GOterms, Enrichment analyses, Multiple testing

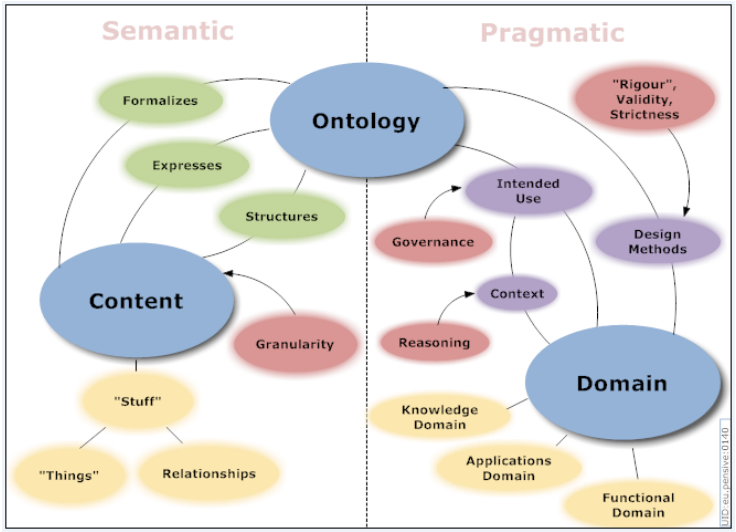
Université Paris-Saclay

November 29, 2023

The Gene Ontology

Ontology

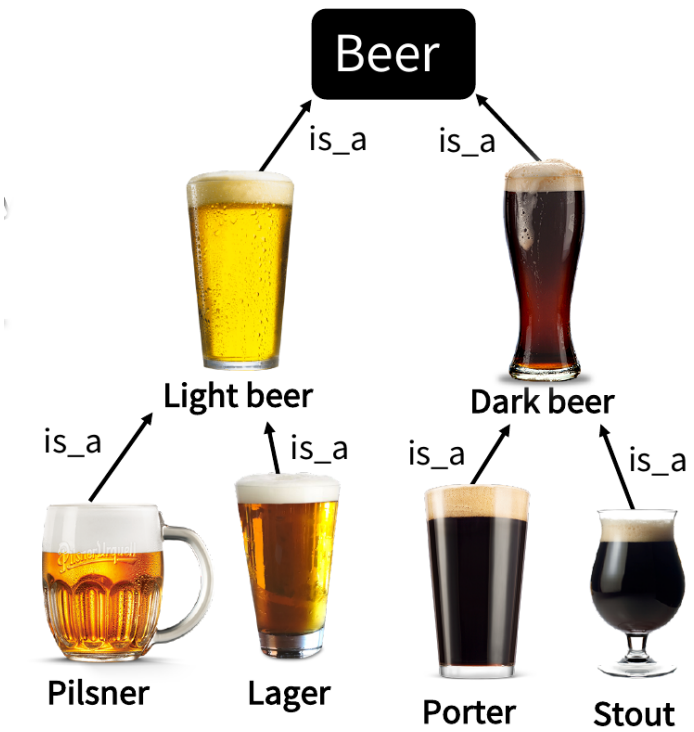
- **Gruber, 1993**: An ontology is a description (like a formal specification of a program) of the concepts and relationships that can formally exist for an agent or a community of agents. This definition is consistent with the usage of ontology as set of concept definitions, but more general. And it is a different sense of the word than its use in philosophy.
  Gruber, T (1993) Toward Principles for the Design of Ontologies Used for Knowledge Sharing. doi:10.1006/ijhc.1995.1081
- **Feilmayr and Woss, 2016**: An ontology is a formal, explicit specification of a shared conceptualization that is characterized by high semantic expressiveness required for increased complexity.
  Feilmayr, Christina; Woss, Wolfram (2016) An analysis of ontologies and their success factors for application to business. Data & Knowledge Engineering. 101: 1-23



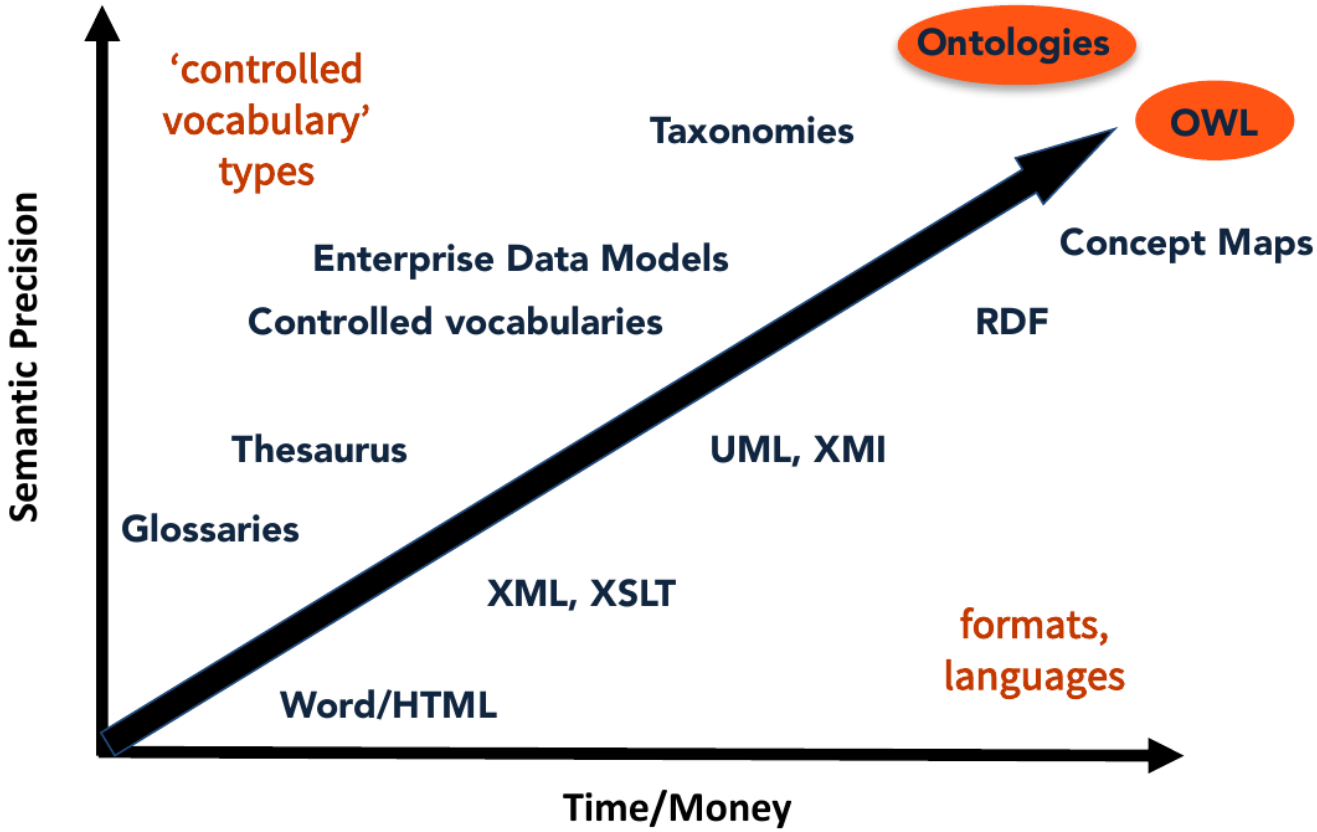
Ontologies

A knowledge classification of a domain, where the relationships between concepts are formally defined and logically related, which allows for computational reasoning

- An ontology is a set of terms, relationships and definitions that capture the knowledge of a certain domain.
- Terms represent a controlled vocabulary, and define the concepts of a domain.
- Terms are linked by relationships, which constitute a semantic network.
- Terms are arranged in a hierarchy
- Ontologies augment natural language annotations and can be more easily processed computationally. (expressed in a knowledge representation language such as RDFS, OBO, or OWL)



Representation languages

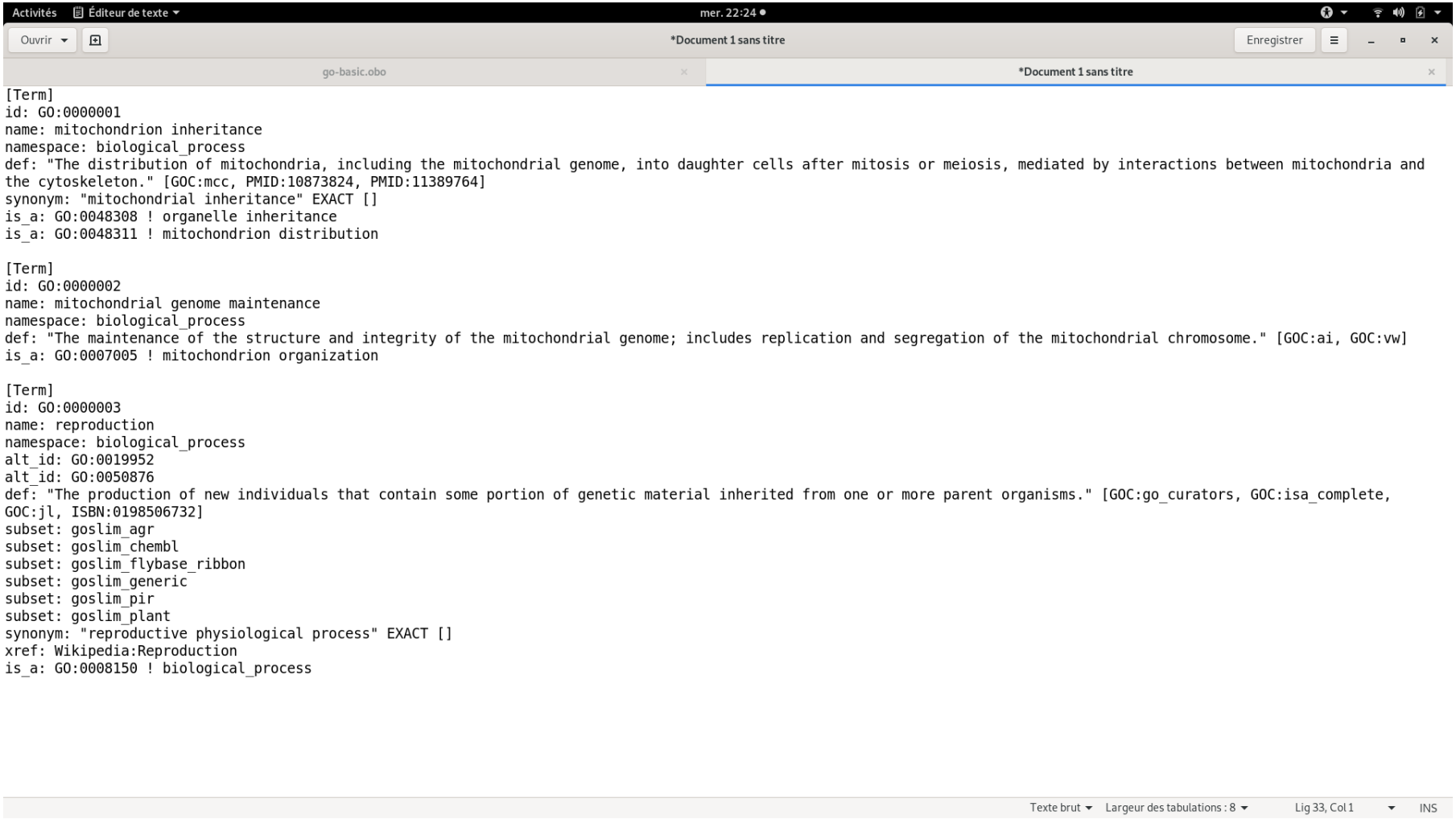


Connectedness
capture logical, non-hierarchical relationships ***across*** the model

Expressivity
formal languages amenable to computational reasoning

Abstraction
not tied to a particular implementation

the OBO knowledge representation language

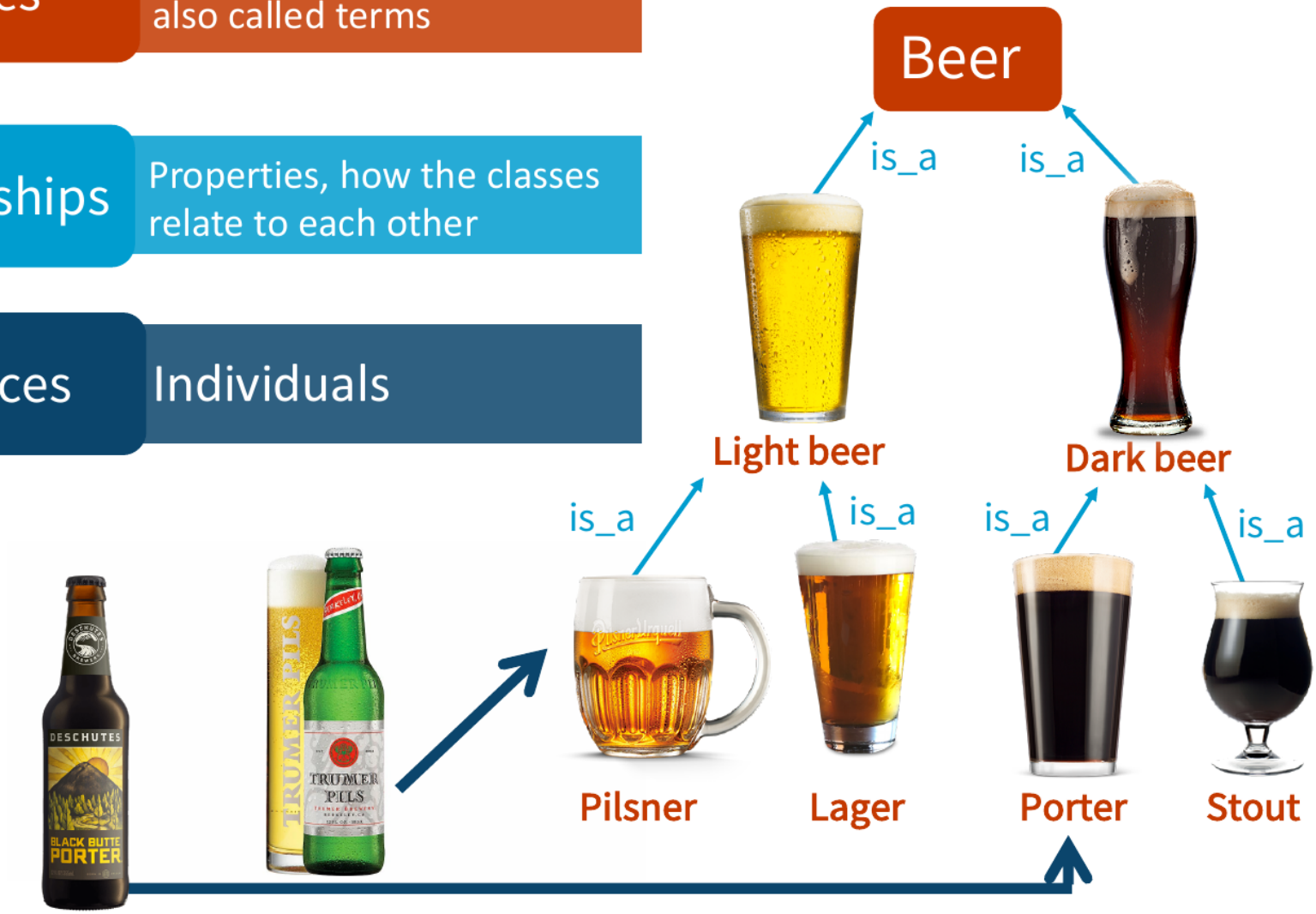


Elements of an ontology

Classes — Concepts in the ontology, also called terms

Relationships — Properties, how the classes relate to each other

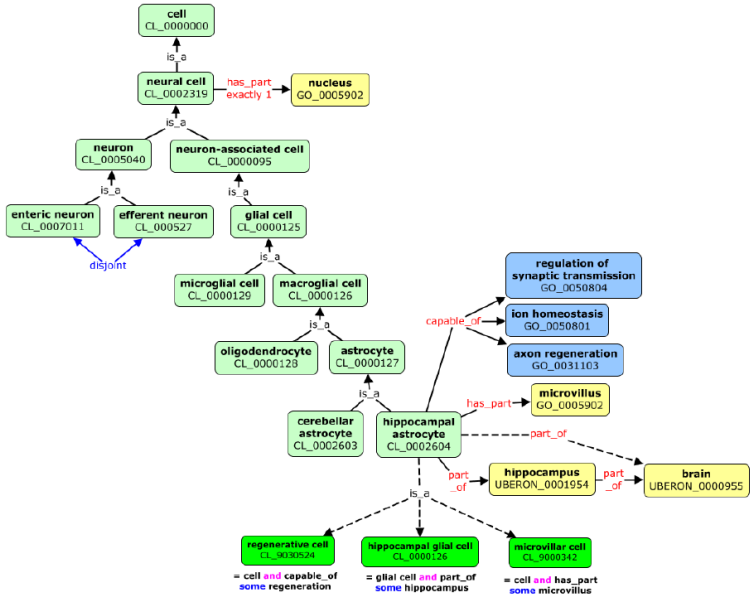
Instances — Individuals



Associative structures

The part_of relationship

- Nothing is a part of itself
- If A is a part of B then the B is not a part of A
- If A is a part of B and B is a part of C then A is a part of C
- The relationship is asymmetrical and transitive



Many associative relationships between classes create network structure.

Famous ontologies

You may have heard of ...

hpo

mondo
MONARCH DISEASE ONTOLOGY

GENEONTOLOGY
Unifying Biology

SNOMED CT
The global language of healthcare

The Sequence Ontology

DISEASE ONTOLOGY

efo Experimental Factor Ontology

Envo

The Open Biological and Biomedical Ontology (OBO) Foundry

`http://www.obofoundry.org/`

Community development of interoperable ontologies for the biological sciences

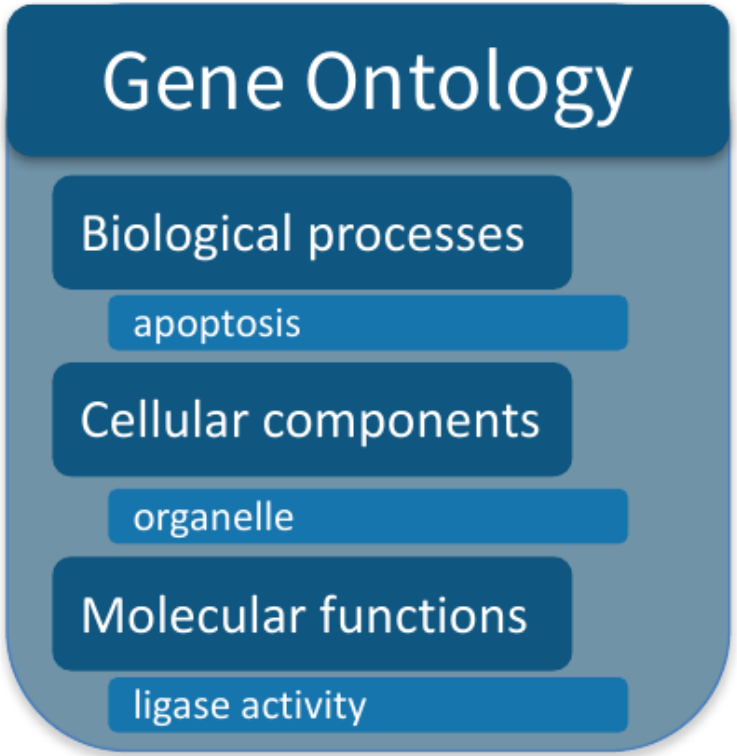
The Gene Ontology

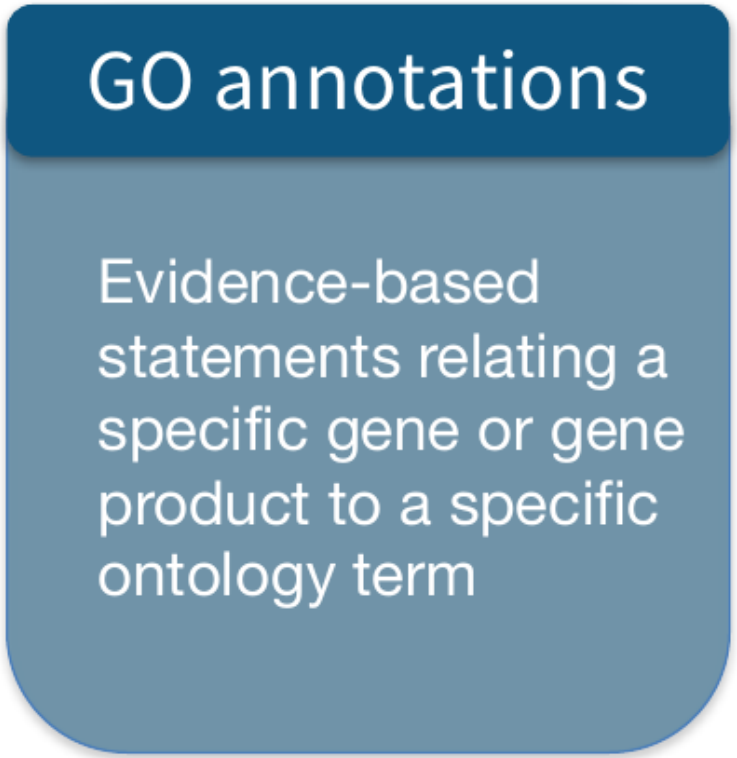
`http://geneontology.org/`

Three hierarchical structures :

- **Molecular function**: elemental activity/task (what)
  (e.g., DNA-binding, polymerase, transcription factor)
  (what a gene does at the biochemical level)
- **Biological process**: goal or objective (why)
  (e.g., mitosis, DNA replication, cell cycle control)
  (A broad biological perspective – not currently a pathway)
- **Cellular component**: location within cellular structures and macromolecular complex (where)
  (e.g., nucleus, ribosome, pre-replication complex)

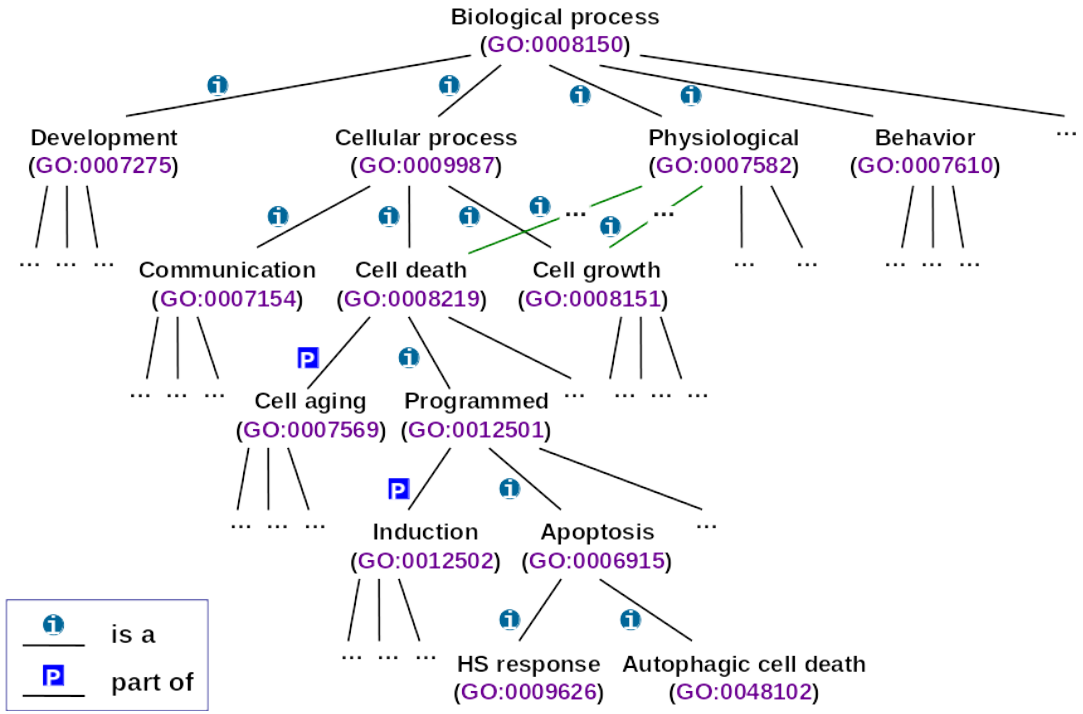
Gene Ontology is associated with experimental evidences





Provides computable knowledge regarding the functions of genes and gene products

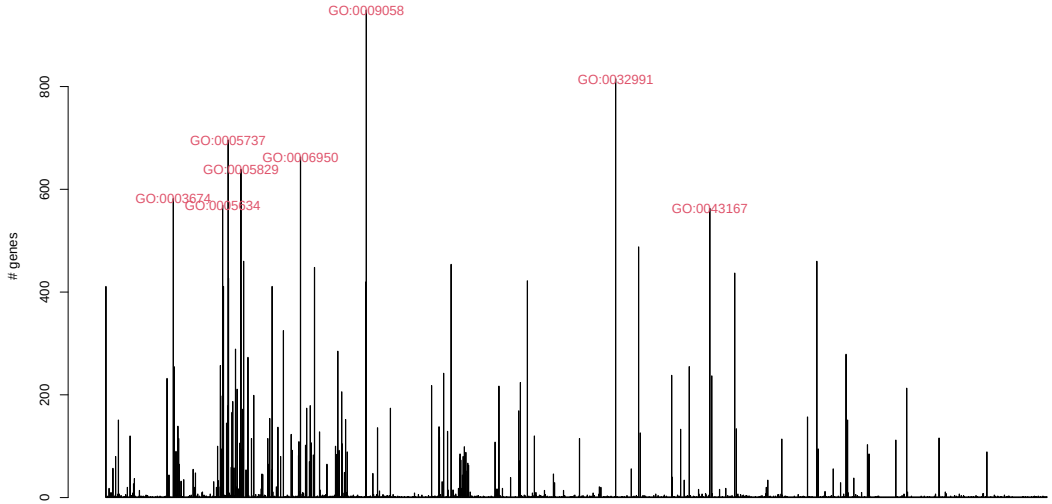
GO structure : Direct Acyclic Graph



A child term may have many parent terms.

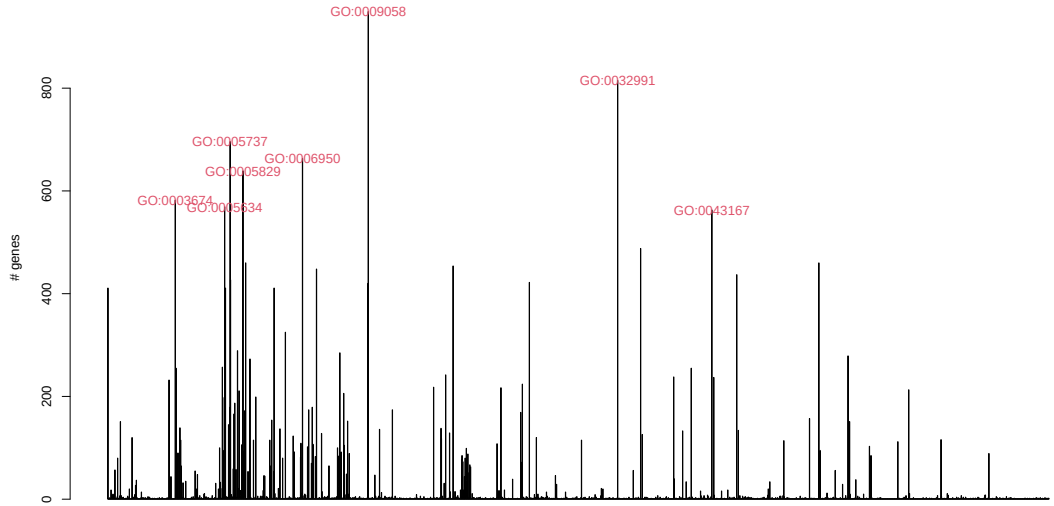
Enrichment analyses

GO annotation of *P. anserina*



- **Total Gene count :** 10803
- **Total Gene count after filtering :**9796
- **Total Annotated genes :**3464
- **Total genes DEGs between CS and SH at 4h :**4598
- **Total Annotated genes DEGs between CS and SH at 4h :**1564

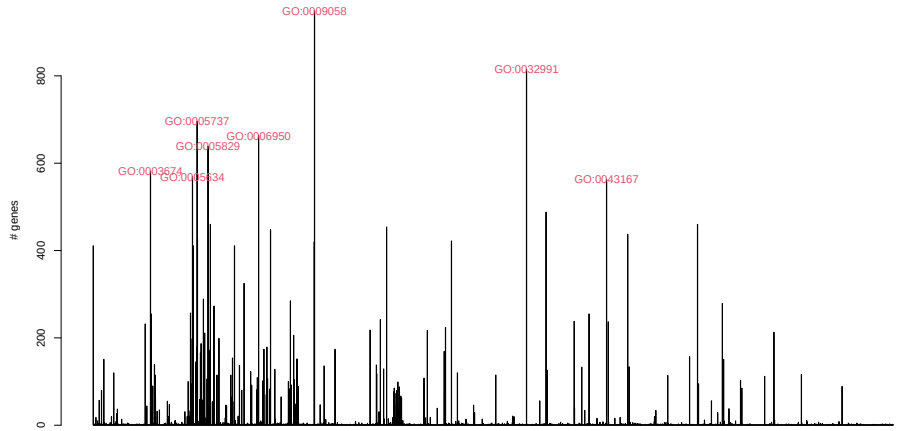
GO annotation of *P. anserina*



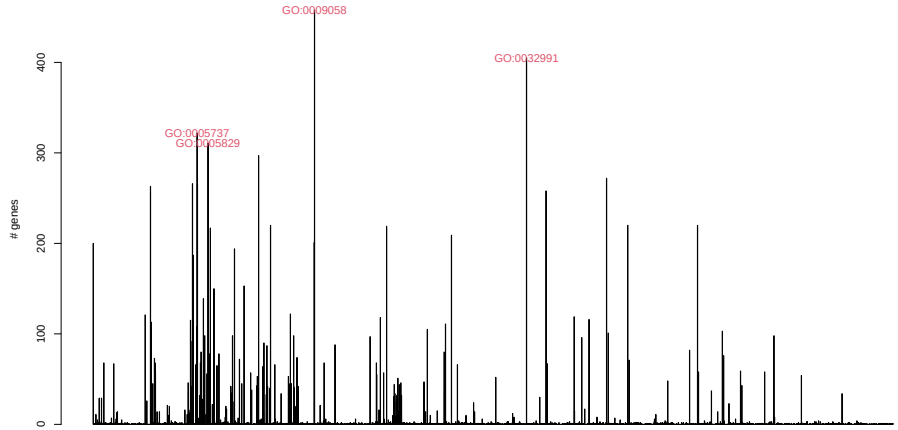
- GO:00109058[Biological_Process/Biosynthetic process]: The chemical reactions and pathways resulting in the formation of substances; typically the energy-requiring part of metabolism in which simpler substances are transformed into more complex ones.
- GO:0032991[cellular_component/Protein containing complex]: A protein complex in this context is meant as a stable set of interacting proteins which can be co-purified by an acceptable method, and where the complex has been shown to exist as an isolated, functional unit in vivo.
- GO:0005737[cellular_component/cytoplasm]: The contents of a cell excluding the plasma membrane and nucleus, but including other subcellular structures.
- GO:0006950[Biological_Process/Response to Stress]: Any process that results in a change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a disturbance in organismal or cellular homeostasis, usually, but not necessarily, exogenous (e.g. temperature, humidity, ionizing radiation).
- GO:0005829: [cellular_cpomponent/cytosol]: The part of the cytoplasm that does not contain organelles but which does contain other particulate matter, such as protein complexes.

Enrichment Analysis

Whole genome



[CS:4h-SH:4h]



Are some GO classes over/under-represented in the DEG set ?

A question of chance...

The Urn

- $A = 3464$ annotated genes
- $K = 325$ occurences of the GOterm $GO : 0006629$ (biological_process/lipid metabolic activity)

The trial

- $n = 1564$ annotated DEGs in the comparison CS-SH at 4h
- $x = 153$ observations of the GOterm $GO : 0006629$

Is GOterm $GO : 0006629$ over-represented in the DEG set ?

Random variables

Random variable : *a random variable (r. v.) is any variable with values depending on the outcome of a random phenomenon.*

The random variable is written as X, and an outcome of this random variable is given as x, which is a particular value taken by the variable in a random selection.

A random variable is characterised by

1. The values it can take, which are called the **support** of the random variable.
2. The probability of finding each value in the population or **probability law**.

Count random variables

Bernouilli variables:

X	Probability
1	$P(X = 1) = p$
0	$P(X = 0) = 1 - p$

Count variable: A count variable Z measures the number of success in the $n - sample$

$$Z = \sum_{i=1}^{n} X_i$$

The probability law of Z depends on the probability of succes p, the sample size n, and the sampling modalities

- $n - sample$ = independent r.v. $\longrightarrow$ Binomial OR Poisson.
- Sampling without replacement $\longrightarrow$ hypergeometrical.
- Stop after r defeats $\longrightarrow$ negative binomial.
- Sampling in a structured population $\longrightarrow$ negative binomial.

Count data : usual laws

Population size A (can be unknown), probability of succes p. Sample size n. Defeat number r.
$Z =$ number of success in the $n - sample$.

Law	parameters	Expectancy	Variance
Binomial	$\mathcal{B}(n,p)$	$n.p$	$n.p.(1-p)$
Poisson	$\mathcal{P}(\lambda = n.p)$	λ	λ
Hypergeometrical	$\mathcal{H}(n,p,A)$	$n.p$	$n.p.(1-p)\frac{A-n}{A-1}$
Negative binomial	$\mathcal{NB}(r,1-p))$	$r\frac{p}{1-p}$	$r\frac{p}{(1-p)^2}$

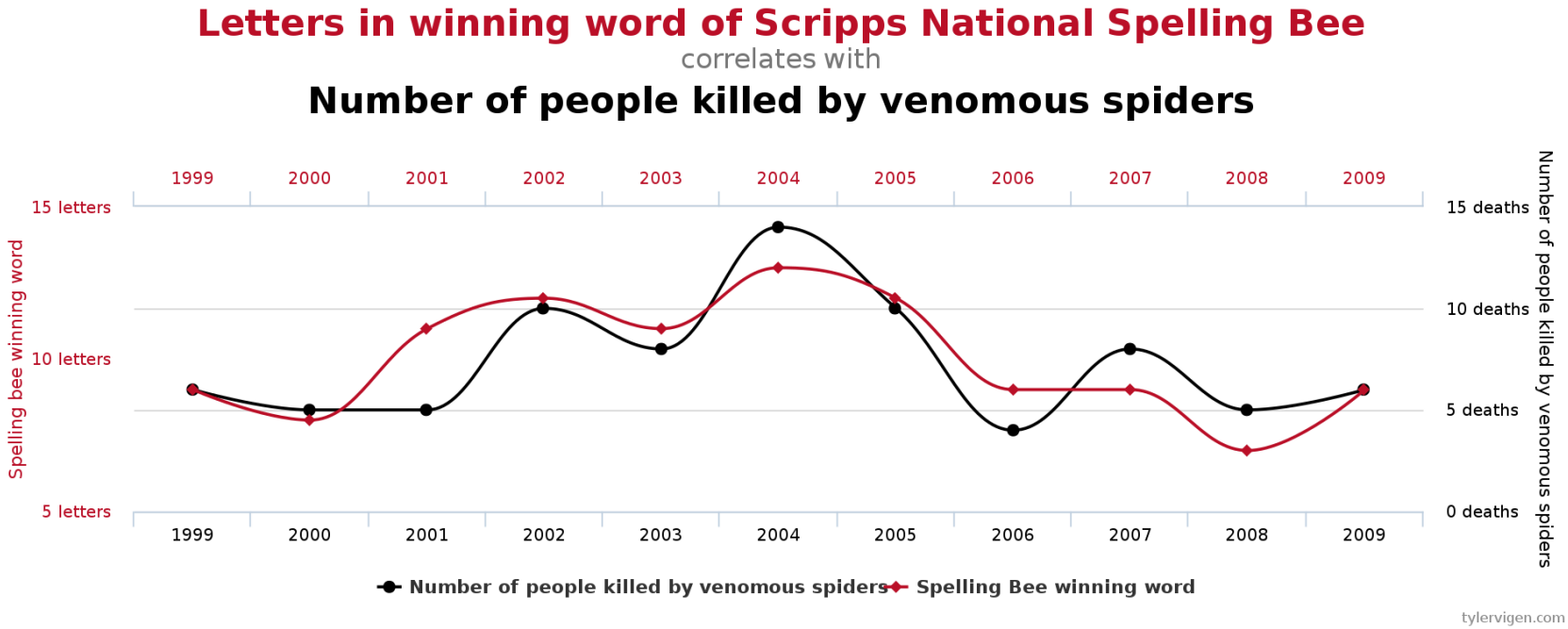
Is GO : 0006629 over-represented in the CS/CH 4h comparison ?

1. **Model**:
   - X: number of occurences of GO : 0006629 among the $n = 1564$ annnotated DEGs.
   - Sampling without replacement. $X \approx \mathcal{H}(A = 3464, p = \frac{K}{3464}, n = 1564)$. K unknown.
   - if GO : 0006629 has no effect, we expect $p = p = \frac{325}{3464}$ as in the urn.
2. **H0/H1 hypotheses**
   - H0: $p = \frac{325}{3464} = 0.094$
   - H1: $p > 0.094$
3. **Test statistics**
   Under the HO hypothesis, $X \approx \mathcal{H}(A = 3464, p = 0.094, n = 1564)$
4. The **pvalue** is computed as the probability that X is lower than 153 under the H0 hypothesis:
   $phyper(153, 325, 3464, 1564, lower.tail = FALSE) = 0.012$
5. The H0 hypothesis is rejected. The GOterm GO : 0006629 is over-represented among the DEGs between the two culture media CS-and SH. Differences between the two culture medi involve lipid metabolism.

Multiple tests

Multiple tests

In statistics, the multiple testing problem occurs when one considers a set of statistical inferences simultaneously or infers a subset of parameters selected based on the observed values. The more inferences are made, the more likely erroneous inferences become.



A simulation

Among $K = 1500$ tests performed in the same factory, the truth was H0 for 1000 tests, and H1 for 500 tests.

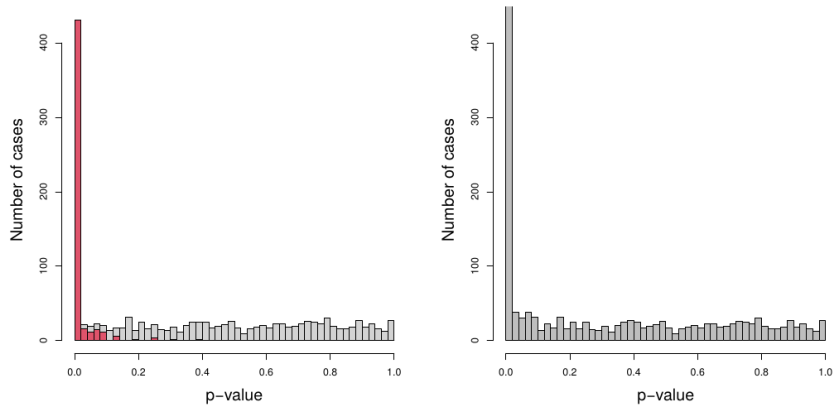


Figure 2.10: **Principle for calculating the *FDR*.** To illustrate the calculation principle, we will again use the example of the test given in chapter 2.2.1. We carry out this test for 1500 machines. For each machine we make $n = 30$ measurements. We carry out a bilateral test with H0: $\mu = 16$ as against H1: $\mu \neq 16$. We will simulate a set of data, by considering that 1000 machines are correctly adjusted, that is, under H0 ($\mu = 16$) and 500 machines are not correctly adjusted, that is, under H1($\mu \neq 16$, we take a mean between 16.01 and 16.2). **Left:** distributions of p-values for the machines under H0 (in white) and under H1 (in red). **Right:** distribution of p-values for all the machines tested. At the level 5%, the calculated FDR is 11.5%, that is to say, 11.5% of the machines for which we rejected H0 are false positives. At the level 1%, we have an FDR of 3.4%. If we want to apply the Bonferroni correction, we need to apply a level of 0.00003 for an overall level of 5% at each test. In this case, we find no false positives. This level is too "stringent", since the calculation made for the Bonferroni correction assumes that all the machines are well adjusted, which is not the case.

Multiple tests correction of the test level

- **Boferoni**: Were all K tests under the H0 hypothesis, an overall α percent risk is reached by performing each test at the $\frac{\alpha}{K}$ level.
- **False Discovery Rate**: Given that the observed pvalue distribution is a mixture between tests under H0 and tests under H1, α_{FDR} is the level at which each test has to be performed to guarantee, overall, a given FDR value, *i.e* the proportion of rejection of the H0 hypothesis when it is true.