A set of terms with their **precise definitions** and **defined relationships** between them.
- The association between gene products and terms, which are used to capture the existing knowledge about each gene
- Gene Ontology is commonly referred to as:
  - GO ontology: describe the set of terms and their hierarchical annotation.
  - GO annotations: describe the set of associations between genes and GO terms.
- Constructed by the Gene Ontology Consortium to produce a controlled vocabulary that can be applied to **all organisms even as knowledge of gene and protein roles in cells is accumulating and changing**.

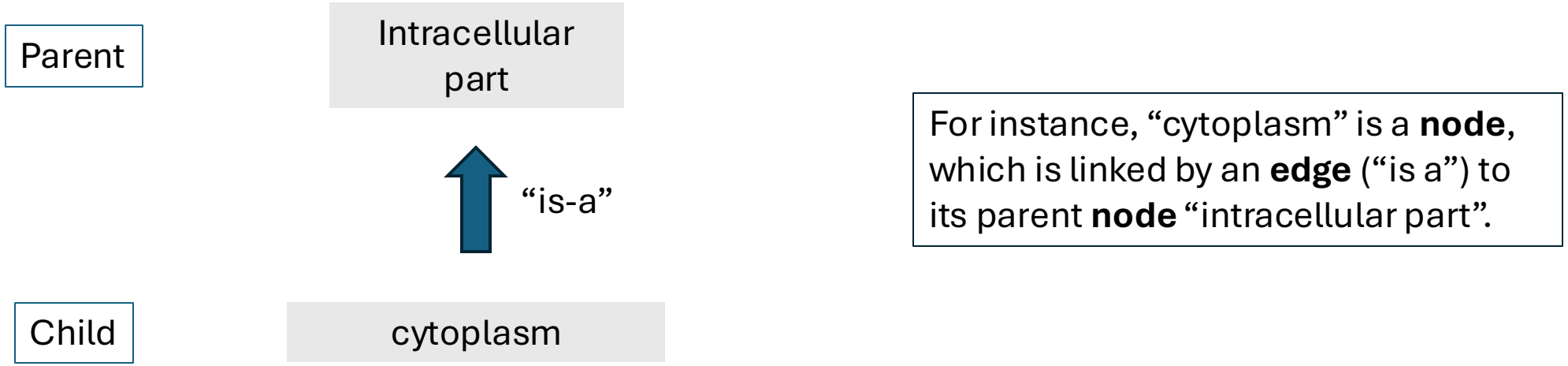
> Nat Genet. 2000 May;25(1):25-9. doi: 10.1038/75556.

**Gene ontology: tool for the unification of biology.**
**The Gene Ontology Consortium**


M Ashburner [1], C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, G Sherlock

# GO structure and data representation

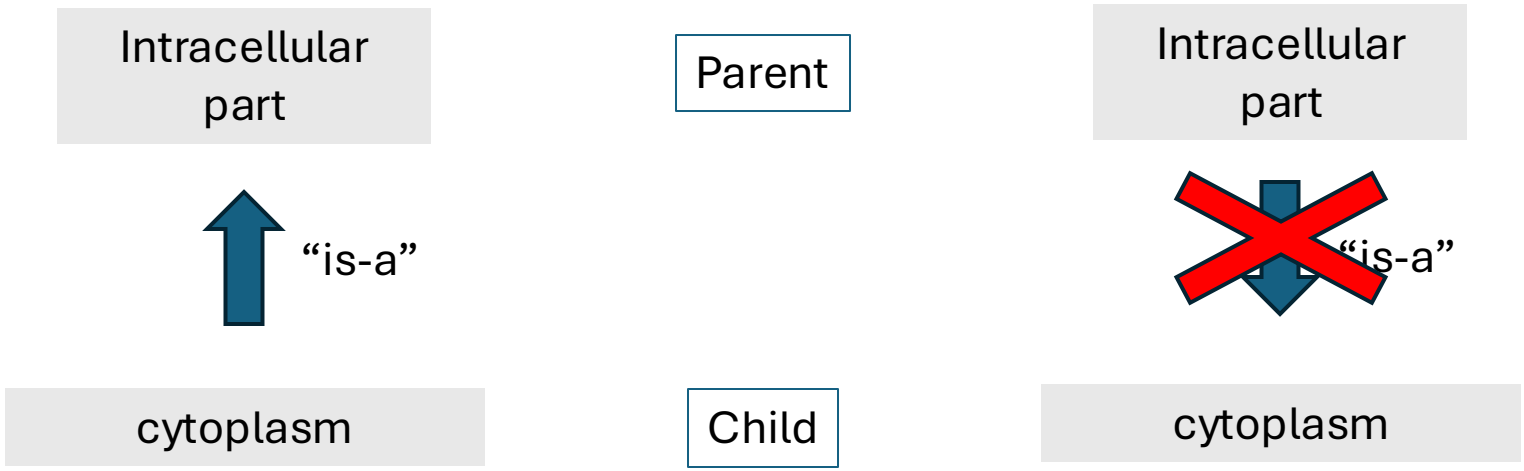
- We have a number of explicitly defined terms for biological objects or events.
- We have terms as **nodes** (also called **vertices**) in a graph, which are linked by **edges** that describe the relationships between nodes.




For instance, "cytoplasm" is a **node**, which is linked by an **edge** ("is a") to its parent **node** "intracellular part".

# GO structure and data representation

- We have a number of explicitly defined terms for biological objects or events.
- We have terms as **nodes** (also called **vertices**) in a graph, which are linked by **edges** that describe the relationships between nodes.
- The graph formed with the nodes is a **directed acyclic graph**:
  - The edges are directed (i.e. there is a source and a destination for each edge)
    - Source: **parent** term
    - Destination: **child** term



# GO structure and data representation

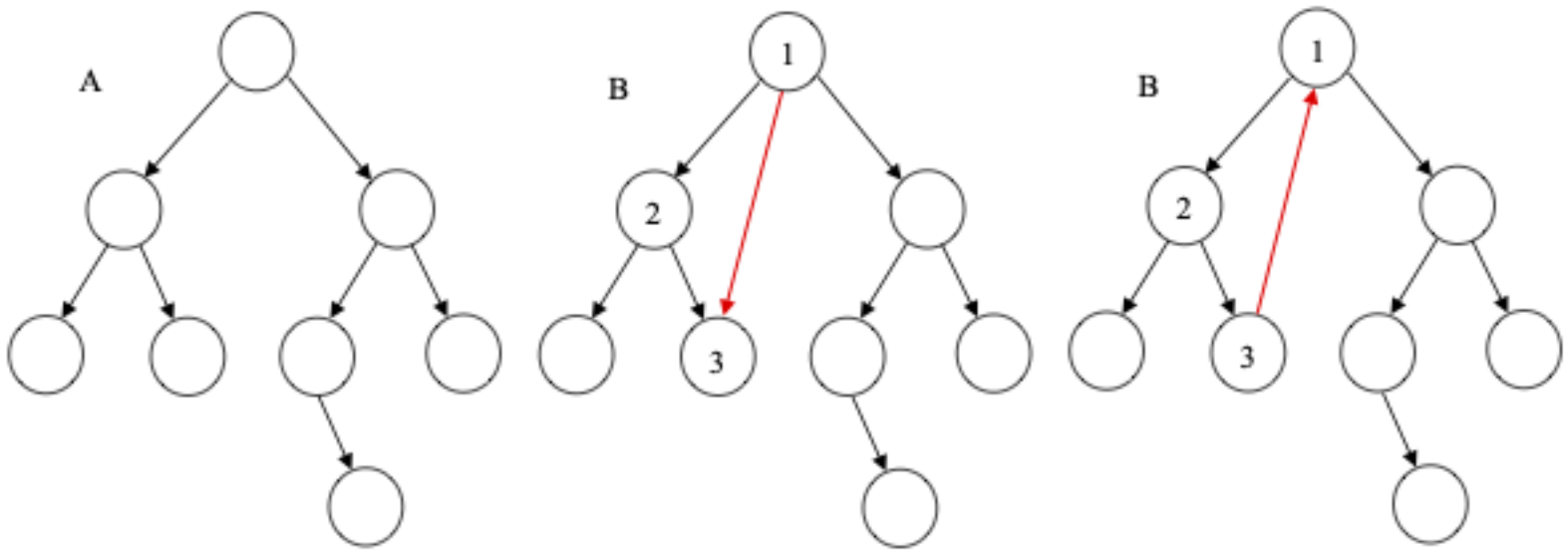
- We have a number of explicitly defined terms for biological objects or events.

- We have terms as **nodes** (also called **vertices**) in a graph, which are linked by **edges** that describe the relationships between nodes.

- The graph formed with the nodes is a **directed acyclic graph**:
  - The edges are directed (i.e. there is a source and a destination for each edge)
    - Source: **parent** term
    - Destination: **child** term

- Directed acyclic graph has no cycles:
  - We can't complete a loop by following directed edges.
  - Two terms cannot be both parents and children of each other
  - At least one node that has no parent: **a root**.

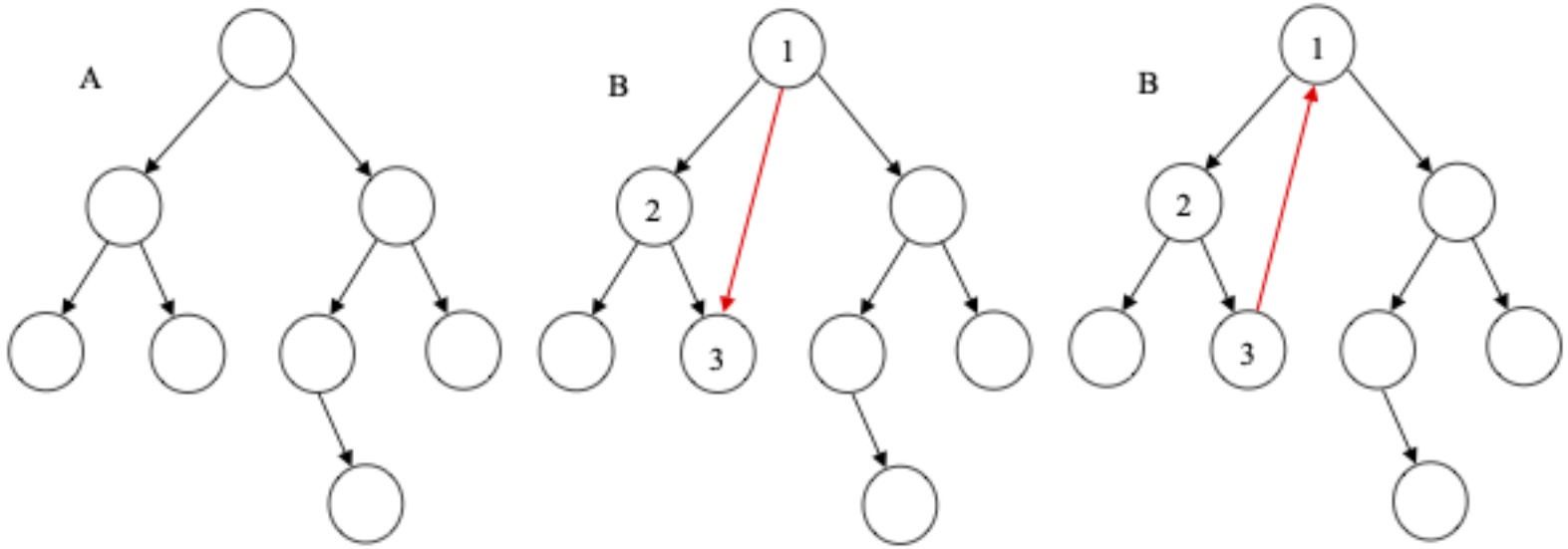
# GO structure and data representation




**A** **B** **C**

Which one is a tree?
Which one is a general graph?
Which one is a directed acyclic graph?

# GO structure and data representation




**A** shows a tree with each node has only one parent.

**B** shows a directed acyclic graph, with node 3 has 2 parents

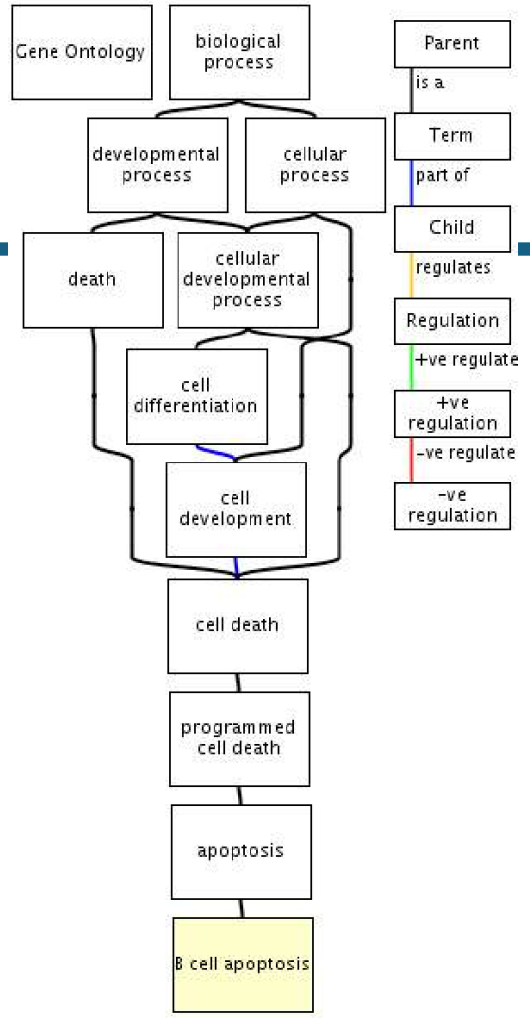
**C** shows a general graph with nodes 1, 2 and 3 forming a loop

Which one is a GO ontology structure?

# Example of Gene Ontology

A way to capture biological knowledge for individual gene products in a written and computable form.

A set of concepts and their relationships to each other arranged in a hierarchy



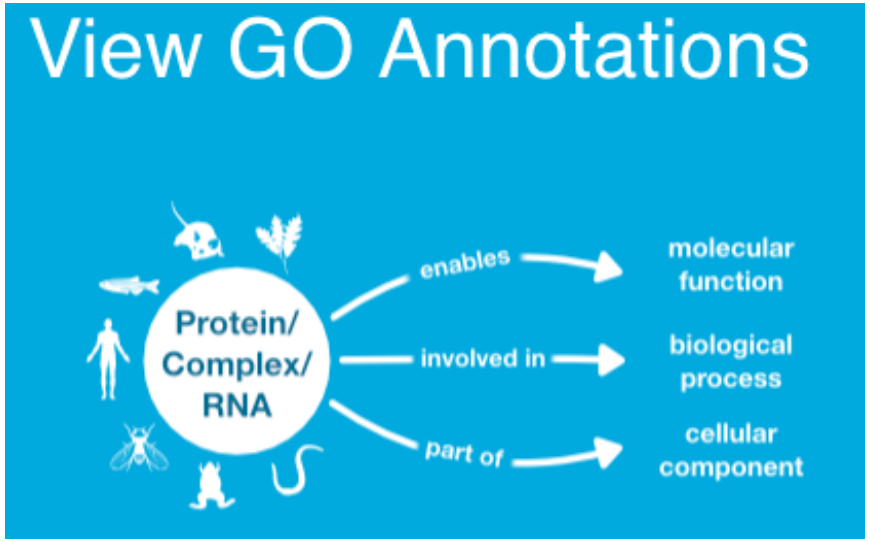

**edges** that describe the relationships between nodes

Less specific concepts

More specific concepts

Source: quickGO at EBI

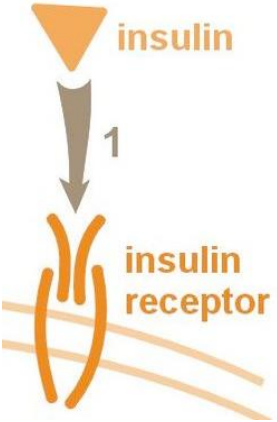
# Gene Ontology




Source: quickGO at EBI

- **Molecular Function** (MF)

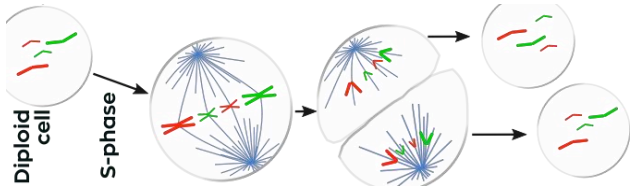
An elemental activity or task or job




Insulin receptor activity

- **Biological Process** (BP)

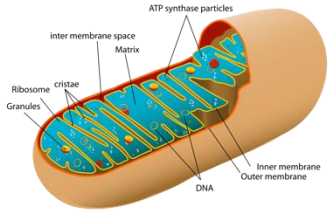
A commonly recognized series of events




Cell division

- **Cellular Component** (CC)

Where a gene product is located




Mitochondrial
Mitochondrial matrix
Mitochondrial inner membrane

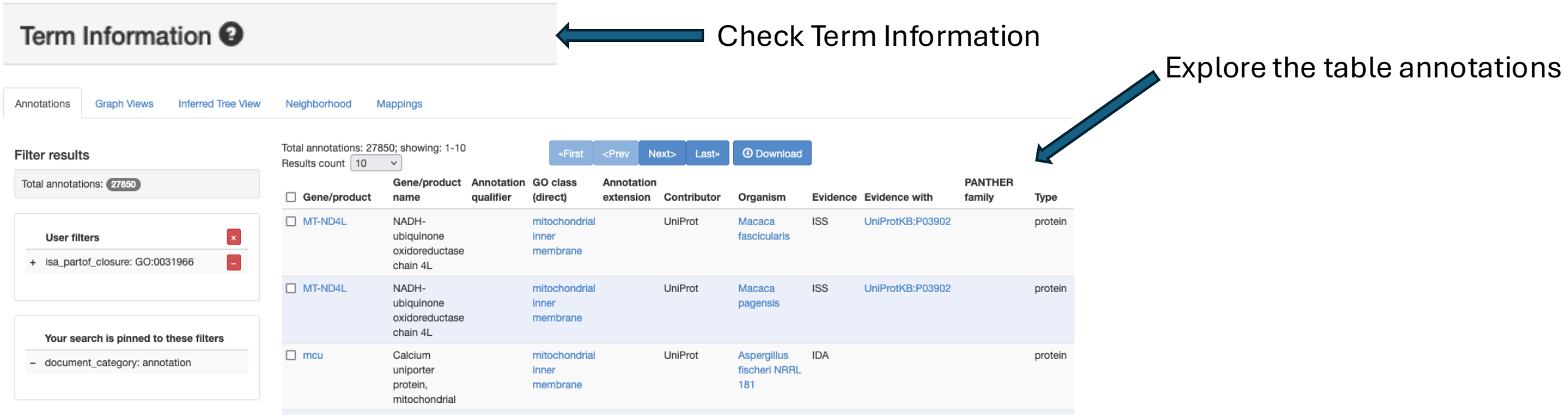
# Gene Ontology

Let's take a look …

https://geneontology.org/docs/ontology-relations/

Click on one of the terms, like:
GO:0031966:mitochondrial membrane

http://amigo.geneontology.org/amigo/term/GO:0031966



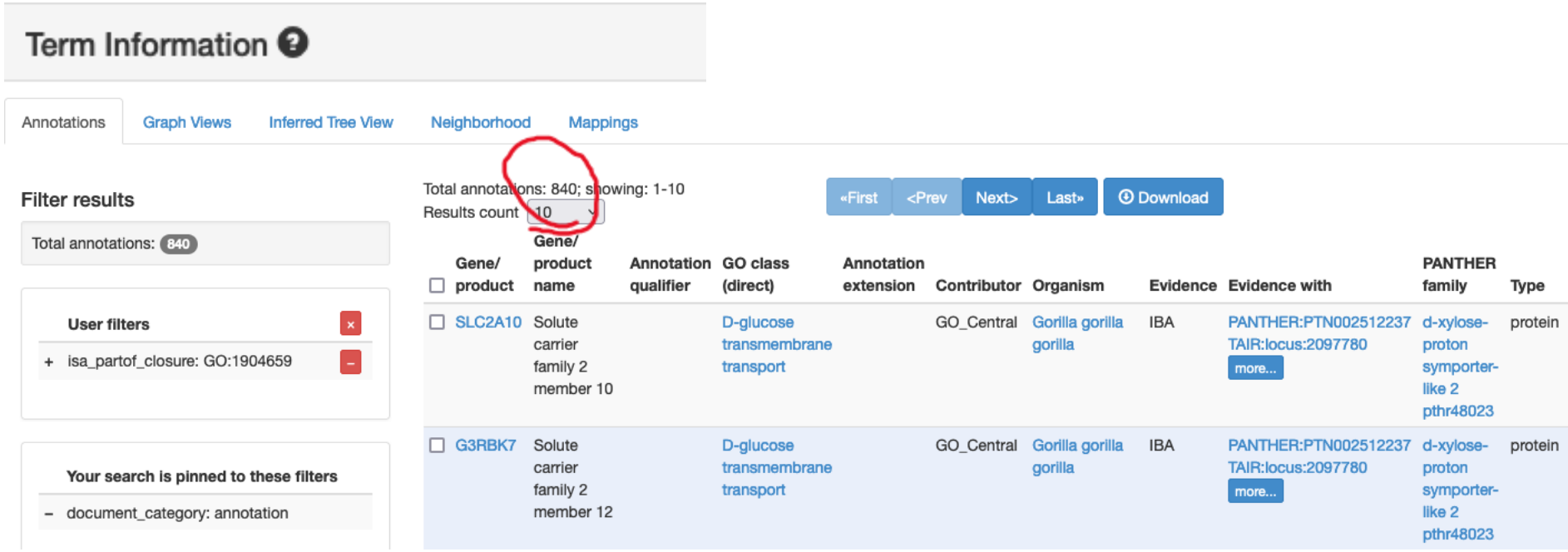
# Gene Ontology

Let's take a look …

https://geneontology.org/docs/ontology-relations/

Click on one of the terms, like:
GO:0031966:mitochondrial membrane

http://amigo.geneontology.org/amigo/term/GO:0031966




October 2024

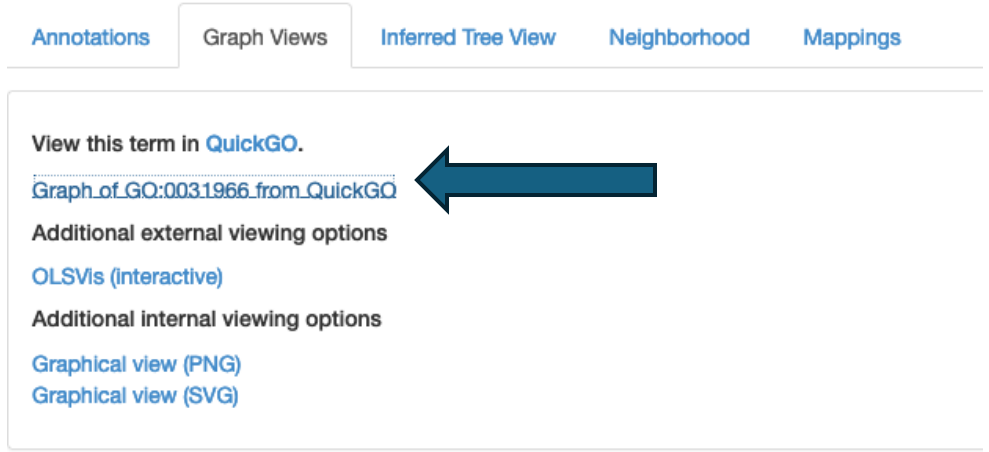
# Gene Ontology

Let's take a look …

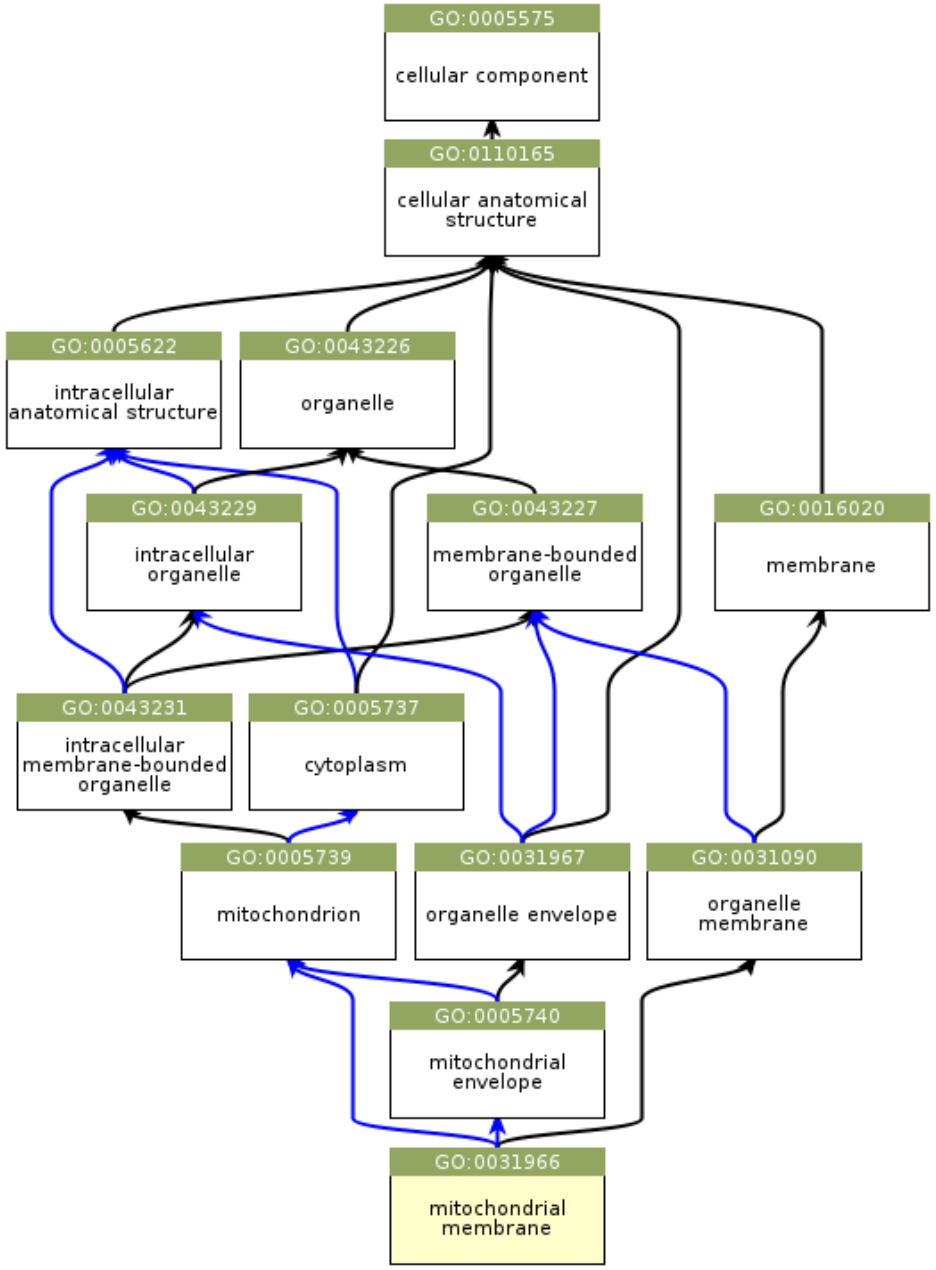
https://geneontology.org/docs/ontology-relations/

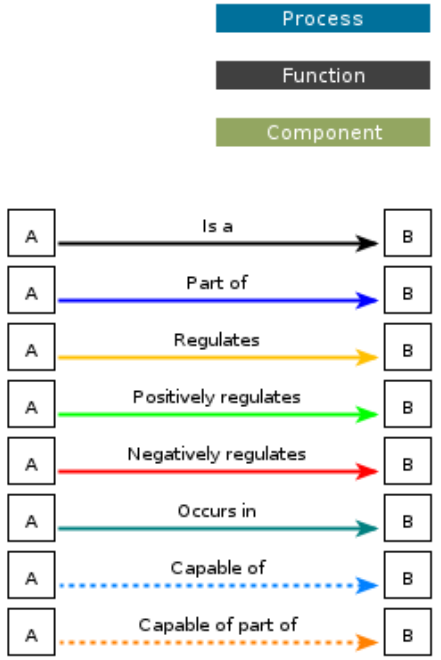
Click on one of the terms, like:
GO:0031966:mitochondrial membrane

We all love graphs. Go to "Graph Views"



Less specific concepts




More specific concepts

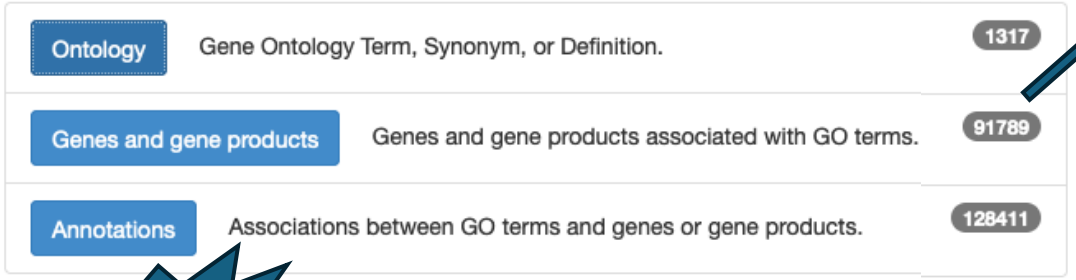
**GO:0031966 mitochondrial membrane**
**Definition:** "Either of the lipid bilayers that surround the mitochondrion and form the mitochondrial envelope."

# Gene Ontology

One more example…

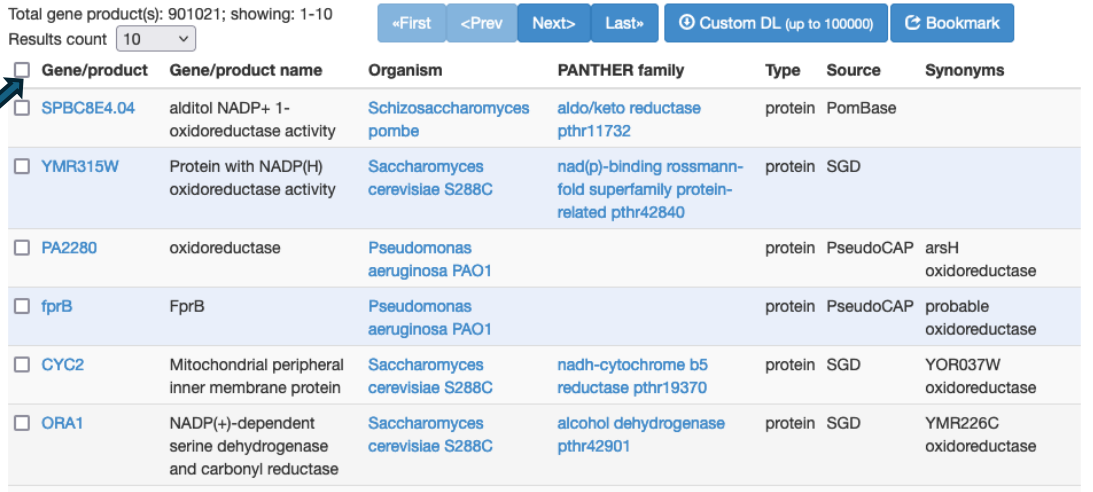
https://geneontology.org/

In the search bar type “oxidoreductase activity”




October 2024

In Genes and Gene products, click on a specific gene

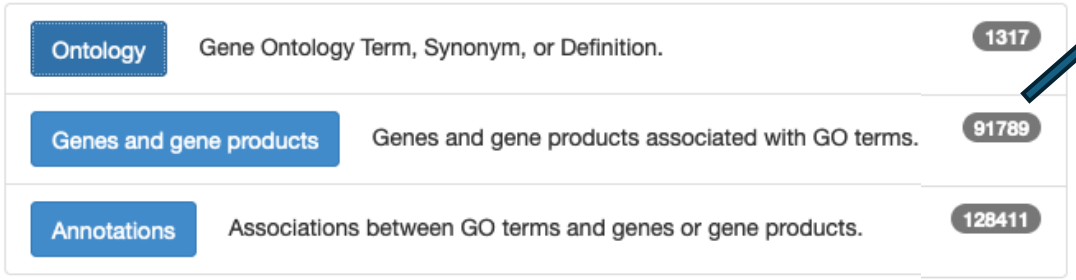



| Gene/product | Gene/product name | Organism | PANTHER family | Type | Source | Synonyms |
|---|---|---|---|---|---|---|
| SPBC8E4.04 | alditol NADP+ 1-oxidoreductase activity | Schizosaccharomyces pombe | aldo/keto reductase pthr11732 | protein | PomBase | |
| YMR315W | Protein with NADP(H) oxidoreductase activity | Saccharomyces cerevisiae S288C | nad(p)-binding rossmann-fold superfamily protein-related pthr42840 | protein | SGD | |
| PA2280 | oxidoreductase | Pseudomonas aeruginosa PAO1 | | protein | PseudoCAP | arsH oxidoreductase |
| fprB | FprB | Pseudomonas aeruginosa PAO1 | | protein | PseudoCAP | probable oxidoreductase |
| CYC2 | Mitochondrial peripheral inner membrane protein | Saccharomyces cerevisiae S288C | nadh-cytochrome b5 reductase pthr19370 | protein | SGD | YOR037W oxidoreductase |
| ORA1 | NADP(+)-dependent serine dehydrogenase and carbonyl reductase | Saccharomyces cerevisiae S288C | alcohol dehydrogenase pthr42901 | protein | SGD | YMR226C oxidoreductase |

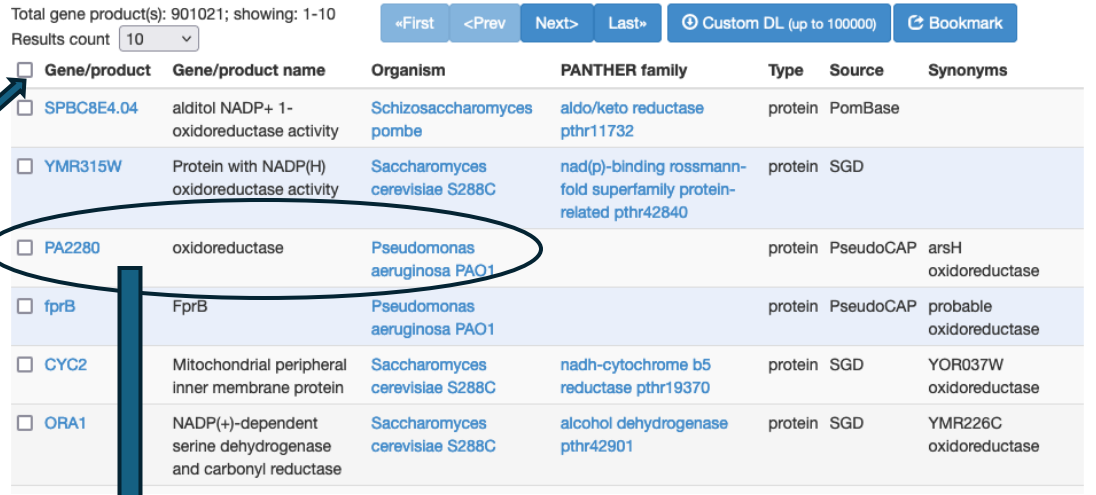
# Gene Ontology

https://geneontology.org/

Search in bar for “oxidoreductase activity”



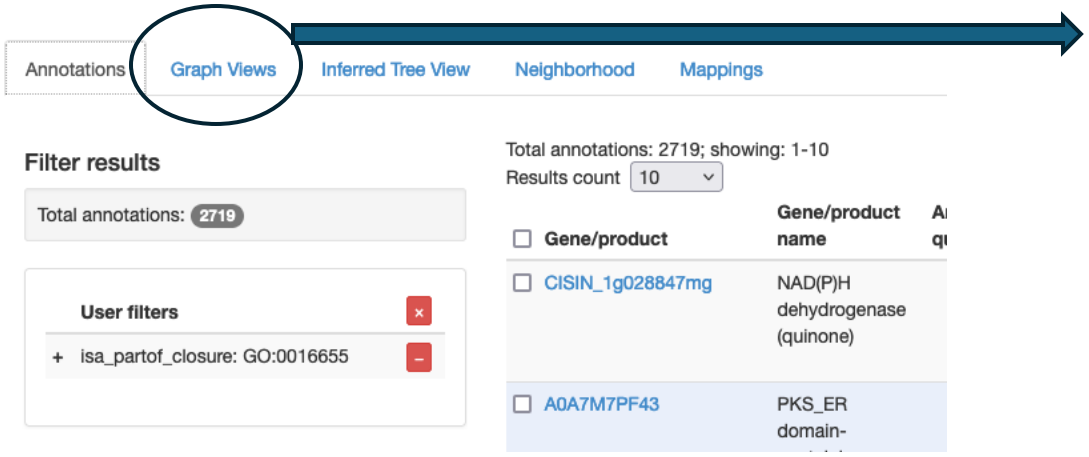

In Genes and Gene products, click on a specific gene



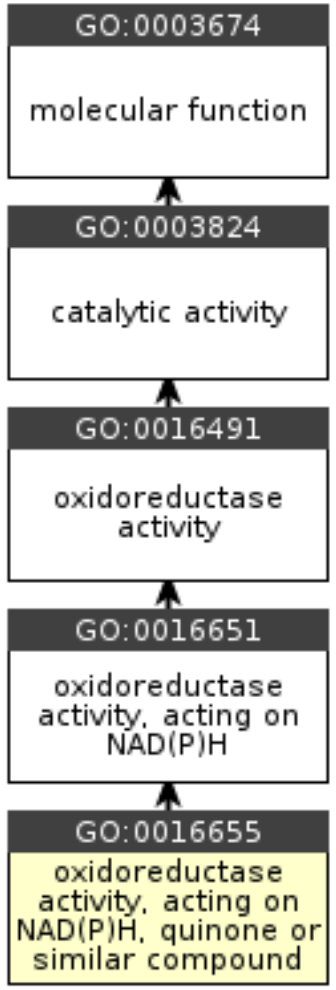

| Gene/ product | Gene/product name | Annotation qualifier | GO class (direct) | Annotation extension | Contributor | Organism | Evidence | Evidence with | PANTHER family | Type | Isoform | Reference | Date |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PA2280 | oxidoreductase | | oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor | | PseudoCAP | Pseudomonas aeruginosa PAO1 | IDA | | | protein | | PMID:22304305 | 20190311 |

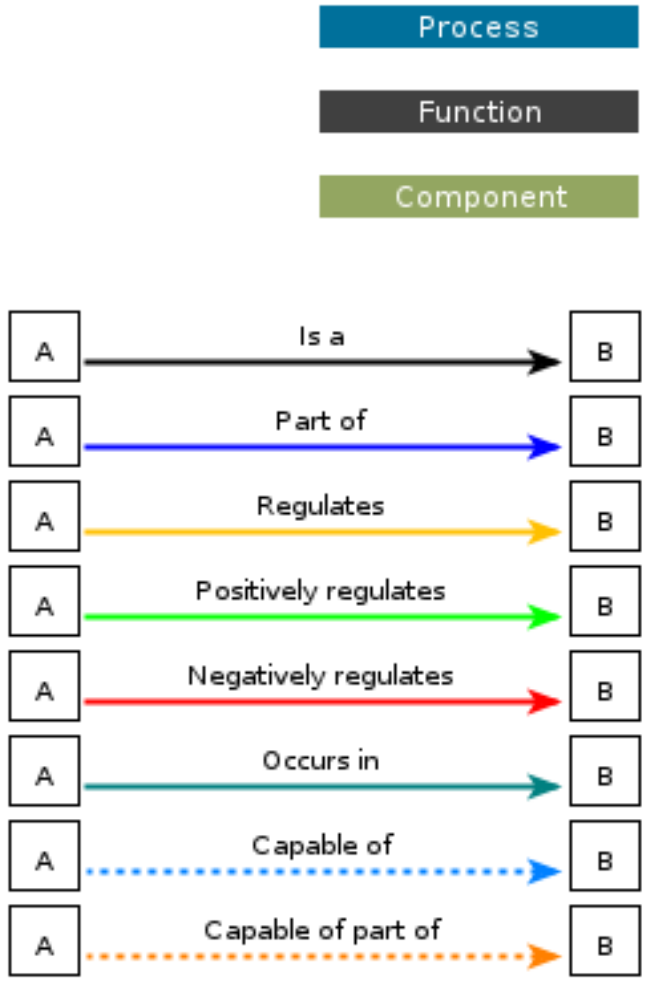
# Gene Ontology

Again, we are at a specific GO term.
With all genes with the GO annotated
Let's see the graph




October 2024




QuickGO - https://www.ebi.ac.uk/QuickGO

# The GO knowledge base is large and dynamic

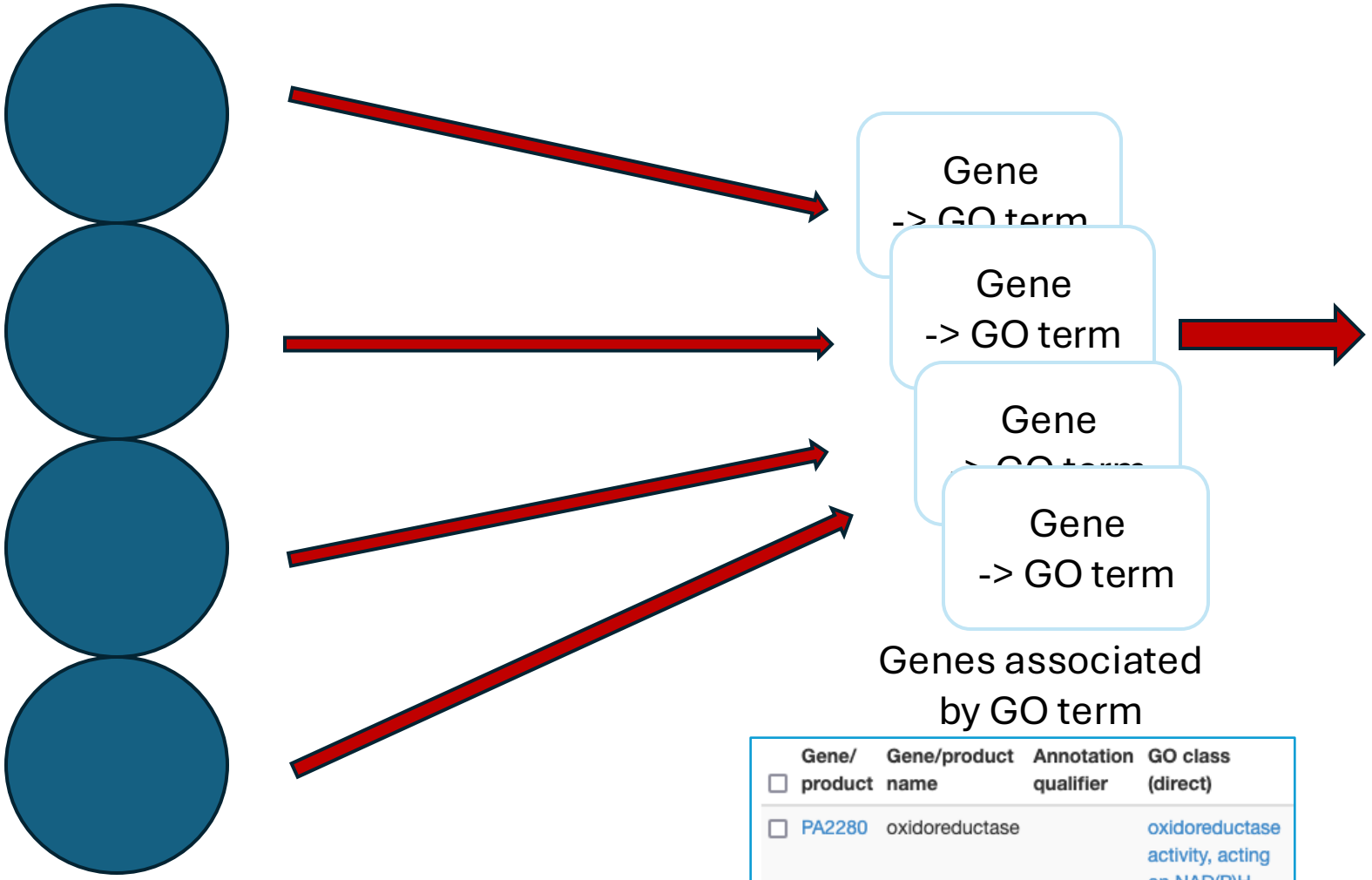

**Table 1.** Changes to GO terms in the past two-year period. The ontology has undergone substantial revision and improvement, with nearly 2,000 terms added or removed.

| GO aspect | Total number of terms | Added terms | Obsoleted terms | Merged terms[a] |
|---|---|---|---|---|
| Molecular function | 11,271 | 315 | 65 | 143 |
| Cellular component | 4,039 | 34 | 19 | 162 |
| Biological process | 27,993 | 217 | 782 | 254 |

[a] Also includes obsoleted terms that have been replaced by another term.

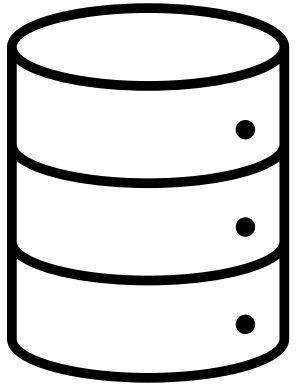
The Gene Ontology Consortium, 2023

# Gene product annotations




Genome/gene and protein databases

Genes associated by GO term

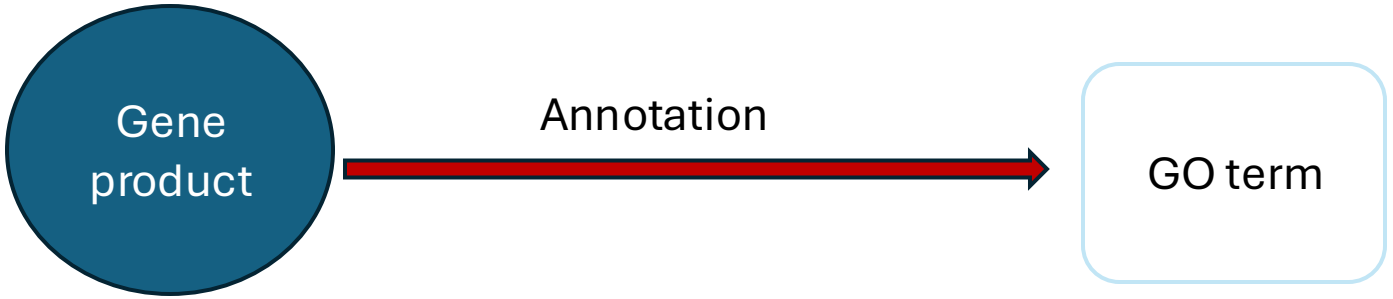


GO database

| Gene/ product | Gene/product name | Annotation qualifier | GO class (direct) |
|---|---|---|---|
| PA2280 | oxidoreductase | | oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor |

# Gene product annotations

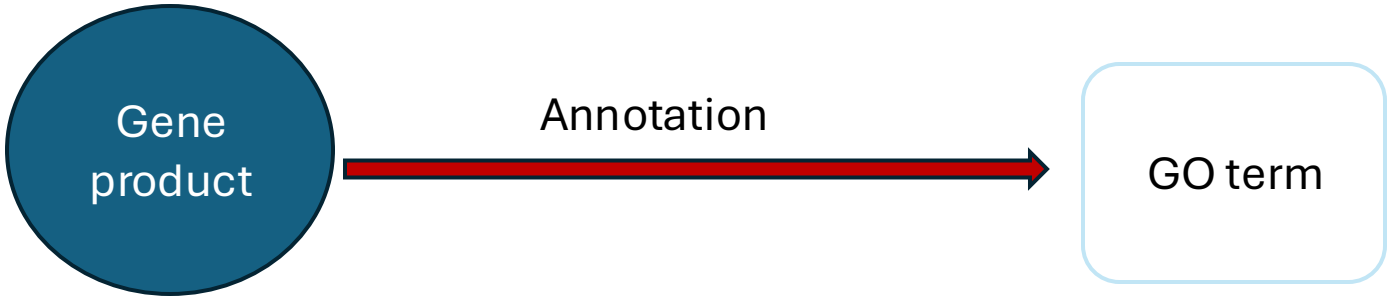
- Gene annotation is the process of associating a gene product (RNA/protein) with a GO term




| Gene/ product | Gene/product name | Annotation qualifier | GO class (direct) | Annotation extension | Contributor | Organism | Evidence | Evidence with | PANTHER family | Type | Isoform | Reference | Date |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PA2280 | oxidoreductase | | oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor | | PseudoCAP | Pseudomonas aeruginosa PAO1 | IDA | | | protein | | PMID:22304305 | 20190311 |

# Gene product annotations

- Gene annotation is the process of associating a gene product (RNA/protein) with a GO term

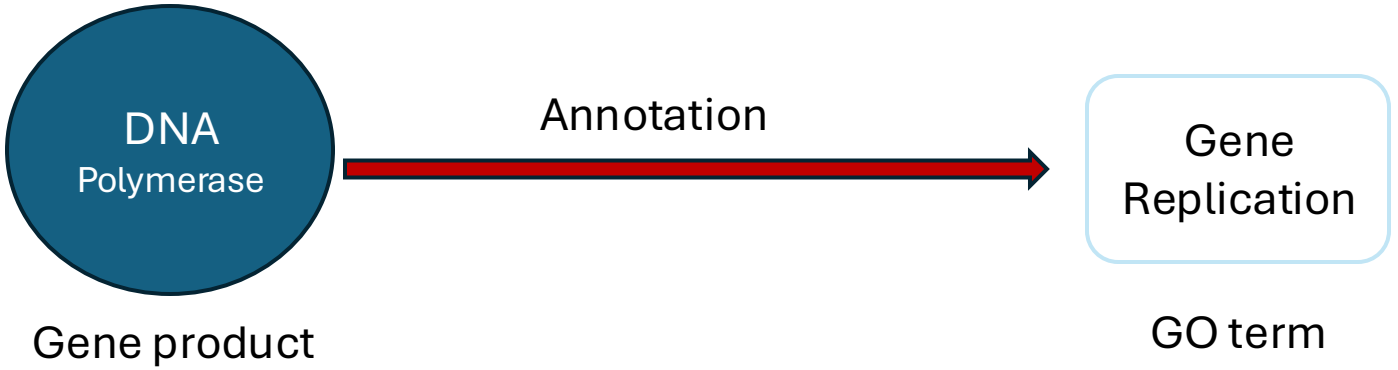



**Electronic annotation**
Computational methods like sequence simulation or genomic models are used to determine GO term associations.
-Fast approach
-Not accurate
-The only source of annotation in non-model organism

**Manual annotation**
Uses primary research or review from published literature to build the annotations
-Labor intensive
-Highly accurate: more specific GO terms
-Essential for creating predictions

# Gene product annotations

- Gene annotation is the process of associating a gene product (RNA/protein) with a GO term




**Electronic annotation**
-Sequence a genome
-Find homologs (BLAST)
-Find GO terms associated with homologs (AmiGO)
-We are making many assumptions (better with closely related species).

**Manual annotation**
-Deduce function by knocking down gene
-Find locations by protein binding assays
-By cloning the gene: In vitro assays

# Biology Analysis with Large Datasets

- If you focus on one gene (traditional molecular biology, genetics), you will try to infer its function by manual annotation. (e.g. Monogenic diseases, antibiotic resistance).

- However, most conditions (in any cell or organism) are controlled by many genes and gene products.

- And that is what happens today's outcomes in biological experiments such as
  - Microarrays
  - Mass Spectrometry / Gas Chromatography
  - High Throughput Sequencing (RNAseq)

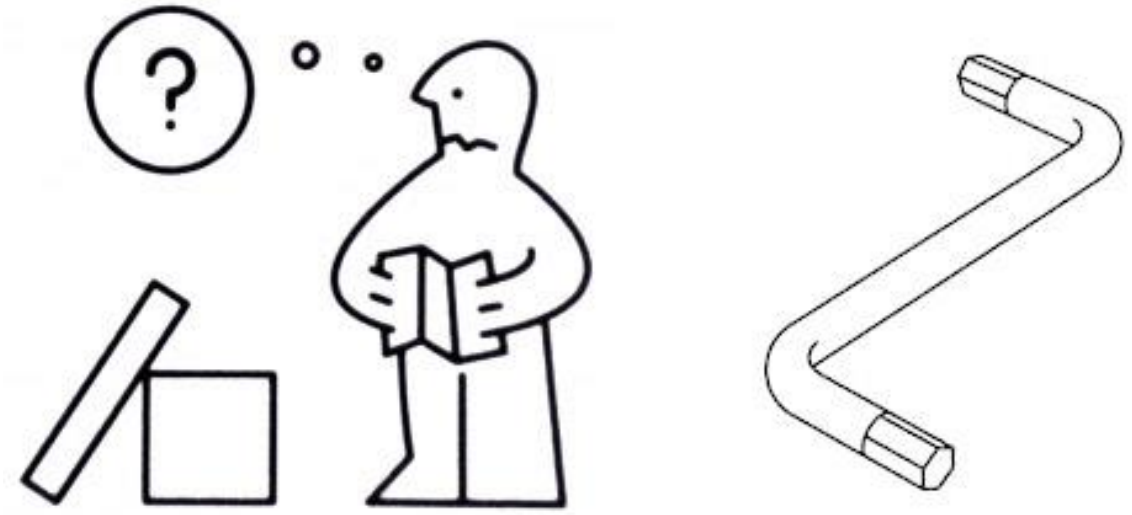
    - ➢ If we have a group of differentially expressed genes in an **experiment/treatment/phenotype/disease**
    - ➢ What are the **Biological Processes, Cellular Components and Molecular Functions** implicated in it?

**Solutions?**

Solutions?

# Gene Ontology analysis

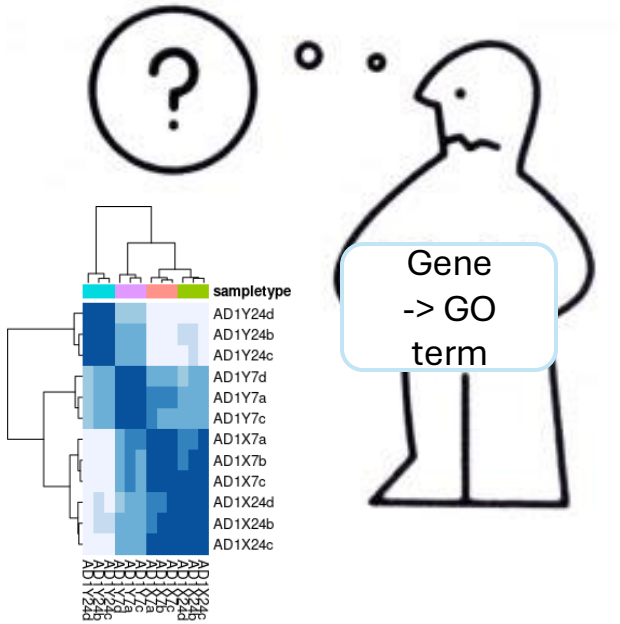
- Knowledgebases, like Gene Ontology, represent and capture out knowledge in biology.
- Analysis of large data sets in the context of GOs may provide a better interpretation.
- Nodes of the GOs that have statistical representation of of a group of genes (Genes of interest) from an experiment may be the important biological concepts.



Solutions?

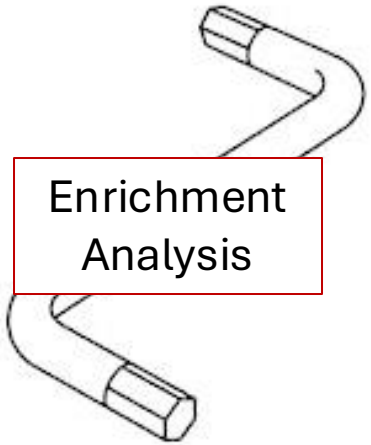
# Gene Ontology analysis

- Knowledgebases, like Gene Ontology, represent and capture out knowledge in biology.
- Analysis of large data sets in the context of GOs may provide a better interpretation.
- Nodes of the GOs that have statistical representation of of a group of genes (genes of interest) from an experiment may be the important biological concepts.



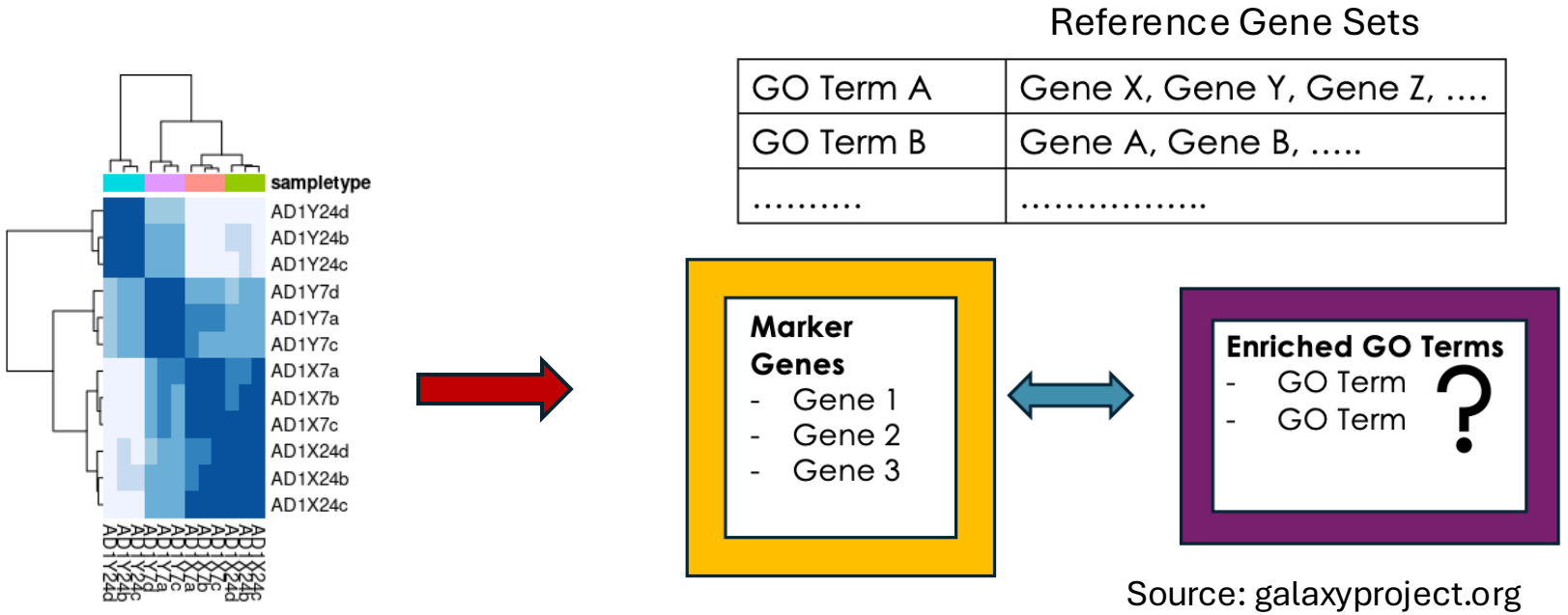






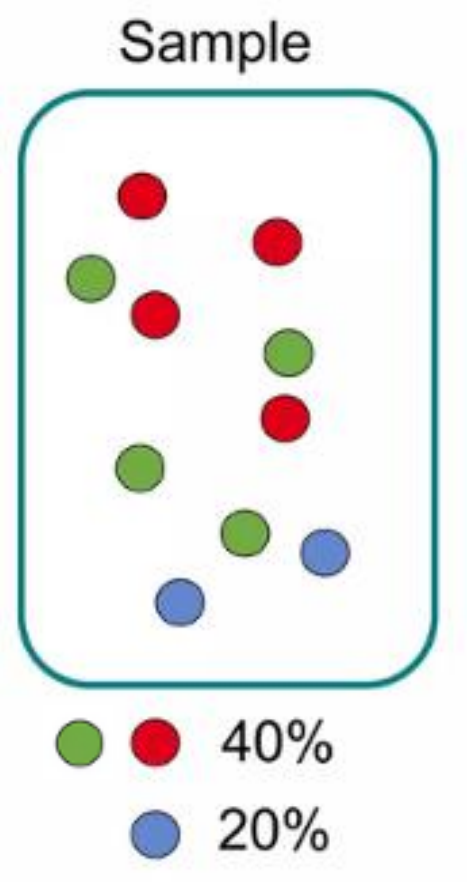
# Enrichment analysis

- Map genes and their states to nodes of GOs.
  - In model organisms…that's easy.
  - In non-model organisms (don't have GOs mapped to annotated genes) -> Electronic annotation.
- Test each node for overrepresentation in associated genes of interest (marker genes) ***versus*** the representation in the control/reference/background.

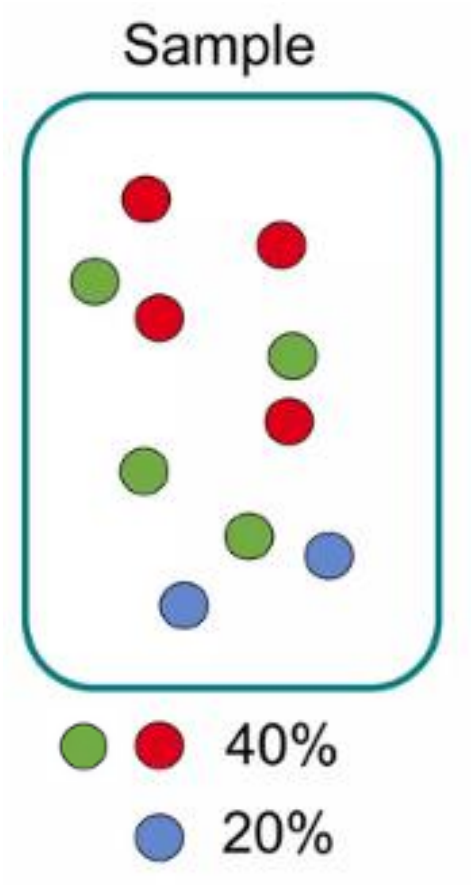


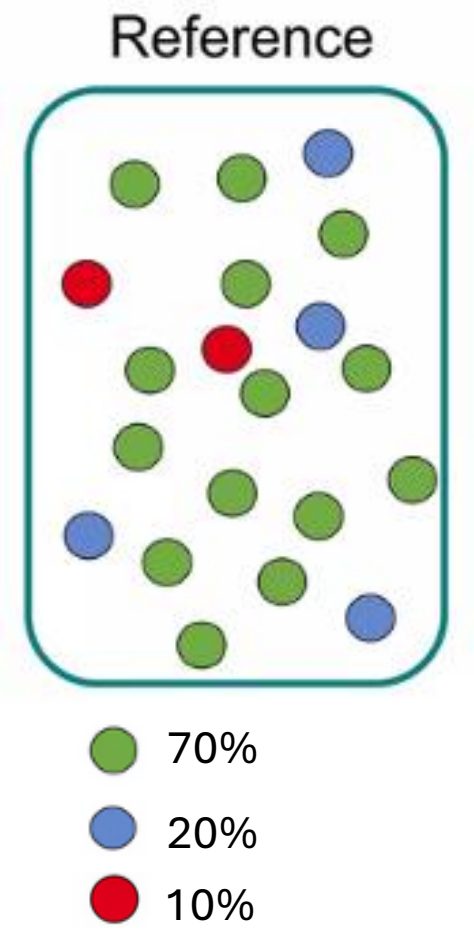

Source: galaxyproject.org

# Enrichment analysis



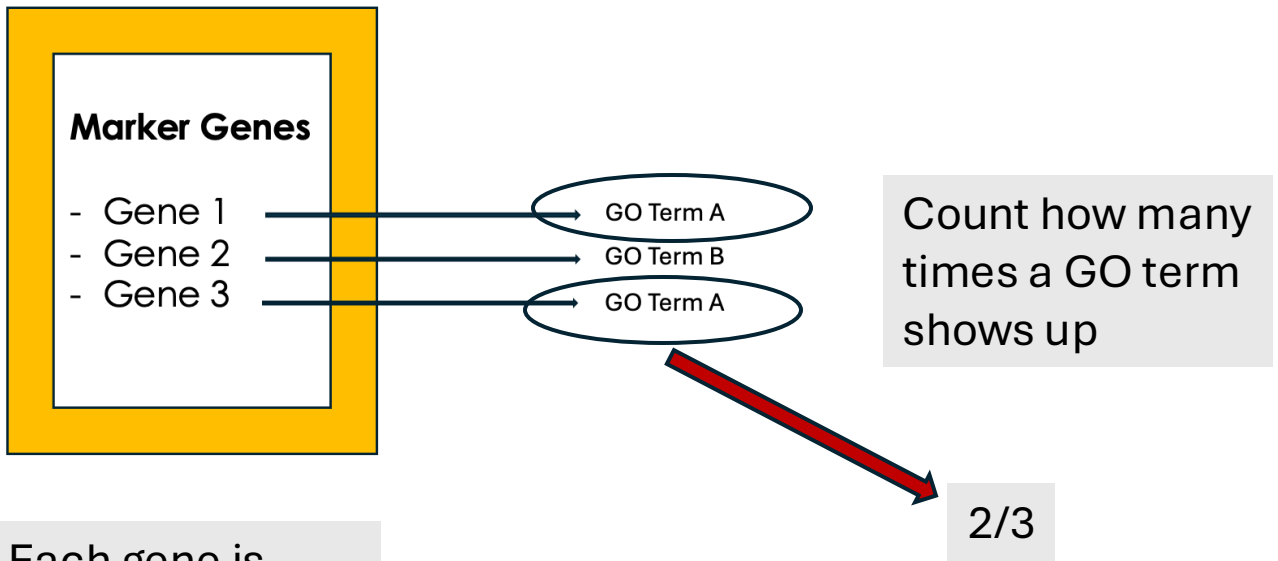
# Enrichment analysis




-> The sample is over-enriched for?

# Enrichment analysis




Count how many times a GO term shows up

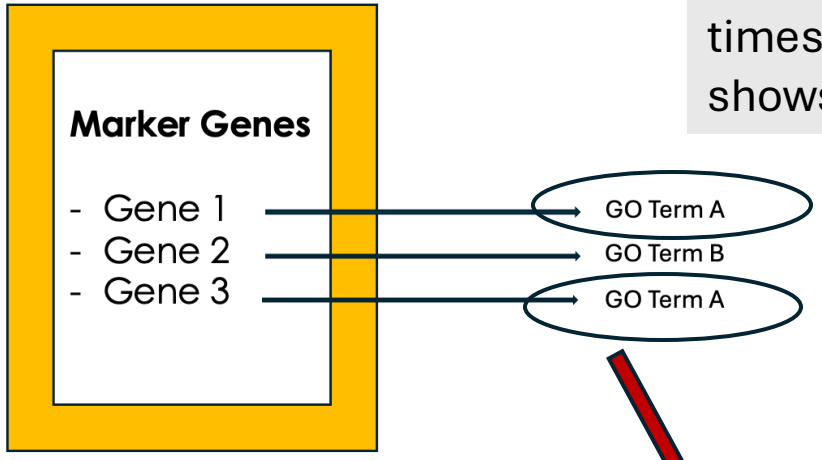
2/3

Each gene is mapped/tagged with GO terms

Source: galaxyproject.org

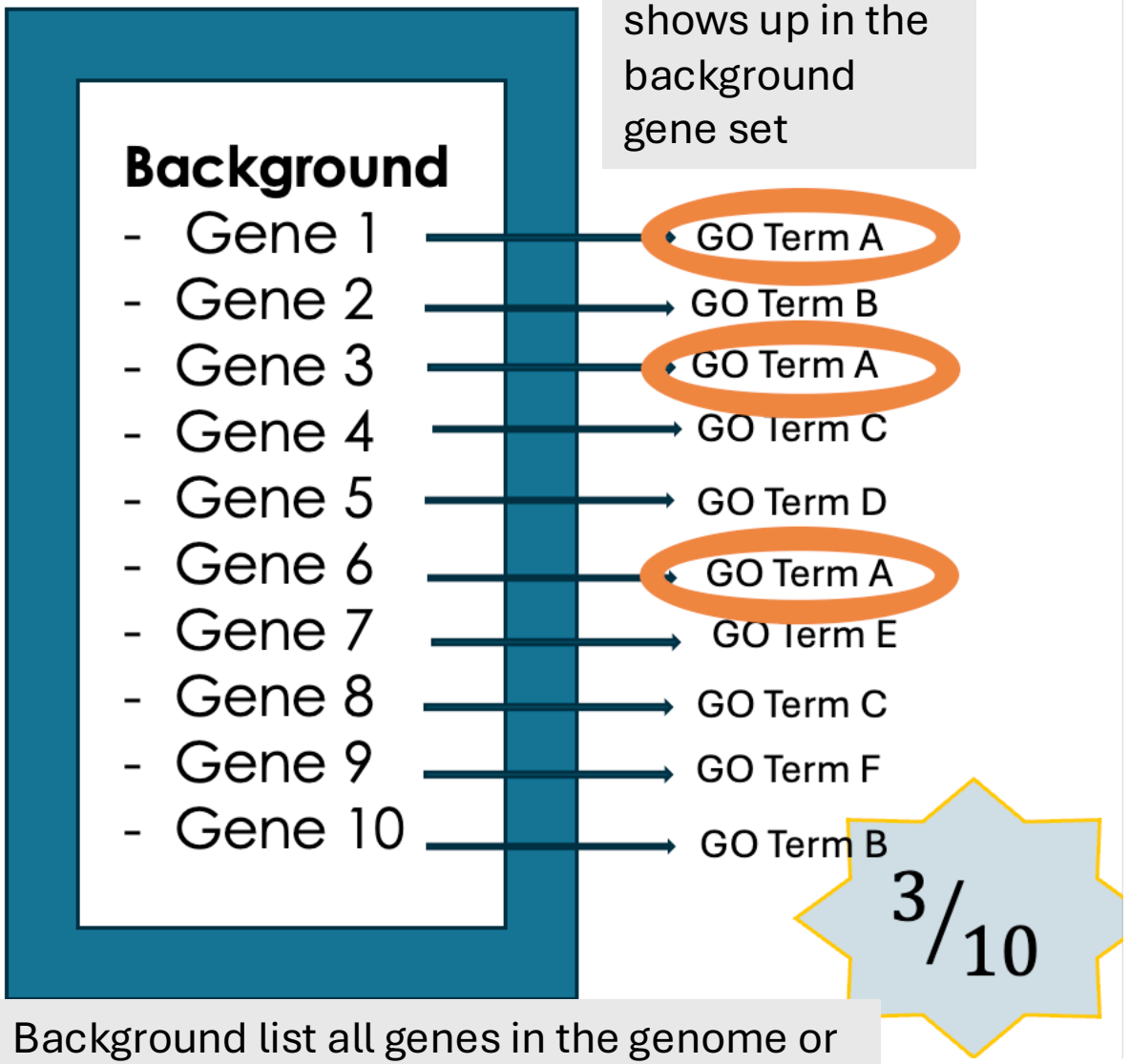
# Enrichment analysis

Count how many times a GO term shows up




Each gene is mapped/tagged with GO terms

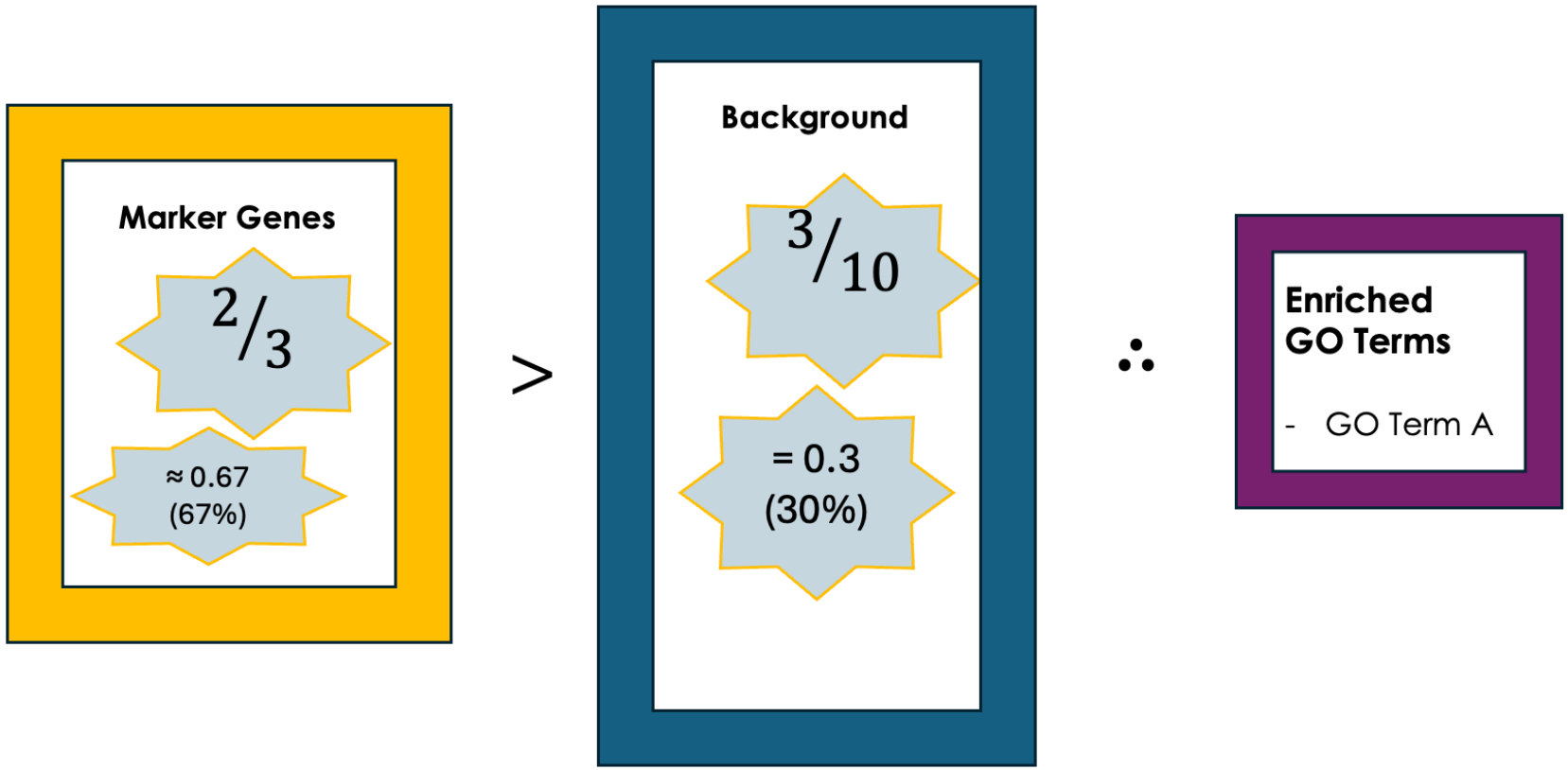

Count how many times a GO term shows up in the background gene set




Background list all genes in the genome or all genes involved in the experiment

Source: galaxyproject.org

# Enrichment analysis




Comparison between background and genes of interest (marked genes) would tell which GO terms are enriched.

Source: galaxyproject.org

# Enrichment analysis

- We use statistical tests to see if the enrichment is statistically significant
  - A GO term overrepresented can be done by chance.

- Statistical tests:

  - Hypergeometric

  - Z-score

  - 2x2 tables:
    - **Fisher's Exact**
    - Chi-square (X2)

# Enrichment analysis

Example:

- Fischer's Exact test

Contingency table

| | GO Term A | Not GO Term A | Total |
|---|---|---|---|
| **Marker Set** | 2 | 1 | 3 |
| **Background Set** | 3 | 7 | 10 |
| **Total** | 5 | 8 | 13 |

$$P = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$$

Is GO term A significantly enriched?

Check it out with one online Fisher's Table calculator:
https://www.graphpad.com/quickcalcs/contingency1/

Source: galaxyproject.org

# Enrichment analysis

- In the context of Fisher's exact test, **false positives** (FPs) can become a significant issue when you're testing for enrichment across a large number of nodes, such as gene expression data.
- Fisher's test checks for the association between two categorical variables (e.g., presence/absence of a GO term in two conditions), and an **alpha level** (often set at 0.05) defines the threshold for statistical significance.
- This alpha level means that, on average, 5% of all tests will result **in false positives simply by chance**.
- Example: Suppose you have 10,000 nodes for which you are using Fisher’s test with an alpha level of 0.05. That means 5% of the test statistically significant can be just due to random chance. So, you expect about 500 of False Positives.

# Enrichment analysis

- RNAseq and microarrays detect many genes with small effects due to their high sensitivity.
- That means more GO terms.
- Solutions:
  - **Multiple testing corrections**: techniques like the Bonferroni correction or False Discovery Rate (FDR) are used to adjust the significance threshold in light of the large number of tests. These techniques aim to reduce the expected proportion of false positives. The **Bonferroni correction** is considered to be the most conservative method to correct for multiple comparisons, meaning that the fewest false positives are returned.
  - **FDR (False Discovery Rate)**: It's more lenient than Bonferroni, allowing more true positives to be reported as significant with the drawback that some false positives may also be reported as such, which is more suitable for gene expression studies where many comparisons are made.
- By applying these corrections, you reduce the chances of flagging a result as significant due to random chance.