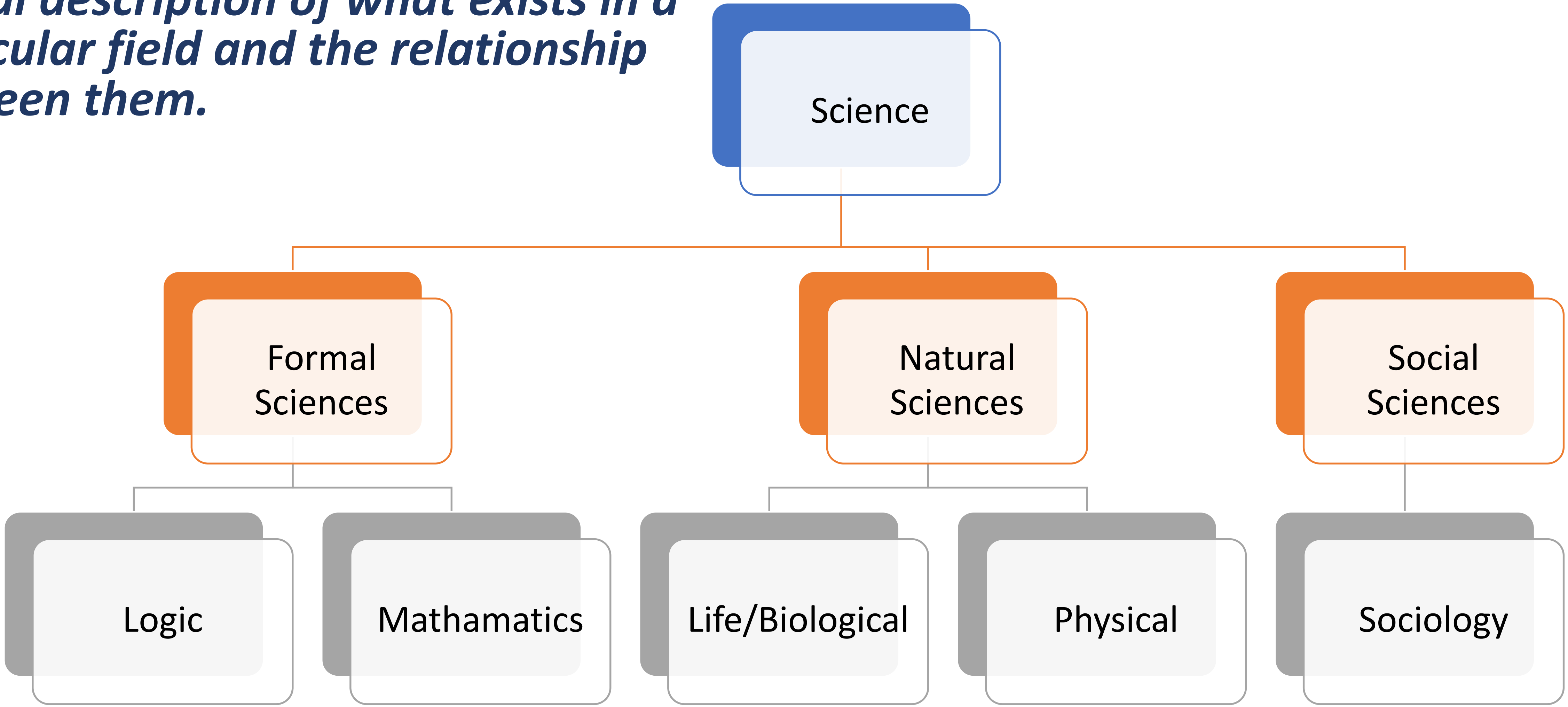
GO Terms & Functional Enrichment Analysis

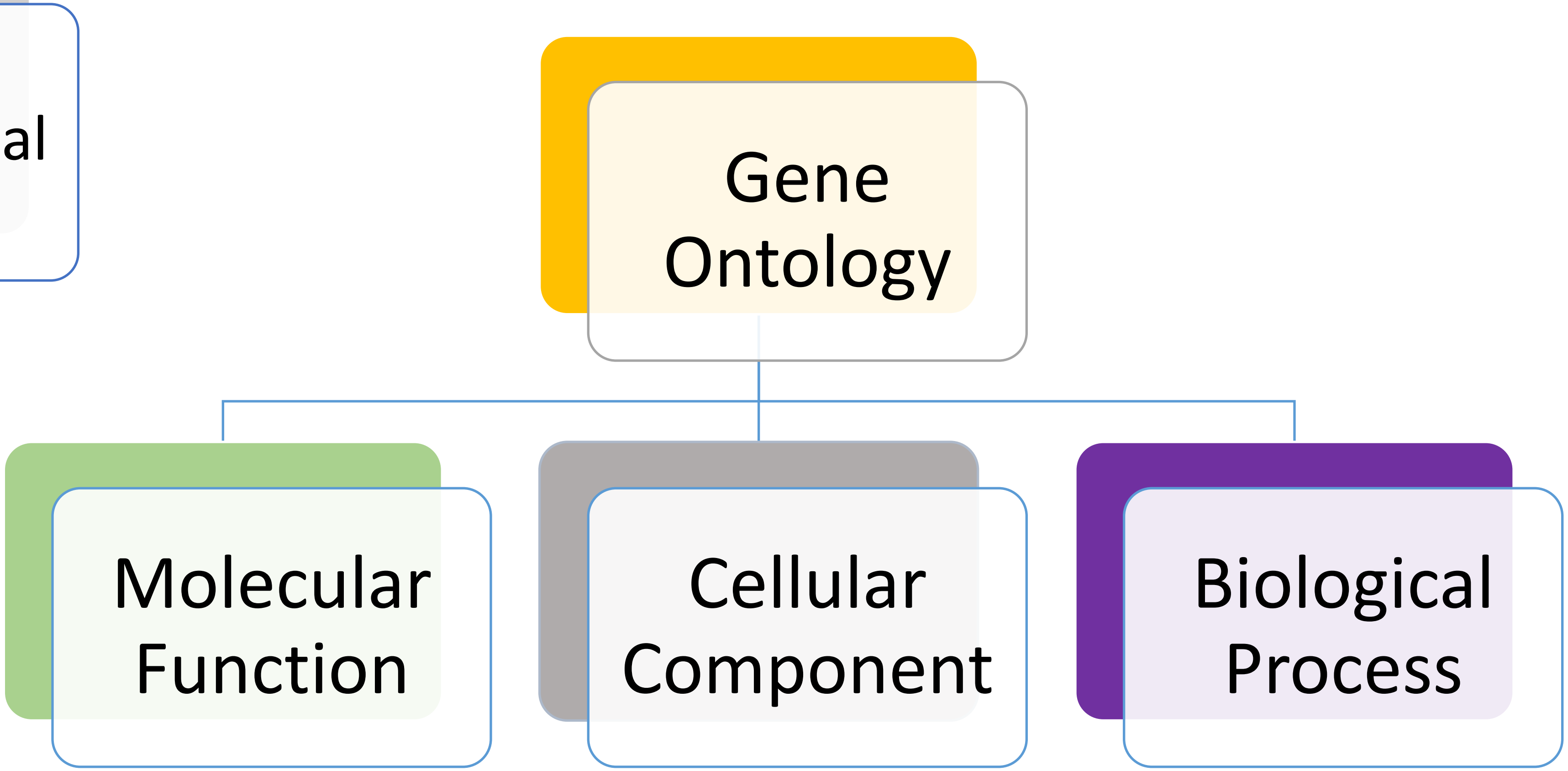
What is functional enrichment?

- Imagine you have a list of a 1000 genes that are upregulated in response to a drug treatment.
- You notice that some of the genes have names that tell you something about their function, like DNA polymerase I or gyrase.
- But did these gene funtions appear by random in your list or are they truly enriched?
- Functional enrichment applies a statistical method (usually a Fisher's exact test) to determine if you have enriched functions in your list compared to the rest of the functions in the whole genome.
- What about the genes that have names that you do not recognize? Would it be nice if you can associate genes with functions in a consistant way? Something like a well described ontology?

What is an ontology? For our purposes when we talk about ontology we mean a formal description of what exists in a particular field and the relationship between them.



Life/Biological



Gene Ontology

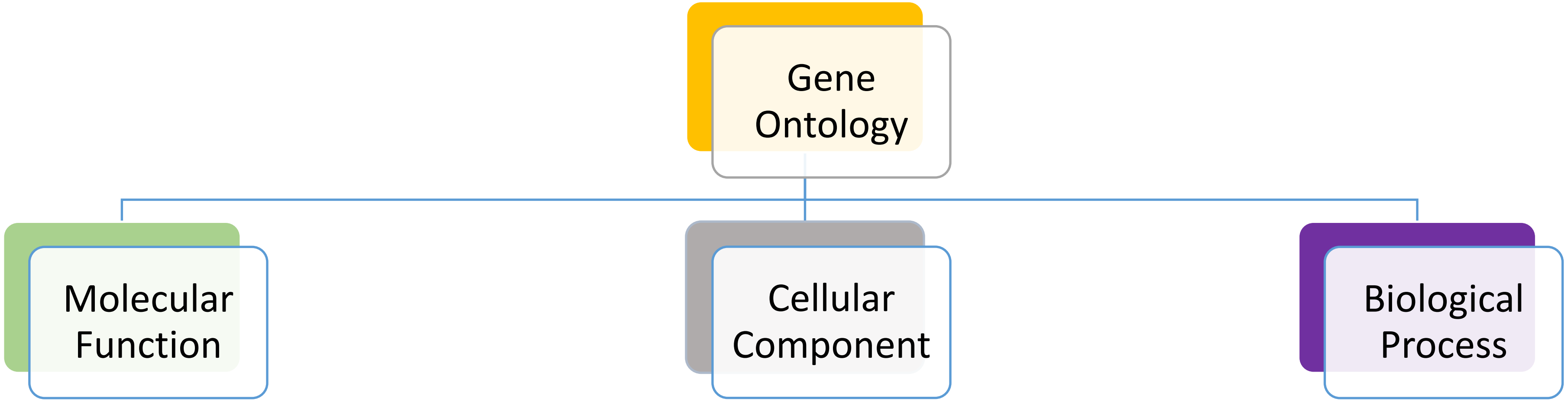
Molecular Function

Cellular Component

Biological Process

The gene ontology describes the knowledge of biological sciences and divides this knowledge up into three broad categories

http://geneontology.org/docs/ontology-documentation/



Gene Ontology

Molecular Function

Cellular Component

Biological Process

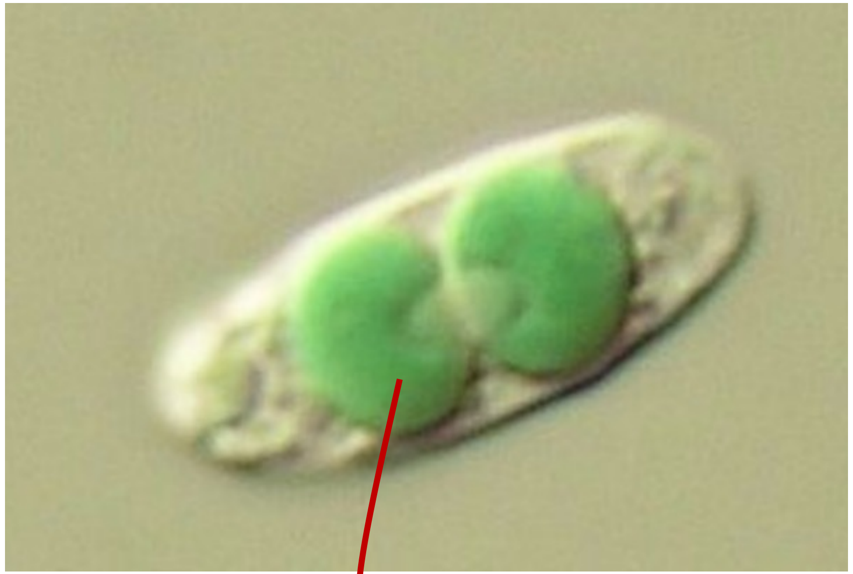
Activites at the molecular level performed by gene products, eg. Toxin activity, catalytic activity of transporter activity

Where a gene product performs its function, eg. Cilium Mitochondrion, plastid, golgi etc…

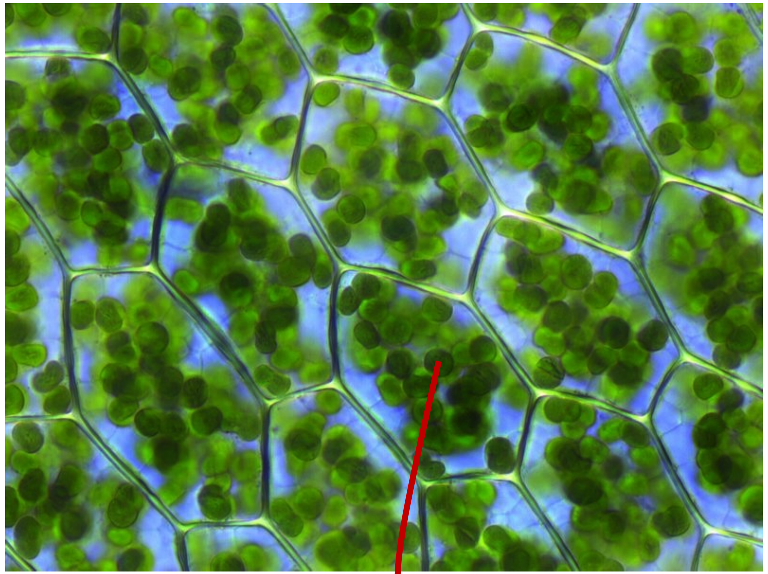
Proceses accomplished by multiple activities, eg pyrimidine biosynthesis

** relationships and hierachies

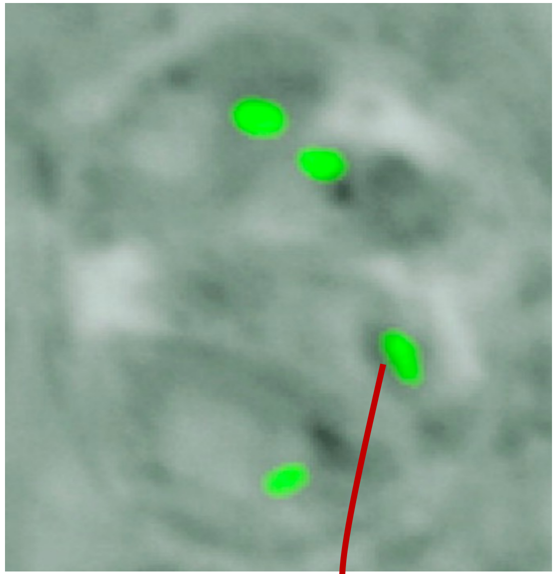
Why is GO ontology useful?



Cyanelle



Chloroplast



Apicoplast

GO:0009536 plastid

http://amigo.geneontology.org/amigo/term/GO:0009536#display-lineage-tab

GO Enrichment:

- Which terms occur more frequently in your list of differentially expressed genes than would be expected by chance based on the frequency in the genome

Has GO term?	Genes in subset	Genes not in subset	Gene in genome
Yes	10 (expect 6)	50 (expect 54)	60
No	90 (expect 94)	850 (expect 846)	940
Total	100	900	1000

- Fisher’s exact test with multiple test correction

Multiple Test Corrections

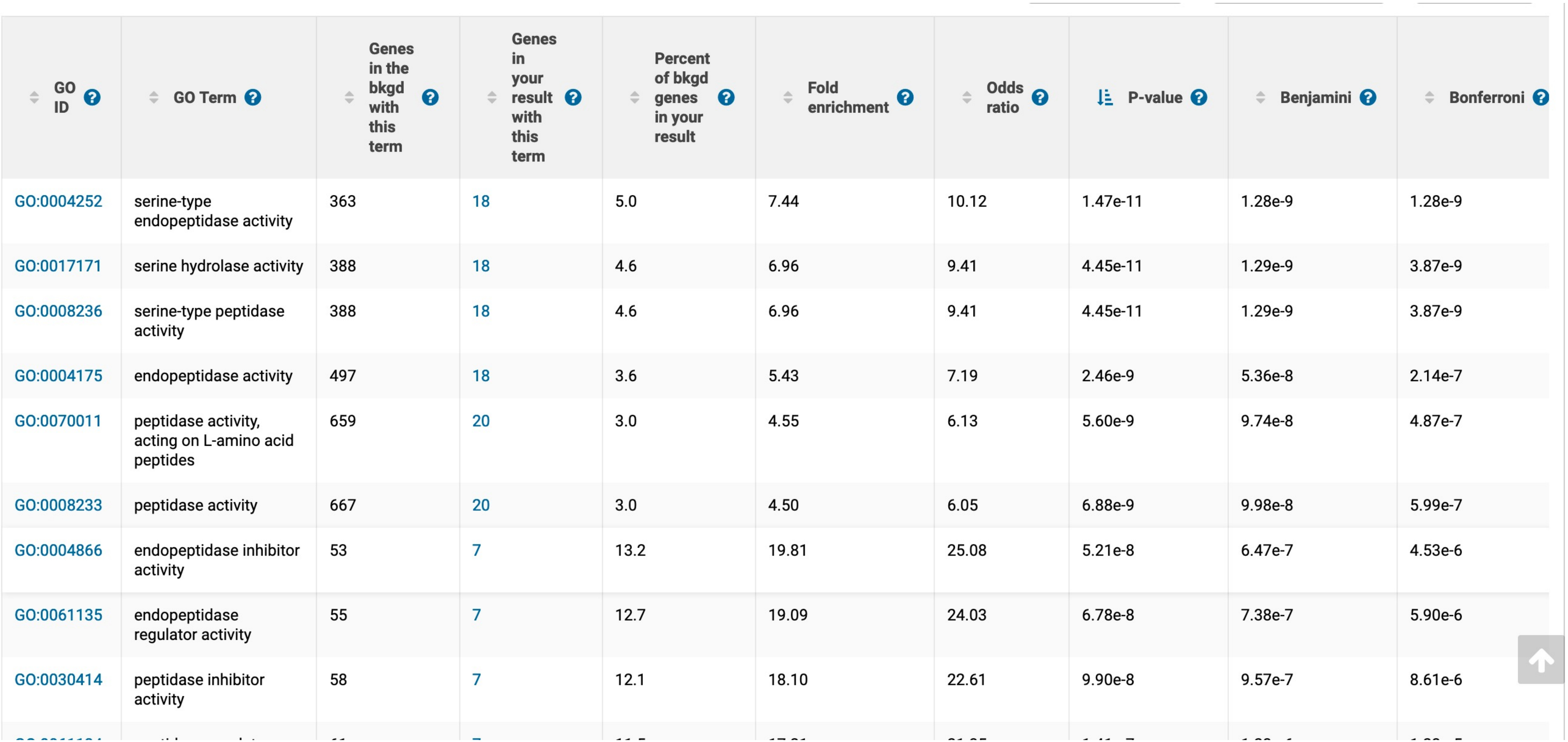
- If we do a statistical test and consider p <= 0.01 as significant, we accept that 1 in 100 results will be false positives
- If we test 10,000 GO terms, we therefore expect 100 terms with p<=0.01 by chance alone
- Multiple test corrections (FDR, adjusted p-value, q-value) adjust the p-values to account for this so you can have more confidence in your results

Some caveats

- GO enrichment relies on the GO term assignments being accurate
    - Always be aware of where they come from

- GO term assignments is not complete. There will be many genes that do not have an assignment
    - What does this mean for your analysis?
    - Enrichment will not tell you anything about genes without an assignment

GO enrichment results in VEuPathDB.org



GO ID	GO Term	Genes in the bkgd with this term	Genes in your result with this term	Percent of bkgd genes in your result	Fold enrichment	Odds ratio	P-value	Benjamini	Bonferroni
GO:0004252	serine-type endopeptidase activity	363	18	5.0	7.44	10.12	1.47e-11	1.28e-9	1.28e-9
GO:0017171	serine hydrolase activity	388	18	4.6	6.96	9.41	4.45e-11	1.29e-9	3.87e-9
GO:0008236	serine-type peptidase activity	388	18	4.6	6.96	9.41	4.45e-11	1.29e-9	3.87e-9
GO:0004175	endopeptidase activity	497	18	3.6	5.43	7.19	2.46e-9	5.36e-8	2.14e-7
GO:0070011	peptidase activity, acting on L-amino acid peptides	659	20	3.0	4.55	6.13	5.60e-9	9.74e-8	4.87e-7
GO:0008233	peptidase activity	667	20	3.0	4.50	6.05	6.88e-9	9.98e-8	5.99e-7
GO:0004866	endopeptidase inhibitor activity	53	7	13.2	19.81	25.08	5.21e-8	6.47e-7	4.53e-6
GO:0061135	endopeptidase regulator activity	55	7	12.7	19.09	24.03	6.78e-8	7.38e-7	5.90e-6
GO:0030414	peptidase inhibitor activity	58	7	12.1	18.10	22.61	9.90e-8	9.57e-7	8.61e-6

Enzyme commission numbers:

systematic and logical nomenclature for enzymes

Numbers of composed of 4 digits:

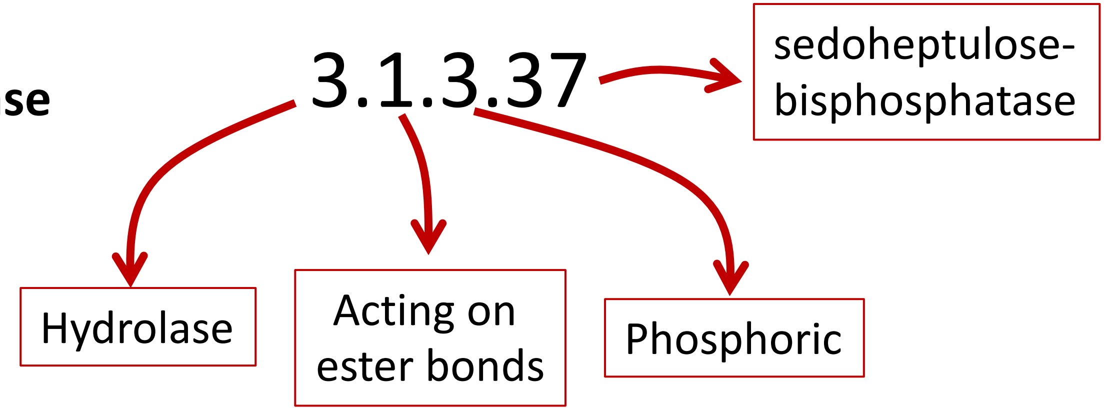
(i) the first number shows to which of the six main divisions (classes) the enzyme belongs,

(ii) the second figure indicates the subclass,

(iii) the third figure gives the sub-subclass,

(iv) the fourth figure is the serial number of the enzyme in its sub-subclass.

Example: **sedoheptulose-1,7-bisphosphatase**



3.1.3.37

sedoheptulose-bisphosphatase

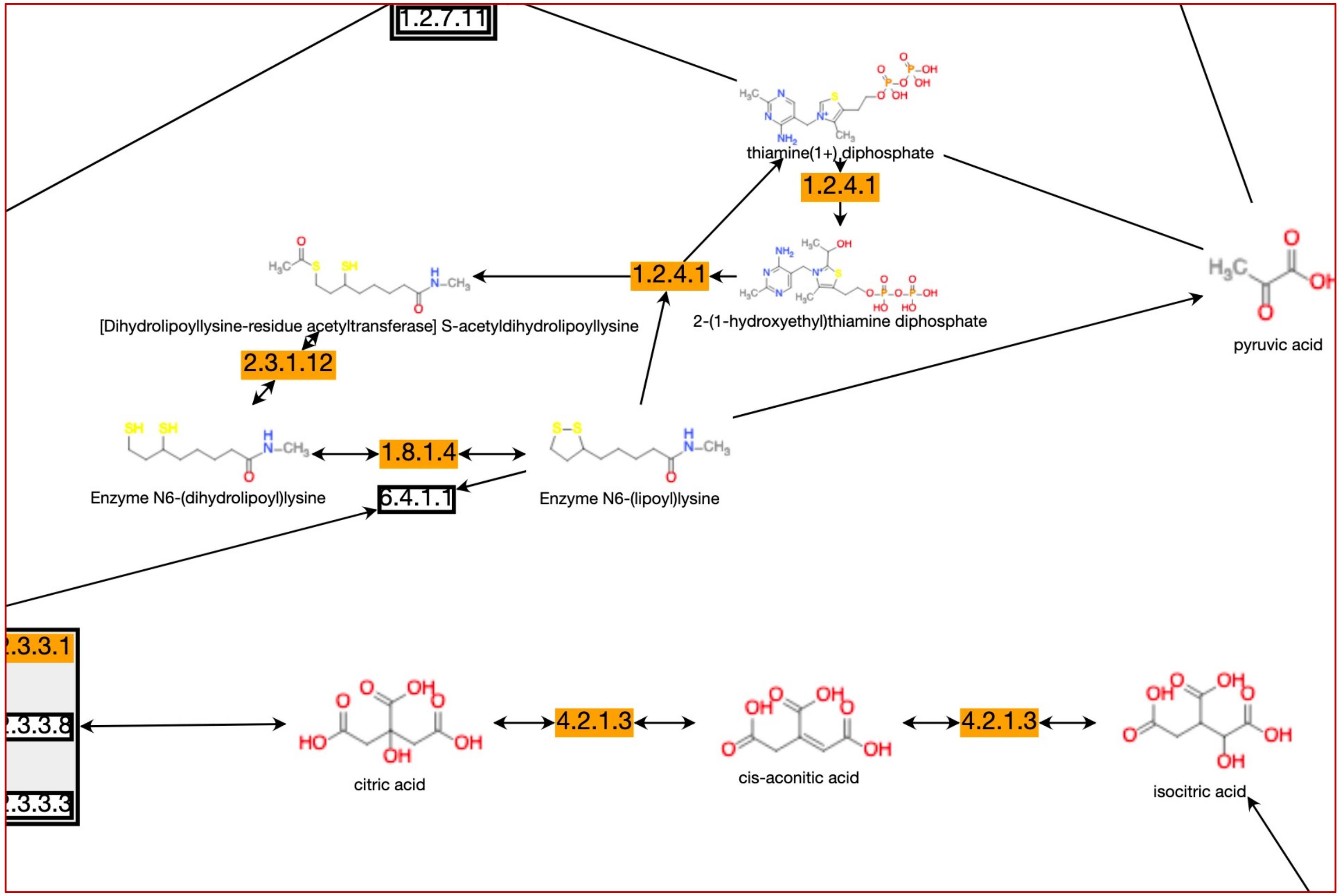
Hydrolase

Acting on ester bonds

Phosphoric

https://www.qmul.ac.uk/sbcs/iubmb/enzyme/

EC numbers can be used to link gene products with pathways



EC numbers and GO terms can be used in enrichment analysis!

For example: Does my list of genes have an over-representation of specific GO terms compared to the rest of the genome?

A standard enrichment method is Fisher’s exact test which is a statistical test used when analyzing contingency tables. Typically used when you have a small sample size. But when you are doing enrichment analysis on a list of genes with the backgroung being the whole genome, your sample size is not small. As a results the P- value you get from a Fisher’s exact test might be misleading.

With a small sample size the a P-value of less than 0.05 is considered significant (5% chance of being wrong/random). But if you are doing an enrichment analysis with all genes in the genome then each gene can be considered a test so the your chances of a type one error becomes higher. As a result you have to correct for this which can be done in different ways including Benjamini-Hochberg false discovery rate (FDR) or Bonferroni adjusted p-value