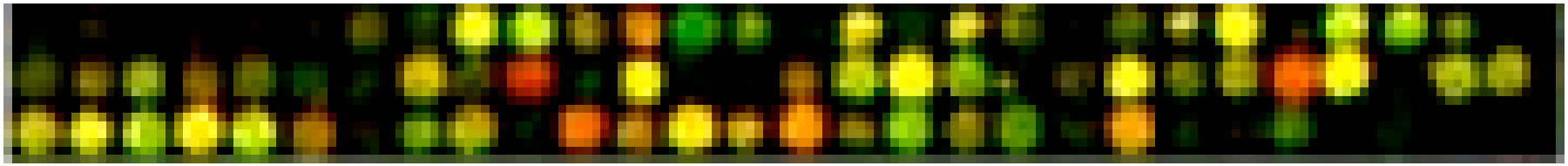
GO::TermFinder



Gavin Sherlock

Department of Genetics

Stanford University

sherlock@genome.stanford.edu

GO::TermFinder includes:

- A way to determine statistically significant GO terms shared by a set of genes
- A module to visualize the results
- A set of software modules to access GO information

Inspiration…

 letter

Systematic determination of genetic network architecture

Saeed Tavazoie[1], Jason D. Hughes[1,2], Michael J. Campbell[3], Raymond J. Cho[4] & George M. Church[1]

- Tavazoie et al, 1999, used the hypergeometric distribution to determine enrichment of MIPS categories in clusters of cell cycle regulated genes.

Hypergeometric Distribution:

We can calculate the probability of observing x of n events as having a particular property, given that in the general population, M of N things have that property, using the ***hypergeometric distribution***, as:

$$P = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{x}}$$

Where, generically

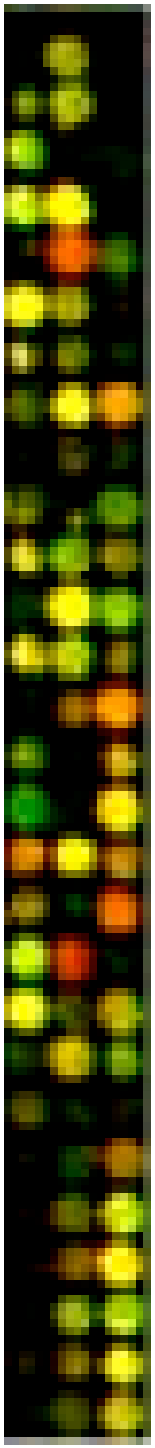
$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

which is the number of permutations by which r ‘things’ can be chosen from a set of n ‘things’.

Calculating a P-value

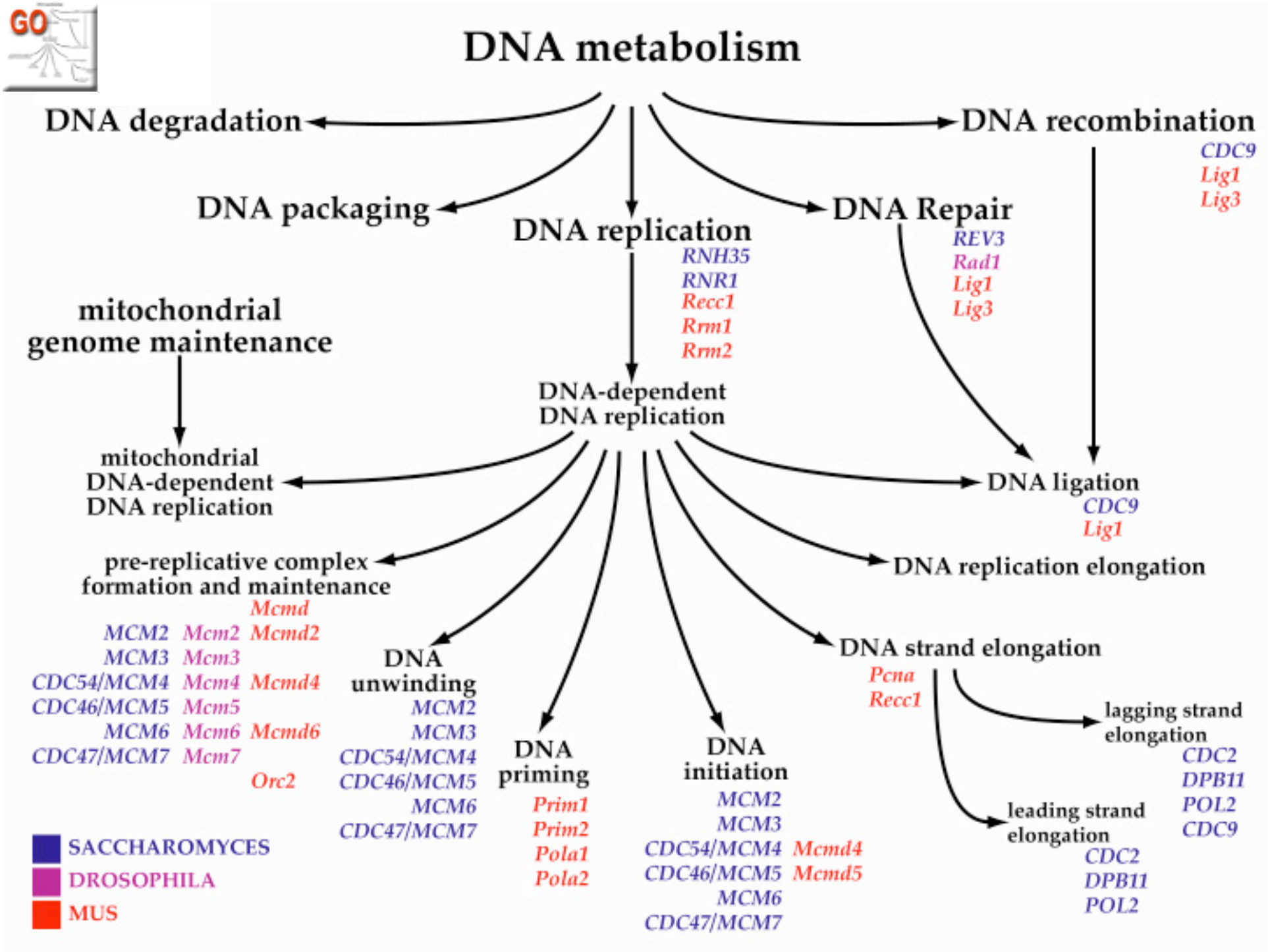
To calculate a P-value, we calculate the probability of having ***at least*** *x* of *n* events:

$$P\_value = 1 - \sum_{i=0}^{x-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{i}}$$



Translating this to GO

- Many analyses result in a list of interesting genes
- Typically biologists can make up a story about any random list
- Look at all GO annotations for the genes in a list, and see if the number of annotations for any is significant


DNA metabolism
DNA degradation
DNA packaging
DNA replication
RNH35
RNR1
Recc1
Rrm1
Rrm2
DNA Repair
REV3
Rad1
Lig1
Lig3
DNA recombination
CDC9
Lig1
Lig3
mitochondrial genome maintenance
mitochondrial DNA-dependent DNA replication
DNA-dependent DNA replication
DNA ligation
CDC9
Lig1
DNA replication elongation
pre-replicative complex formation and maintenance
Mcmd
MCM2 Mcm2 Mcmd2
MCM3 Mcm3
CDC54/MCM4 Mcm4 Mcmd4
CDC46/MCM5 Mcm5
MCM6 Mcm6 Mcmd6
CDC47/MCM7 Mcm7
Orc2
DNA unwinding
MCM2
MCM3
CDC54/MCM4
CDC46/MCM5
MCM6
CDC47/MCM7
DNA priming
Prim1
Prim2
Pola1
Pola2
DNA initiation
MCM2
MCM3
CDC54/MCM4 Mcmd4
CDC46/MCM5 Mcmd5
MCM6
CDC47/MCM7
DNA strand elongation
Pcna
Recc1
lagging strand elongation
CDC2
DPB11
POL2
CDC9
leading strand elongation
CDC2
DPB11
POL2
SACCHAROMYCES
DROSOPHILA
MUS

Multiple hypothesis correction

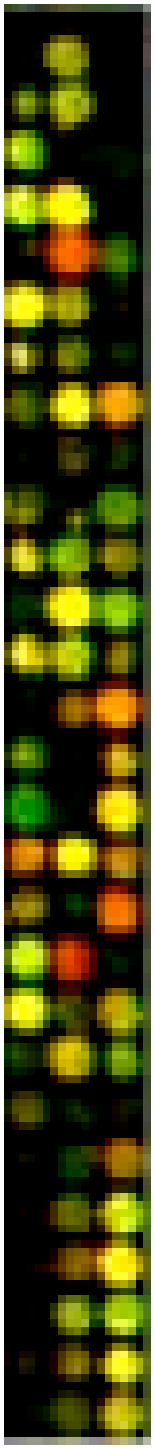
- If we choose a P-value cutoff of 0.05, we have a 1 in 20 chance of falsely picking something as significant that is not.
- If we test multiple hypotheses (GO nodes), each one has a 1 in 20 chance of being wrong. Thus if we test 10 nodes, we have a 0.4 chance of falsely picking one as significant.

Multiple hypothesis correction (continued)

- Correct for multiple hypotheses to keep the overall chance of picking a false positive at 1 in 20.
- Bonferroni correction simply divides the alpha value by the number of hypotheses - assumes independence, which is not the case for our GO nodes.

Correction (continued)

- Use simulation to determine a p-value.
- Turns out Bonferroni is not conservative enough
- Why?
- Should all nodes be corrected equally?

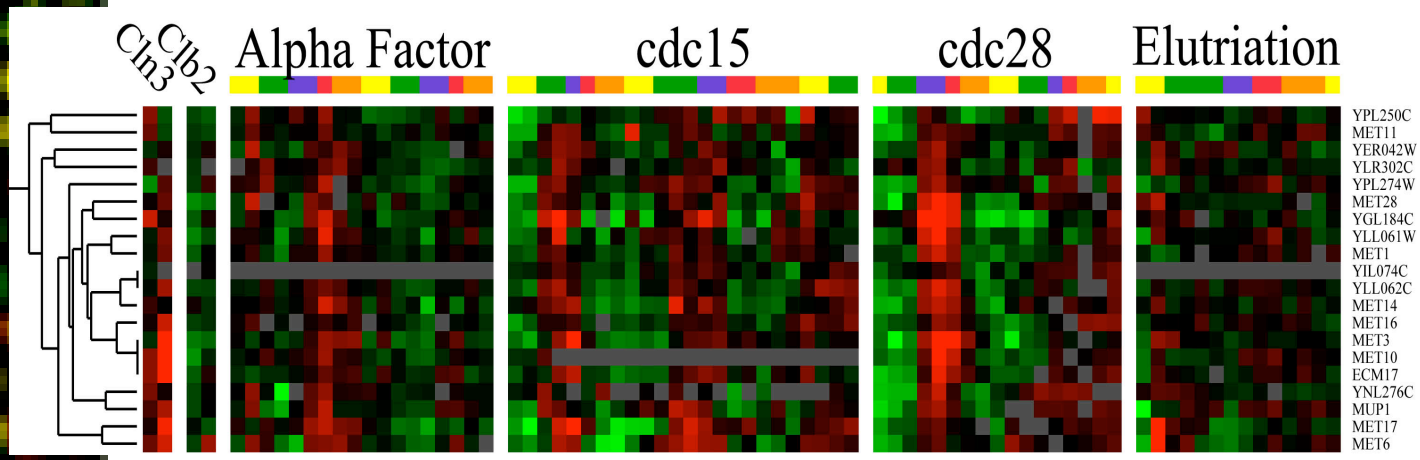


Using GO::TermFinder to look at microarray data

- Our general assumption is guilt by association:

  i.e. genes with similar expression patterns are more likely to participate in the same biological process.
- So let's take this assumption and exploit the Gene Ontology to examine our expression clusters:

Methionine Cluster



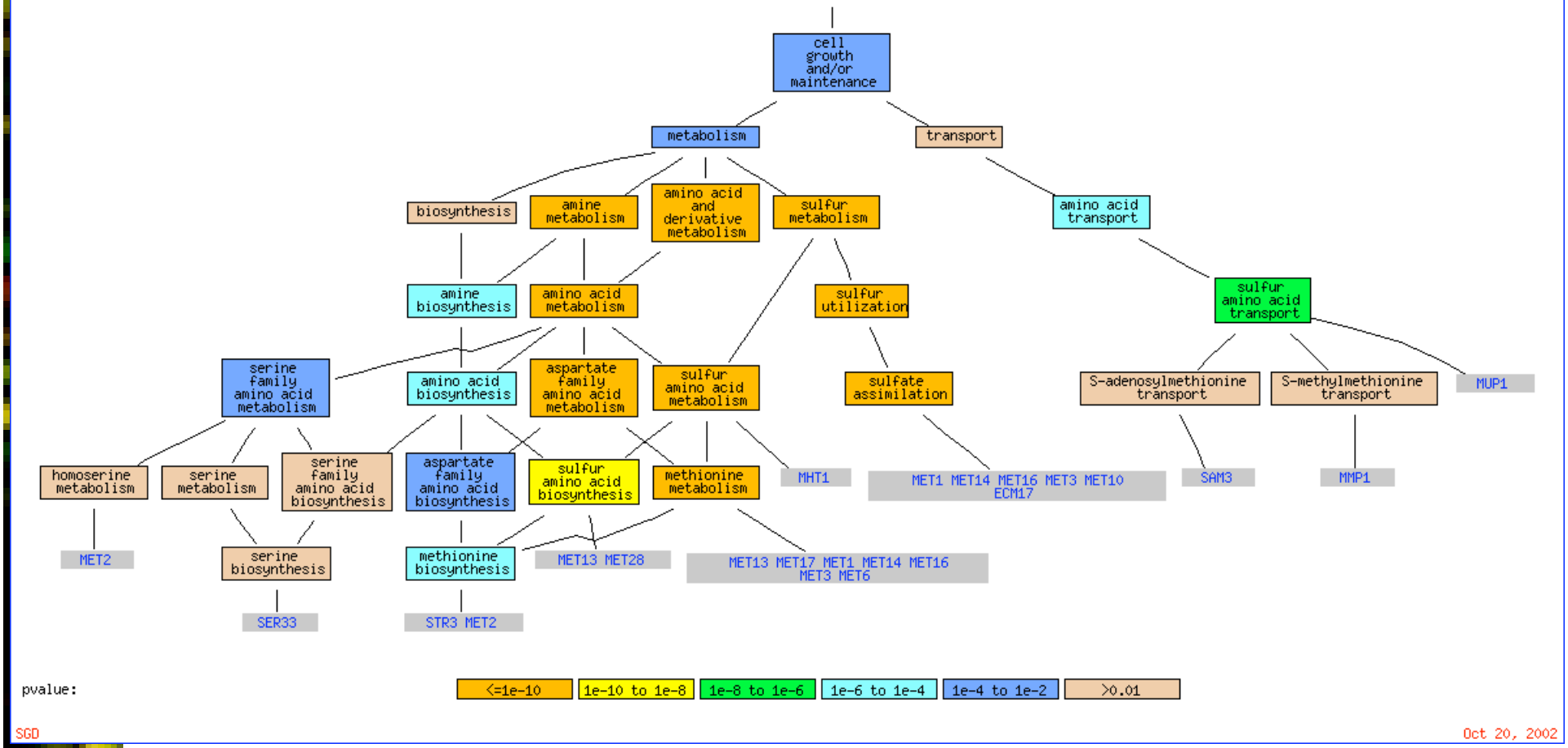
YPL250C
MET11
YER042W
YLR302C
YPL274W
MET28
YGL184C
YLL061W
MET1
YIL074C
YLL062C
MET14
MET16
MET3
MET10
ECM17
YNL276C
MUP1
MET17
MET6

Visualization module

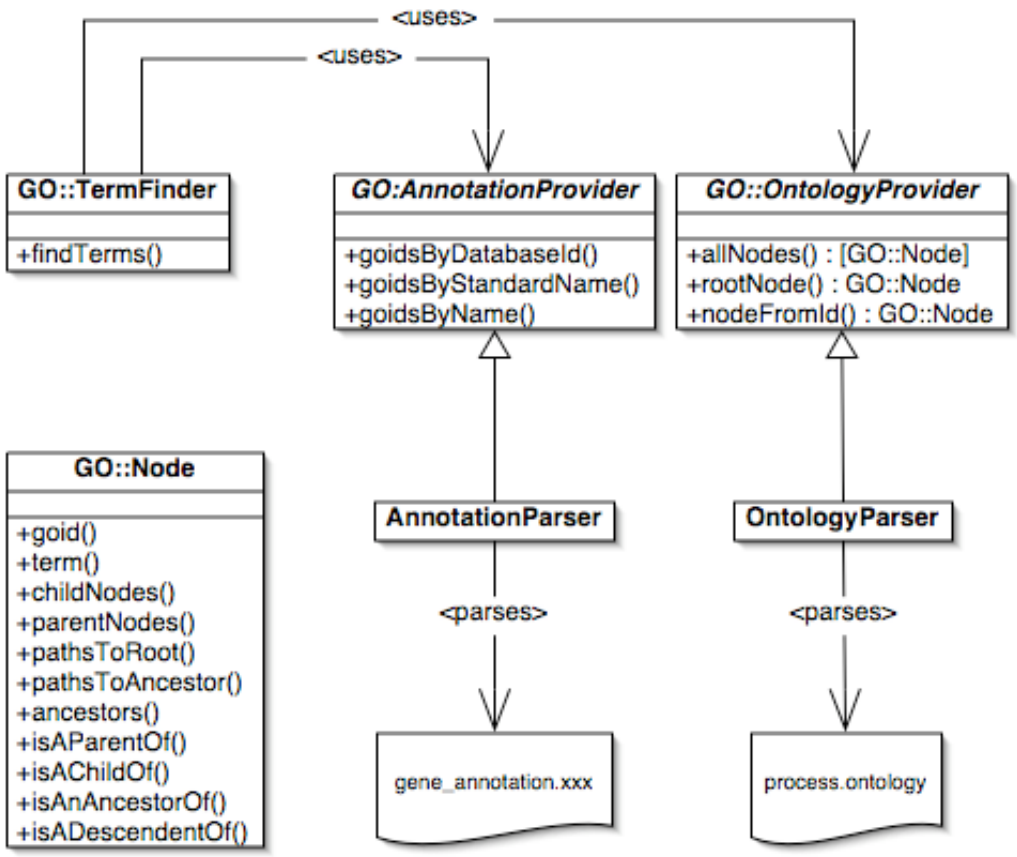
- Developed by SGD
- Takes output from GO::TermFinder
- Uses Perl interface to AT & T GraphViz tool for graph layout

GO Annotations

- sulfur metabolism : 1.77e-26 (13/19 vs 33/6911)
- methionine metabolism : 8.08e-19 (9/19 vs 19/6911)

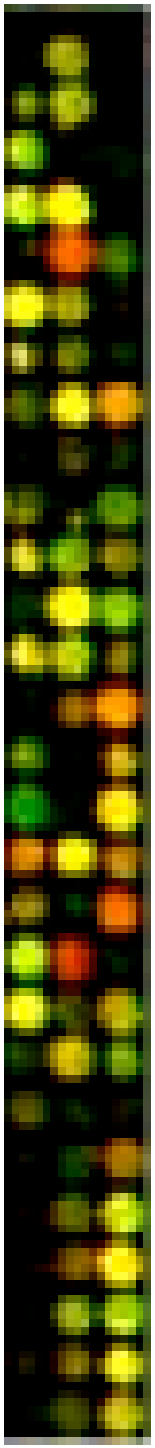


The API


<uses>
<uses>
GO::TermFinder
+findTerms()
GO:AnnotationProvider
+goidsByDatabaseId()
+goidsByStandardName()
+goidsByName()
GO::OntologyProvider
+allNodes() : [GO::Node]
+rootNode() : GO::Node
+nodeFromId() : GO::Node
GO::Node
+goid()
+term()
+childNodes()
+parentNodes()
+pathsToRoot()
+pathsToAncestor()
+ancestors()
+isAParentOf()
+isAChildOf()
+isAAncestorOf()
+isADescendentOf()
AnnotationParser
OntologyParser
<parses>
<parses>
gene_annotation.xxx
process.ontology

Included example tools

- ancestors.pl
- children.pl
- termFinderClient.pl
- analyze.pl
- batchGOView.pl



Acknowledgements

- Ellie Boyle
- Shuai Weng
- Jeremy Gollub
- Heng Jin
- Mike Cherry
- David Botstein

GO::TermFinder URL

- Full source code available under the MIT license from:

  http://search.cpan.org/dist/GO-TermFinder/