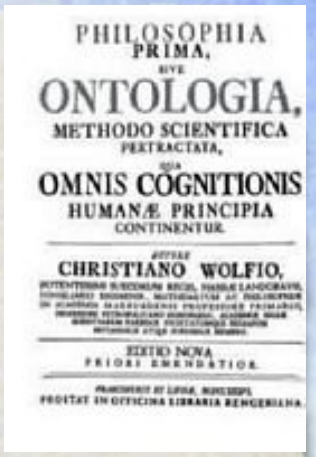
Gene Ontology

Shifra Ben-Dor
Weizmann Institute of Science

Outline

- What is GO (Gene Ontology)?
- What tools do we use to work with it?
- (Combination of GO with other analyses)

What is Ontology?

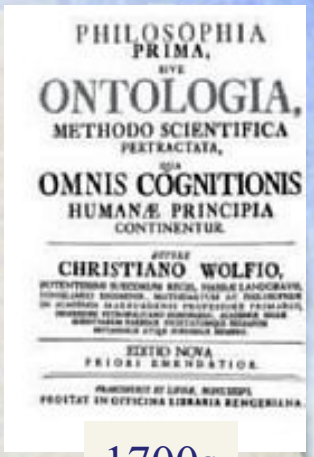

PHILOSOPHIA
PRIMA,
SIVE
ONTOLOGIA,
METHODO SCIENTIFICA
PERTRACTATA,
QUA
OMNIS COGNITIONIS
HUMANÆ PRINCIPIA
CONTINENTUR.
AUTORE
CHRISTIANO WOLFIO,
EDITIO NOVA
PRIORI EMENDATIOR.

1700s

Oxford English Dictionary

1. a. Philos. The science or study of being; that branch of metaphysics concerned with the nature or essence of being or existence.

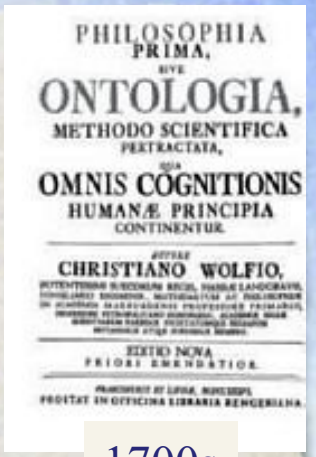
What is Ontology?


PHILOSOPHIA
PRIMA,
SIVE
ONTOLOGIA,
METHODO SCIENTIFICA
PERTRACTATA,
QUA
OMNIS COGNITIONIS
HUMANÆ PRINCIPIA
CONTINENTUR.
AUCTORE
CHRISTIANO WOLFIO,
EDITIO NOVA
PRIORI EMENDATIOR.

1700s

Ontology (from the Greek…) is the philosophical study of the nature of being, existence or reality in general, as well as of the basic categories of being and their relations. Traditionally listed as a part of the major branch of philosophy known as metaphysics, ontology deals with questions concerning what entities exist or can be said to exist, and how such entities can be grouped, related within a hierarchy, and subdivided according to similarities and differences.

What is Ontology?


PHILOSOPHIA
PRIMA,
SIVE
ONTOLOGIA,
METHODO SCIENTIFICA
PERTRACTATA,
QUA
OMNIS COGNITIONIS
HUMANÆ PRINCIPIA
CONTINENTUR.
AUCTORE
CHRISTIANO WOLFIO,
EDITIO NOVA
PRIORI EMENDATIOR.

1700s

Ontology (from the Greek…) is the philosophical study of the nature of being, existence or reality in general, as well as of the basic categories of being and their relations. Traditionally listed as a part of the major branch of philosophy known as metaphysics, ontology deals with questions concerning what entities exist or can be said to exist, and how such entities can be grouped, related within a hierarchy, and subdivided according to similarities and differences.

What is a Gene ?

So what is Gene Ontology?

- Unfortunately, not an ontology of genes, but rather of gene products
- It is an attempt to classify gene products using a structured language (controlled vocabulary) to give a consistent description of characteristics inherent to them.

GENEONTOLOGY

Unifying Biology

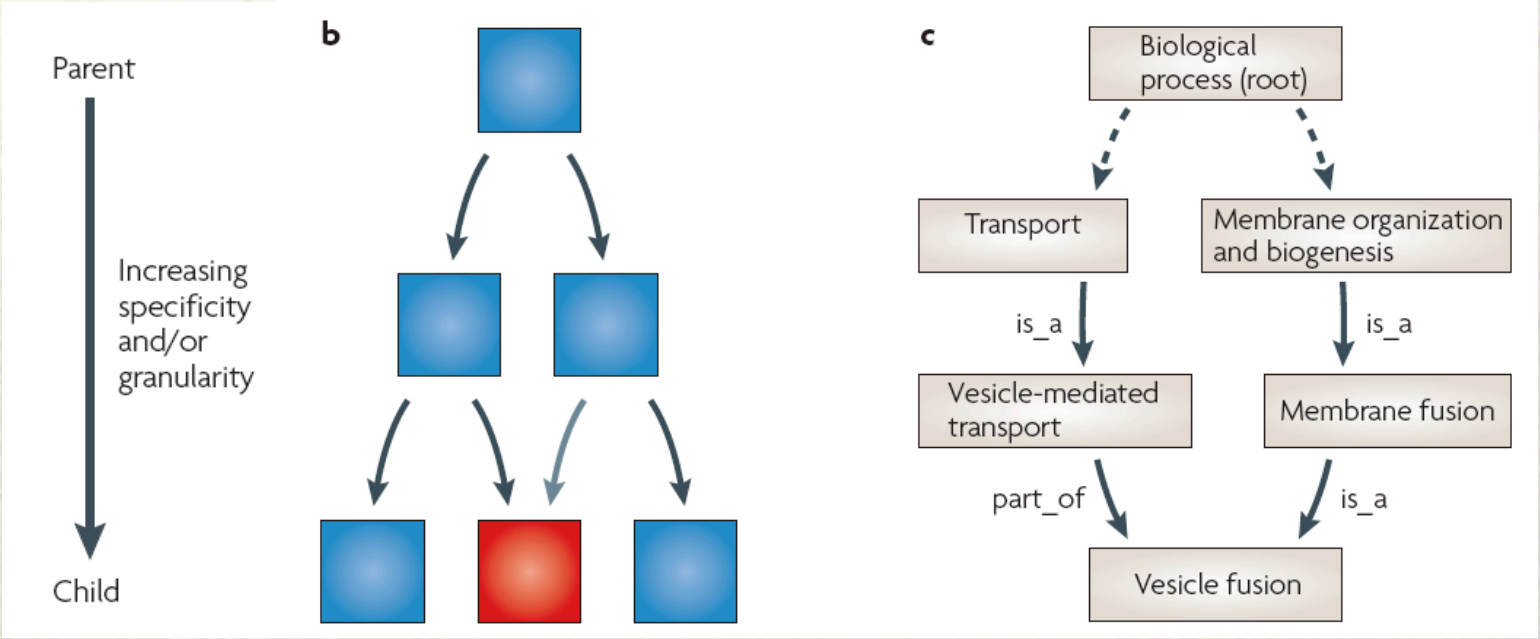
The Gene Ontology project is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases.

The project provides the controlled vocabulary of terms and gene product annotations from consortium members.

GENEONTOLOGY
Unifying Biology

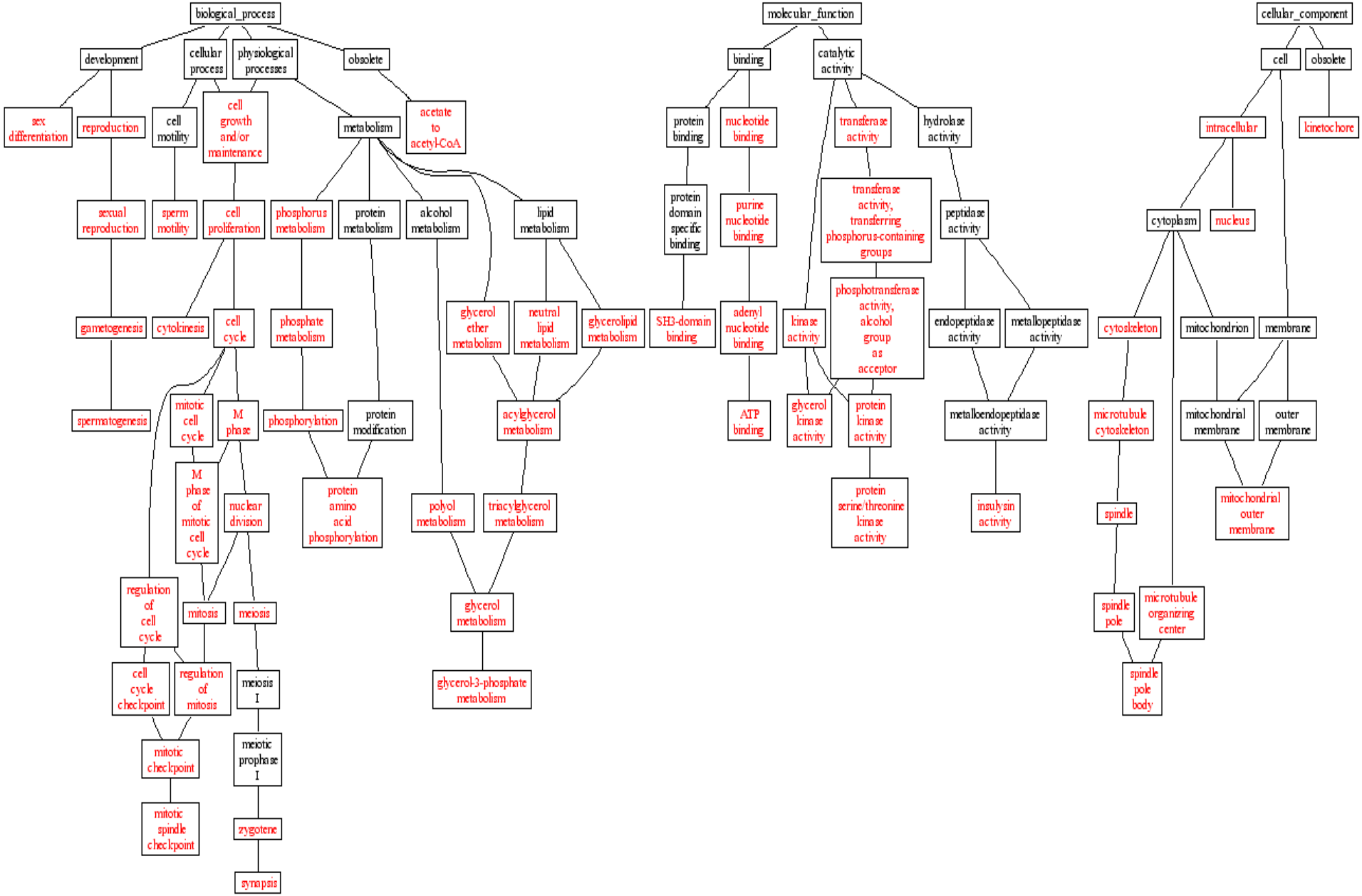
- Gene ontology is an annotation system which tries to describe attributes of gene products (what does it do? where? how?)
- It represents a unified consistent system, i.e. terms occur only once, and there is a dictionary of allowed words, which is consistent across species
- Furthermore, terms are related to each other: the hierarchy goes from very general terms to very detailed ones

Gene ontology is represented as a directed acyclic graph (DAG)



Taken from: Nature Reviews Genetics 9:509-515 (2008)

- A child can have more than one parent (parents are closer to the root and are more general, children are further from the root and more specific)
- There are no cycles - there is a root
- It is a directed graph
- You can skip levels in the graph

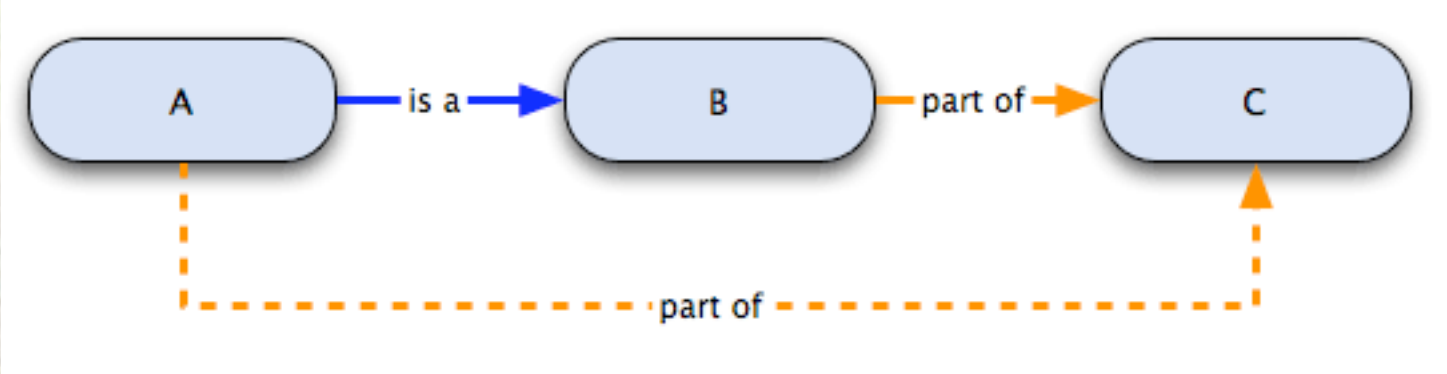

biological_process
development
cellular process
physiological processes
obsolete
sex differentiation
reproduction
cell motility
cell growth and/or maintenance
metabolism
acetate to acetyl-CoA
sexual reproduction
sperm motility
cell proliferation
phosphorus metabolism
protein metabolism
alcohol metabolism
lipid metabolism
gametogenesis
cytokinesis
cell cycle
phosphate metabolism
glycerol ether metabolism
neutral lipid metabolism
glycerolipid metabolism
spermatogenesis
mitotic cell cycle
M phase
phosphorylation
protein modification
acylglycerol metabolism
M phase of mitotic cell cycle
nuclear division
protein amino acid phosphorylation
polyol metabolism
triacylglycerol metabolism
regulation of cell cycle
mitosis
meiosis
glycerol metabolism
cell cycle checkpoint
regulation of mitosis
meiosis I
glycerol-3-phosphate metabolism
mitotic checkpoint
meiotic prophase I
mitotic spindle checkpoint
zygotene
synapsis
molecular_function
binding
catalytic activity
protein binding
nucleotide binding
transferase activity
hydrolase activity
protein domain specific binding
purine nucleotide binding
transferase activity, transferring phosphorus-containing groups
peptidase activity
SH3-domain binding
adenyl nucleotide binding
kinase activity
phosphotransferase activity, alcohol group as acceptor
endopeptidase activity
metallopeptidase activity
ATP binding
glycerol kinase activity
protein kinase activity
metalloendopeptidase activity
protein serine/threonine kinase activity
insulysin activity
cellular_component
cell
obsolete
intracellular
kinetochore
cytoplasm
nucleus
cytoskeleton
mitochondrion
membrane
microtubule cytoskeleton
mitochondrial membrane
outer membrane
spindle
mitochondrial outer membrane
spindle pole
microtubule organizing center
spindle pole body

Ontology Relations

- Just as the ontology terms are defined, so are the relationships between them (the arrows). The terms are linked by three relationships:
  - is_a
  - part_of
  - regulates, positively regulates, negatively regulates

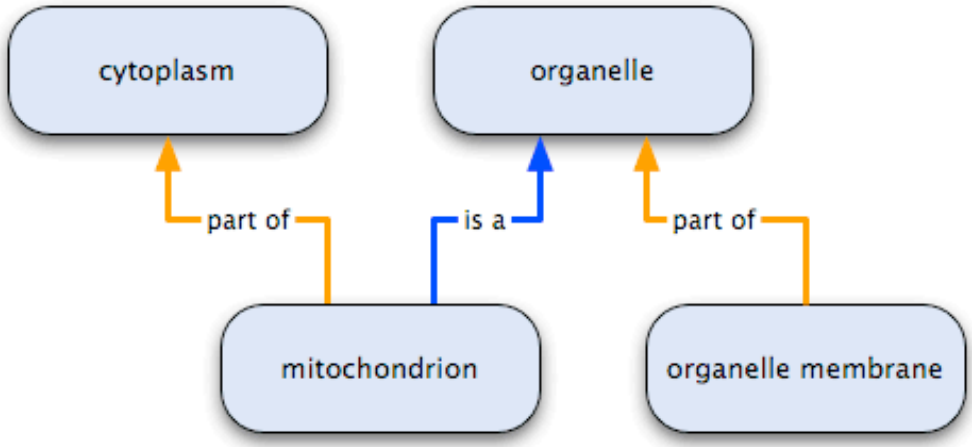
Ontology Relations

- is_a is a simple class-subclass relationship, for example, nuclear chromosome is_a chromosome.
- part_of is slightly more complex; C part_of D means that whenever C is present, it is always a part of D. An example would be nucleus part_of cell; nuclei are always part of a cell, but not all cells have nuclei.



A dotted line means an inferred relationship, e.g. one that has not been expressly stated

Taken from http://www.geneontology.org/

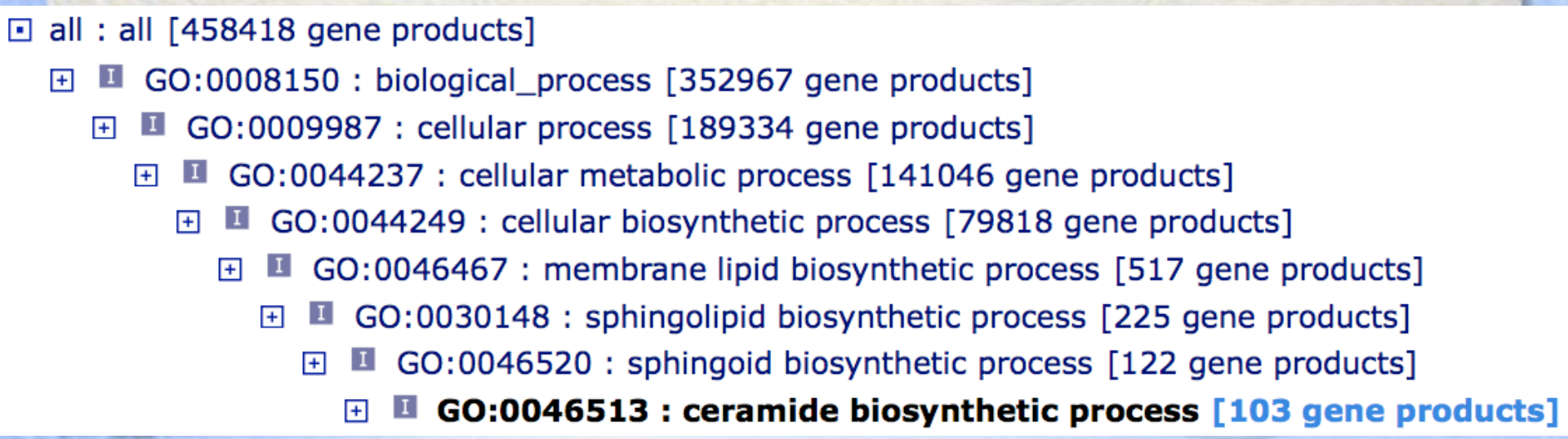


mitochondrion has two parents: it *is an* organelle and it is *part of* the cytoplasm;
organelle has two children: mitochondrion *is an* organelle, and organelle membrane is *part of* organelle

Taken from http://www.geneontology.org/

Ontology Structure

Every GO term must obey “the true path rule”: if the child term describes the gene product, then all its parent terms must also apply to that gene product.



GO has 3 major divisions (roots)

- Biological Process
- Molecular Function
- Cellular Component

Biological Process

A biological process is series of events accomplished by one or more ordered 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. It can be difficult to distinguish between a biological process and a molecular function, but the general rule is that a process must have more than one distinct steps.

Biological Process

A biological process is not equivalent to a pathway; at present, GO does not try to represent the dynamics or dependencies that would be required to fully describe a pathway.

Molecular Function

Molecular function describes activities, such as catalytic or binding activities, that occur at the molecular level. GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where or 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, transporter activity, or binding; examples of narrower functional terms are adenylate cyclase activity or Toll receptor binding.

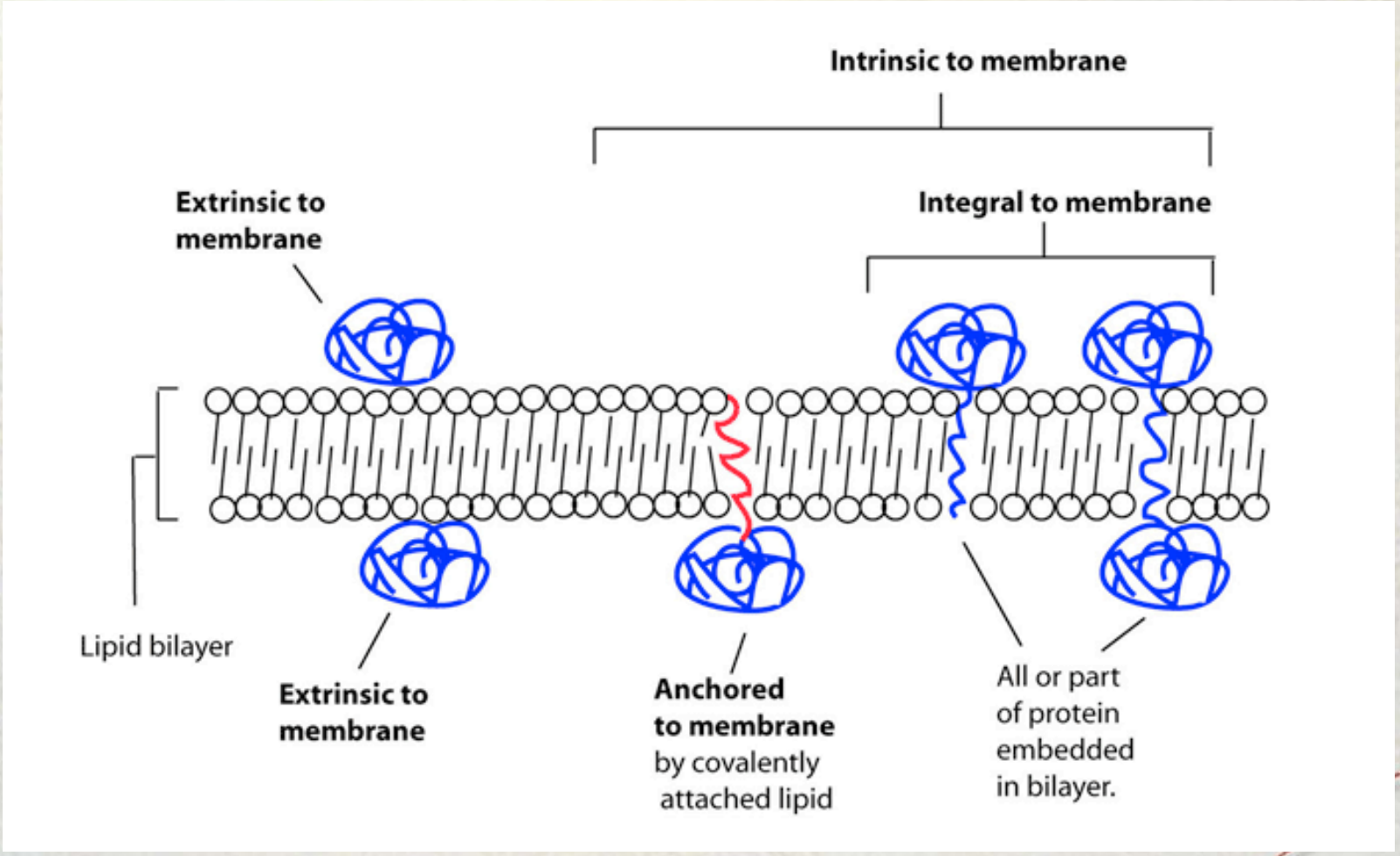
Molecular Function

It is easy to confuse a gene product name with its molecular function, and for that reason many GO molecular functions are appended with the word "activity".

Cellular Component

A cellular component is just that, a component of a cell, but with the proviso that it is part of some larger object; this may be an anatomical structure (e.g. rough endoplasmic reticulum or nucleus) or a gene product group (e.g. ribosome, proteasome or a protein dimer).

Cellular Component



Available GO Information

Current ontology statistics, as of June 9, 2019:

44,990 terms

29,701 biological_process

4,213 cellular_component

11,076 molecular_function

~2000 obsolete terms (not counted above)

What is not GO?

- Gene products: e.g. cytochrome c is not in the ontologies, but attributes of cytochrome c, such as oxidoreductase activity, are
- Processes, functions or components that are unique to mutants or diseases: e.g. oncogenesis
- Attributes of sequence such as intron/exon parameters
- Protein domains or structural features
- Protein-protein interactions
- Environment, evolution and expression
- It is not **complete,** it is done "by hand" by curators

Annotation

- What connects the GO terms to specific gene products
- Annotation is carried out by curators in a range of bioinformatics database resource groups. These groups then contribute their data to the central GO repository for storage and redistribution.
- There are two general principles: first, annotations should be attributed to a source; second, each annotation should indicate the evidence on which it is based.

Evidence Codes

Evidence code	Evidence code description	Source of evidence	Manually checked	Current number of annotations*
IDA	Inferred from direct assay	Experimental	Yes	71,050
IEP	Inferred from expression pattern	Experimental	Yes	4,598
IGI	Inferred from genetic interaction	Experimental	Yes	8,311
IMP	Inferred from mutant phenotype	Experimental	Yes	61,549
IPI	Inferred from physical interaction	Experimental	Yes	17,043
ISS	Inferred from sequence or structural similarity	Computational	Yes	196,643
RCA	Inferred from reviewed computational analysis	Computational	Yes	103,792
IGC	Inferred from genomic context	Computational	Yes	4
IEA	Inferred from electronic annotation	Computational	No	15,687,382
IC	Inferred by curator	Indirectly derived from experimental or computational evidence made by a curator	Yes	5,167
TAS	Traceable author statement	Indirectly derived from experimental or computational evidence made by the author of the published article	Yes	44,564
NAS	Non-traceable author statement	No 'source of evidence' statement given	Yes	25,656
ND	No biological data available	No information available	Yes	132,192
NR	Not recorded	Unknown	Yes	1,185

*October 2007 release

Evidence codes — not all annotations are created equal

Taken from: Nature Reviews Genetics 9:509-515 (2008)

GO Pitfalls

- Not complete
- Computational annotations
- NOT qualifier
- Splice variants
- Identifier flagged as ‘obsolete’

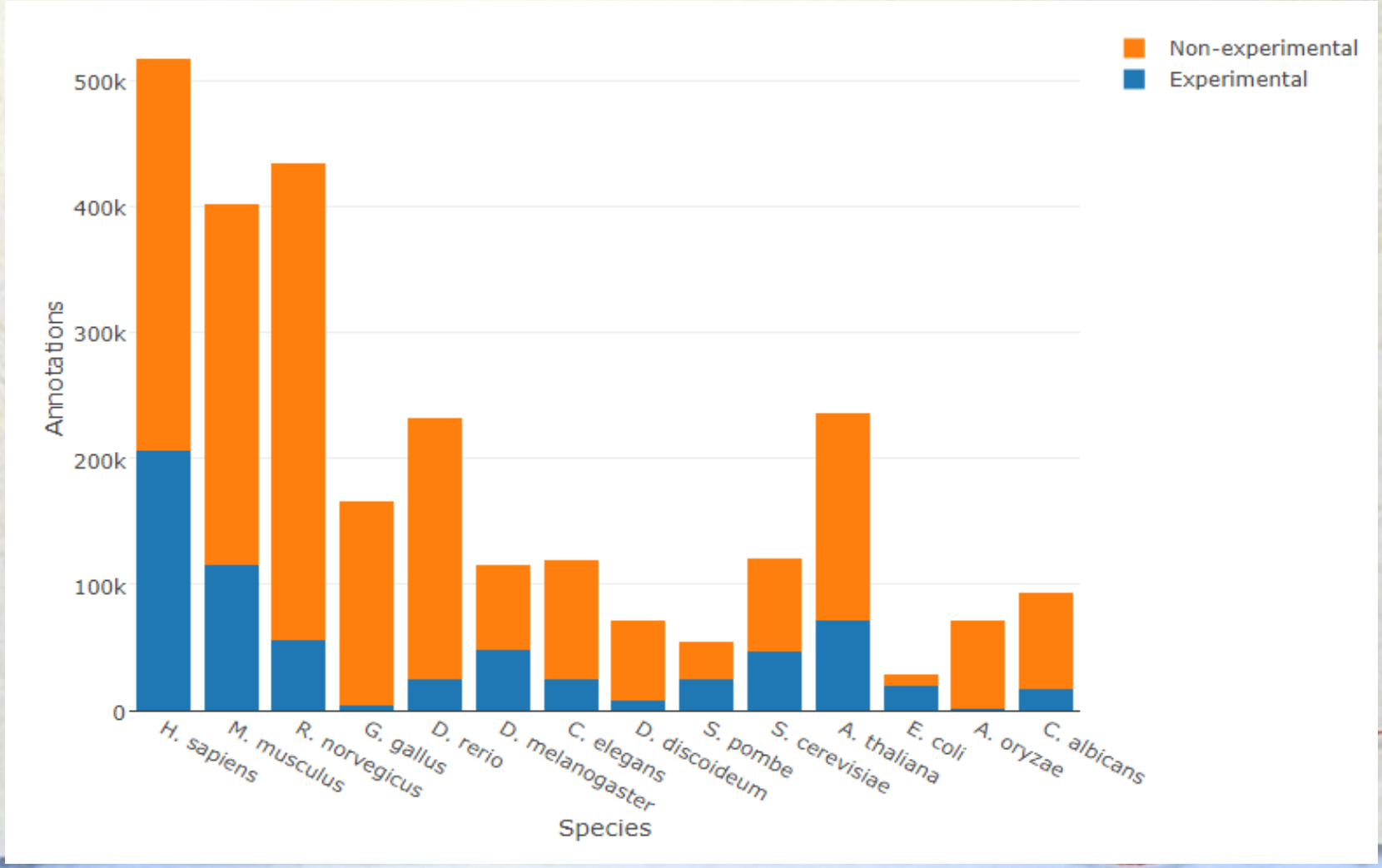
GO Pitfalls

- Not complete
- Computational annotations
- NOT qualifier
- Splice variants
- Identifier flagged as ‘obsolete’

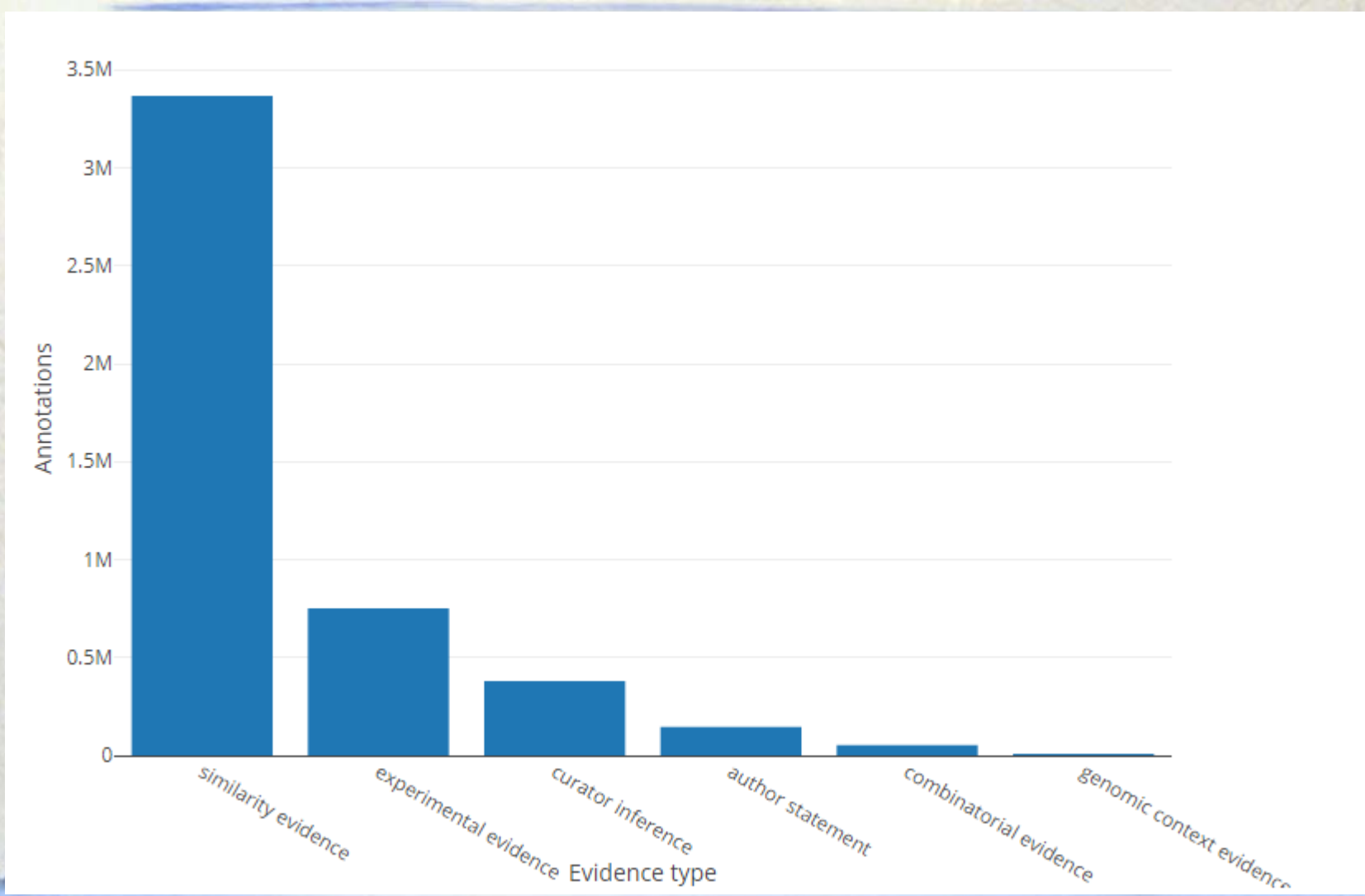
GO Pitfalls

- Not complete
- Computational annotations
- NOT qualifier
- Splice variants
- Identifier flagged as ‘obsolete’

Type of annotation per species



Type of annotation per evidence



GO Pitfalls

- Not complete
- Computational annotations
- NOT qualifier
- Splice Variants
- Identifier flagged as ‘obsolete’

NOT annotations in the gene ontology (GO) database

Table 3 | **NOT annotations in the gene ontology (GO) database***

Contributing database	Number of NOT annotations
CGD	11
Dictybase	76
FlyBase	246
GeneDB_Spombe	83
UniProt	148
AgBase	3
HGNC	41
MGI	217
RGD	21
SGD	88
TAIR	127
ZFIN	37

*As of 12 November 2007. CGD, *Candida* Genome Database; HGNC, HUGO Gene

Qualifiers:
contributes_to
colocalizes_with
NOT

Annotation qualifiers — to be or not to be is crucial for GO

GO Pitfalls

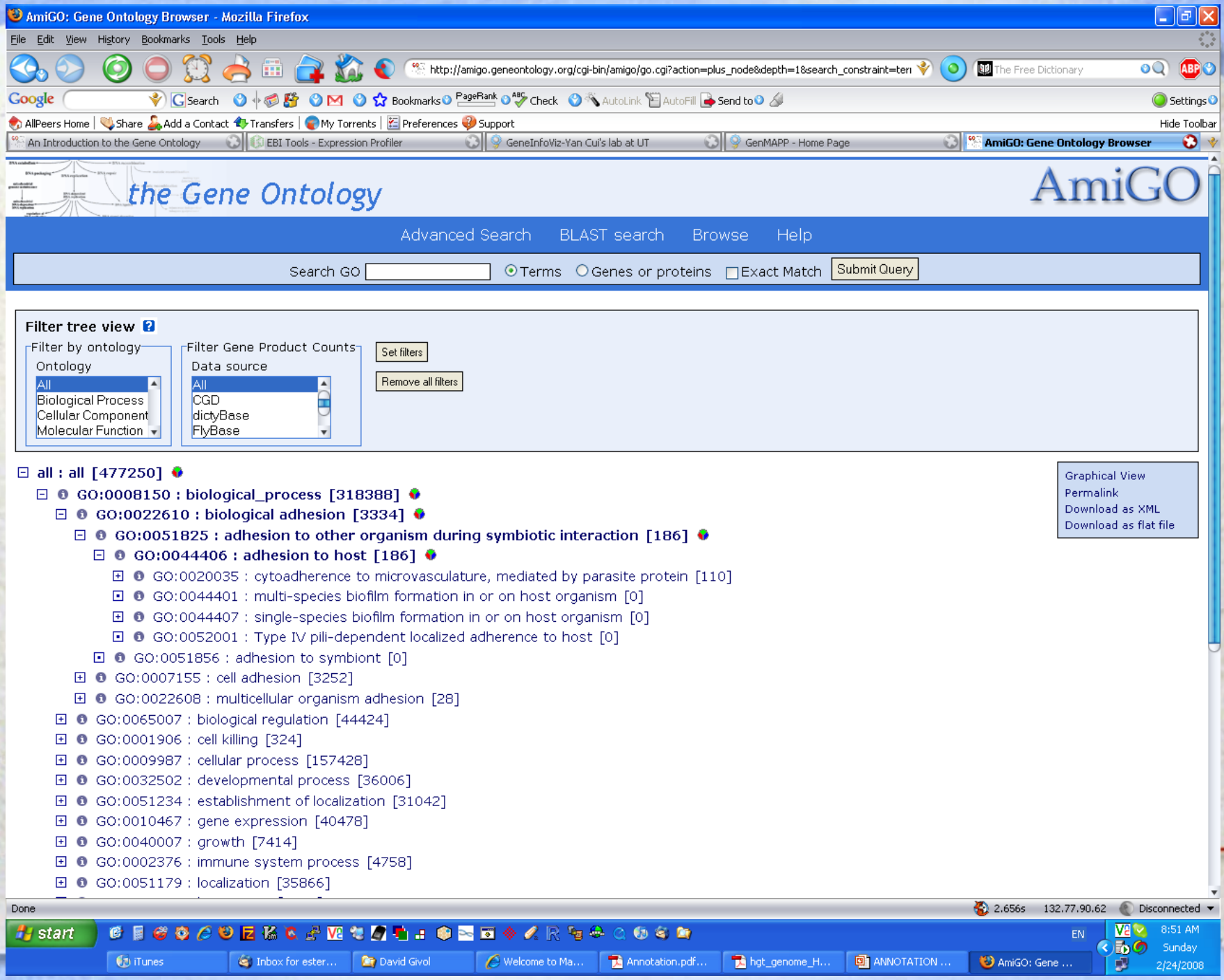
- Not complete
- Computational annotations
- NOT qualifier
- Splice Variants
- Identifier flagged as ‘obsolete’

Splice Variants

- GO annotation is related to gene products, not proteins, so the defining unit is the gene
- If you have different splice variants that have opposite effects, you will have opposing annotation for the same gene, for example BCLX – the long form is anti-apoptotic, the short form is pro-apoptotic, but they are from the same gene...

GO Pitfalls

- Not complete
- Computational annotations
- NOT qualifier
- Splice Variants
- Identifier flagged as ‘obsolete’

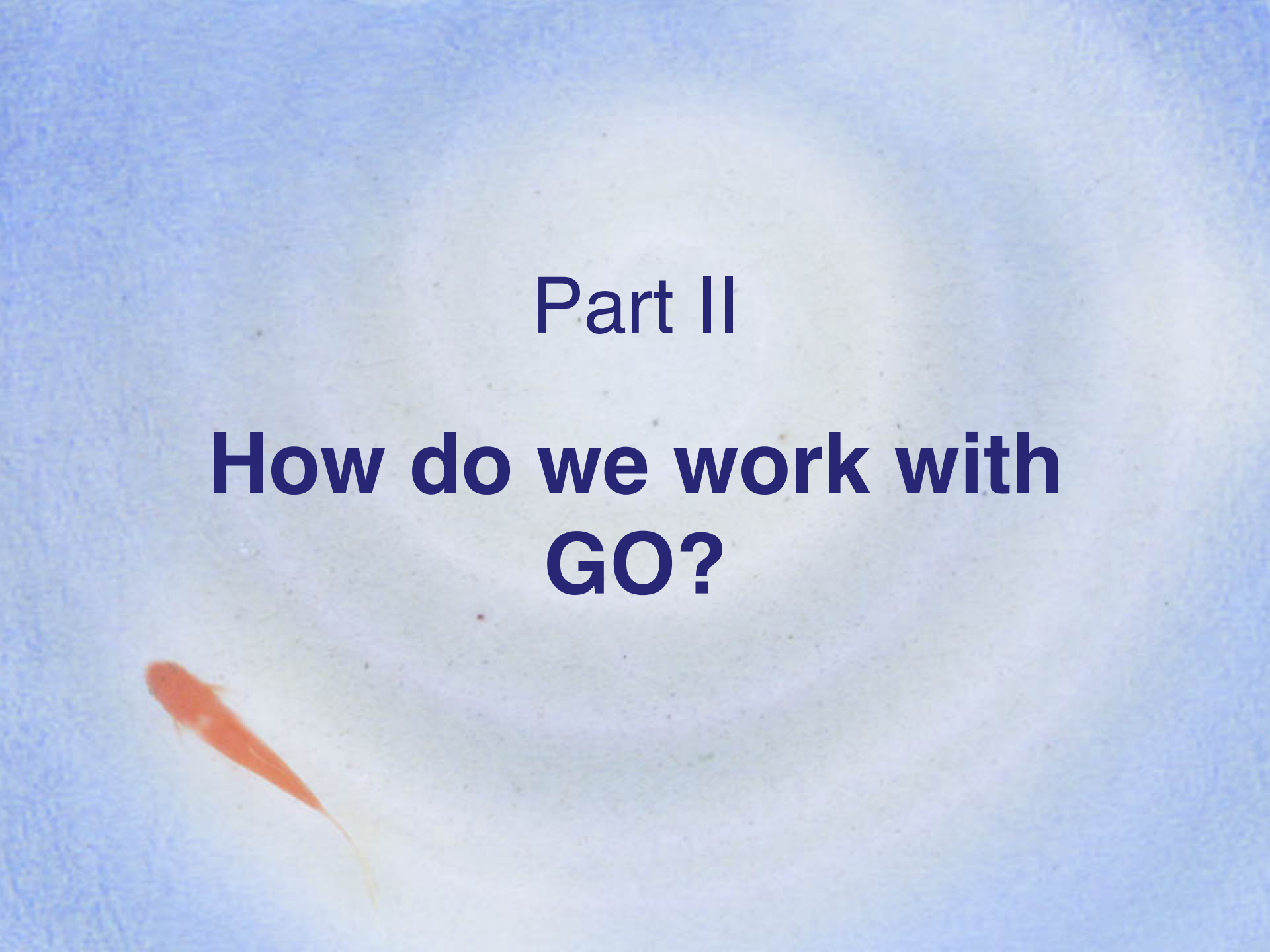

AmiGO: Gene Ontology Browser - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://amigo.geneontology.org/cgi-bin/amigo/go.cgi?action=plus_node&depth=1&search_constraint=ter
The Free Dictionary
Google Search Bookmarks PageRank Check AutoLink AutoFill Send to Settings
AllPeers Home Share Add a Contact Transfers My Torrents Preferences Support Hide Toolbar
An Introduction to the Gene Ontology
EBI Tools - Expression Profiler
GeneInfoViz-Yan Cui's lab at UT
GenMAPP - Home Page
AmiGO: Gene Ontology Browser
the Gene Ontology
AmiGO
Advanced Search BLAST search Browse Help
Search GO
Terms
Genes or proteins
Exact Match
Submit Query
Filter tree view
Filter by ontology
Ontology
All
Biological Process
Cellular Component
Molecular Function
Filter Gene Product Counts
Data source
All
CGD
dictyBase
FlyBase
Set filters
Remove all filters
all : all [477250]
GO:0008150 : biological_process [318388]
GO:0022610 : biological adhesion [3334]
GO:0051825 : adhesion to other organism during symbiotic interaction [186]
GO:0044406 : adhesion to host [186]
GO:0020035 : cytoadherence to microvasculature, mediated by parasite protein [110]
GO:0044401 : multi-species biofilm formation in or on host organism [0]
GO:0044407 : single-species biofilm formation in or on host organism [0]
GO:0052001 : Type IV pili-dependent localized adherence to host [0]
GO:0051856 : adhesion to symbiont [0]
GO:0007155 : cell adhesion [3252]
GO:0022608 : multicellular organism adhesion [28]
GO:0065007 : biological regulation [44424]
GO:0001906 : cell killing [324]
GO:0009987 : cellular process [157428]
GO:0032502 : developmental process [36006]
GO:0051234 : establishment of localization [31042]
GO:0010467 : gene expression [40478]
GO:0040007 : growth [7414]
GO:0002376 : immune system process [4758]
GO:0051179 : localization [35866]
Graphical View
Permalink
Download as XML
Download as flat file
Done
2.656s 132.77.90.62 Disconnected
start
EN
8:51 AM
Sunday
2/24/2008
iTunes
Inbox for ester...
David Givol
Welcome to Ma...
Annotation.pdf...
hgt_genome_H...
ANNOTATION ...
AmiGO: Gene ...

THANKS TO:

- Dr. Esti Feldmesser, for slides, ideas, and encouragement
- GO consortium website
- Nature Genetics Review article (reference given on earlier slides and on the webpage)

Part II

How do we work with GO?



When do we work with GO?

- Trying to make sense of high throughput experiments:
  - Microarrays
  - Deep Sequencing
  - High throughput RNAi
  - Proteomics
  - ….

Functional Genomics: Find the Biological Meaning

- Take a list of "interesting" genes and find their biological meaning
- Requires a reference set of "biological knowledge"
- Linking between genes and biological function:
  - Gene ontology: GO
  - Pathways databases

- All of these techniques give us large gene lists that we have to make sense of
- We have to do some preparation before we can start with GO
- First we have to define the groups we'd like to look at, which isn't a simple task…

Why can't we just use fold change?

- Fold change tells us about individual genes, but does not give us a sense of the "big picture" or the underlying biology
- To see what is changing on a systemic level (system can also be an organelle or a cell…) we have to look at the results as a whole, as best we can

Rankings are notoriously unstable

The scores of 30.000 genes typically form an almost continuous spectrum with little or no outliers.

The difference in score between genes that are several hundred ranks apart are so small that they can not be reproduced

Choosing Gene Lists

- What is your biological question?
- Are you looking at a time course?
- Case/Control?
- Upregulated and Downregulated?
  - Separately or Together?
- Is the fold change important?

Testing Groups of Genes

- Predefined gene groups provide more biological knowledge
- More meaningful interpretation in biological context
- Number of gene sets to be investigated is smaller than number of individual genes
- Useful for validation of published gene groups. Example: Does a gene signature have predictive value?

What is overrepresentation? (enrichment)

- It is a measure of how much a group of gene products is found in our data set
- It requires some type of background measure, as a basis for comparison
- What we look at is how many we have (observed) as opposed to how many we would expect to see at random, given our background.

Background

- The choice of an appropriate background is critical to get meaningful results
- You should use the set used in the experiment, if possible. If its a microarray, then use all the genes on the array, not all the genes in the genome. If its deep sequencing, then all the genes detected by the method you used

Problems working with large data sets

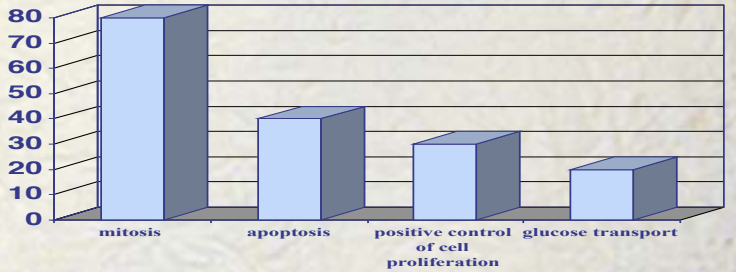
- The more comparisons we make, the more there is a chance that we will get random hits
- We have to correct for the ‘randomness’ factor when we decide if our results are significant or not
- Statistical significance doesn’t necessarily mean biological significance

What methods are used to correct for multiple tests?

- Bonferroni
- Benjamini
- FDR
- ...

Using GO in Practice

How likely is it that your differentially regulated genes fall into a particular category by chance?



microarray
1000 genes
→
experiment
→
100 genes differentially regulated
→
mitosis – 80/100
apoptosis – 40/100
p. ctrl. cell prol. – 30/100
glucose transp. – 20/100

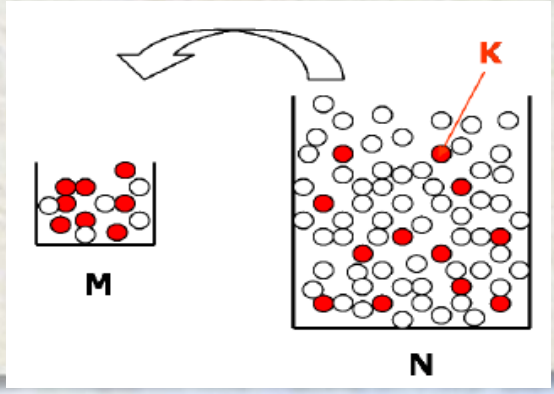
Using GO in Practice

However, when you look at the distribution of all genes on the microarray:

Process	Genes on array	# expected in 100 random genes	occurred
mitosis	800/1000	80	80
apoptosis	400/1000	40	40
p. ctrl. cell prol.	100/1000	10	30
glucose transp.	50/1000	5	20

Group testing: Hypergeometric Test

- Define differentially expressed genes (t-statistic p-values)
- Adjust for multiple testing and choose a cutoff to define a list of interesting genes
- Given N genes on the microarray and M genes in a gene group, what is the probability of having x number of genes from K interesting genes in this group?



Fisher's Exact Test

- The hypergeometric test is equivalent to Fisher's exact test

	$\in$ gene group	$\notin$ gene group	
$\in$ DE genes	x	$K - x$	K
$\notin$ DE genes	$M - x$	$(N - M) - (K - x)$	$N - K$
	M	$N - M$	N

- Fisher-test and similar tests based on gene counts are very often used in Gene Ontology analysis

Fisher's Exact Test

Example:

N = 20000 genes on the microarray
K = 300 differentially expressed genes
M = 100 genes in a gene group of interest

Treatment 1

Diff Expr.	apoptosis	not apoptosis	total
Yes	3	297	300
No	97	19603	19700
	100	19900	20000

could be random
p-value = 0.19

Treatment 2

apoptosis	not apoptosis	total
6	294	300
94	19606	19700
100	19900	20000

not likely random
p-value = 0.004

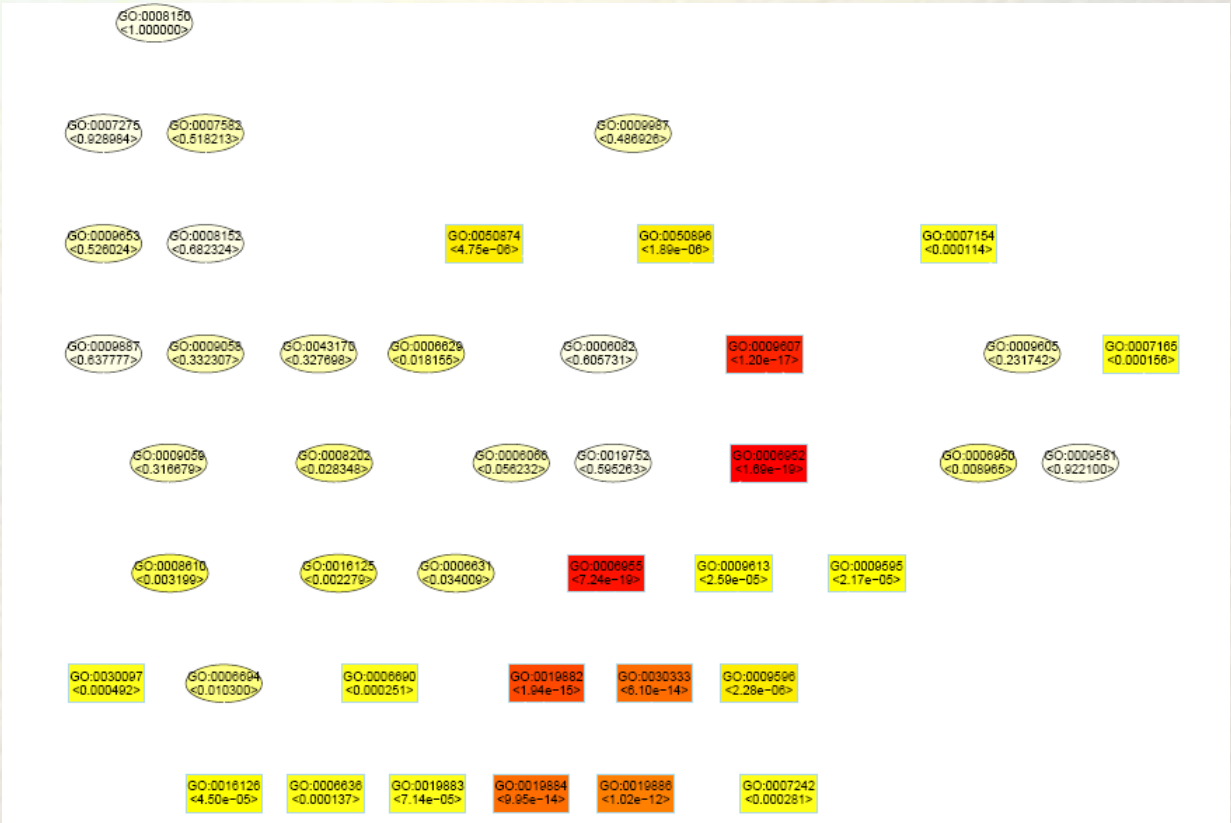
Term-for-term

- The most common type of analysis
- Each term is considered independently of its neighbors in the GO tree
- Compares observed to expected and calculates significance

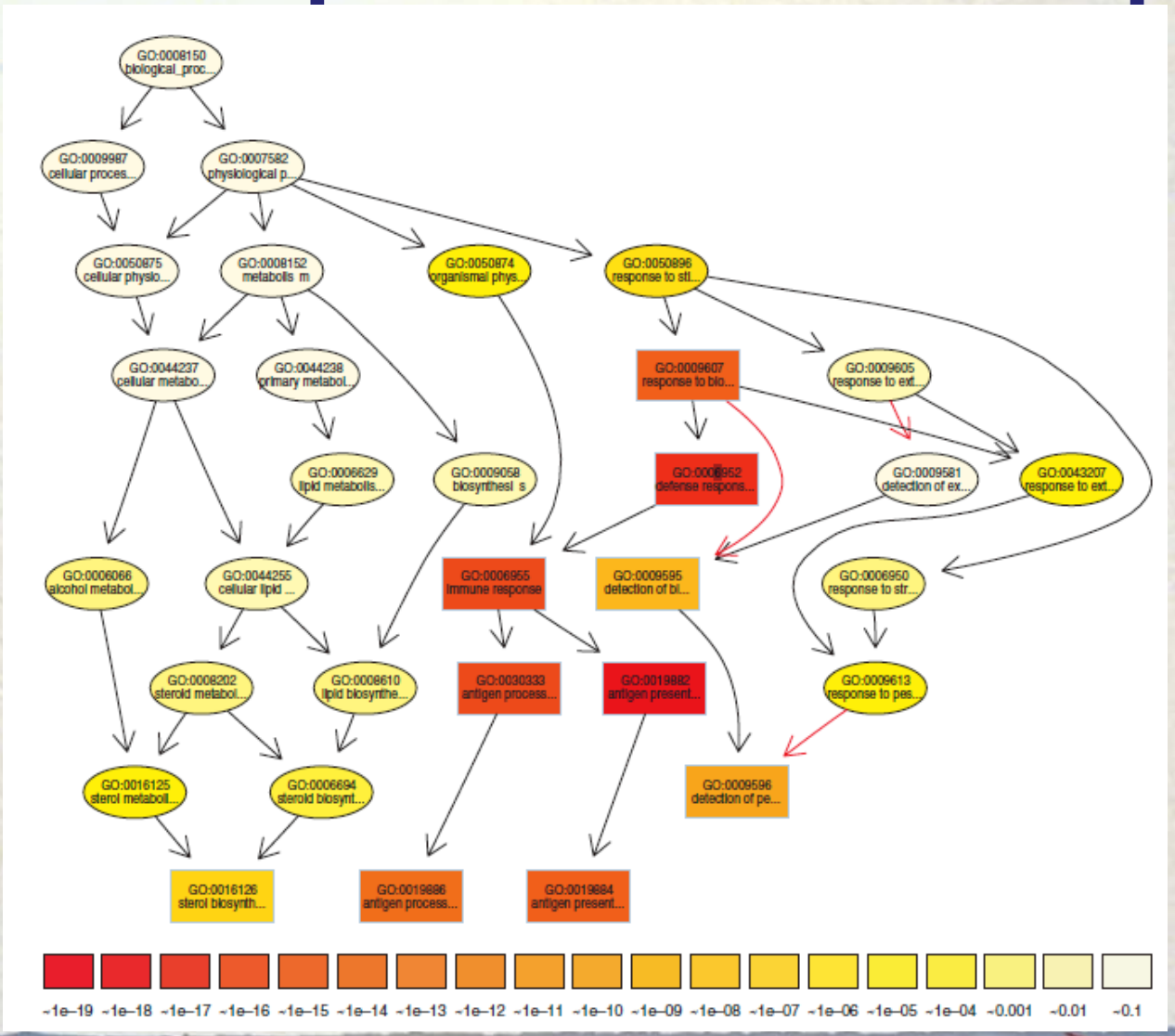
But GO terms aren't really independent

- Due to the true-path rule, each GO term includes all of the terms in all of its descendants
- This causes an over-weighting of the terms closer to the root in the tree

GO Independence Assumption



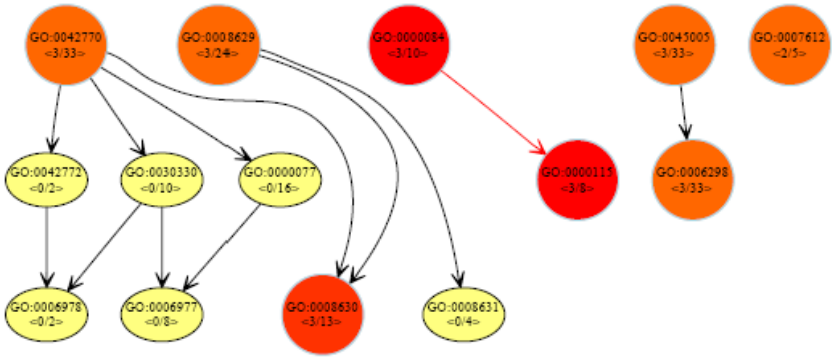
GO Independence Assumption



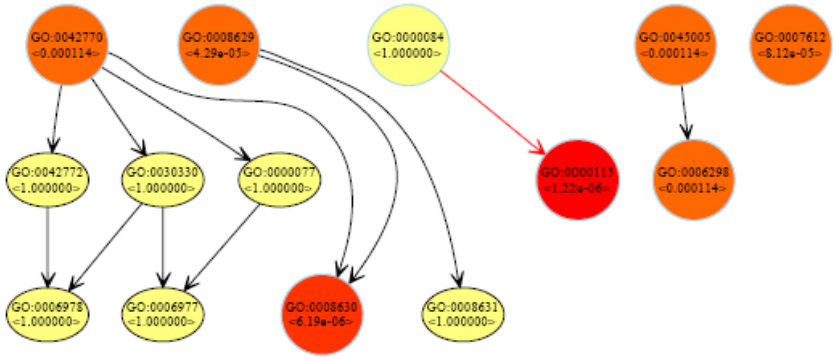
Elim and weight methods

The main idea: Test how enriched node x is if we do not consider the genes from its significant children

1. Perform the analysis from bottom to top. This assures that all children of node x were investigated before node x itself
2. The p-value for the children of node x is computed using Fisher's exact test
3. If one of the children of the node x is found significant, we remove all the genes mapped to this node, from x (elim)
4. If one of the children of the node x is found significant, we lower the weight of the genes that are common to both on the higher level (weight)



classic method



elim method

The elim methods tends to be a bit too stringent, unless you want to know the most significant nodes – in which case it cleans up most of the background

For most purposes, the weight method is better

Algorithm review

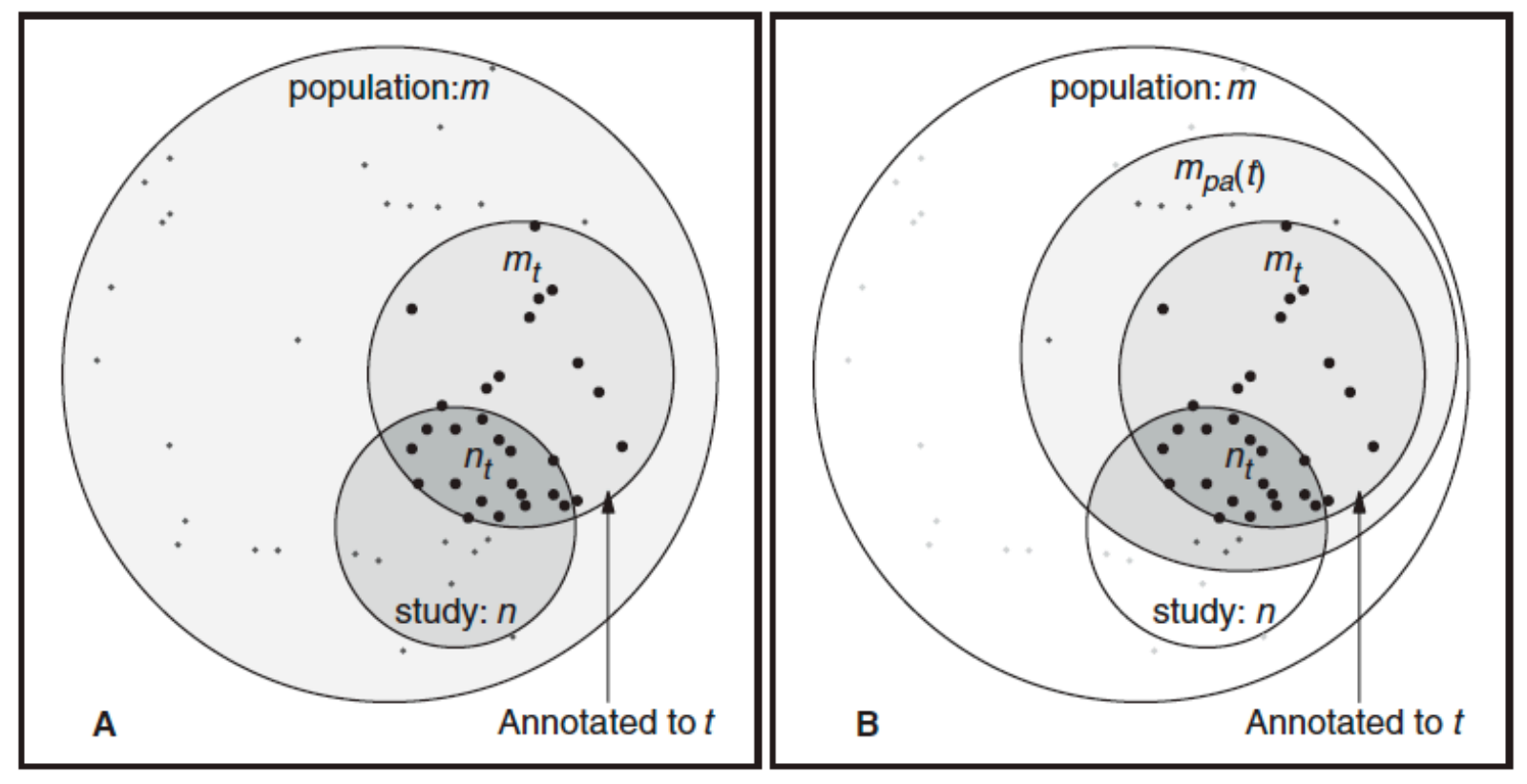
- classic **algorithm**
  - Calculate significance of each GO term independently.
  - Adjust pvalues for multiple testing (Bonferroni, FDR, etc.).
  - Kolmogorov-Smirnov test can easily be used in this case
- elim **algorithm**
  - Nodes are processed bottom-up in the GO graph.
  - It iteratively removes the genes annotated to significant GO terms from more general GO terms.
  - Intuitive and simple to interpret.
- weight **algorithm**
  - The genes obtain weights that denote the gene relevance in the significant nodes.
  - To decide if a GO term u better represents the interesting genes, the enrichment score of node u is compared with the scores of its children.
  - Children with a better score than u better represent the interesting genes; their significance is increased
  - Children with a lower score than u have their significance reduced.

Alexa A, Rahnenführer J, Lengauer T. Bioinformatics. 2006 Jul 1;22(13):1600-7

Parent-Child Methods (intersection and union)

- This is another attempt to correct the dependencies of GO terms
- It takes into account the parents of the term (as opposed to all the neighbors)
- It can clean up downwards, not just upwards (when the descendants have almost the same genes as the parents)

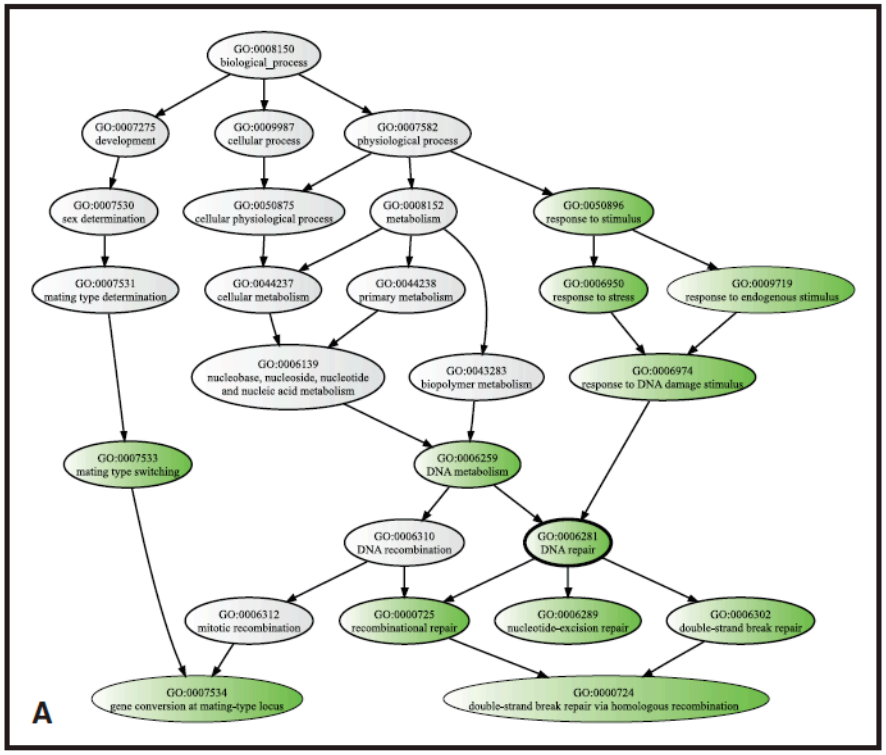
Parent-Child

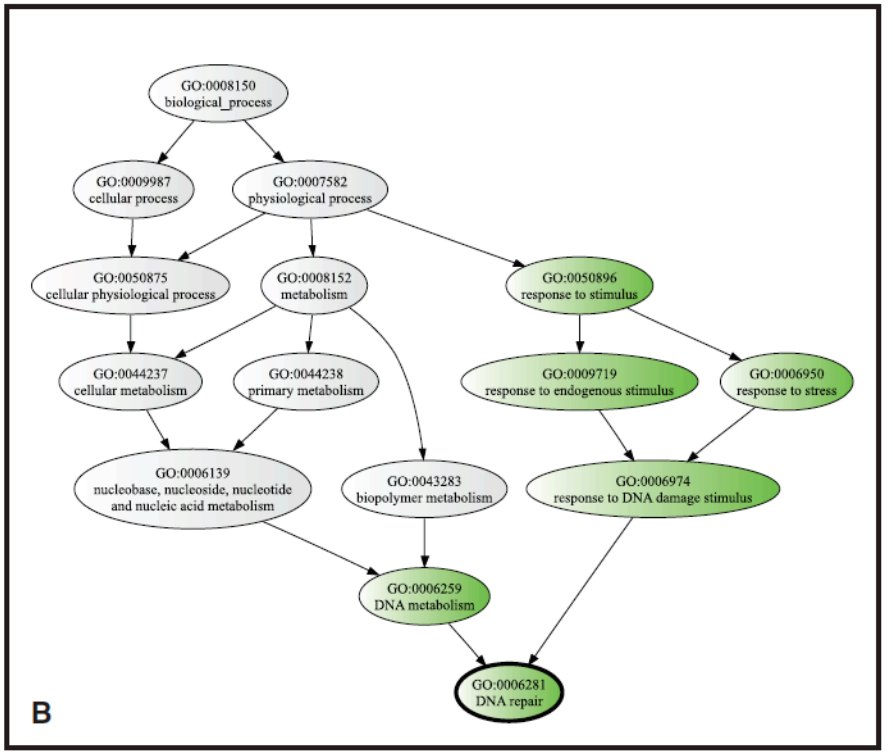


A) term-for-term B) parent-child intersection

The parent-child methods measure overrepresentation of a term *t* in the context of annotations to the parents of the term

Parent-child intersection method





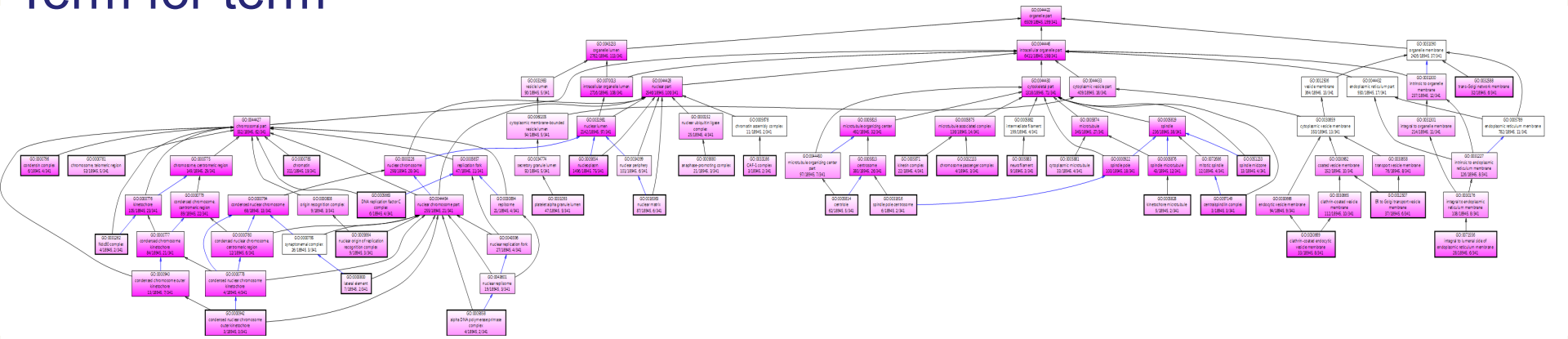
Artificial overrepresentation of the GO term DNA repair.

A) term-for-term

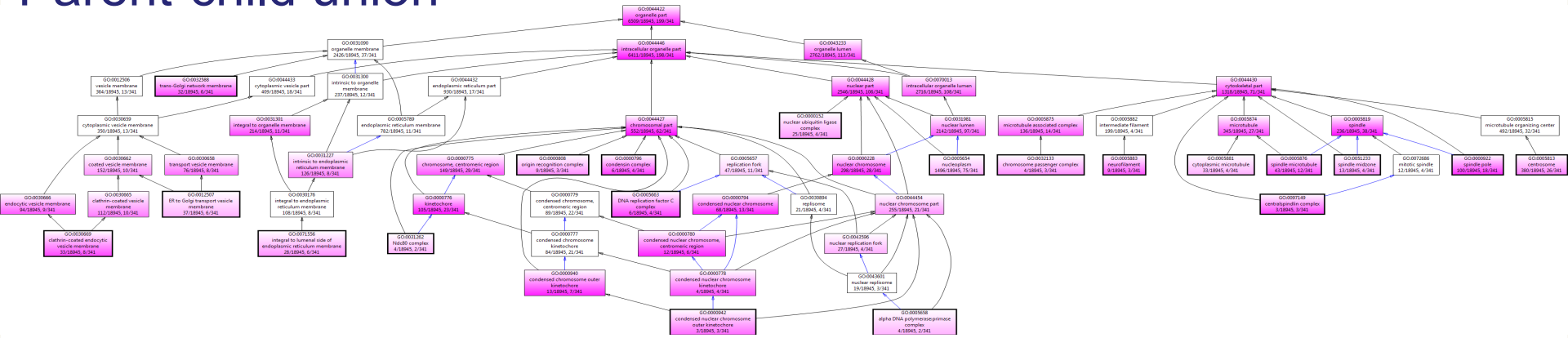
B) parent-child-intersection

Grossmann S, Bauer S, Robinson PN, Vingron M. Bioinformatics. 2007 Nov 15;23(22):3024-31

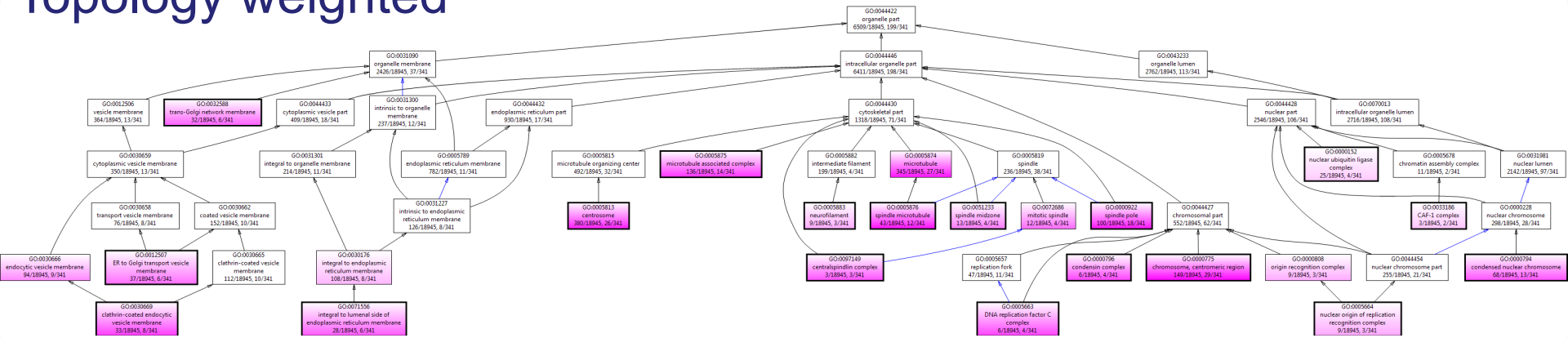
Term for term



Parent-child union



Topology weighted



GO analysis summary

Advantages

- If you just get a list of somehow interesting genes and want to assess biological background, tests based on gene counts are the only way to go

Problems

- Loss of information because of two separated steps
- Small but consistent differential expression is not detected
- Dividing genes into differentially and non-differentially expressed genes is artificial
- No clear way of defining x: p-value correction and choice of a cutoff are crucial

So what method should I use?

- First of all, it depends on your biological question
- Because there are issues with random hits, its always better to run more than one and compare – what remains consistent (even if the p-value changes) is more likely to be biologically relevant

What program should I use?

- As they are all built a bit differently, we recommend running more than one, and once again, comparing the results.
- Some of our favorites are:
  - Ontologizer
  - WebGestalt
  - DAVID

Keep in mind

- The final results are completely dependent on the initial choices of dataset and background!!
- GO is incomplete
- Not all programs will recognize all identifiers
- Weakly changing genes may not be statistically significant, but may be highly biologically significant

Keep in mind

- It is hard to differentiate between primary and secondary effect
- Size is important. The bigger the input gene list, the more of a chance that you'll get something significant. For smaller groups of genes significance will be lower. The same is true for GO – if you can search against part of the tree, instead of all (particular levels for example) it will increase significance, because it cuts down the number of comparisons (Slim, Fat)

Thanks...

- once again, to Dr. Esti Feldmesser for slides and help!