Getting to know Blast2GO

Functional annotation: from sequences to functional labels

Outline

Concepts on Functional Annotation:

Biological Databases

Blast2GO annotation strategy

The Blast2GO annotation framework:

Annotation steps, Modulation of annotation intensity, Export/Import Functions, Sequence Selection, Additional Tools

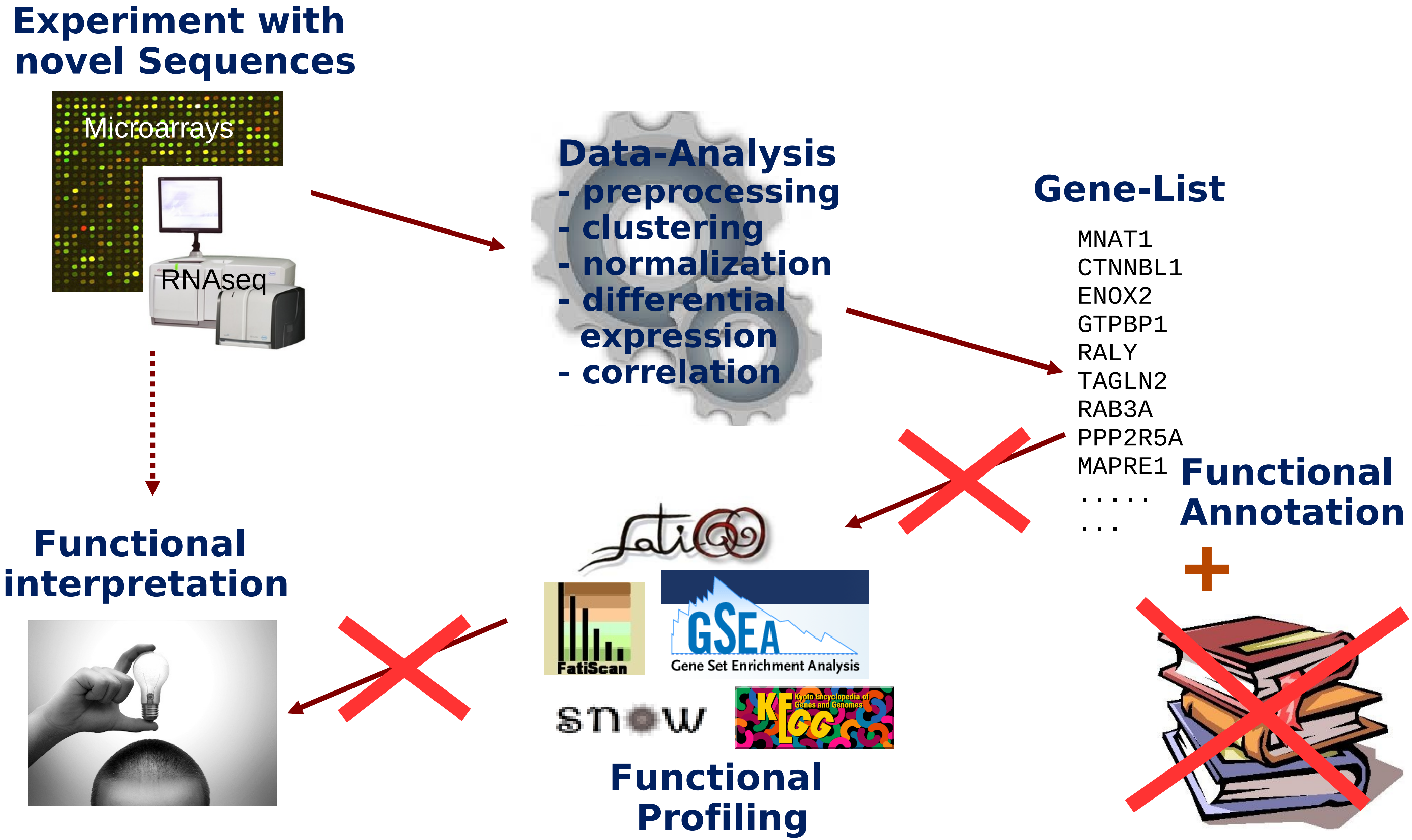
---

Data Mining: Visualization Techniques

Combined Graph, Charts and Pies

Hands on: Blast2GO annotation excercise

Why Blast2GO?



Experiment with novel Sequences

Data-Analysis

- **preprocessing**
- **clustering**
- **normalization**
- **differential expression**
- **correlation**

Gene-List

```
MNAT1
CTNNBL1
ENOX2
GTPBP1
RALY
TAGLN2
RAB3A
PPP2R5A
MAPRE1
.....
...
```

Functional Annotation

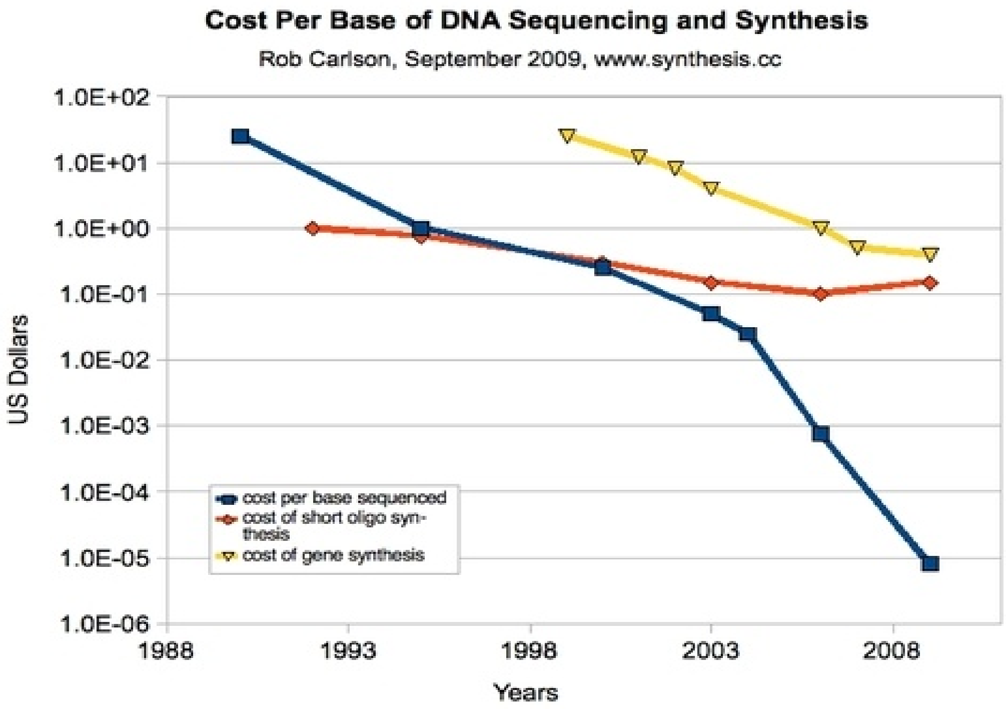
+

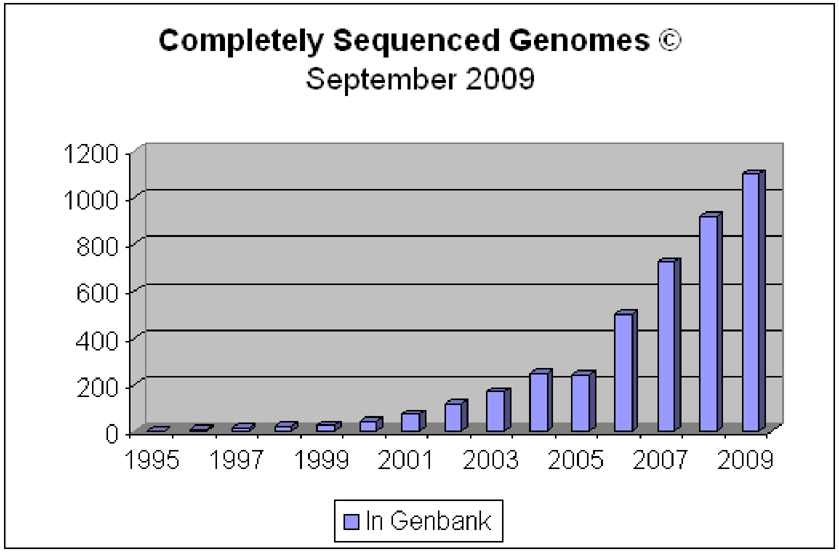
Functional Profiling

FatiGO
FatiScan
GSEA
Gene Set Enrichment Analysis
snow
KEGG
Kyoto Encyclopedia of Genes and Genomes

Functional interpretation

Why automated functional annotation?

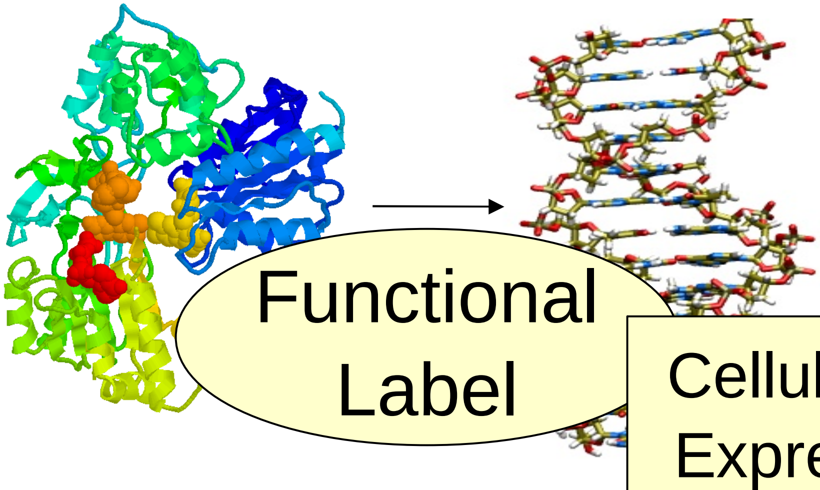




We can not characterize experimentally all these sequence data at the current growth rate:
We need automated annotation tools to achieve an equivalent throughput

What is functional annotation?

The function is on the protein



But frequently we annotate nt-sequences

Functional Label

Cellular Rol
Expression
Interactions
.....

Controlled Vocabulary

High throughput

Accessible

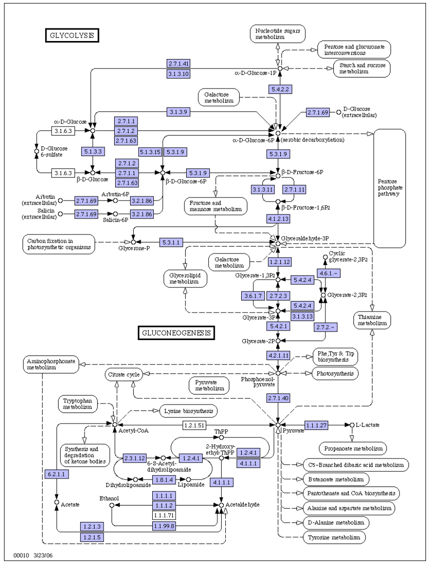
Functional Vocabularies

the Gene Ontology

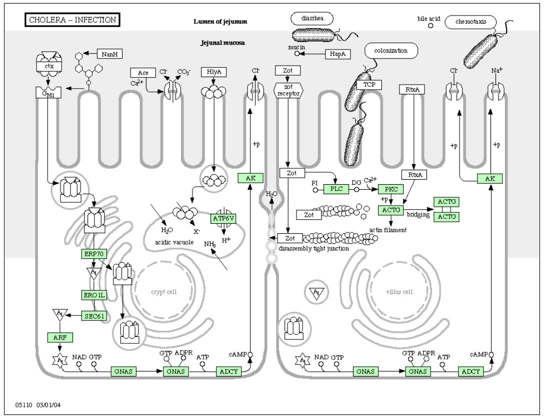
Molecular Function
Biological Process
Cellular Component

KEGG Kyoto Encyclopedia of Genes and Genomes

Metabolic pathways

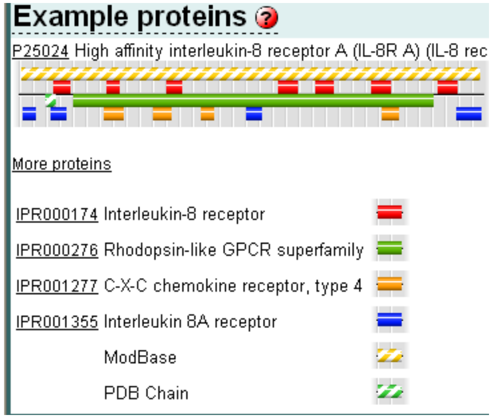


KEGG orthologues



InterPro prosite

Functional motifs



Example proteins

P25024 High affinity interleukin-8 receptor A (IL-8R A) (IL-8 rec

More proteins

IPR000174	Interleukin-8 receptor	🟥
IPR000276	Rhodopsin-like GPCR superfamily	🟩
IPR001277	C-X-C chemokine receptor, type 4	🟧
IPR001355	Interleukin 8A receptor	🟦
	ModBase	▨
	PDB Chain	▨

The Gene Ontology

✓ Project developed by the Gene Ontology Consortium

✓ Provides a controlled vocabulary to describe gene and gene product attributes in any organism

✓ Lastest version more than 22.000 terms

✓ Includes both the development of the Ontology and the maintenance of a Database of annotations

http://www.geneontology.org

Gene Ontology

The three categories of GO

Molecular Function

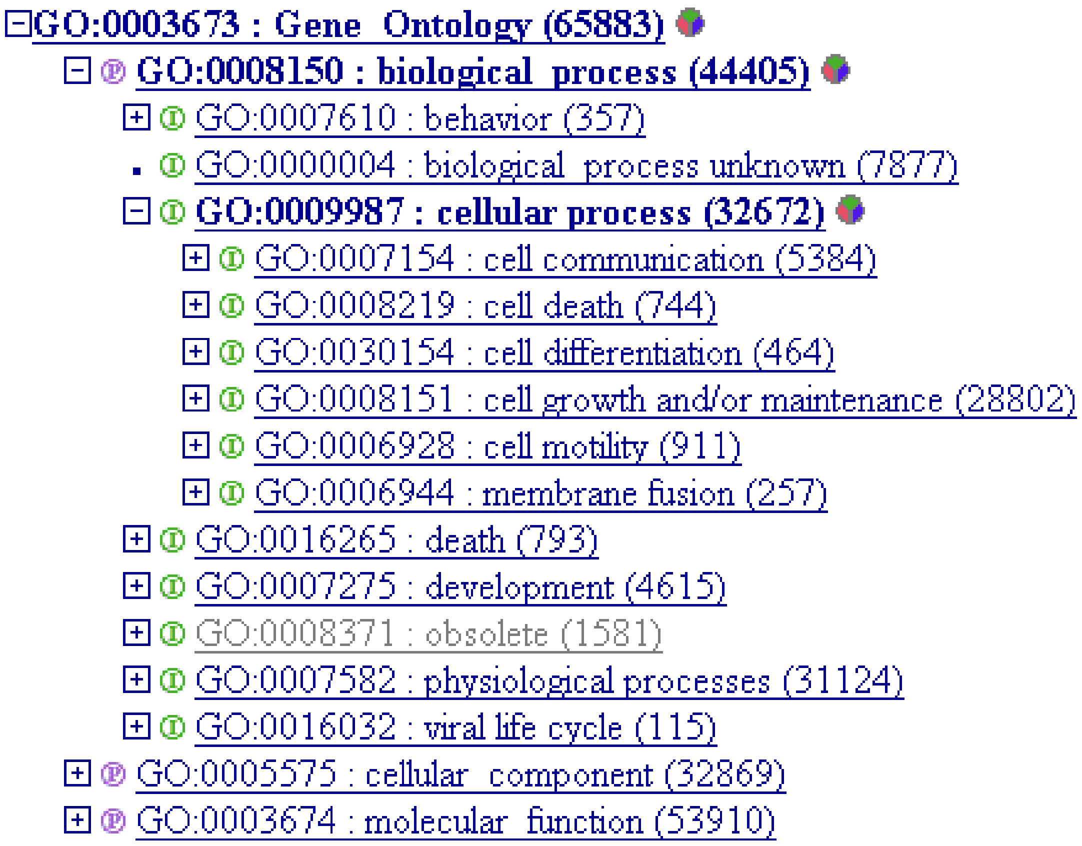
the tasks performed by individual gene products; examples are *transcription factor* and *DNA helicase*

Biological Process

broad biological goals, such as *mitosis* or *purine metabolism*, that are accomplished by ordered assemblies of molecular functions

Cellular Component

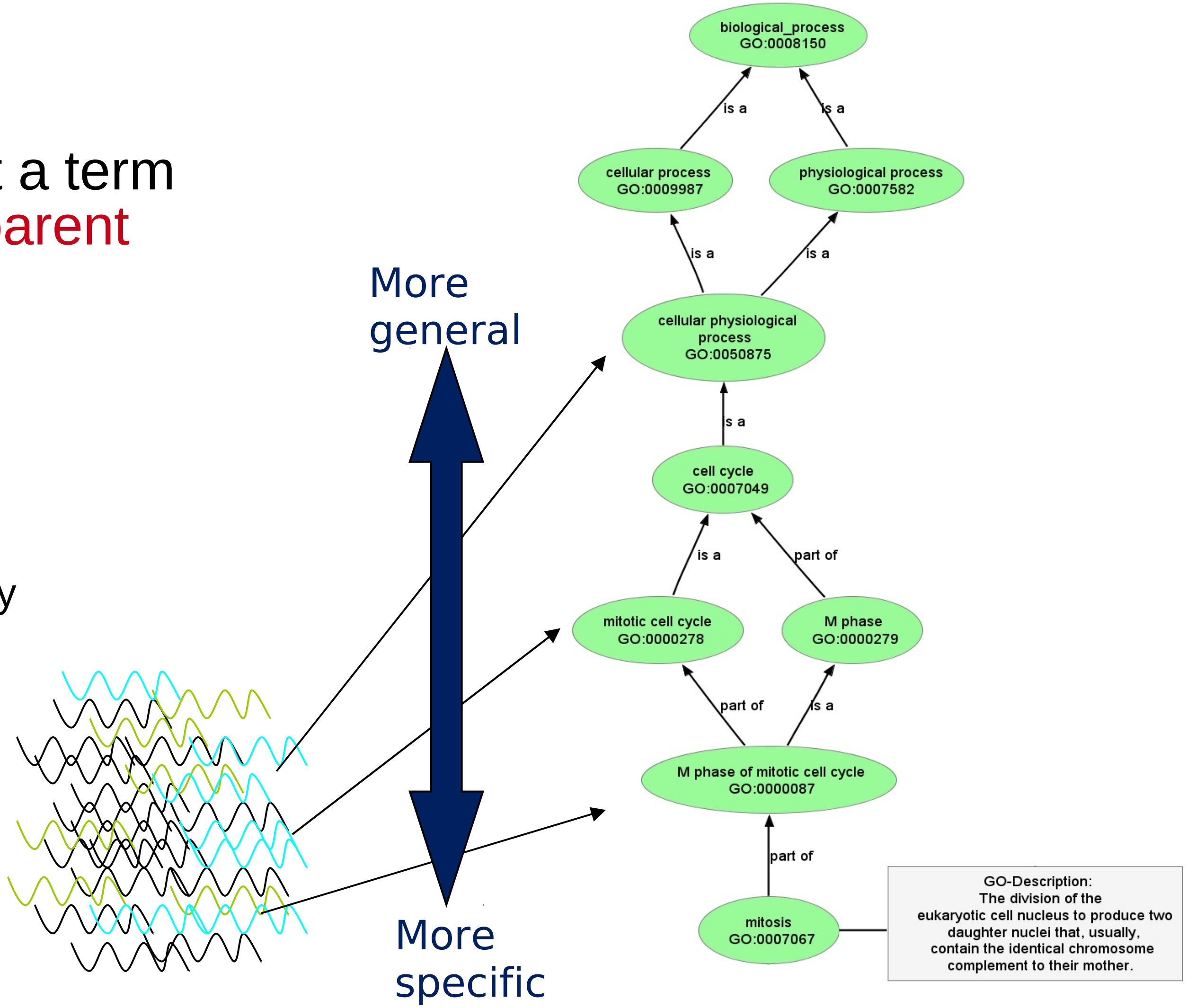
subcellular structures, locations, and macromolecular complexes; examples include *nucleus*, *telomere*, and *origin recognition complex*



- ⊟ **GO:0003673 : Gene Ontology (65883)**
  - ⊟ Ⓟ **GO:0008150 : biological process (44405)**
    - ⊞ Ⓘ GO:0007610 : behavior (357)
    - ▪ Ⓘ GO:0000004 : biological process unknown (7877)
    - ⊟ Ⓘ **GO:0009987 : cellular process (32672)**
      - ⊞ Ⓘ GO:0007154 : cell communication (5384)
      - ⊞ Ⓘ GO:0008219 : cell death (744)
      - ⊞ Ⓘ GO:0030154 : cell differentiation (464)
      - ⊞ Ⓘ GO:0008151 : cell growth and/or maintenance (28802)
      - ⊞ Ⓘ GO:0006928 : cell motility (911)
      - ⊞ Ⓘ GO:0006944 : membrane fusion (257)
    - ⊞ Ⓘ GO:0016265 : death (793)
    - ⊞ Ⓘ GO:0007275 : development (4615)
    - ⊞ Ⓘ GO:0008371 : obsolete (1581)
    - ⊞ Ⓘ GO:0007582 : physiological processes (31124)
    - ⊞ Ⓘ GO:0016032 : viral life cycle (115)
  - ⊞ Ⓟ GO:0005575 : cellular component (32869)
  - ⊞ Ⓟ GO:0003674 : molecular function (53910)

The Ontology

- Annotations are given to te most specific (low) level
- True path rule: annotation at a term implies annotation to all its parent terms
- Annotation is given with an Evidence Code:
  - **IDA**: inferred by direct assay
  - **TAS**: traceable author statement
  - **ISS**: infered by sequence similarity
  - **IEA**: electronic annotation
  - ….



More general
More specific
biological_process GO:0008150
cellular process GO:0009987
physiological process GO:0007582
cellular physiological process GO:0050875
cell cycle GO:0007049
mitotic cell cycle GO:0000278
M phase GO:0000279
M phase of mitotic cell cycle GO:0000087
mitosis GO:0007067
is a
part of

GO-Description:
The division of the eukaryotic cell nucleus to produce two daughter nuclei that, usually, contain the identical chromosome complement to their mother.

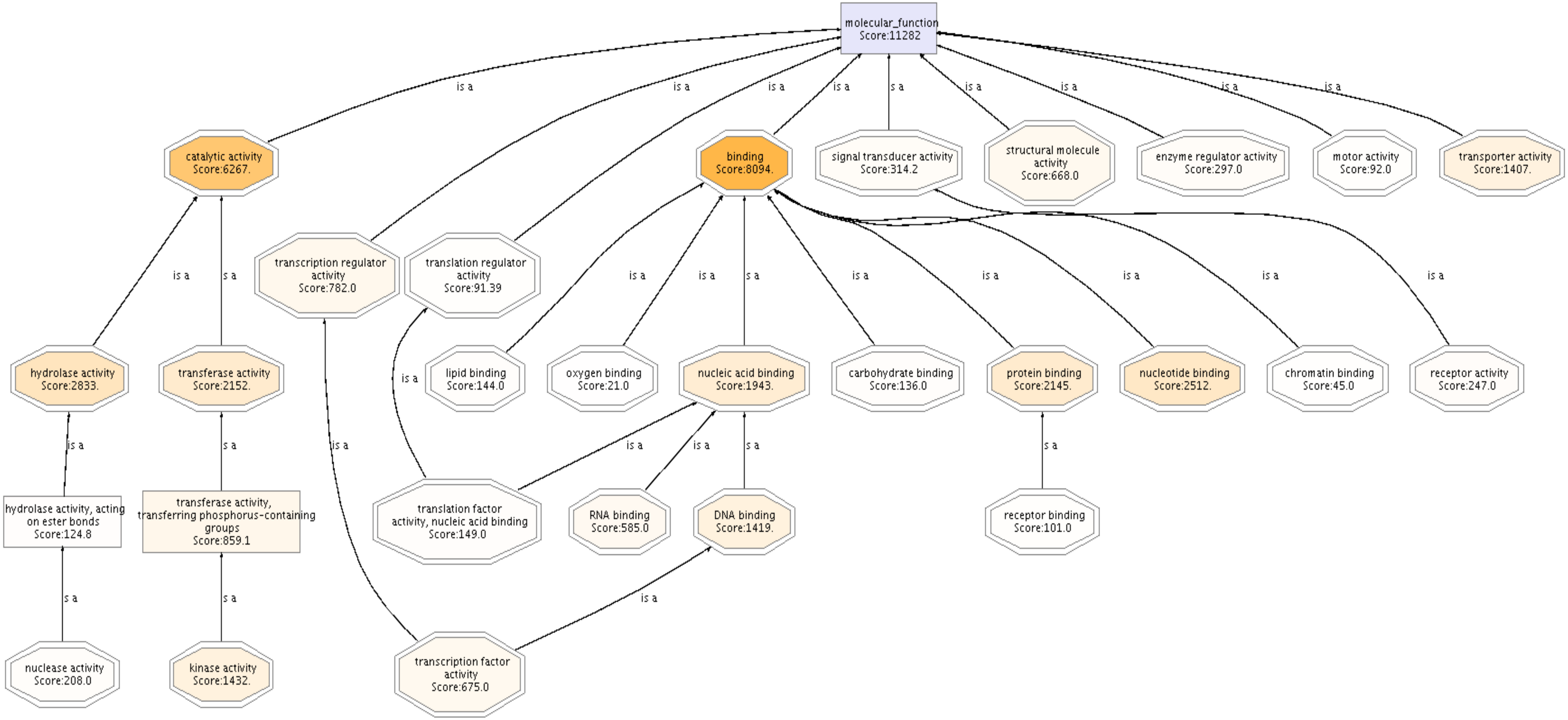
The Gene Ontology Database (GOA)

http://www.geneontology.org/GO.current.annotations.shtml

- There is a collaborating institution per organism to provide annotations
- Most of the GOA annotations come from UniProt
- Most of the annotations are electronic annotations

Some numbers of the actual version (jan11):
Terms: 31.794, mf(8.912), bp(20.092), cc(2.790)
Genes: 8.864.425
Annotations: 52.571.310

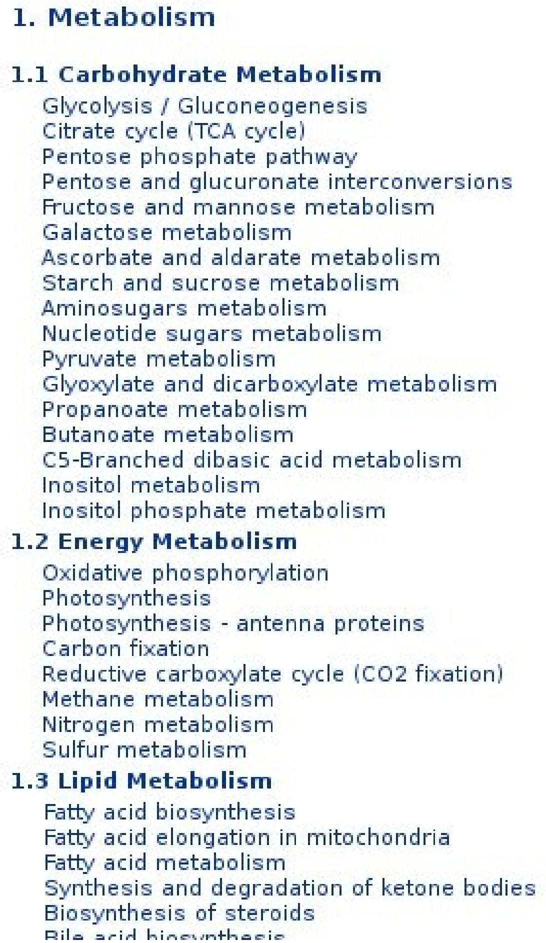
The GO has a DAG structure



molecular_function Score:11282
is a
is a
is a
is a
s a
is a
is a
is a
is a
catalytic activity Score:6267.
binding Score:8094.
signal transducer activity Score:314.2
structural molecule activity Score:668.0
enzyme regulator activity Score:297.0
motor activity Score:92.0
transporter activity Score:1407.
transcription regulator activity Score:782.0
translation regulator activity Score:91.39
is a
s a
is a
is a
s a
is a
is a
is a
is a
is a
hydrolase activity Score:2833.
transferase activity Score:2152.
is a
lipid binding Score:144.0
oxygen binding Score:21.0
nucleic acid binding Score:1943.
carbohydrate binding Score:136.0
protein binding Score:2145.
nucleotide binding Score:2512.
chromatin binding Score:45.0
receptor activity Score:247.0
is a
s a
is a
is a
is a
s a
s a
hydrolase activity, acting on ester bonds Score:124.8
transferase activity, transferring phosphorus-containing groups Score:859.1
translation factor activity, nucleic acid binding Score:149.0
RNA binding Score:585.0
DNA binding Score:1419.
receptor binding Score:101.0
s a
s a
is a
nuclease activity Score:208.0
kinase activity Score:1432.
transcription factor activity Score:675.0

molecular_function - combined Graph

KEGG pathways



1. Metabolism

1.1 Carbohydrate Metabolism

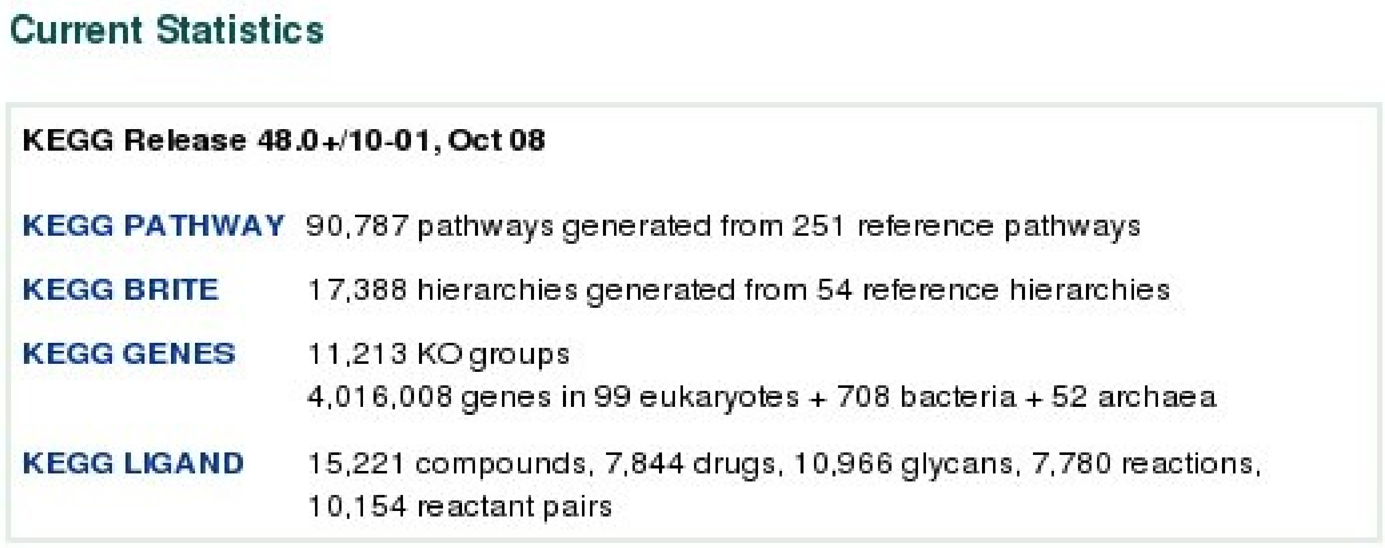
- Glycolysis / Gluconeogenesis
- Citrate cycle (TCA cycle)
- Pentose phosphate pathway
- Pentose and glucuronate interconversions
- Fructose and mannose metabolism
- Galactose metabolism
- Ascorbate and aldarate metabolism
- Starch and sucrose metabolism
- Aminosugars metabolism
- Nucleotide sugars metabolism
- Pyruvate metabolism
- Glyoxylate and dicarboxylate metabolism
- Propanoate metabolism
- Butanoate metabolism
- C5-Branched dibasic acid metabolism
- Inositol metabolism
- Inositol phosphate metabolism

1.2 Energy Metabolism

- Oxidative phosphorylation
- Photosynthesis
- Photosynthesis - antenna proteins
- Carbon fixation
- Reductive carboxylate cycle (CO2 fixation)
- Methane metabolism
- Nitrogen metabolism
- Sulfur metabolism

1.3 Lipid Metabolism

- Fatty acid biosynthesis
- Fatty acid elongation in mitochondria
- Fatty acid metabolism
- Synthesis and degradation of ketone bodies
- Biosynthesis of steroids
- Bile acid biosynthesis



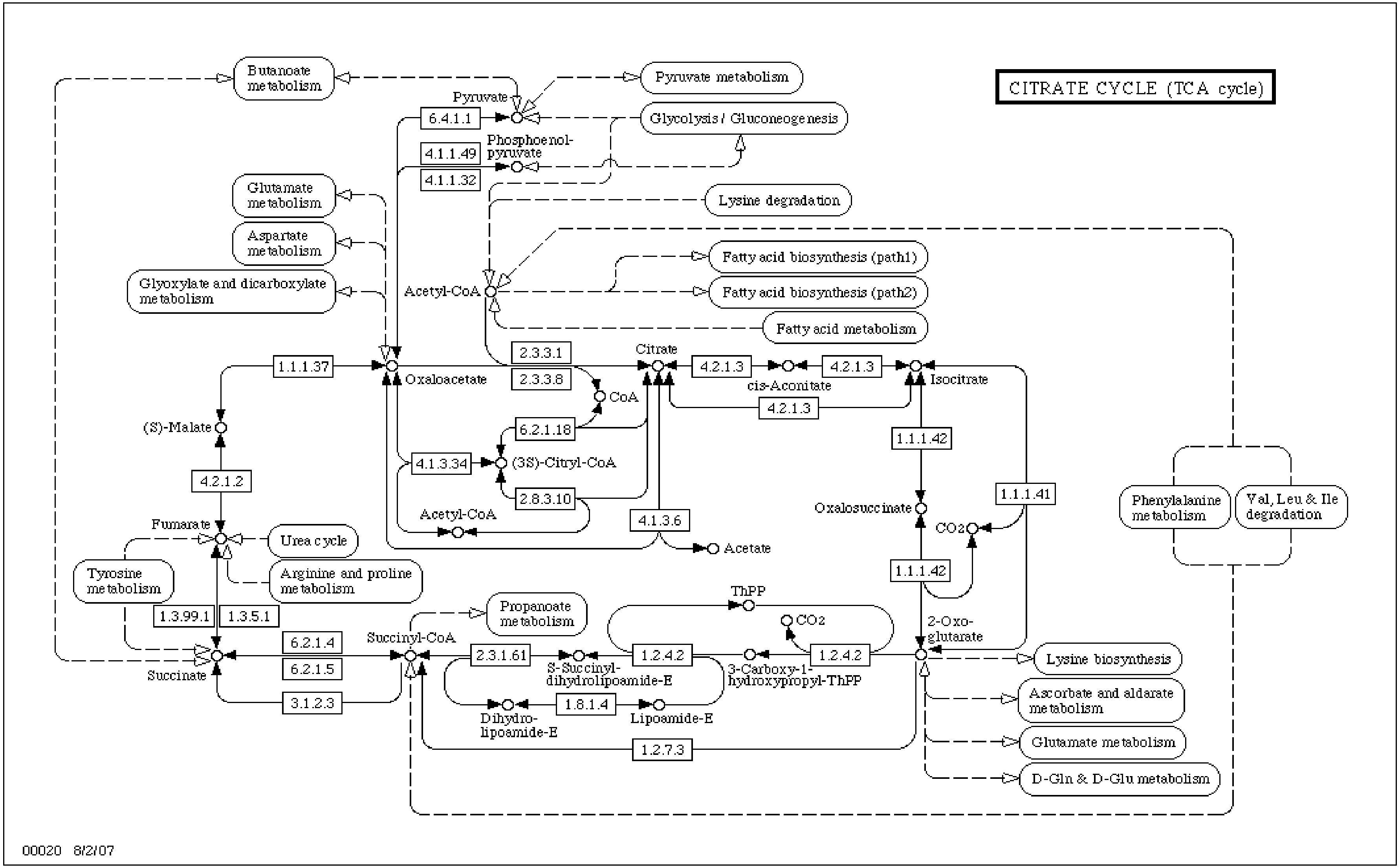
Current Statistics

KEGG Release 48.0+/10-01, Oct 08

KEGG PATHWAY	90,787 pathways generated from 251 reference pathways
KEGG BRITE	17,388 hierarchies generated from 54 reference hierarchies
KEGG GENES	11,213 KO groups 4,016,008 genes in 99 eukaryotes + 708 bacteria + 52 archaea
KEGG LIGAND	15,221 compounds, 7,844 drugs, 10,966 glycans, 7,780 reactions, 10,154 reactant pairs

http://www.genome.jp/kegg/

KEGG pathways



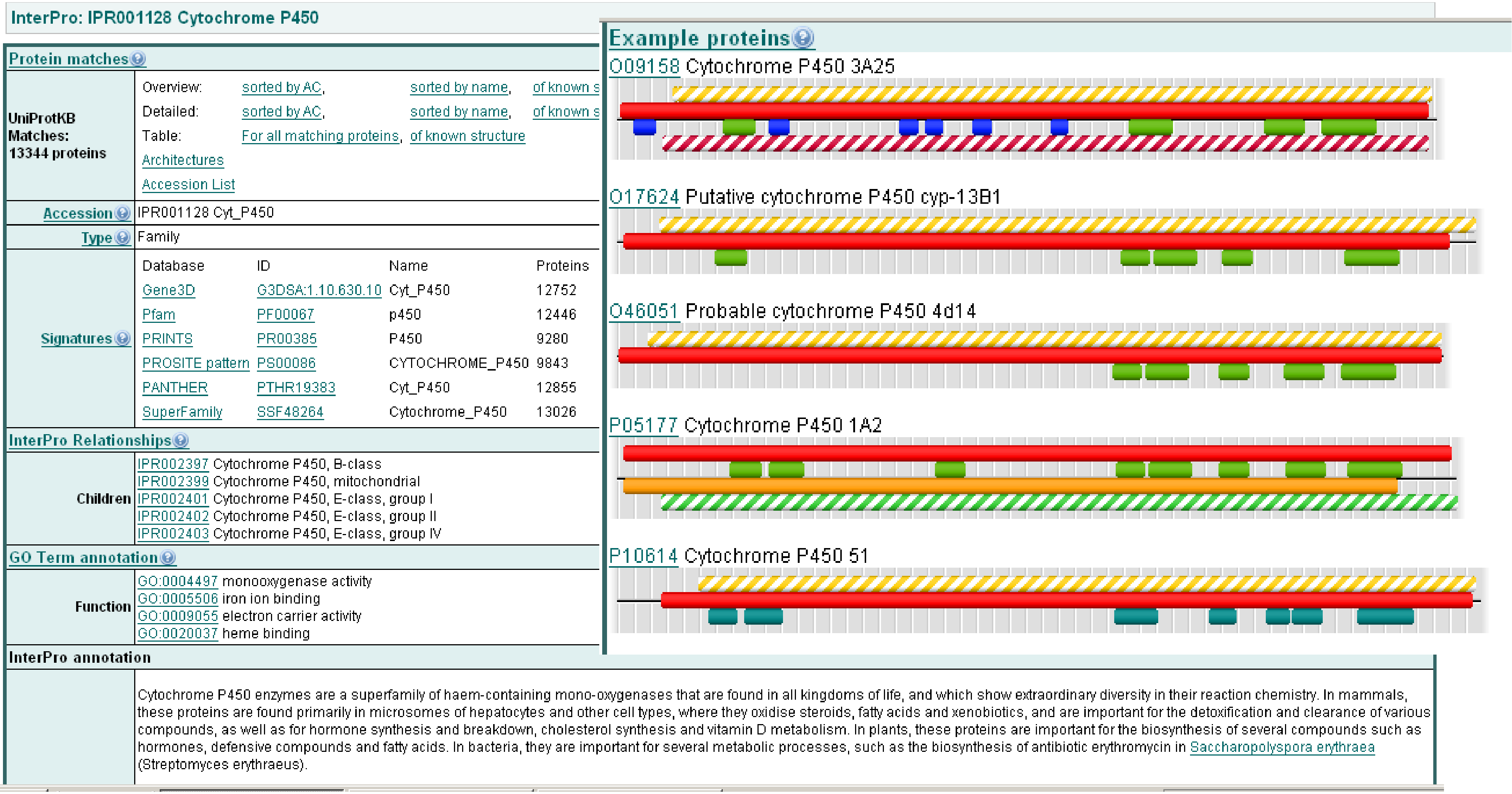
InterPro

http://www.ebi.ac.uk/interpro/databases.html

UniProt
Universal Protein Resource
prosite
Pfam
PRINTS
Protein Fingerprints Database
ProDom
SMART
TIGR
tigr fams
PANTHER
Classification System
PIRSF
Superfamily
Gene3D
MSD
CATH
SCOP
SWISS-MODEL
ModBase

- ✓ Collection of databases with functional annotation of protein motifs
- ✓ Functional vocabulary at UniProt
- ✓ There is an equivalence table between GO and InterPro

InterPro Scan



InterPro: IPR001128 Cytochrome P450

Protein matches

UniProtKB Matches: 13344 proteins	Overview:	sorted by AC,	sorted by name,	of known s
	Detailed:	sorted by AC,	sorted by name,	of known s
	Table:	For all matching proteins,	of known structure	
	Architectures			
	Accession List			

Accession IPR001128 Cyt_P450

Type Family

Signatures

Database	ID	Name	Proteins
Gene3D	G3DSA:1.10.630.10	Cyt_P450	12752
Pfam	PF00067	p450	12446
PRINTS	PR00385	P450	9280
PROSITE pattern	PS00086	CYTOCHROME_P450	9843
PANTHER	PTHR19383	Cyt_P450	12855
SuperFamily	SSF48264	Cytochrome_P450	13026

InterPro Relationships

Children
IPR002397 Cytochrome P450, B-class
IPR002399 Cytochrome P450, mitochondrial
IPR002401 Cytochrome P450, E-class, group I
IPR002402 Cytochrome P450, E-class, group II
IPR002403 Cytochrome P450, E-class, group IV

GO Term annotation

Function
GO:0004497 monooxygenase activity
GO:0005506 iron ion binding
GO:0009055 electron carrier activity
GO:0020037 heme binding

Example proteins

O09158 Cytochrome P450 3A25

O17624 Putative cytochrome P450 cyp-13B1

O46051 Probable cytochrome P450 4d14

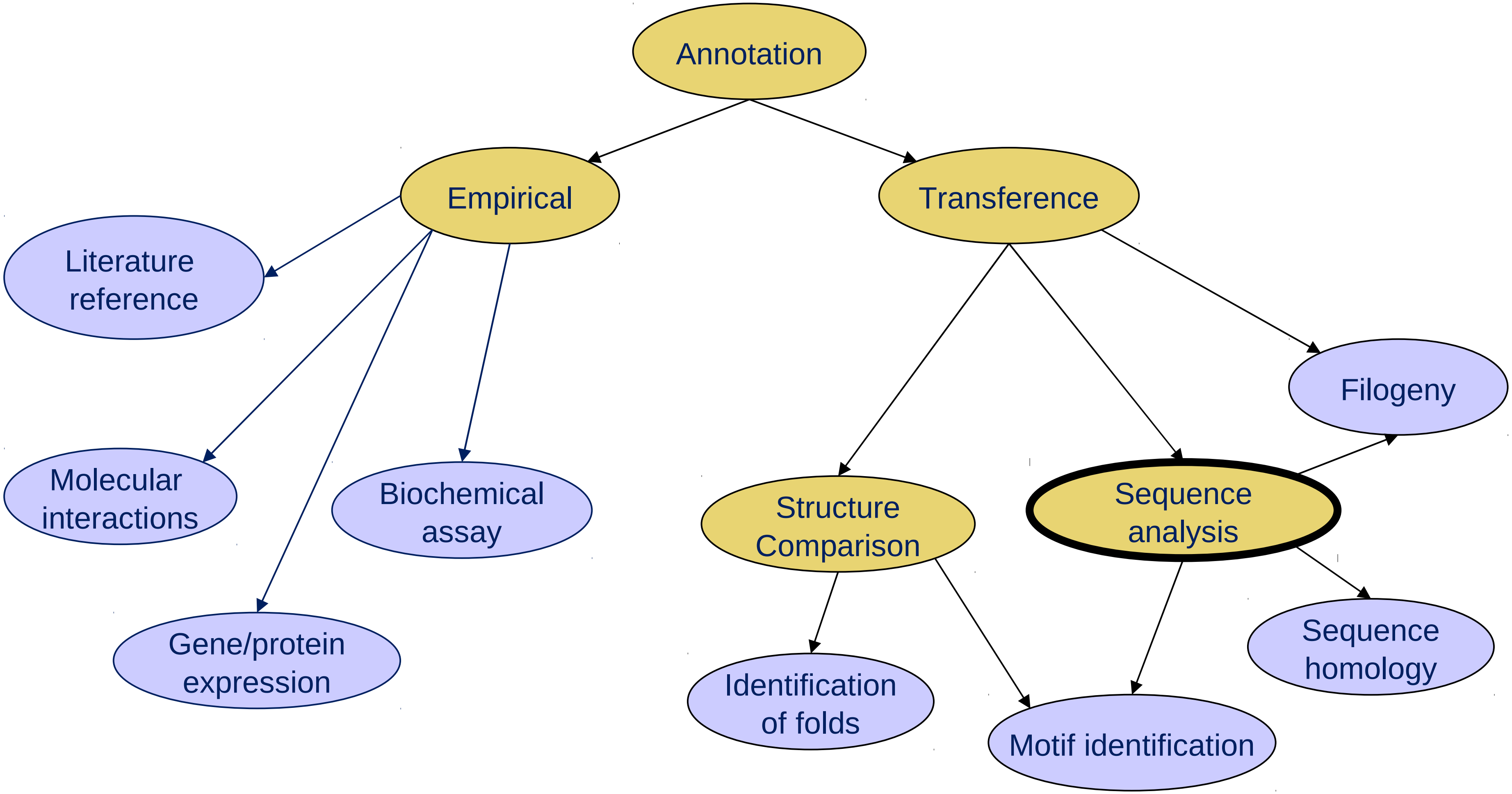
P05177 Cytochrome P450 1A2

P10614 Cytochrome P450 51

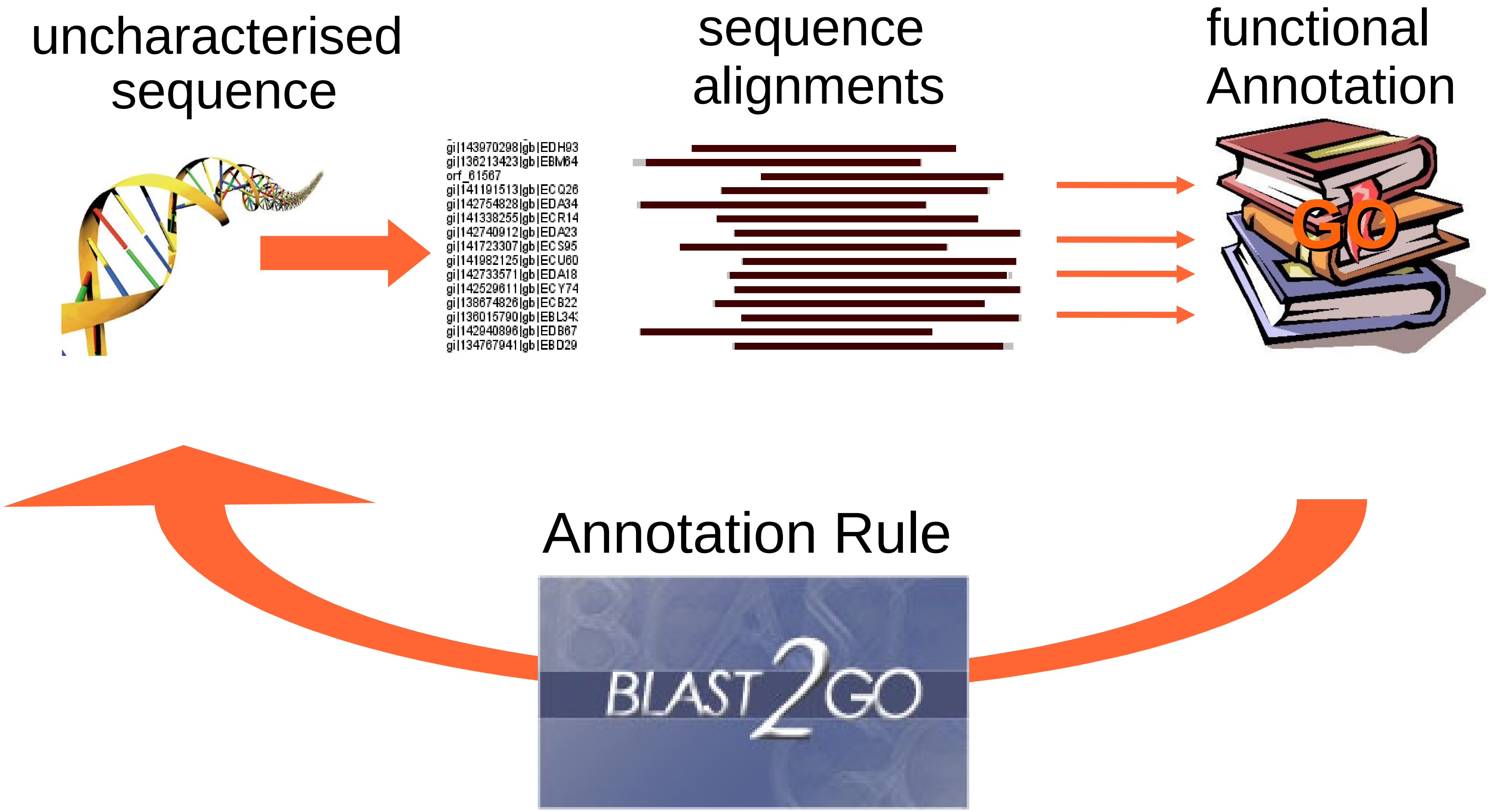
InterPro annotation

Cytochrome P450 enzymes are a superfamily of haem-containing mono-oxygenases that are found in all kingdoms of life, and which show extraordinary diversity in their reaction chemistry. In mammals, these proteins are found primarily in microsomes of hepatocytes and other cell types, where they oxidise steroids, fatty acids and xenobiotics, and are important for the detoxification and clearance of various compounds, as well as for hormone synthesis and breakdown, cholesterol synthesis and vitamin D metabolism. In plants, these proteins are important for the biosynthesis of several compounds such as hormones, defensive compounds and fatty acids. In bacteria, they are important for several metabolic processes, such as the biosynthesis of antibiotic erythromycin in Saccharopolyspora erythraea (Streptomyces erythraeus).

Functional assignment



Function transfer



uncharacterised sequence

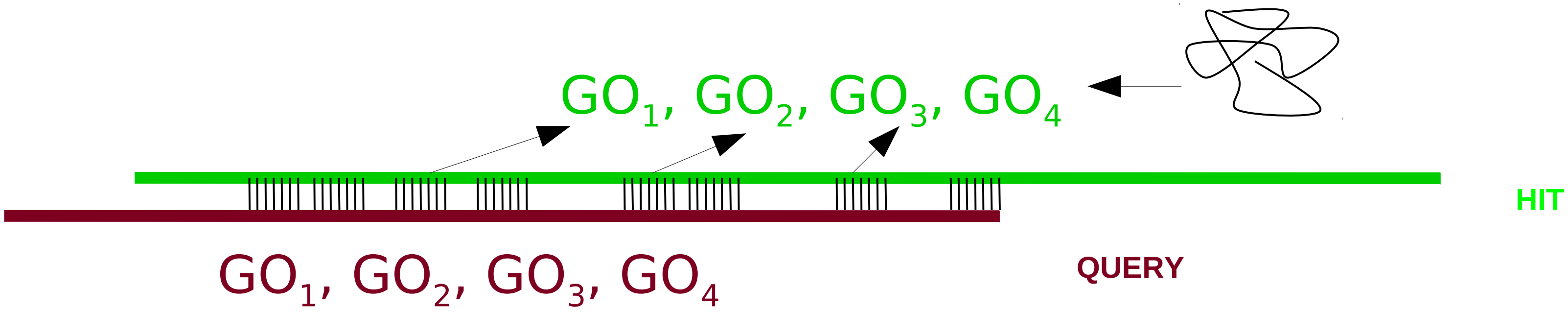
sequence alignments

functional Annotation

```
gi|143970298|gb|EDH93
gi|136213423|gb|EBM64
orf_61567
gi|141191513|gb|ECQ26
gi|142754828|gb|EDA34
gi|141338255|gb|ECR14
gi|142740912|gb|EDA23
gi|141723307|gb|ECS95
gi|141982125|gb|ECU60
gi|142733571|gb|EDA18
gi|142529611|gb|ECY74
gi|138674826|gb|ECB22
gi|136015790|gb|EBL343
gi|142940896|gb|EDB67
gi|134767941|gb|EBD29
```

Annotation Rule

Annotation by similarity: concerns



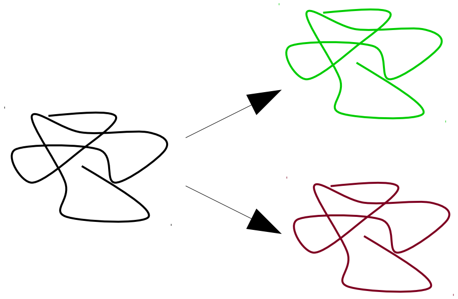
GO_1, GO_2, GO_3, GO_4

HIT

GO_1, GO_2, GO_3, GO_4

QUERY

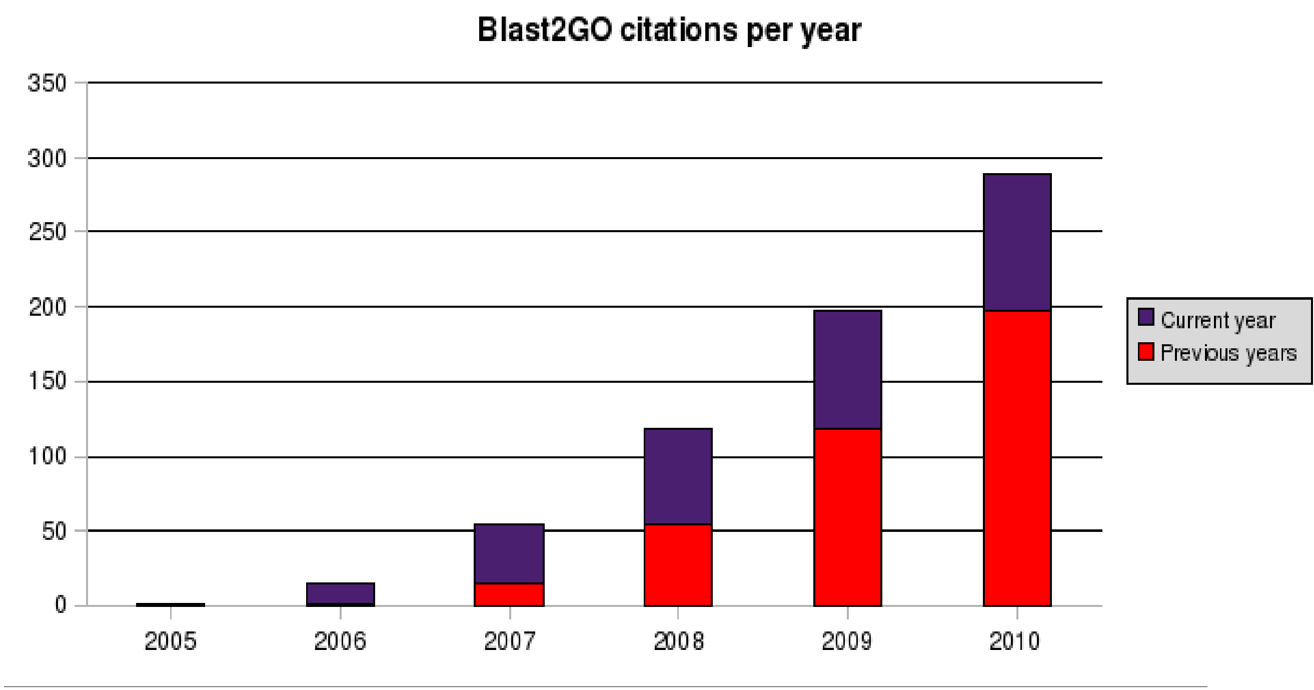
- ✓ Level of homology (~ from 40-60% is possible)
- ✓ The overlap query and hit sequences (not much a problem)
- ✓ The domain or structure function association
- ✓ The paralog problem: genes with similar sequences might have different functional specifications
- ✓ The evidence for the original annotation
- ✓ Balance between quality and quantity: depends on the use

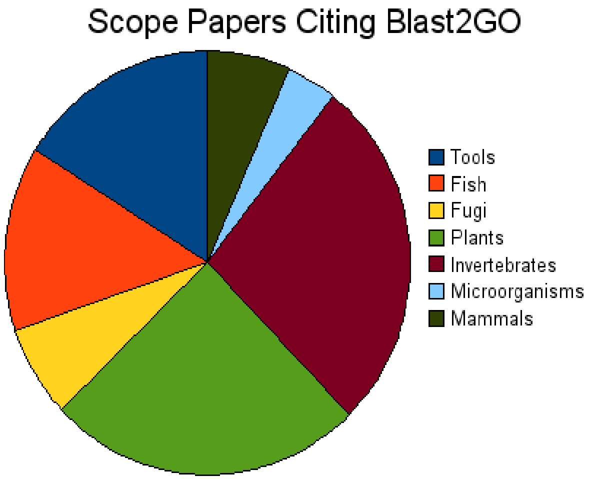


Blast2GO

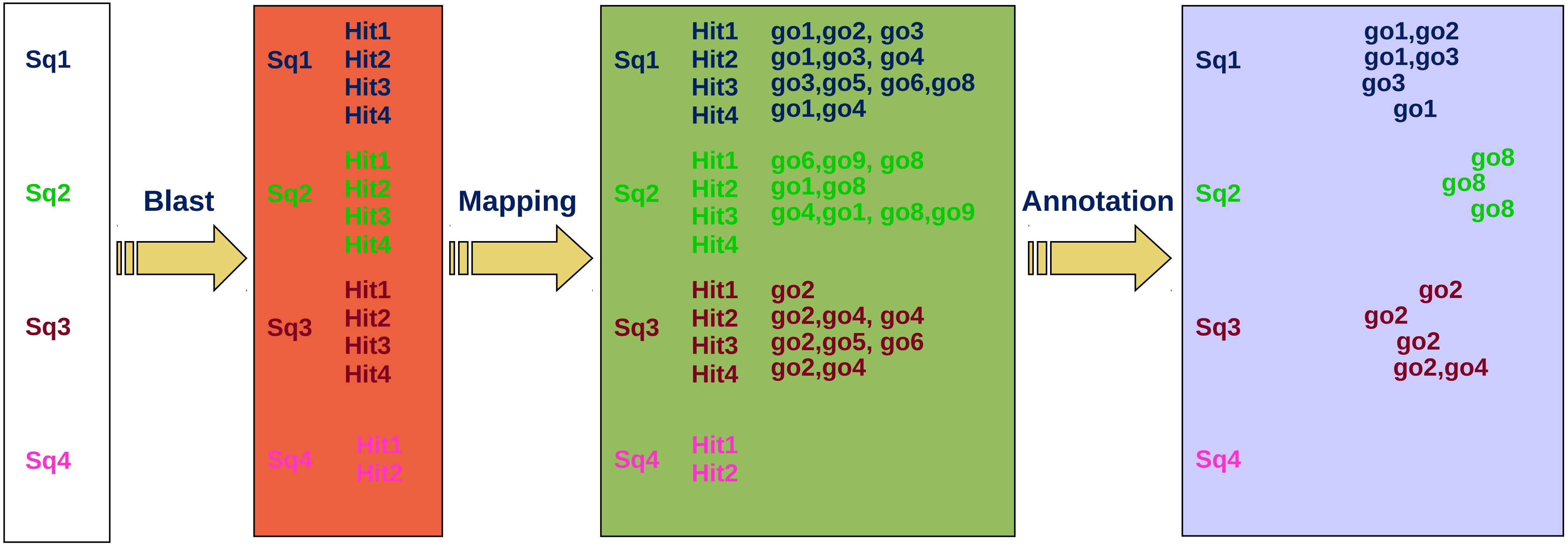
- ✓ Suite for functional annotation and data mining on functional data
  - o Considerations for **annotation**
    - Simlarity
    - Length of the overlap
    - Percentage of hit sequence spanned by the overlap
    - Evidence original annotation
    - Blast hits and motif hits
    - Refinement by additional methods
  - o Visualization:
    - Annotation charts
    - **Knowledge discovery on the DAG**
- ✓ Desktop Java application

Scientific Impact





Blast2GO Annotation strategy



Sq1
Sq2
Sq3
Sq4

Blast

Sq1 Hit1 Hit2 Hit3 Hit4
Sq2 Hit1 Hit2 Hit3 Hit4
Sq3 Hit1 Hit2 Hit3 Hit4
Sq4 Hit1 Hit2

Mapping

Sq1
Hit1 go1,go2, go3
Hit2 go1,go3, go4
Hit3 go3,go5, go6,go8
Hit4 go1,go4

Sq2
Hit1 go6,go9, go8
Hit2 go1,go8
Hit3 go4,go1, go8,go9
Hit4

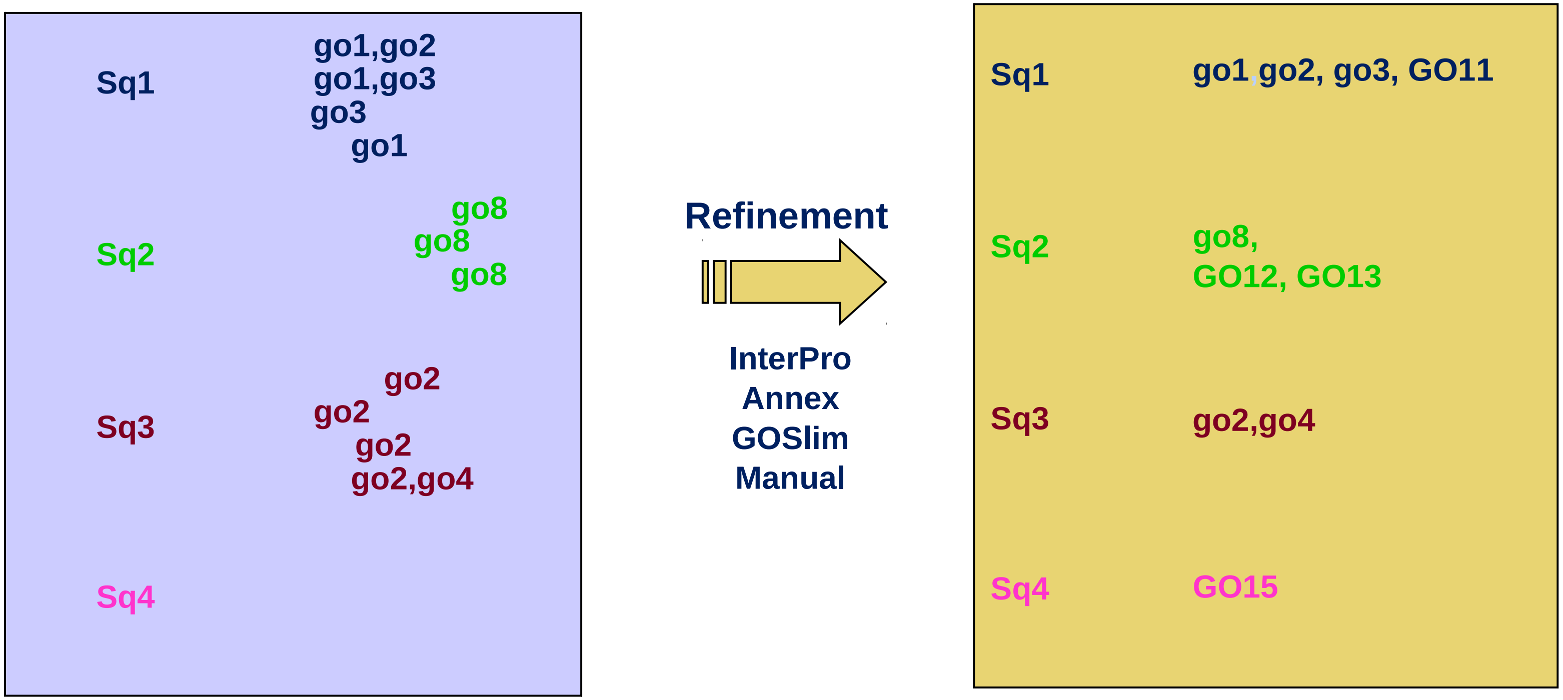
Sq3
Hit1 go2
Hit2 go2,go4, go4
Hit3 go2,go5, go6
Hit4 go2,go4

Sq4
Hit1
Hit2

Annotation

Sq1 go1,go2 go1,go3 go3 go1
Sq2 go8 go8 go8
Sq3 go2 go2 go2 go2,go4
Sq4

Blast2GO Annotation Strategy



Sq1 go1,go2
go1,go3
go3
go1

Sq2 go8
go8
go8

Sq3 go2
go2
go2
go2,go4

Sq4

Refinement

InterPro
Annex
GOSlim
Manual

Sq1 go1 go2, go3, GO11

Sq2 go8,
GO12, GO13

Sq3 go2,go4

Sq4 GO15

Blast2GO annotation rule

Similarity requirement

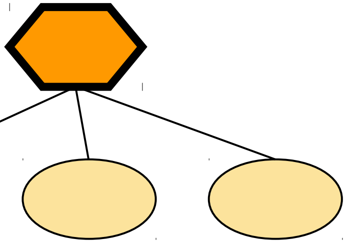
$$sim = \frac{\sum positives_{hsp}}{\sum alignmentlength_{hsp}}$$

Quality of annotation source

EC	weight
IC	1
TAS	1
IDA	1
IMP	0.9
IGI	0.9
IPI	0.9
ISS	0.8
IEP	0.8
NAS	0.7
IEA	0.7
ND	0.5
NR	0.5
RCA	0.5

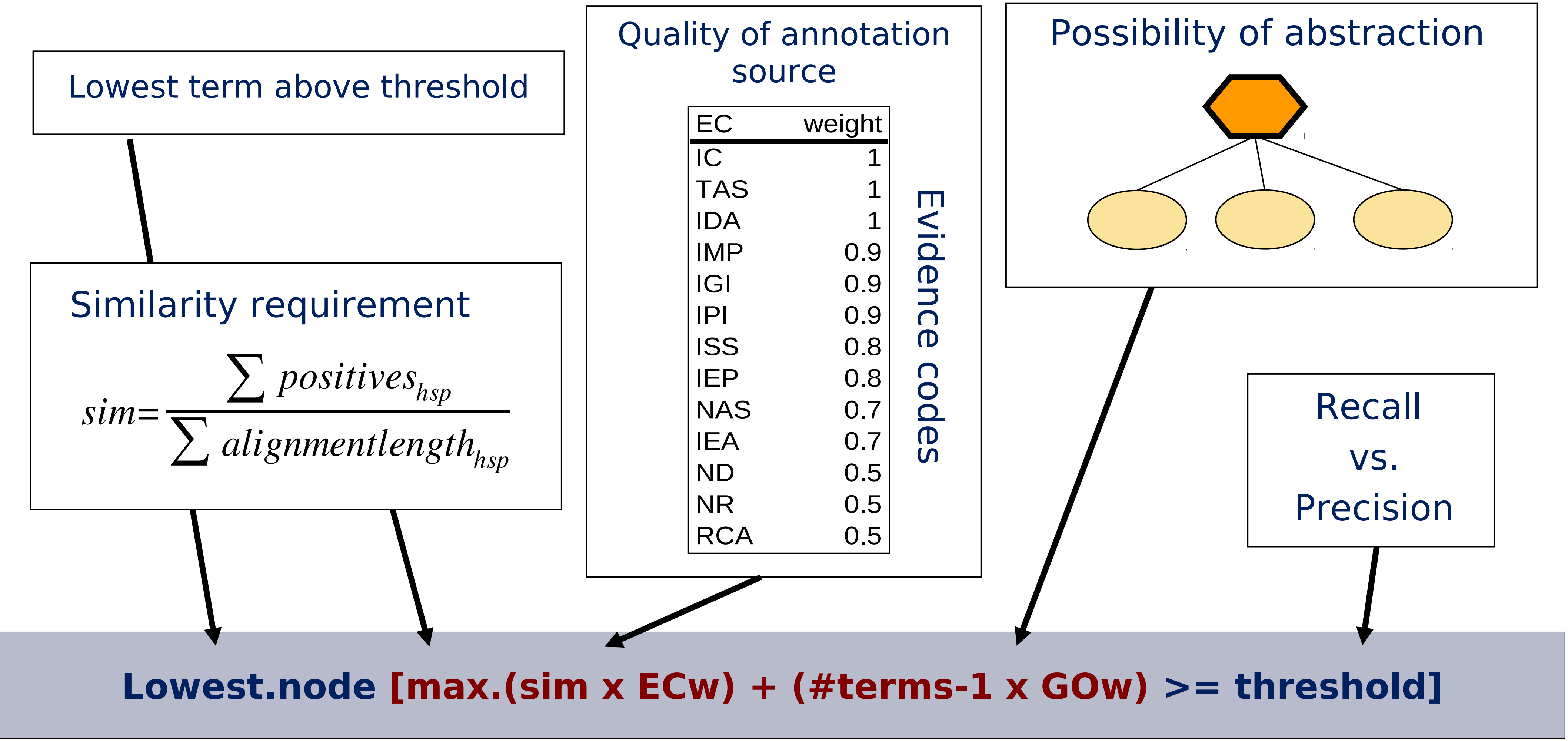
Evidence codes

Possibility of abstraction



[max.(sim x ECw) + (#terms-1 x GOw)]

Blast2GO annotation rule



Lowest term above threshold

Similarity requirement

$$sim = \frac{\sum positives_{hsp}}{\sum alignmentlength_{hsp}}$$

Quality of annotation source

EC	weight
IC	1
TAS	1
IDA	1
IMP	0.9
IGI	0.9
IPI	0.9
ISS	0.8
IEP	0.8
NAS	0.7
IEA	0.7
ND	0.5
NR	0.5
RCA	0.5

Evidence codes

Possibility of abstraction

Recall
vs.
Precision

Lowest.node [max.(sim x ECw) + (#terms-1 x GOw) >= threshold]

Blast2GO annotation rule

Lowest.node [max.(sim x ECw) + (#terms-1 x GOw) >= threshold]

- When I have a GO with ECw = 1 and I do not allow abstraction (GOw = 0), then the Annotation Score = %similarity
- If the ECw < 1 my similarity requirement is higher to obtain the same Annotation Score
- If I allow abstraction GOw > 0, then with less similarity I can obtain the required Annotation Score at a parent node

Outline

Concepts on Functional Annotation:

Biological Databases

Blast2GO annotation strategy

The Blast2GO annotation framework:

Annotation steps, Modulation of annotation intensity, Export/Import Functions, Sequence Selection, Additional Tools

----

Data Mining: Visualization Techniques

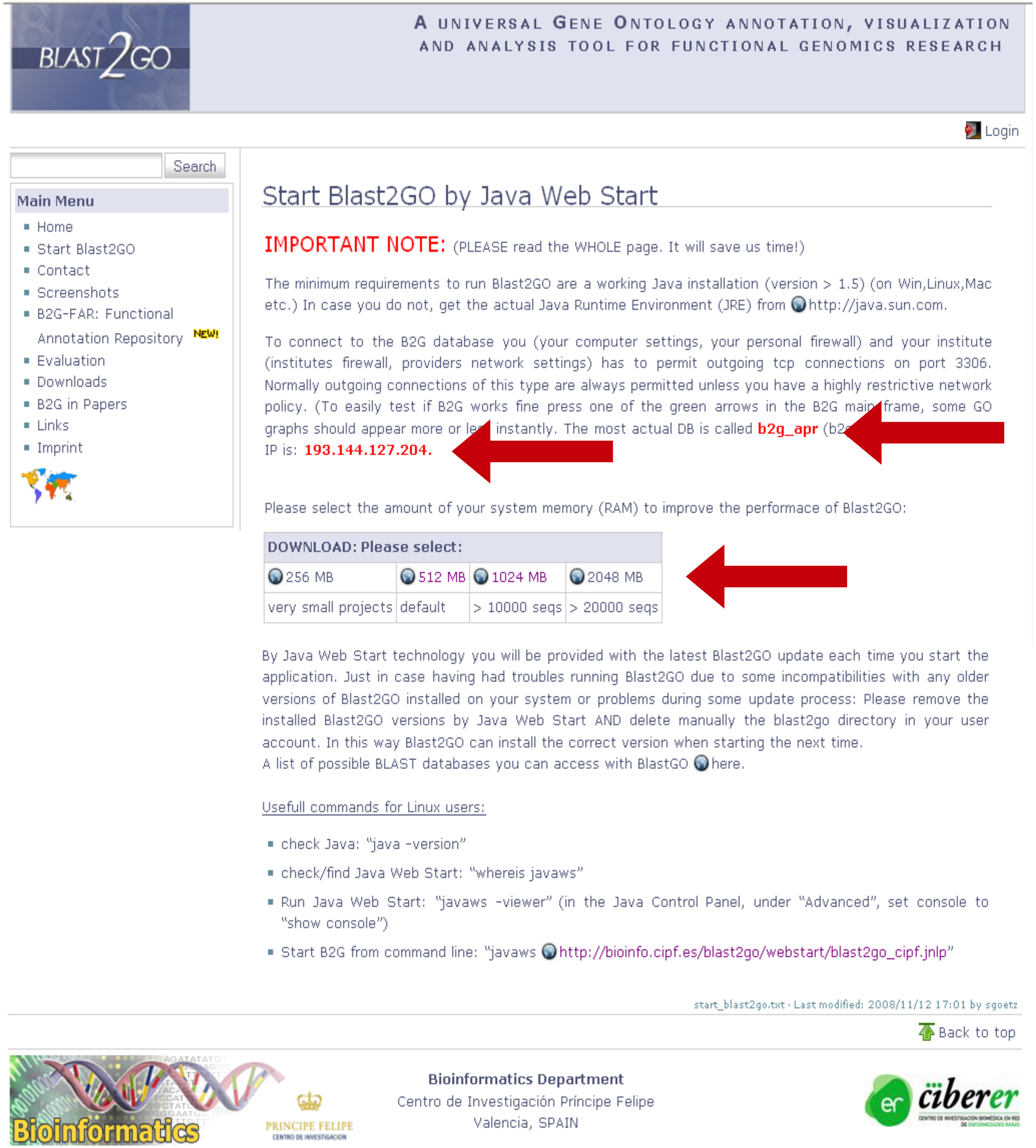
Combined Graph, Charts and Pies

Hands on: Blast2GO annotation excercise

Start Blast2GO

www.blast2go.org

- Desktop application
- Java WebStart
- Internet connection



A UNIVERSAL GENE ONTOLOGY ANNOTATION, VISUALIZATION AND ANALYSIS TOOL FOR FUNCTIONAL GENOMICS RESEARCH

Start Blast2GO by Java Web Start

IMPORTANT NOTE: (PLEASE read the WHOLE page. It will save us time!)

The minimum requirements to run Blast2GO are a working Java installation (version > 1.5) (on Win,Linux,Mac etc.) In case you do not, get the actual Java Runtime Environment (JRE) from http://java.sun.com.

To connect to the B2G database you (your computer settings, your personal firewall) and your institute (institutes firewall, providers network settings) has to permit outgoing tcp connections on port 3306. Normally outgoing connections of this type are always permitted unless you have a highly restrictive network policy. (To easily test if B2G works fine press one of the green arrows in the B2G main frame, some GO graphs should appear more or le instantly. The most actual DB is called **b2g_apr** (b2 IP is: **193.144.127.204.**

Please select the amount of your system memory (RAM) to improve the performace of Blast2GO:

DOWNLOAD: Please select:			
256 MB	512 MB	1024 MB	2048 MB
very small projects	default	> 10000 seqs	> 20000 seqs

By Java Web Start technology you will be provided with the latest Blast2GO update each time you start the application. Just in case having had troubles running Blast2GO due to some incompatibilities with any older versions of Blast2GO installed on your system or problems during some update process: Please remove the installed Blast2GO versions by Java Web Start AND delete manually the blast2go directory in your user account. In this way Blast2GO can install the correct version when starting the next time.
A list of possible BLAST databases you can access with BlastGO here.

Usefull commands for Linux users:

- check Java: "java -version"
- check/find Java Web Start: "whereis javaws"
- Run Java Web Start: "javaws -viewer" (in the Java Control Panel, under "Advanced", set console to "show console")
- Start B2G from command line: "javaws http://bioinfo.cipf.es/blast2go/webstart/blast2go_cipf.jnlp"

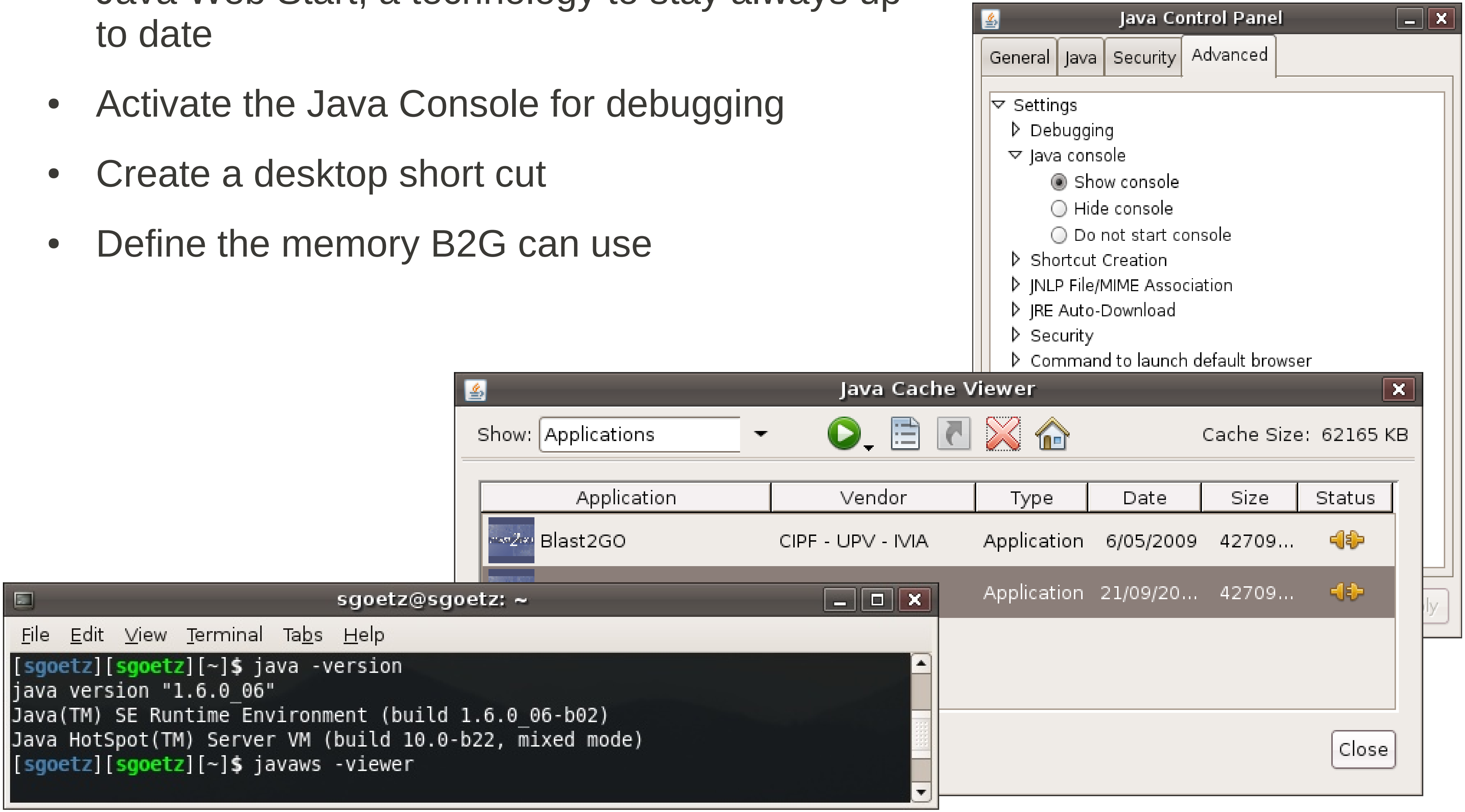
start_blast2go.txt · Last modified: 2008/11/12 17:01 by sgoetz

Bioinformatics Department
Centro de Investigación Príncipe Felipe
Valencia, SPAIN

ciberer
CENTRO DE INVESTIGACIÓN BIOMÉDICA EN RED DE ENFERMEDADES RARAS

Java and Java Web Start

- SUNs Java Runtime Environment (1.6)
- Java Web Start, a technology to stay always up to date
- Activate the Java Console for debugging
- Create a desktop short cut
- Define the memory B2G can use



Java Control Panel

- Settings
  - Debugging
  - Java console
    - ● Show console
    - ○ Hide console
    - ○ Do not start console
  - Shortcut Creation
  - JNLP File/MIME Association
  - JRE Auto-Download
  - Security
  - Command to launch default browser

Java Cache Viewer

Show: Applications

Cache Size: 62165 KB

Application	Vendor	Type	Date	Size	Status
Blast2GO	CIPF - UPV - IVIA	Application	6/05/2009	42709...	
		Application	21/09/20...	42709...	

sgoetz@sgoetz: ~

```
[sgoetz][sgoetz][~]$ java -version
java version "1.6.0_06"
Java(TM) SE Runtime Environment (build 1.6.0_06-b02)
Java HotSpot(TM) Server VM (build 10.0-b22, mixed mode)
[sgoetz][sgoetz][~]$ javaws -viewer
```

Input data

(in FASTA format, AA or nt)

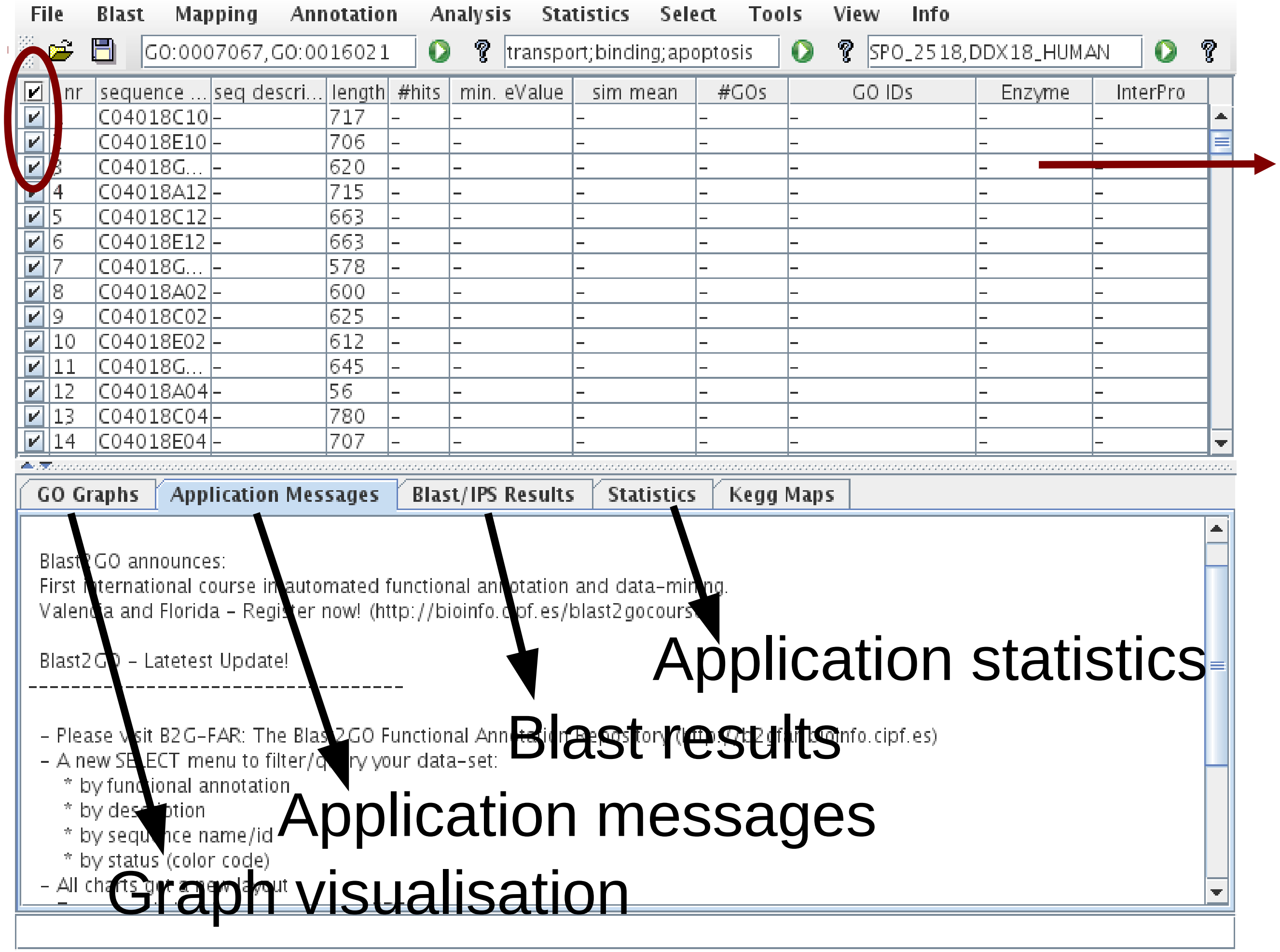
```
>my_favourite_species_seq1 | still unknown
gtgatggaaaagaaaagttttgttatcgtcgacgcatatgggtttctttttcgcgcgtattatgcgctgcctggattaagcacctcatacaattttcctgtaggaggtgtatatggttt
tataaacatacttttgaaacatctctctttccacgatgcagattatttagttgtggtatttgattcggggtcgaaaaattttcgtcacactatgtattccgaatacaaaactaatcgc
cctaaagcaccagaggatctgtcactacaatgtgctccgctacgtgaggctgttgaagcgtttaatattgtaagtgaagaagtgcttaactacgaagcagacgacgtaata
gctacactctgtacaaaatatgcatctagtaatgttggagtgagaatactgtcagcagataaggatttactacaactcctaaatgataatgttcaagtttacgaccctataaaa
agcagatacctcaccaatgaatacgttttagaaaaatttggtgtttcatcagataagttgcatattgatacggttgcatcgagttataatgagaaaattattctcagctaagctgt
acaccgtttattacacactcgaaaggccgttag
>my_favourite_species_seq2 | no clue
ttgttagctaaaaaggaagactttcacacctttggtaatggtgttggctctgctggaacaggtggagttgtagtttctgcatccatgttgtctgcggatttttcaaatcttagagaag
agatagcagcggttagtacggctggtgcagattggttacacattgatgtgatggatgggtgcttcgtccccagtttgactatgggtcctgtggtgatttccggcattaggaaatg
tacaaatatgtttcttgatgtgcatttgatgattaatcgcccaggcgatcatctgaagagtgtggtagatgctggagctgataagatagagcacattcgcaagatgatagagg
aaagctcatcaaccgcgaaaatcgctgttgatggtggtgtttcaacggataatgcccgggctgttatcgaggcaggtgcgaatatactcgttgttggaacggcgctgtttgct
gctgacgatatgagtaaagttgtaagaactttaaaatcattttaa
>my_favourite_species_seq3 | just sequenced
gtgggactgctcatccctgtaggcagggtggctattttttgtgtaaaggcagtctttcatagtcttgtaccgccatactatctatggataactacaaagcagttttttgaggtgtggt
ttttctctcttcctatagtagcagttacatctttgtttacgggaggcgcgttagcccttcaggataccctcgtgggaagcgctaaagtatcagggtaatggagtttttactcctgca
agatgtaatagagggtctggtaaaagctgtatcgtttgggctggtaatttcgctagttgggtgttacaacgggtatcactgtgagataggcgcaaggggtgtaggaacagcg
acaacaaaaacttcggtagcagcttctatgctcataattttgttaaactatataattactgttttttacgcgta
>my_favourite_species_seq4 | we will see soon...
atgtacgctgtatctctttcaaatttgcatgtctctttcaacaacaaggaggttttgaaaggtgttgacttggacatagcatgggggattccctggttatactgggagaatctggt
agtggaaagtctgtactaacaaaggttgtattgggtctaatagtgccccaagagggaagtgttactgtagatggcaccaatattcttgagaataggcagggcatcaagaat
tttagtgttttgtttcaaaactgtgcgttatttgacagtcttacgatttgggaaaatgtagtattcaatttccgtaggaggcttcgtttagataaggataatgccaaggctttggcttta
cggggattggagcttgtgggattggacgccagtgtaatgaacgtgtatcctgtggagctatcaggcgggatgaaaaagcgcgtagctttggcaagagctattataggtagt
cccaaaattctaattttggatgagccaacttcgggattggatcctataatgtcttcagtggt
```

Blast2GO Application

(1) Blast

(2) Mapping

(3) Annotation



GO:0007067,GO:0016021

transport;binding;apoptosis

SPO_2518,DDX18_HUMAN

☑	nr	sequence ...	seq descri...	length	#hits	min. eValue	sim mean	#GOs	GO IDs	Enzyme	InterPro
☑	1	C04018C10	-	717	-	-	-	-	-	-	-
☑	2	C04018E10	-	706	-	-	-	-	-	-	-
☑	3	C04018G...	-	620	-	-	-	-	-	-	-
☑	4	C04018A12	-	715	-	-	-	-	-	-	-
☑	5	C04018C12	-	663	-	-	-	-	-	-	-
☑	6	C04018E12	-	663	-	-	-	-	-	-	-
☑	7	C04018G...	-	578	-	-	-	-	-	-	-
☑	8	C04018A02	-	600	-	-	-	-	-	-	-
☑	9	C04018C02	-	625	-	-	-	-	-	-	-
☑	10	C04018E02	-	612	-	-	-	-	-	-	-
☑	11	C04018G...	-	645	-	-	-	-	-	-	-
☑	12	C04018A04	-	56	-	-	-	-	-	-	-
☑	13	C04018C04	-	780	-	-	-	-	-	-	-
☑	14	C04018E04	-	707	-	-	-	-	-	-	-

Blast2GO announces:
First International course in automated functional annotation and data-mining.
Valencia and Florida - Register now! (http://bioinfo.cipf.es/blast2gocourse)

Blast2GO - Latetest Update!

- Please visit B2G-FAR: The Blast2GO Functional Annotation Repository (http://b2gfar.bioinfo.cipf.es)
- A new SELECT menu to filter/query your data-set:
  * by functional annotation
  * by description
  * by sequence name/id
  * by status (color code)
- All charts got a new layout
-

Main Sequence Table

Any operation will only affect to selected sequences!!!!

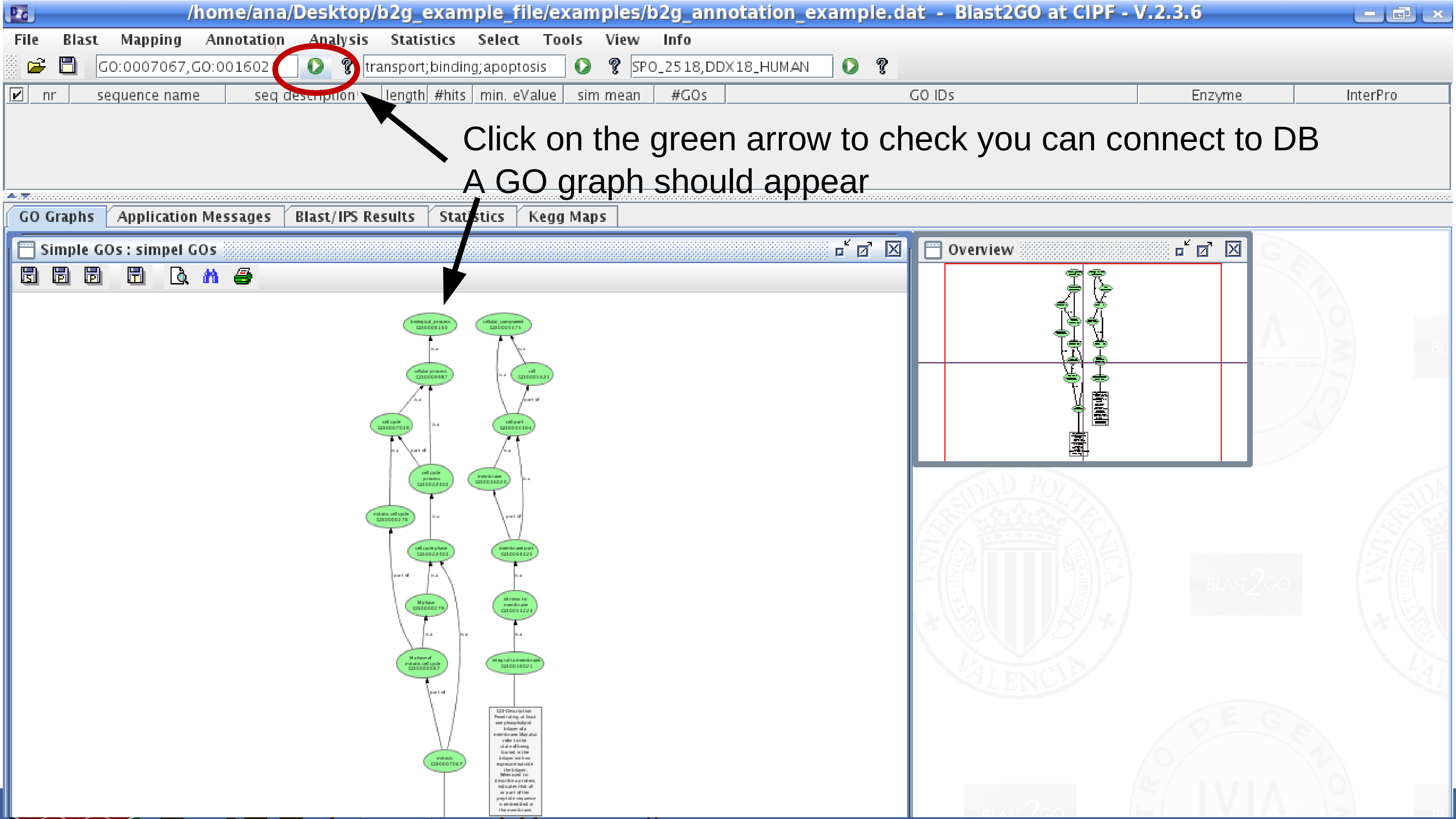
Application statistics

Blast results

Application messages

Graph visualisation

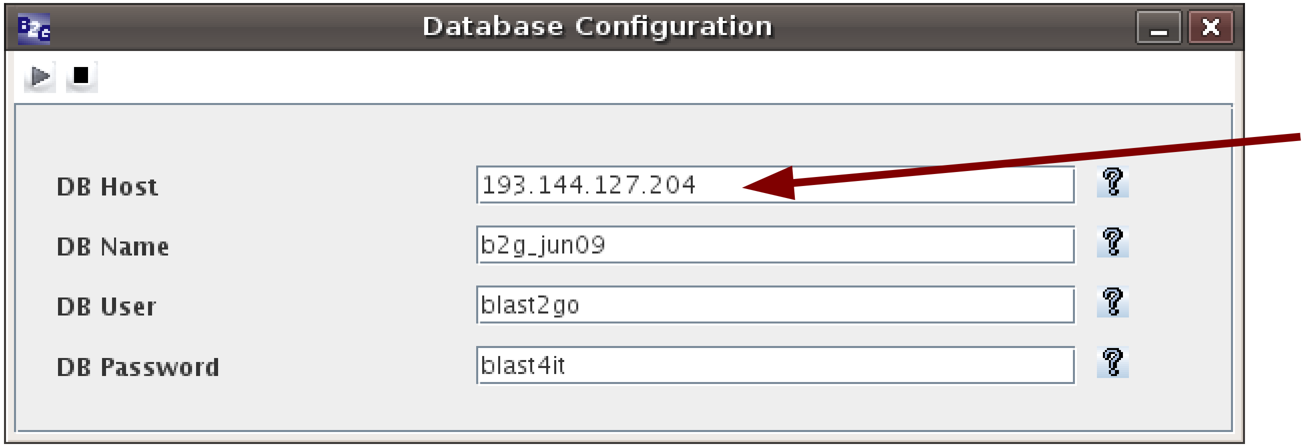
The First Check



Click on the green arrow to check you can connect to DB

A GO graph should appear

Database configuration



DB Host 193.144.127.204

DB Name b2g_jun09

DB User blast2go

DB Password blast4it

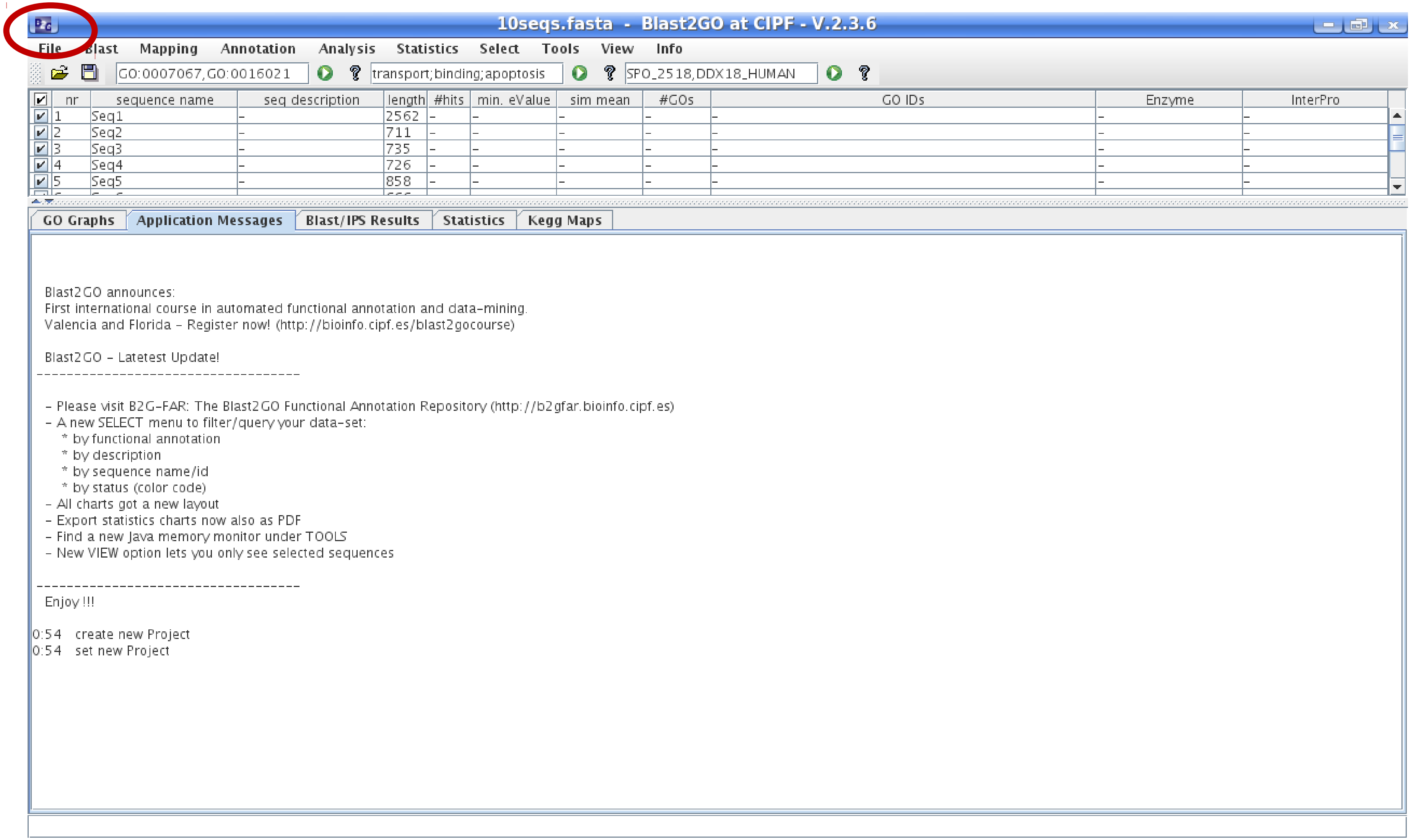
FOR TODAY: mem20

Open port 3306 (mysql) for outgoing connections at your institute

Configure/check personal firewalls

Actual settings can be found at www.blast2go.org

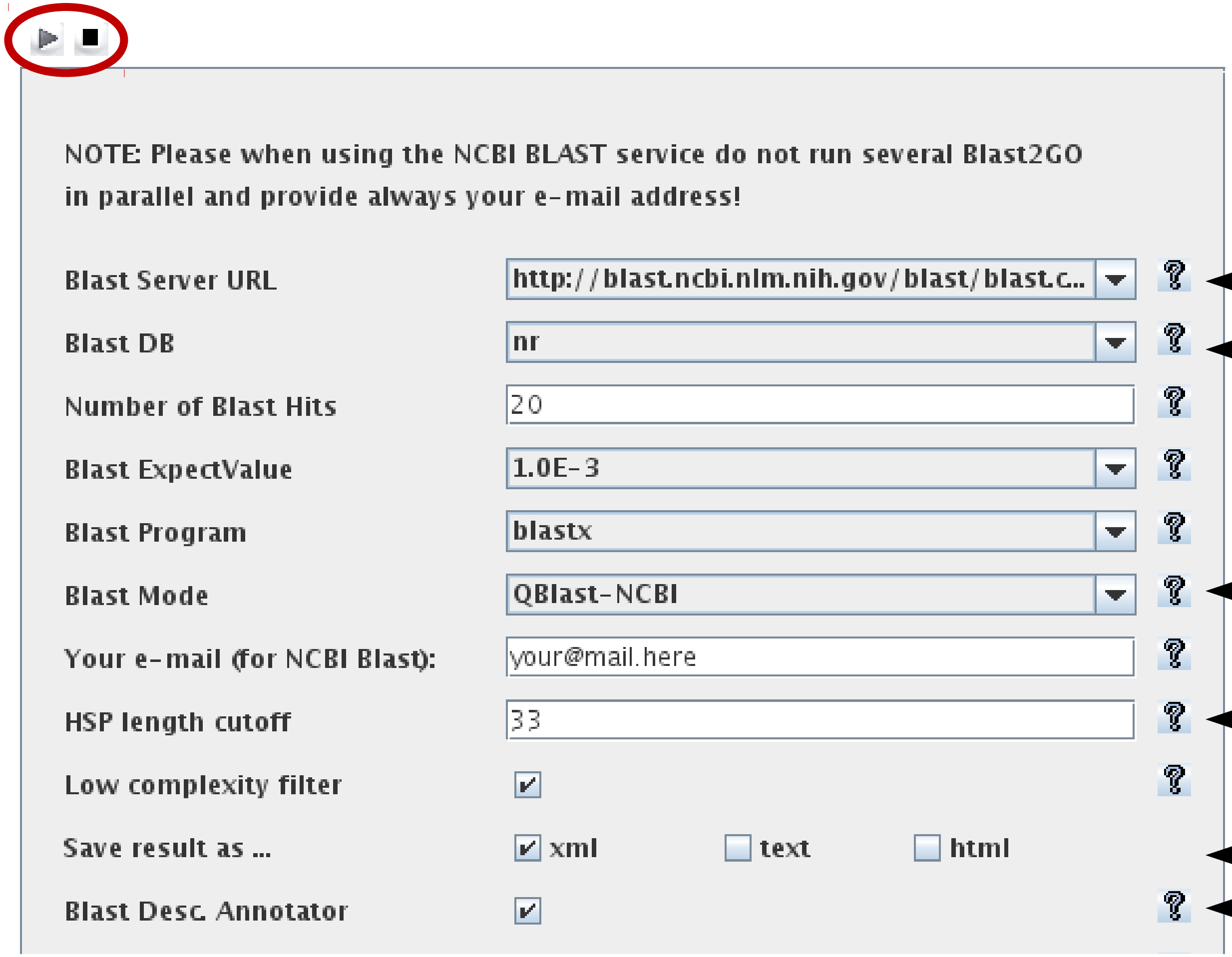
Load Sequences



☑	nr	sequence name	seq description	length	#hits	min. eValue	sim mean	#GOs	GO IDs	Enzyme	InterPro
☑	1	Seq1	-	2562	-	-	-	-	-	-	-
☑	2	Seq2	-	711	-	-	-	-	-	-	-
☑	3	Seq3	-	735	-	-	-	-	-	-	-
☑	4	Seq4	-	726	-	-	-	-	-	-	-
☑	5	Seq5	-	858	-	-	-	-	-	-	-

```
 Blast2GO announces:
 First international course in automated functional annotation and data-mining.
 Valencia and Florida - Register now! (http://bioinfo.cipf.es/blast2gocourse)

 Blast2GO - Latetest Update!
-----------------------------------

 - Please visit B2G-FAR: The Blast2GO Functional Annotation Repository (http://b2gfar.bioinfo.cipf.es)
 - A new SELECT menu to filter/query your data-set:
    * by functional annotation
    * by description
    * by sequence name/id
    * by status (color code)
 - All charts got a new layout
 - Export statistics charts now also as PDF
 - Find a new Java memory monitor under TOOLS
 - New VIEW option lets you only see selected sequences

-----------------------------------
 Enjoy !!!

0:54   create new Project
0:54   set new Project
```

Run BLAST search



NOTE: Please when using the NCBI BLAST service do not run several Blast2GO in parallel and provide always your e-mail address!

Blast Server URL	http://blast.ncbi.nlm.nih.gov/blast/blast.c...	
Blast DB	nr	
Number of Blast Hits	20	
Blast ExpectValue	1.0E-3	
Blast Program	blastx	
Blast Mode	QBlast-NCBI	
Your e-mail (for NCBI Blast):	your@mail.here	
HSP length cutoff	33	
Low complexity filter	☑	
Save result as ...	☑ xml	□ text □ html
Blast Desc. Annotator	☑	

BLAST against NCBI or locally

Choose different DBs

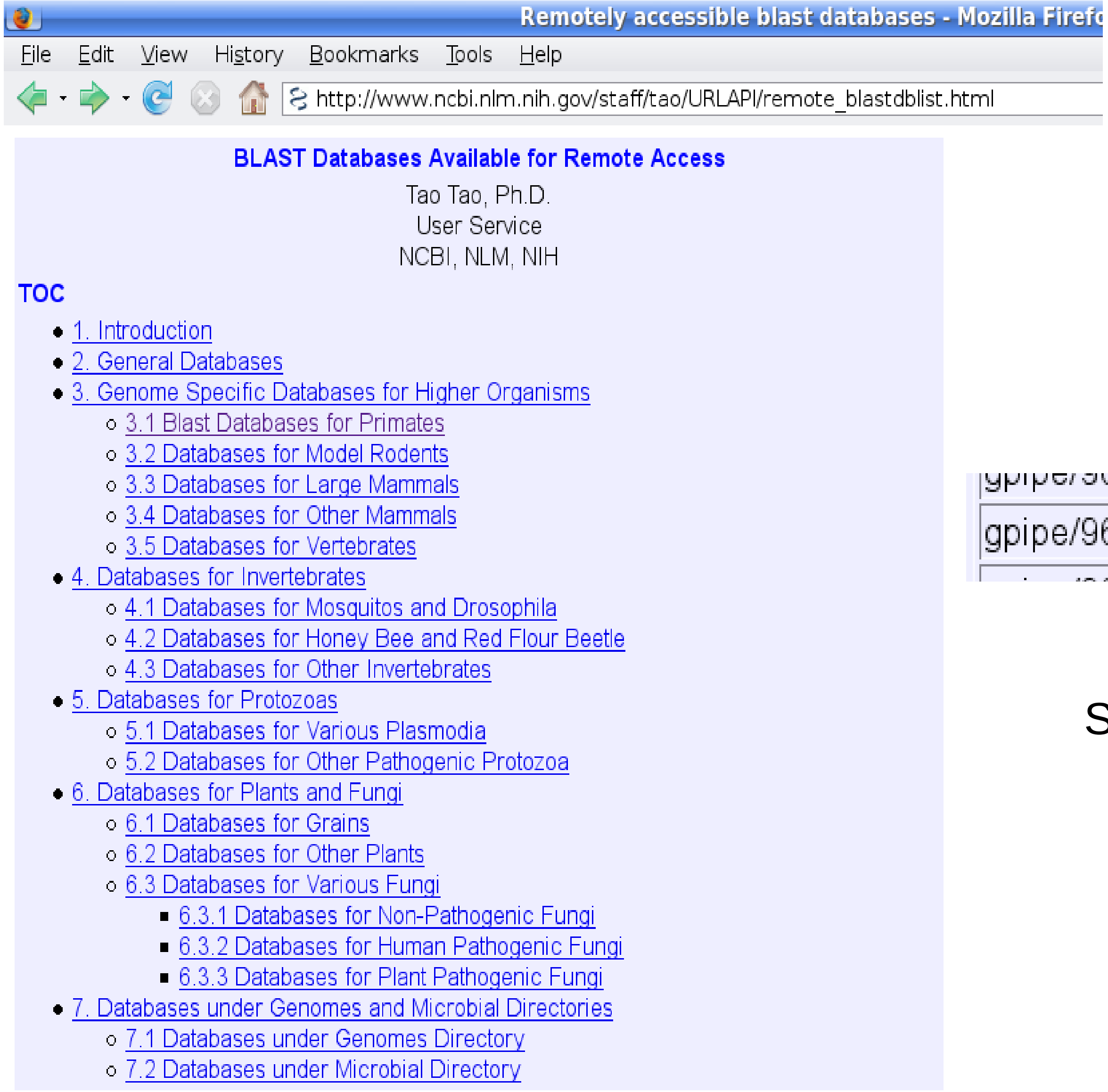
In combination with URL

Limit to query-hit overlap

Recommended to save as XML

Text mining on BLAST hit description

Choose other DB at NCBI



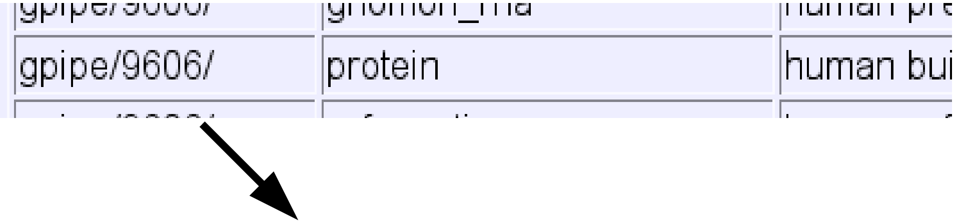
http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/remote_blastdblist.html

BLAST Databases Available for Remote Access

Tao Tao, Ph.D.
User Service
NCBI, NLM, NIH

TOC

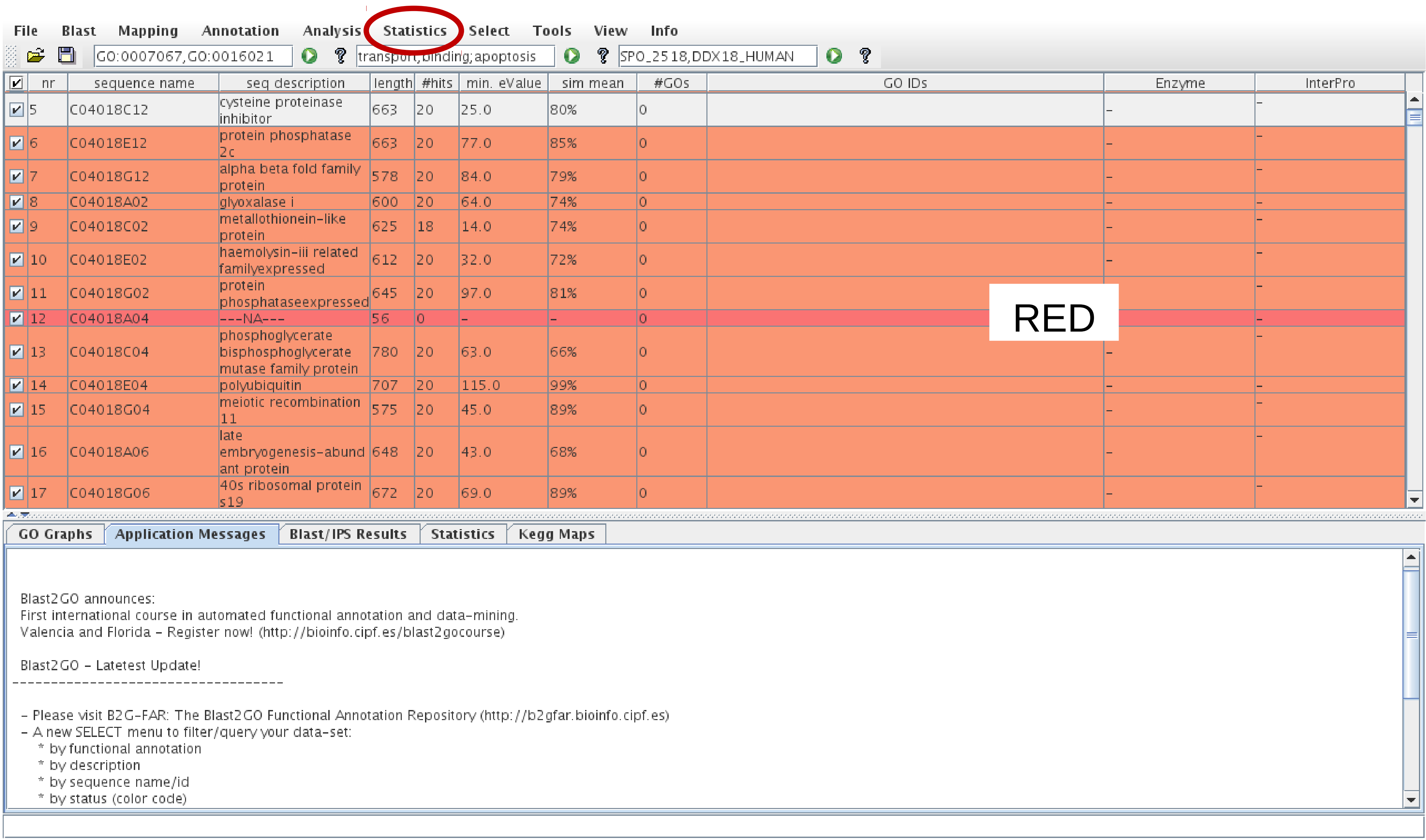
- 1. Introduction
- 2. General Databases
- 3. Genome Specific Databases for Higher Organisms
    - 3.1 Blast Databases for Primates
    - 3.2 Databases for Model Rodents
    - 3.3 Databases for Large Mammals
    - 3.4 Databases for Other Mammals
    - 3.5 Databases for Vertebrates
- 4. Databases for Invertebrates
    - 4.1 Databases for Mosquitos and Drosophila
    - 4.2 Databases for Honey Bee and Red Flour Beetle
    - 4.3 Databases for Other Invertebrates
- 5. Databases for Protozoas
    - 5.1 Databases for Various Plasmodia
    - 5.2 Databases for Other Pathogenic Protozoa
- 6. Databases for Plants and Fungi
    - 6.1 Databases for Grains
    - 6.2 Databases for Other Plants
    - 6.3 Databases for Various Fungi
        - 6.3.1 Databases for Non-Pathogenic Fungi
        - 6.3.2 Databases for Human Pathogenic Fungi
        - 6.3.3 Databases for Plant Pathogenic Fungi
- 7. Databases under Genomes and Microbial Directories
    - 7.1 Databases under Genomes Directory
    - 7.2 Databases under Microbial Directory



gpipe/9606/	protein	human bui

Set at blast2go.properties file

BLAST Results



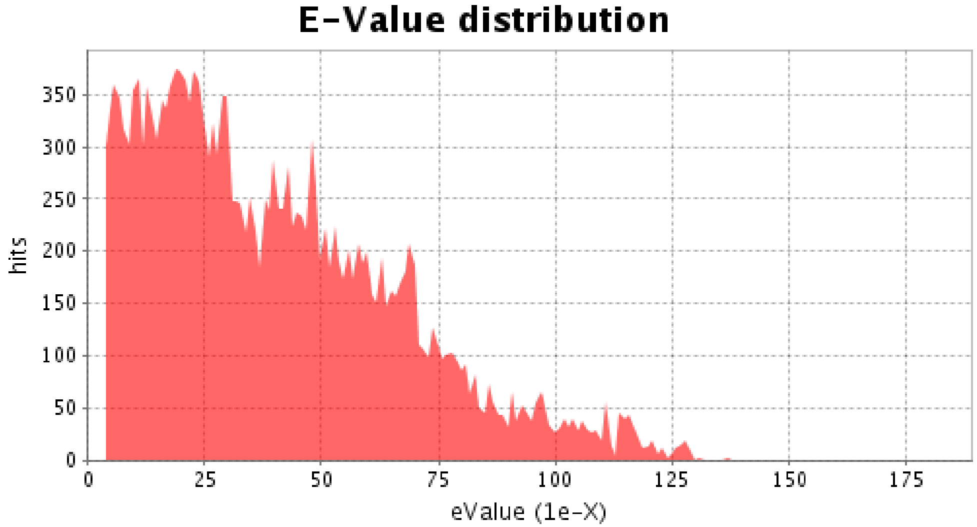
☑	nr	sequence name	seq description	length	#hits	min. eValue	sim mean	#GOs	GO IDs	Enzyme	InterPro
☑	5	C04018C12	cysteine proteinase inhibitor	663	20	25.0	80%	0		–	–
☑	6	C04018E12	protein phosphatase 2c	663	20	77.0	85%	0		–	–
☑	7	C04018G12	alpha beta fold family protein	578	20	84.0	79%	0		–	–
☑	8	C04018A02	glyoxalase i	600	20	64.0	74%	0		–	–
☑	9	C04018C02	metallothionein-like protein	625	18	14.0	74%	0		–	–
☑	10	C04018E02	haemolysin-iii related familyexpressed	612	20	32.0	72%	0		–	–
☑	11	C04018G02	protein phosphataseexpressed	645	20	97.0	81%	0		–	–
☑	12	C04018A04	---NA---	56	0	–	–	0		–	–
☑	13	C04018C04	phosphoglycerate bisphosphoglycerate mutase family protein	780	20	63.0	66%	0		–	–
☑	14	C04018E04	polyubiquitin	707	20	115.0	99%	0		–	–
☑	15	C04018G04	meiotic recombination 11	575	20	45.0	89%	0		–	–
☑	16	C04018A06	late embryogenesis-abundant protein	648	20	43.0	68%	0		–	–
☑	17	C04018G06	40s ribosomal protein s19	672	20	69.0	89%	0		–	–

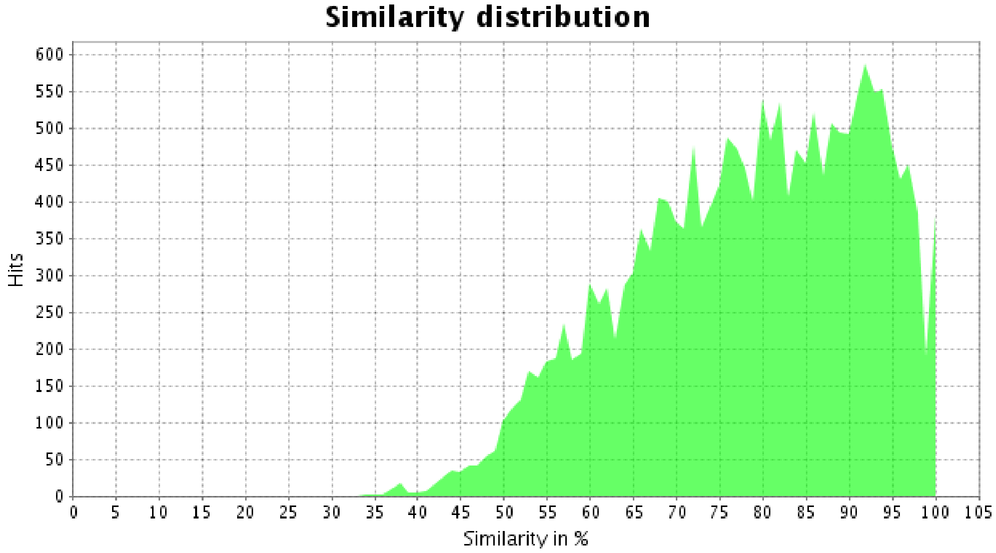
Blast2GO announces:
First international course in automated functional annotation and data-mining.
Valencia and Florida - Register now! (http://bioinfo.cipf.es/blast2gocourse)

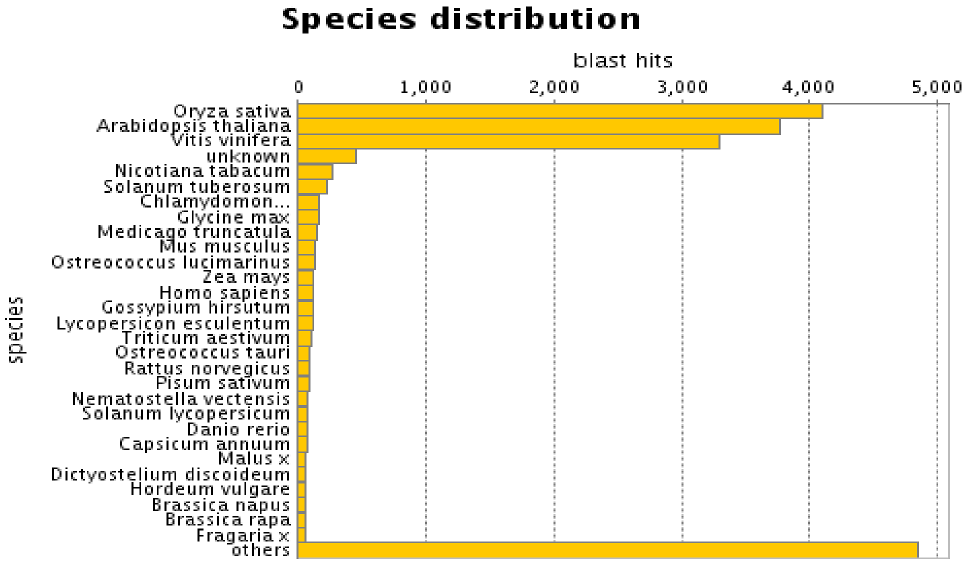
Blast2GO - Latetest Update!

- Please visit B2G-FAR: The Blast2GO Functional Annotation Repository (http://b2gfar.bioinfo.cipf.es)
- A new SELECT menu to filter/query your data-set:
    * by functional annotation
    * by description
    * by sequence name/id
    * by status (color code)

Blast Distribution Charts

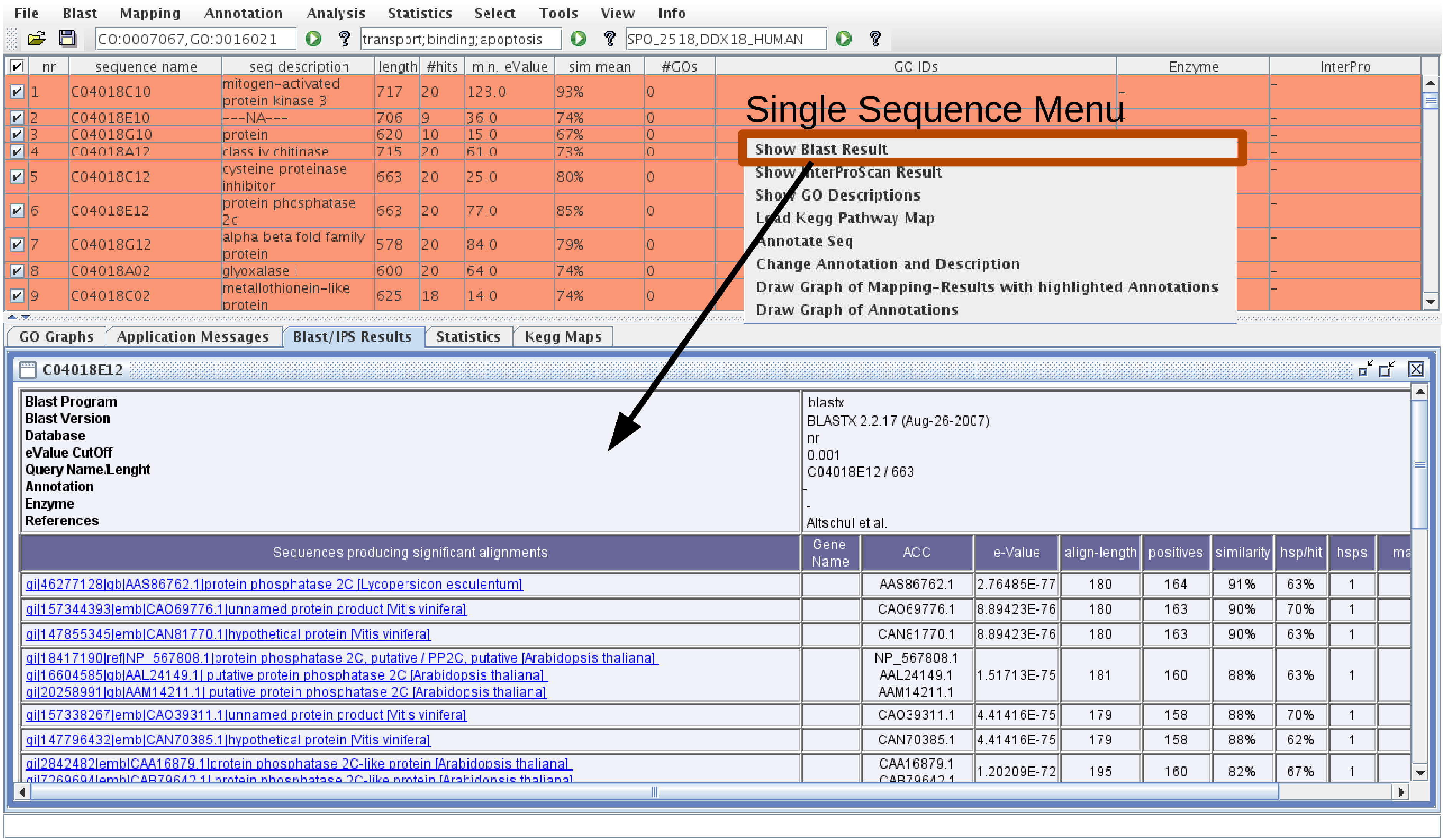






Evaluate the similarity of
your sequences with public DBs

Single Sequence Menu

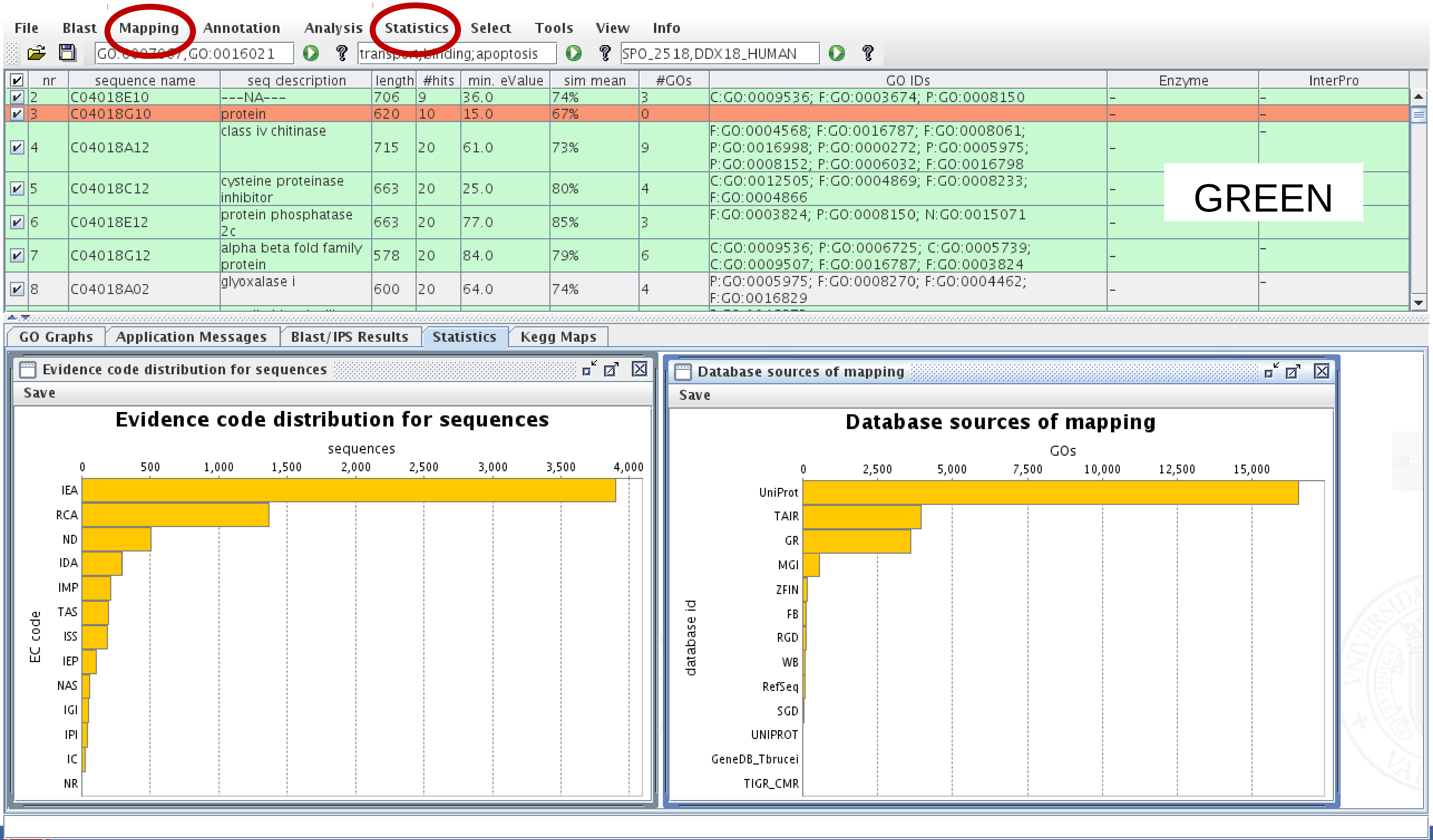


☑	nr	sequence name	seq description	length	#hits	min. eValue	sim mean	#GOs	GO IDs	Enzyme	InterPro
☑	1	C04018C10	mitogen-activated protein kinase 3	717	20	123.0	93%	0			-
☑	2	C04018E10	---NA---	706	9	36.0	74%	0			-
☑	3	C04018G10	protein	620	10	15.0	67%	0			-
☑	4	C04018A12	class iv chitinase	715	20	61.0	73%	0			-
☑	5	C04018C12	cysteine proteinase inhibitor	663	20	25.0	80%	0			-
☑	6	C04018E12	protein phosphatase 2c	663	20	77.0	85%	0			-
☑	7	C04018G12	alpha beta fold family protein	578	20	84.0	79%	0			-
☑	8	C04018A02	glyoxalase i	600	20	64.0	74%	0			-
☑	9	C04018C02	metallothionein-like protein	625	18	14.0	74%	0			-

Blast Program	blastx
Blast Version	BLASTX 2.2.17 (Aug-26-2007)
Database	nr
eValue CutOff	0.001
Query Name/Lenght	C04018E12 / 663
Annotation	-
Enzyme	-
References	Altschul et al.

Sequences producing significant alignments	Gene Name	ACC	e-Value	align-length	positives	similarity	hsp/hit	hsps	ma
gi|46277128|gb|AAS86762.1|protein phosphatase 2C [Lycopersicon esculentum]		AAS86762.1	2.76485E-77	180	164	91%	63%	1	
gi|157344393|emb|CAO69776.1|unnamed protein product [Vitis vinifera]		CAO69776.1	8.89423E-76	180	163	90%	70%	1	
gi|147855345|emb|CAN81770.1|hypothetical protein [Vitis vinifera]		CAN81770.1	8.89423E-76	180	163	90%	63%	1	
gi|18417190|ref|NP_567808.1|protein phosphatase 2C, putative / PP2C, putative [Arabidopsis thaliana] gi|16604585|gb|AAL24149.1| putative protein phosphatase 2C [Arabidopsis thaliana] gi|20258991|gb|AAM14211.1| putative protein phosphatase 2C [Arabidopsis thaliana]		NP_567808.1 AAL24149.1 AAM14211.1	1.51713E-75	181	160	88%	63%	1	
gi|157338267|emb|CAO39311.1|unnamed protein product [Vitis vinifera]		CAO39311.1	4.41416E-75	179	158	88%	70%	1	
gi|147796432|emb|CAN70385.1|hypothetical protein [Vitis vinifera]		CAN70385.1	4.41416E-75	179	158	88%	62%	1	
gi|2842482|emb|CAA16879.1|protein phosphatase 2C-like protein [Arabidopsis thaliana] gi|7269694|emb|CAB79642.1| protein phosphatase 2C-like protein [Arabidopsis thaliana]		CAA16879.1 CAB79642.1	1.20209E-72	195	160	82%	67%	1	

Mapping Results



☑	nr	sequence name	seq description	length	#hits	min. eValue	sim mean	#GOs	GO IDs	Enzyme	InterPro
☑	2	C04018E10	---NA---	706	9	36.0	74%	3	C:GO:0009536; F:GO:0003674; P:GO:0008150	-	-
☑	3	C04018G10	protein	620	10	15.0	67%	0		-	-
☑	4	C04018A12	class iv chitinase	715	20	61.0	73%	9	F:GO:0004568; F:GO:0016787; F:GO:0008061; P:GO:0016998; P:GO:0000272; P:GO:0005975; P:GO:0008152; P:GO:0006032; F:GO:0016798	-	-
☑	5	C04018C12	cysteine proteinase inhibitor	663	20	25.0	80%	4	C:GO:0012505; F:GO:0004869; F:GO:0008233; F:GO:0004866	-	
☑	6	C04018E12	protein phosphatase 2c	663	20	77.0	85%	3	F:GO:0003824; P:GO:0008150; N:GO:0015071	-	
☑	7	C04018G12	alpha beta fold family protein	578	20	84.0	79%	6	C:GO:0009536; P:GO:0006725; C:GO:0005739; C:GO:0009507; F:GO:0016787; F:GO:0003824	-	-
☑	8	C04018A02	glyoxalase i	600	20	64.0	74%	4	P:GO:0005975; F:GO:0008270; F:GO:0004462; F:GO:0016829	-	-

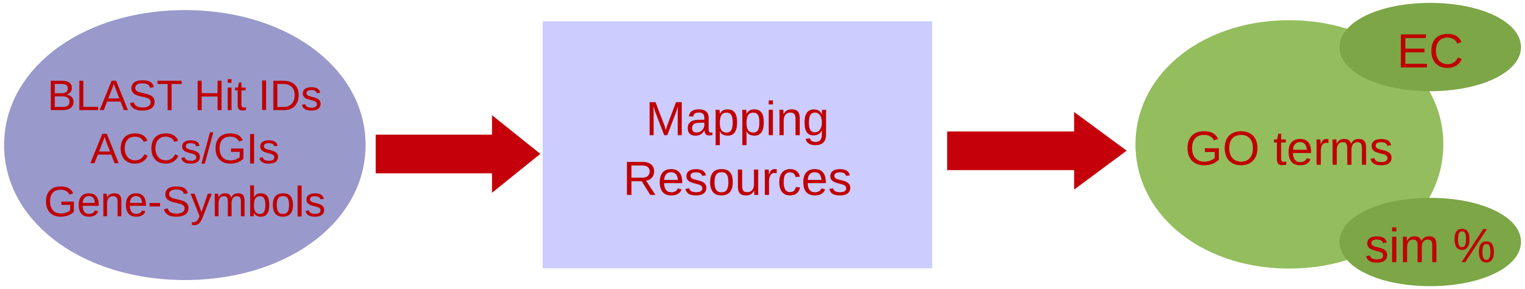
GREEN

Resources for mapping

Gene Ontology Database

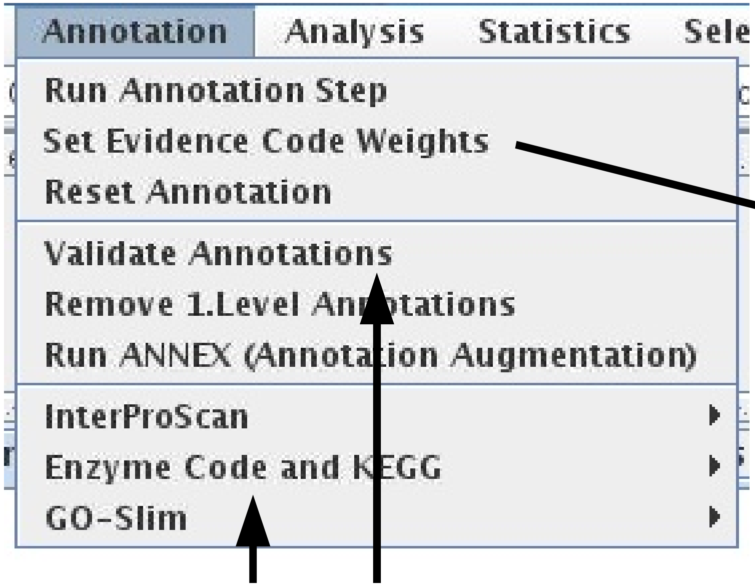
NCBI data-files:
- gene2accession (4 079 414 entries)
- gene_info (1 635 614 entries)

Protein Information Resource (PIR):
- Non-Redundant Reference Protein Database including PSD, UniProt, Swiss-Prot, TrEMBL, RefSeq, GenPept and PDB



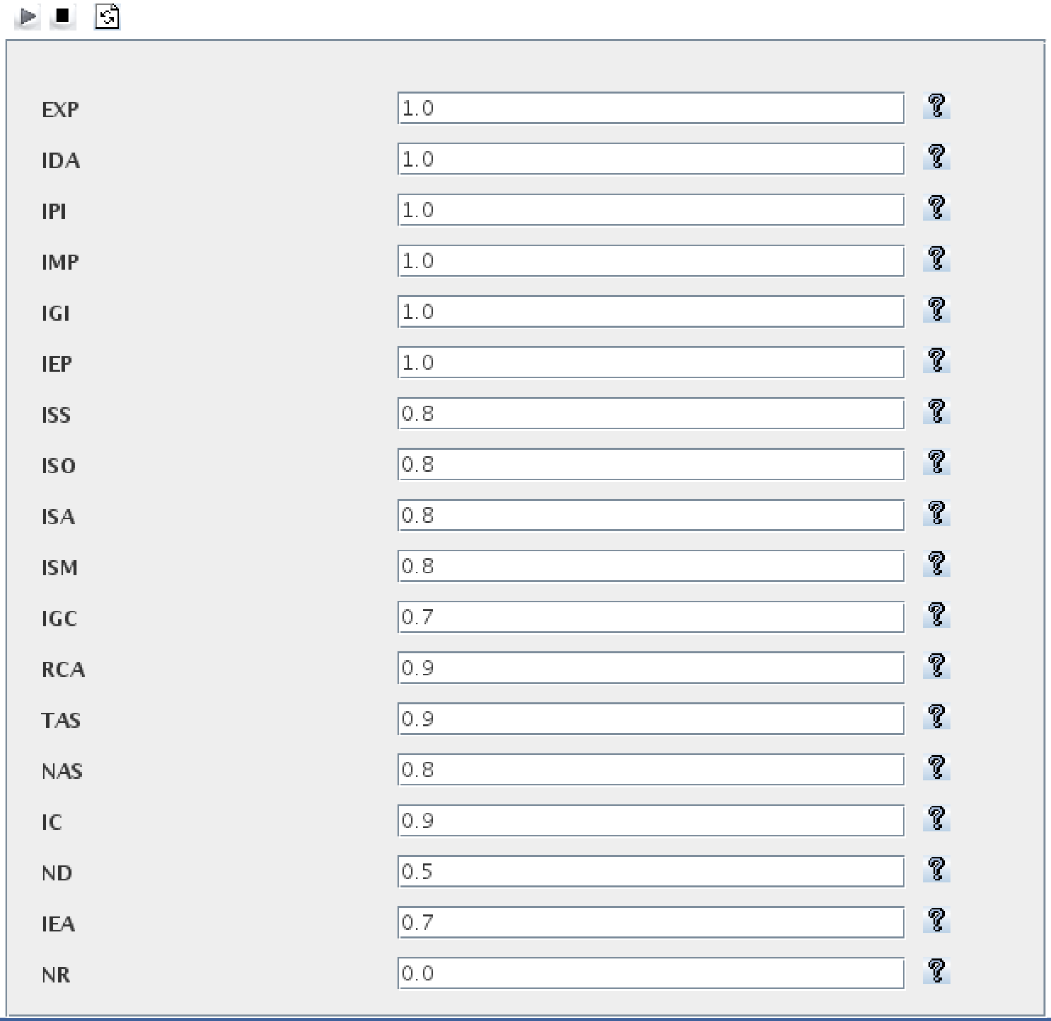
Annotation Menu

BLAST based annotation



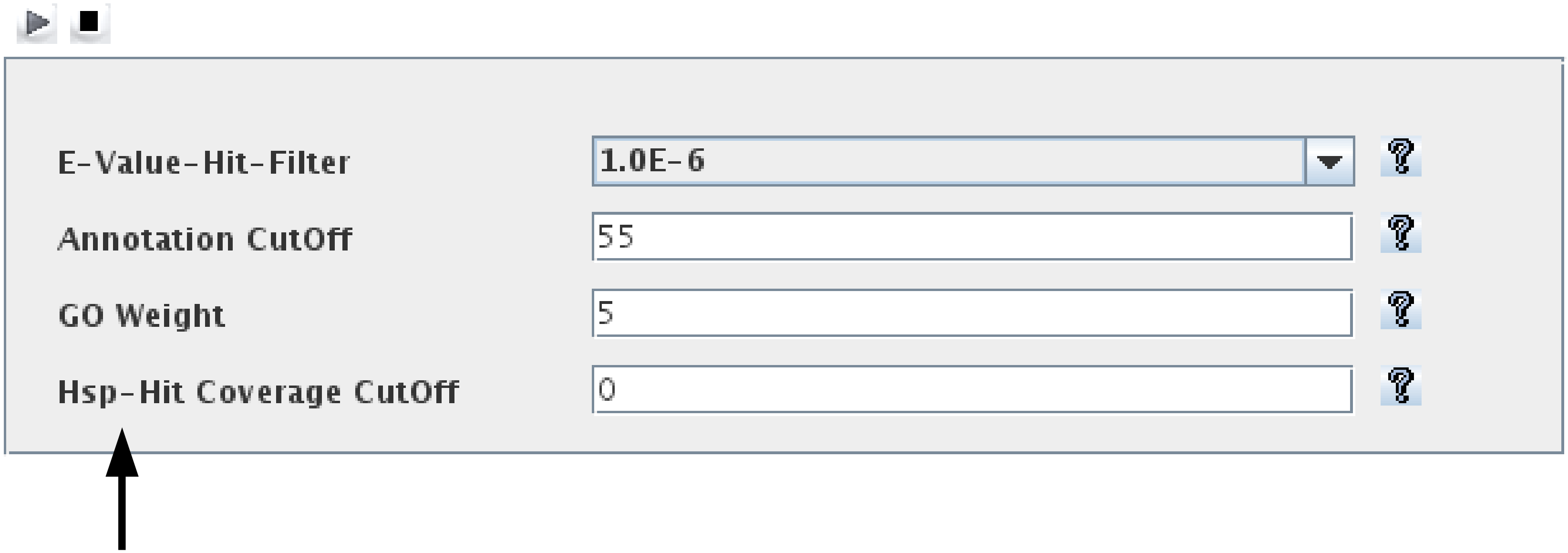
Other Annotation modes

Validation and Annex



EXP	1.0
IDA	1.0
IPI	1.0
IMP	1.0
IGI	1.0
IEP	1.0
ISS	0.8
ISO	0.8
ISA	0.8
ISM	0.8
IGC	0.7
RCA	0.9
TAS	0.9
NAS	0.8
IC	0.9
ND	0.5
IEA	0.7
NR	0.0

Annotation

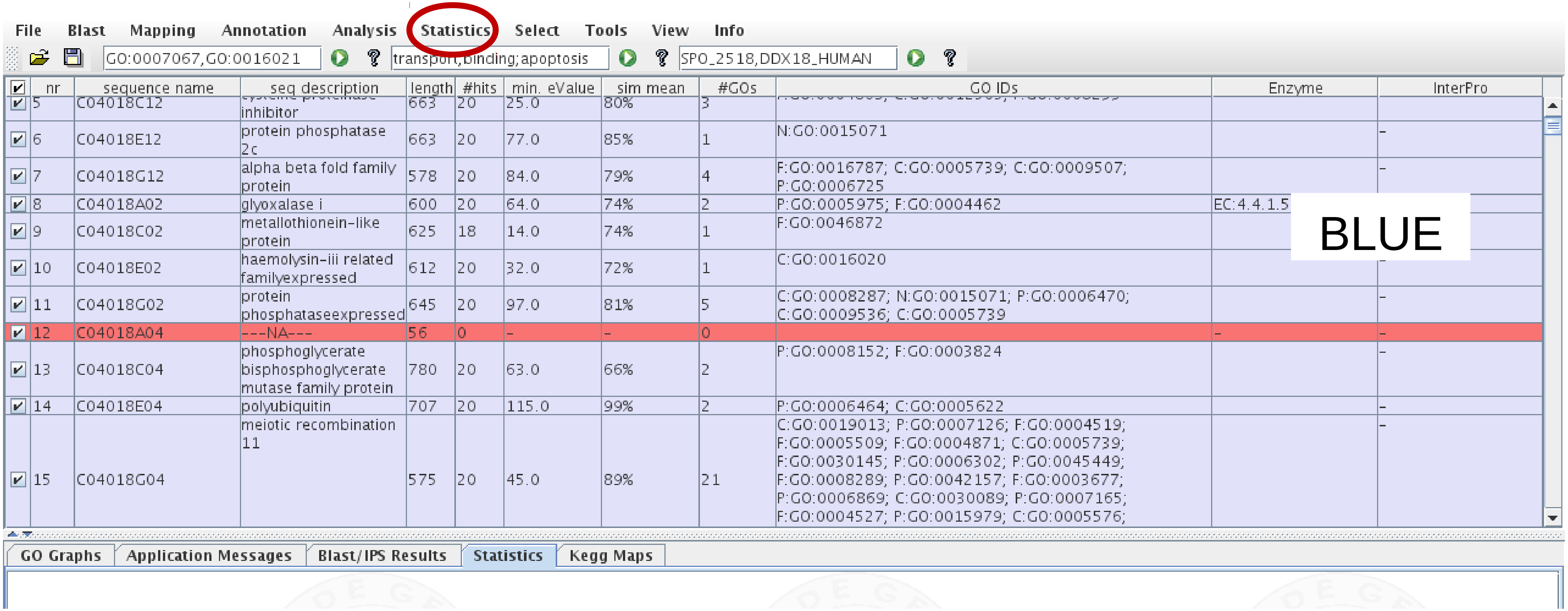


E-Value-Hit-Filter	1.0E-6
Annotation CutOff	55
GO Weight	5
Hsp-Hit Coverage CutOff	0

Allows to set a minimum percentage of the HIT sequence which should be expand by the QUERY sequence

This helps to avoid the problem of cis-annotation

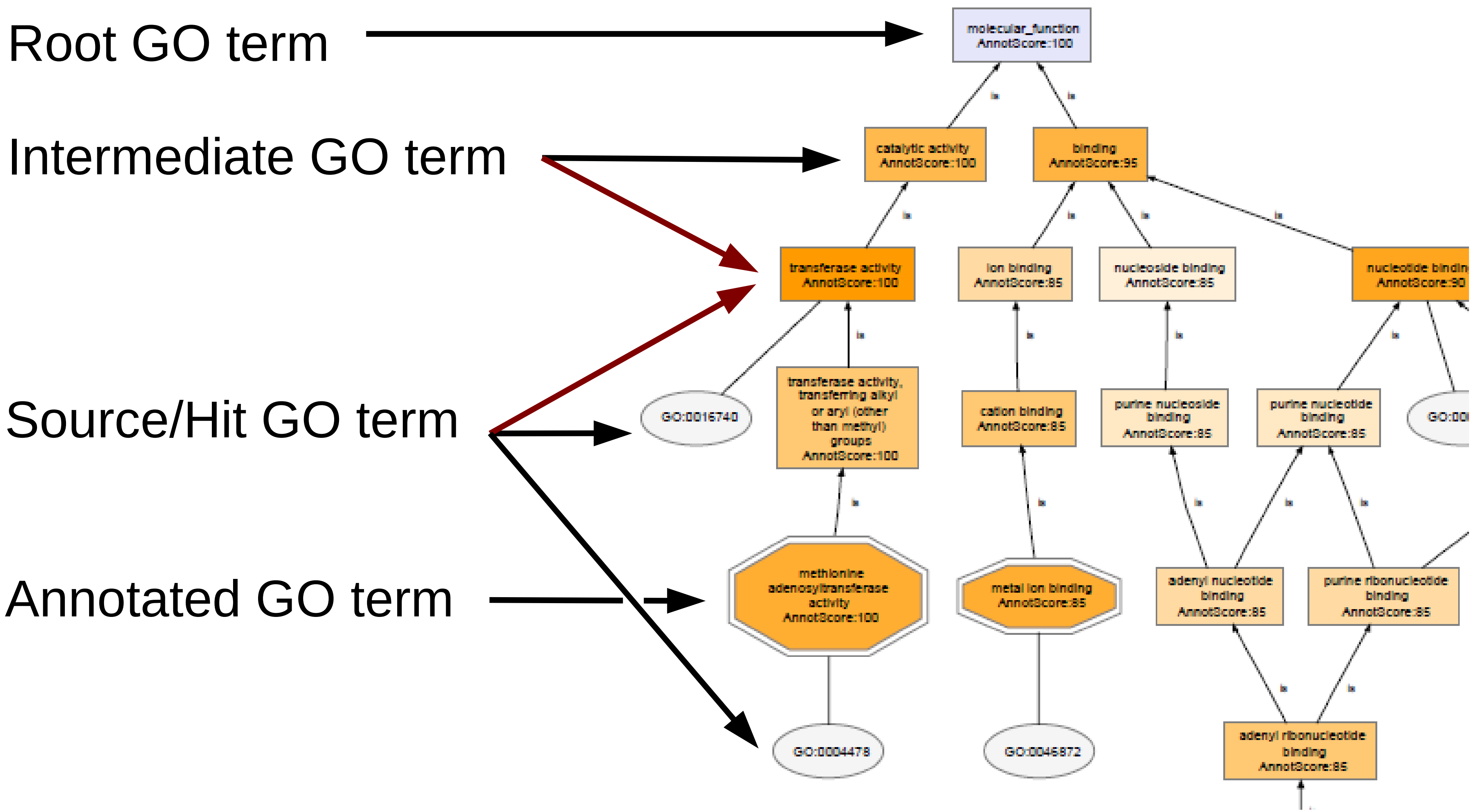
Annotation Result



☑	nr	sequence name	seq description	length	#hits	min. eValue	sim mean	#GOs	GO IDs	Enzyme	InterPro
☑	5	C04018C12	cysteine proteinase inhibitor	663	20	25.0	80%	3			
☑	6	C04018E12	protein phosphatase 2c	663	20	77.0	85%	1	N:GO:0015071		-
☑	7	C04018G12	alpha beta fold family protein	578	20	84.0	79%	4	F:GO:0016787; C:GO:0005739; C:GO:0009507; P:GO:0006725		-
☑	8	C04018A02	glyoxalase i	600	20	64.0	74%	2	P:GO:0005975; F:GO:0004462	EC:4.4.1.5	
☑	9	C04018C02	metallothionein-like protein	625	18	14.0	74%	1	F:GO:0046872		
☑	10	C04018E02	haemolysin-iii related familyexpressed	612	20	32.0	72%	1	C:GO:0016020		-
☑	11	C04018G02	protein phosphataseexpressed	645	20	97.0	81%	5	C:GO:0008287; N:GO:0015071; P:GO:0006470; C:GO:0009536; C:GO:0005739		-
☑	12	C04018A04	---NA---	56	0	-	-	0		-	-
☑	13	C04018C04	phosphoglycerate bisphosphoglycerate mutase family protein	780	20	63.0	66%	2	P:GO:0008152; F:GO:0003824		-
☑	14	C04018E04	polyubiquitin	707	20	115.0	99%	2	P:GO:0006464; C:GO:0005622		-
☑	15	C04018G04	meiotic recombination 11	575	20	45.0	89%	21	C:GO:0019013; P:GO:0007126; F:GO:0004519; F:GO:0005509; F:GO:0004871; C:GO:0005739; F:GO:0030145; P:GO:0006302; P:GO:0045449; F:GO:0008289; P:GO:0042157; F:GO:0003677; P:GO:0006869; C:GO:0030089; P:GO:0007165; F:GO:0004527; P:GO:0015979; C:GO:0005576;		-

BLUE

Graph Visualization



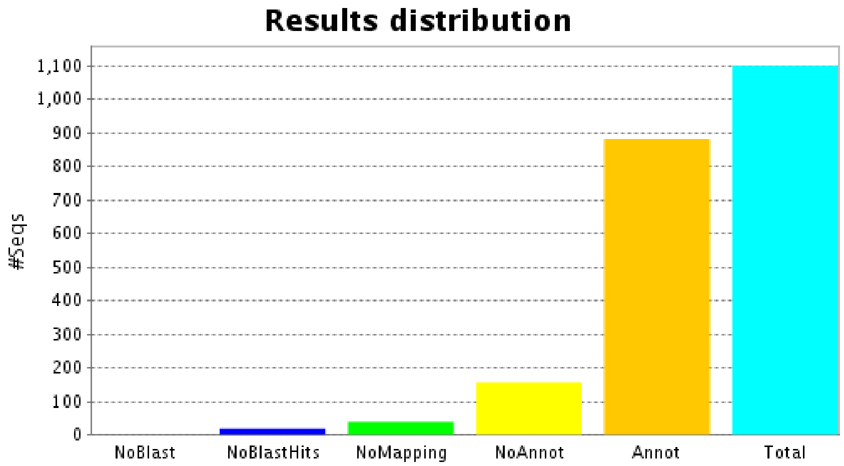
Root GO term

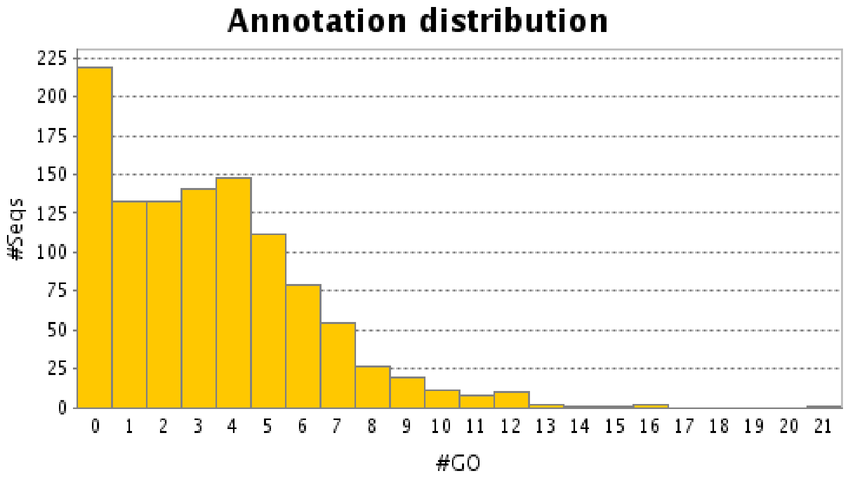
Intermediate GO term

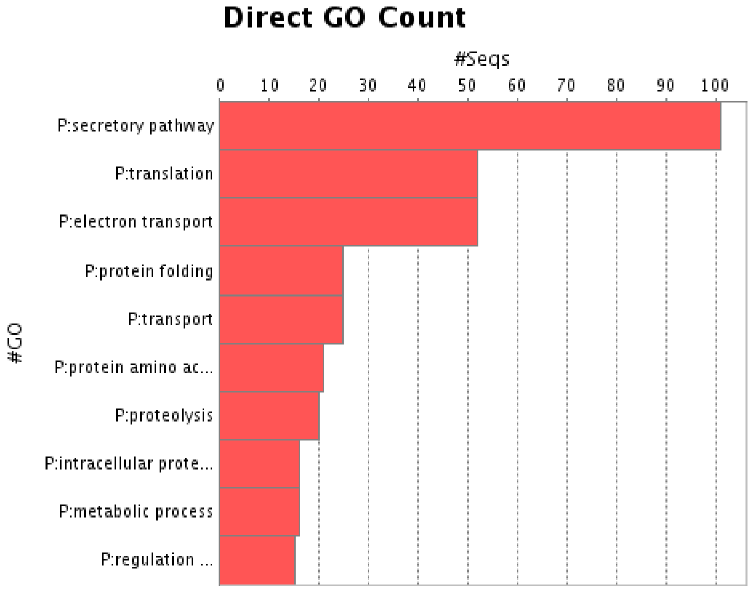
Source/Hit GO term

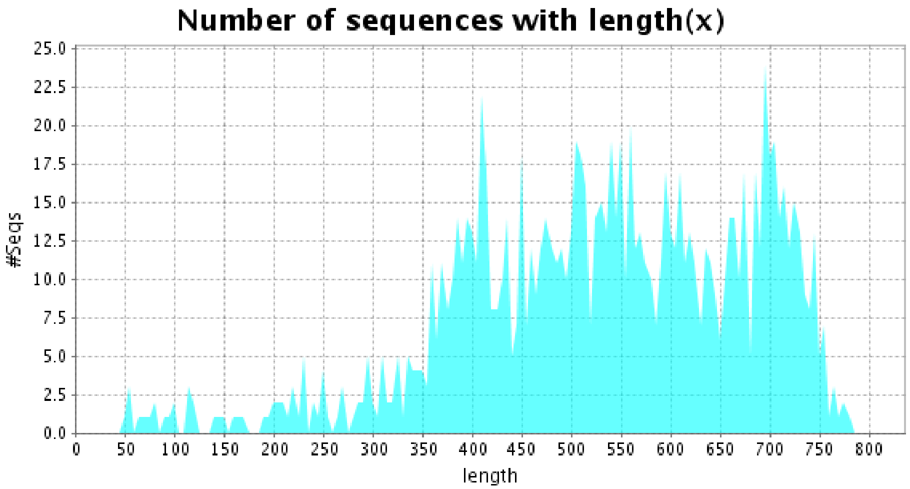
Annotated GO term

Annotation Charts

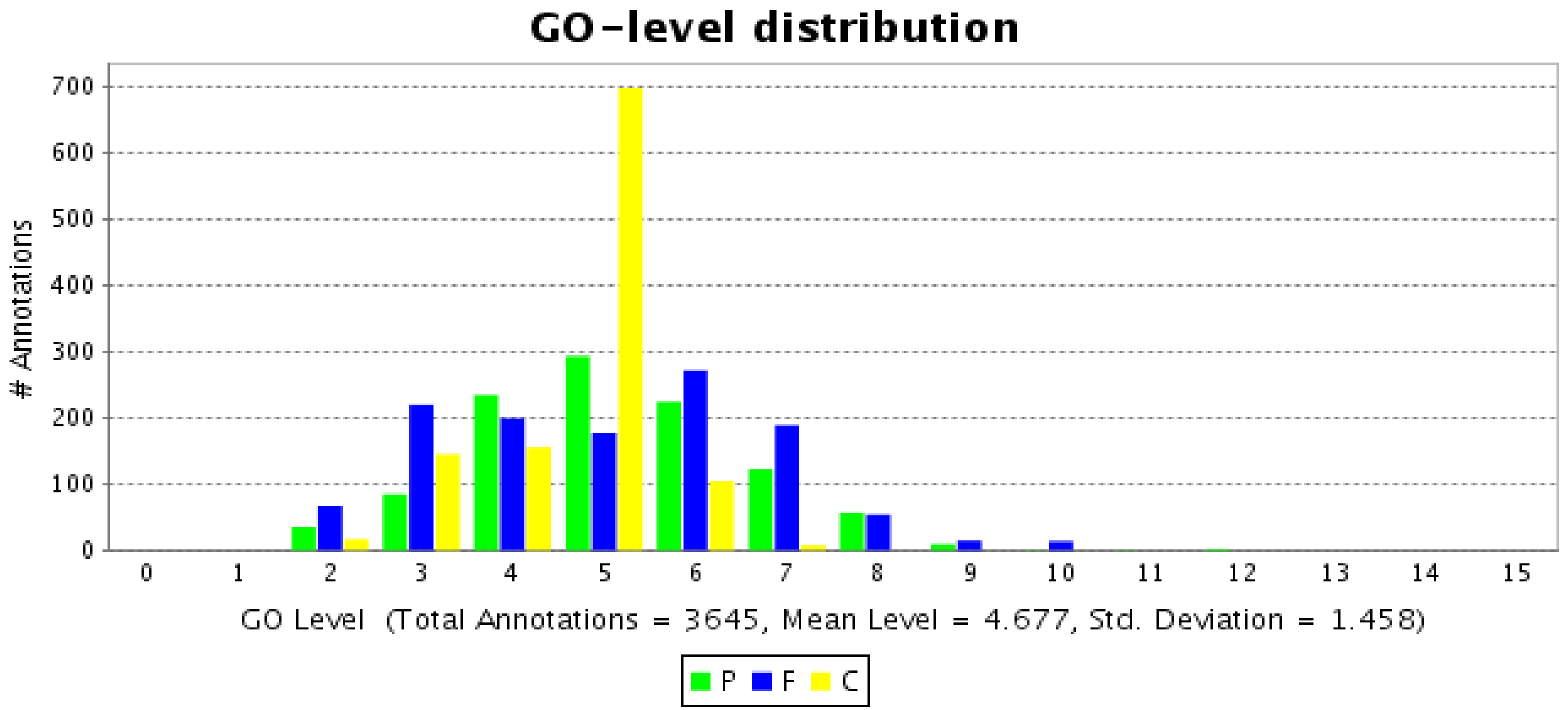

Results distribution


Annotation distribution


Direct GO Count

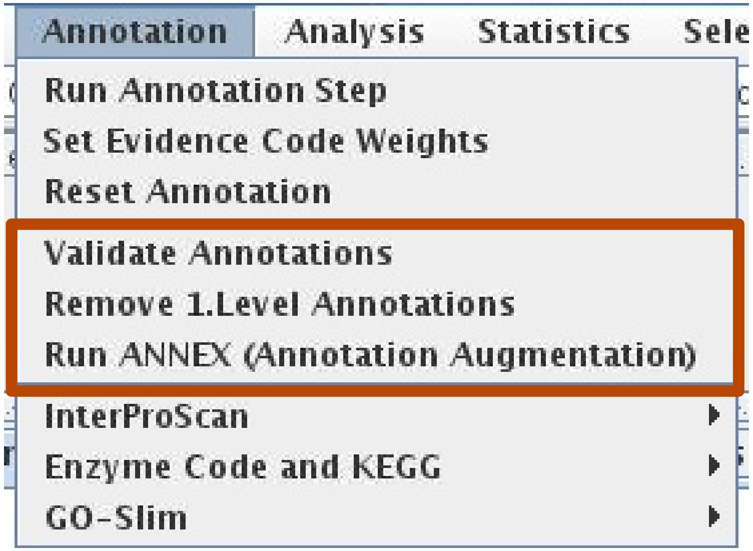

Number of sequences with length(x)

Annotation Charts

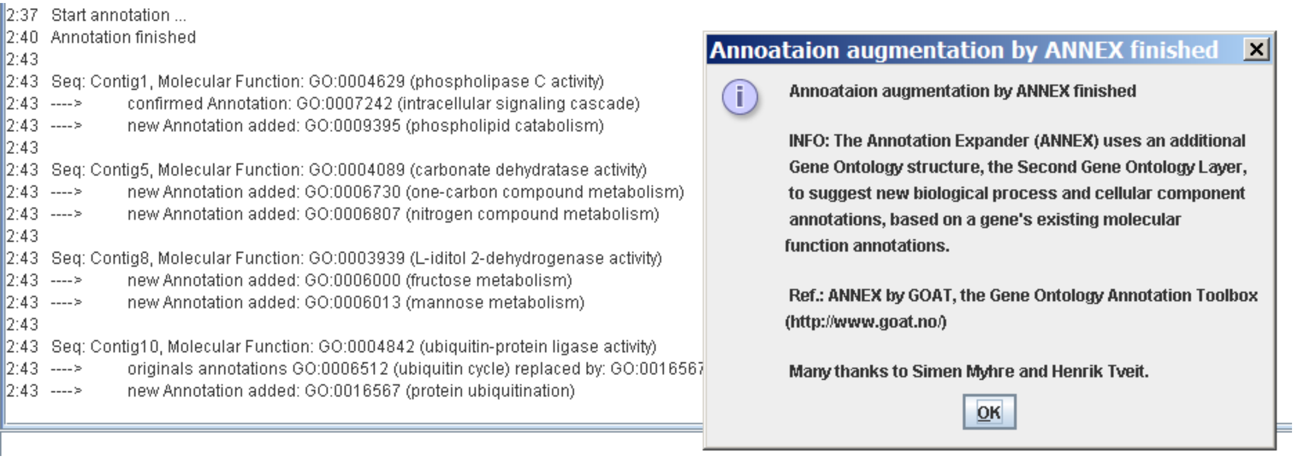


Commonly, level 5 is the most abundant specificity level in the Gene Ontology

Additional Annotation: ANNEX



Recovers implicit biological process and cellular component GO terms based on molecular function annotations



```
2:37  Start annotation ...
2:40  Annotation finished
2:43
2:43  Seq: Contig1, Molecular Function: GO:0004629 (phospholipase C activity)
2:43  ---->       confirmed Annotation: GO:0007242 (intracellular signaling cascade)
2:43  ---->       new Annotation added: GO:0009395 (phospholipid catabolism)
2:43
2:43  Seq: Contig5, Molecular Function: GO:0004089 (carbonate dehydratase activity)
2:43  ---->       new Annotation added: GO:0006730 (one-carbon compound metabolism)
2:43  ---->       new Annotation added: GO:0006807 (nitrogen compound metabolism)
2:43
2:43  Seq: Contig8, Molecular Function: GO:0003939 (L-iditol 2-dehydrogenase activity)
2:43  ---->       new Annotation added: GO:0006000 (fructose metabolism)
2:43  ---->       new Annotation added: GO:0006013 (mannose metabolism)
2:43
2:43  Seq: Contig10, Molecular Function: GO:0004842 (ubiquitin-protein ligase activity)
2:43  ---->       originals annotations GO:0006512 (ubiquitin cycle) replaced by: GO:0016567
2:43  ---->       new Annotation added: GO:0016567 (protein ubiquitination)
```

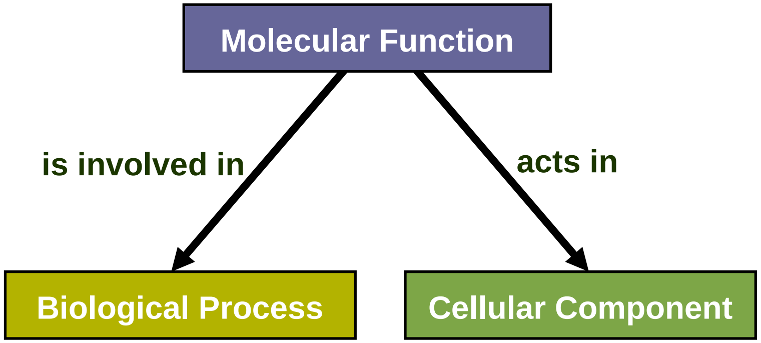
Annoataion augmentation by ANNEX finished

Annoataion augmentation by ANNEX finished

INFO: The Annotation Expander (ANNEX) uses an additional Gene Ontology structure, the Second Gene Ontology Layer, to suggest new biological process and cellular component annotations, based on a gene's existing molecular function annotations.

Ref.: ANNEX by GOAT, the Gene Ontology Annotation Toolbox (http://www.goat.no/)

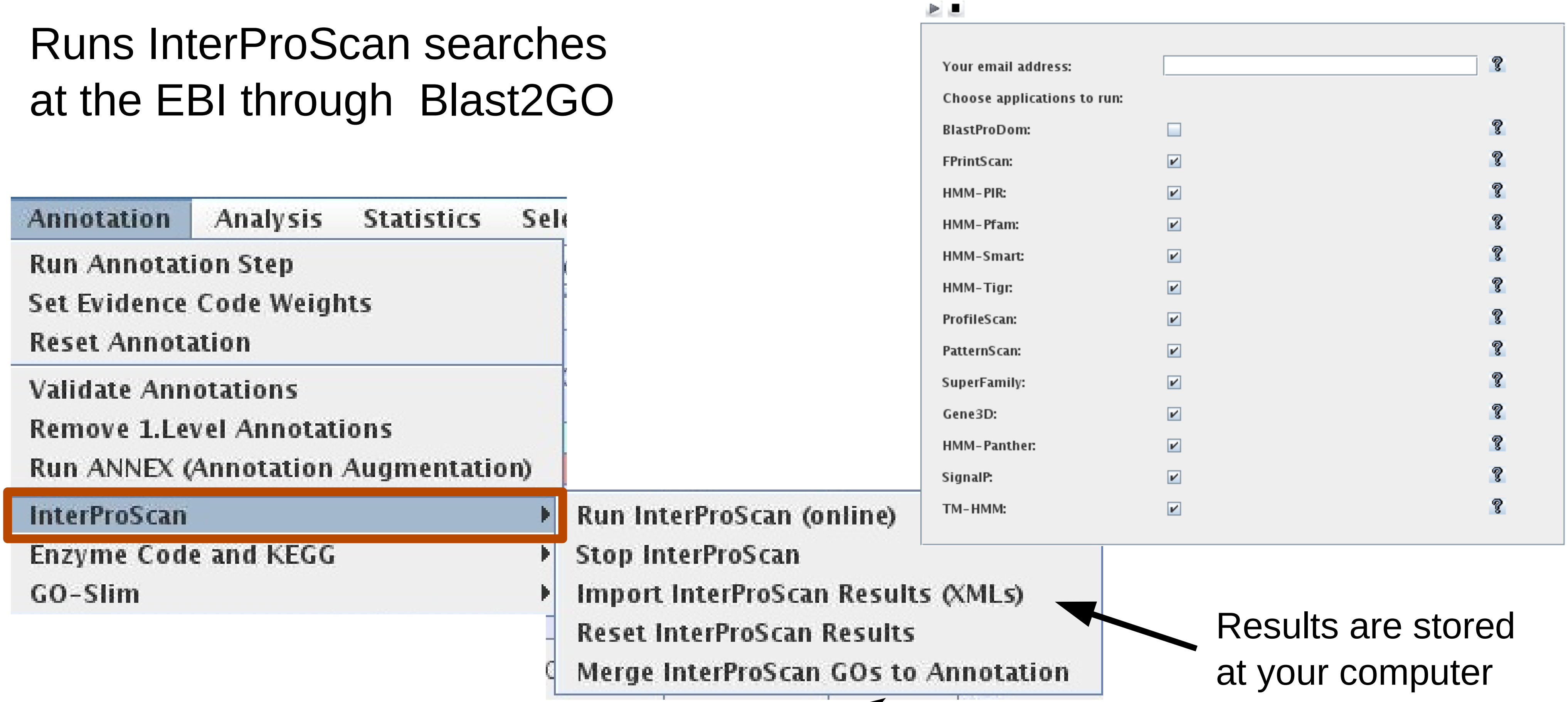
Many thanks to Simen Myhre and Henrik Tveit.



Myhre *et al*, Bioinformatics 2006

Additional Annotation: InterProScan

Runs InterProScan searches at the EBI through Blast2GO


Annotation
Analysis
Statistics
Sel
Run Annotation Step
Set Evidence Code Weights
Reset Annotation
Validate Annotations
Remove 1.Level Annotations
Run ANNEX (Annotation Augmentation)
InterProScan
Enzyme Code and KEGG
GO-Slim
Run InterProScan (online)
Stop InterProScan
Import InterProScan Results (XMLs)
Reset InterProScan Results
Merge InterProScan GOs to Annotation

Your email address:

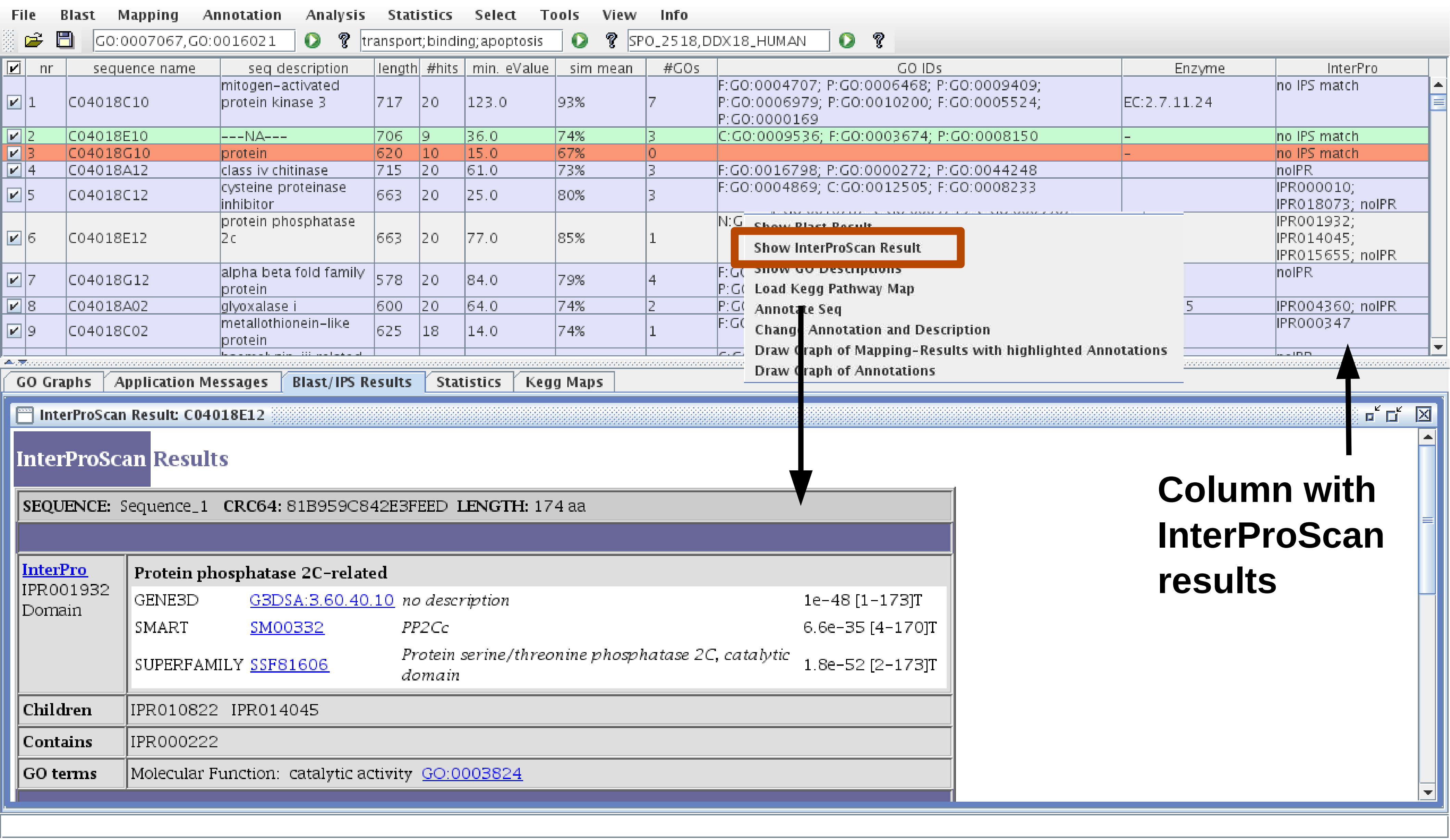
Choose applications to run:

BlastProDom:	□
FPrintScan:	☑
HMM-PIR:	☑
HMM-Pfam:	☑
HMM-Smart:	☑
HMM-Tigr:	☑
ProfileScan:	☑
PatternScan:	☑
SuperFamily:	☑
Gene3D:	☑
HMM-Panther:	☑
SignalP:	☑
TM-HMM:	☑

Results are stored at your computer as XML files. You can upload them later

Once you have completed your InterPro annotation, results can be transformed to GO terms and merged to Blast annotation

InterProScan Results



☑	nr	sequence name	seq description	length	#hits	min. eValue	sim mean	#GOs	GO IDs	Enzyme	InterPro
☑	1	C04018C10	mitogen-activated protein kinase 3	717	20	123.0	93%	7	F:GO:0004707; P:GO:0006468; P:GO:0009409; P:GO:0006979; P:GO:0010200; F:GO:0005524; P:GO:0000169	EC:2.7.11.24	no IPS match
☑	2	C04018E10	---NA---	706	9	36.0	74%	3	C:GO:0009536; F:GO:0003674; P:GO:0008150	-	no IPS match
☑	3	C04018G10	protein	620	10	15.0	67%	0		-	no IPS match
☑	4	C04018A12	class iv chitinase	715	20	61.0	73%	3	F:GO:0016798; P:GO:0000272; P:GO:0044248		noIPR
☑	5	C04018C12	cysteine proteinase inhibitor	663	20	25.0	80%	3	F:GO:0004869; C:GO:0012505; F:GO:0008233		IPR000010; IPR018073; noIPR
☑	6	C04018E12	protein phosphatase 2c	663	20	77.0	85%	1	N:G		IPR001932; IPR014045; IPR015655; noIPR
☑	7	C04018G12	alpha beta fold family protein	578	20	84.0	79%	4	F:GO P:GO		noIPR
☑	8	C04018A02	glyoxalase i	600	20	64.0	74%	2	P:GO	5	IPR004360; noIPR
☑	9	C04018C02	metallothionein-like protein	625	18	14.0	74%	1	F:GO		IPR000347

InterProScan Result: C04018E12

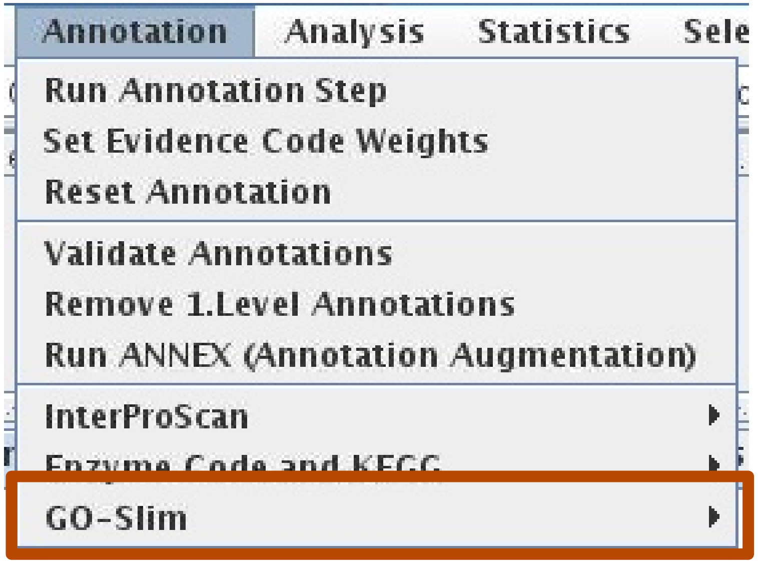
InterProScan Results

SEQUENCE: Sequence_1 **CRC64:** 81B959C842E3FEED **LENGTH:** 174 aa

InterPro IPR001932 Domain	**Protein phosphatase 2C-related**			
	GENE3D	G3DSA:3.60.40.10	*no description*	1e-48 [1-173]T
	SMART	SM00332	*PP2Cc*	6.6e-35 [4-170]T
	SUPERFAMILY	SSF81606	*Protein serine/threonine phosphatase 2C, catalytic domain*	1.8e-52 [2-173]T
Children	IPR010822 IPR014045			
Contains	IPR000222			
GO terms	Molecular Function: catalytic activity GO:0003824			

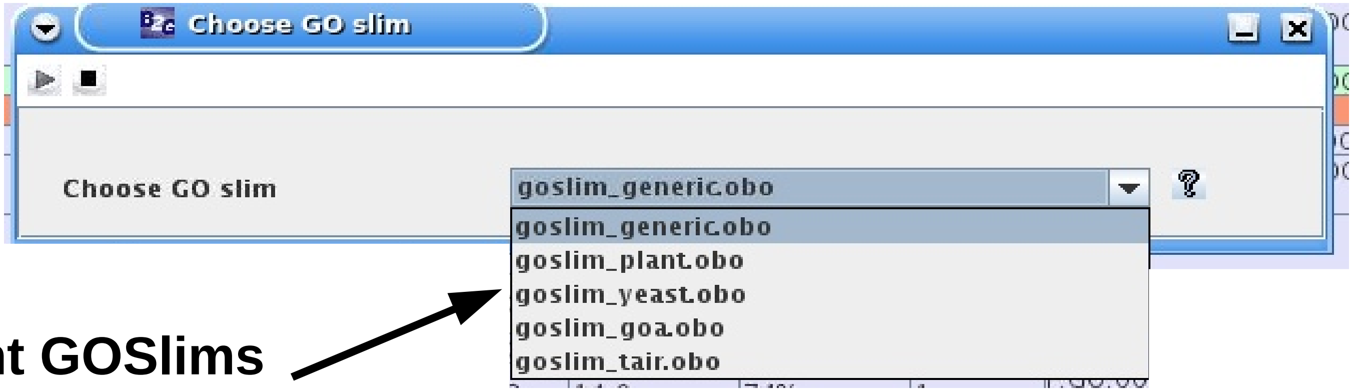
Column with InterProScan results

Additional Annotation: GOSlim



GOSlim is a reduction of the Gene Ontology to a more reduced vocabulary → Helps to summarize information

After GOSlim transformation sequences get YELLOW

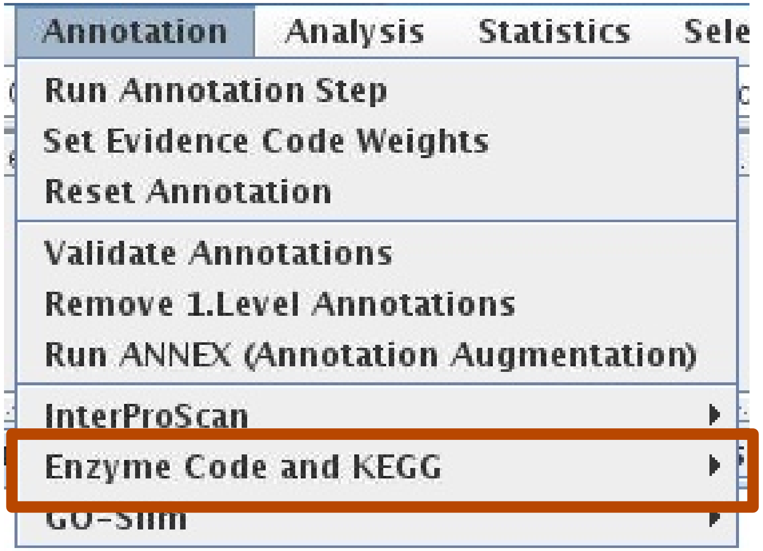


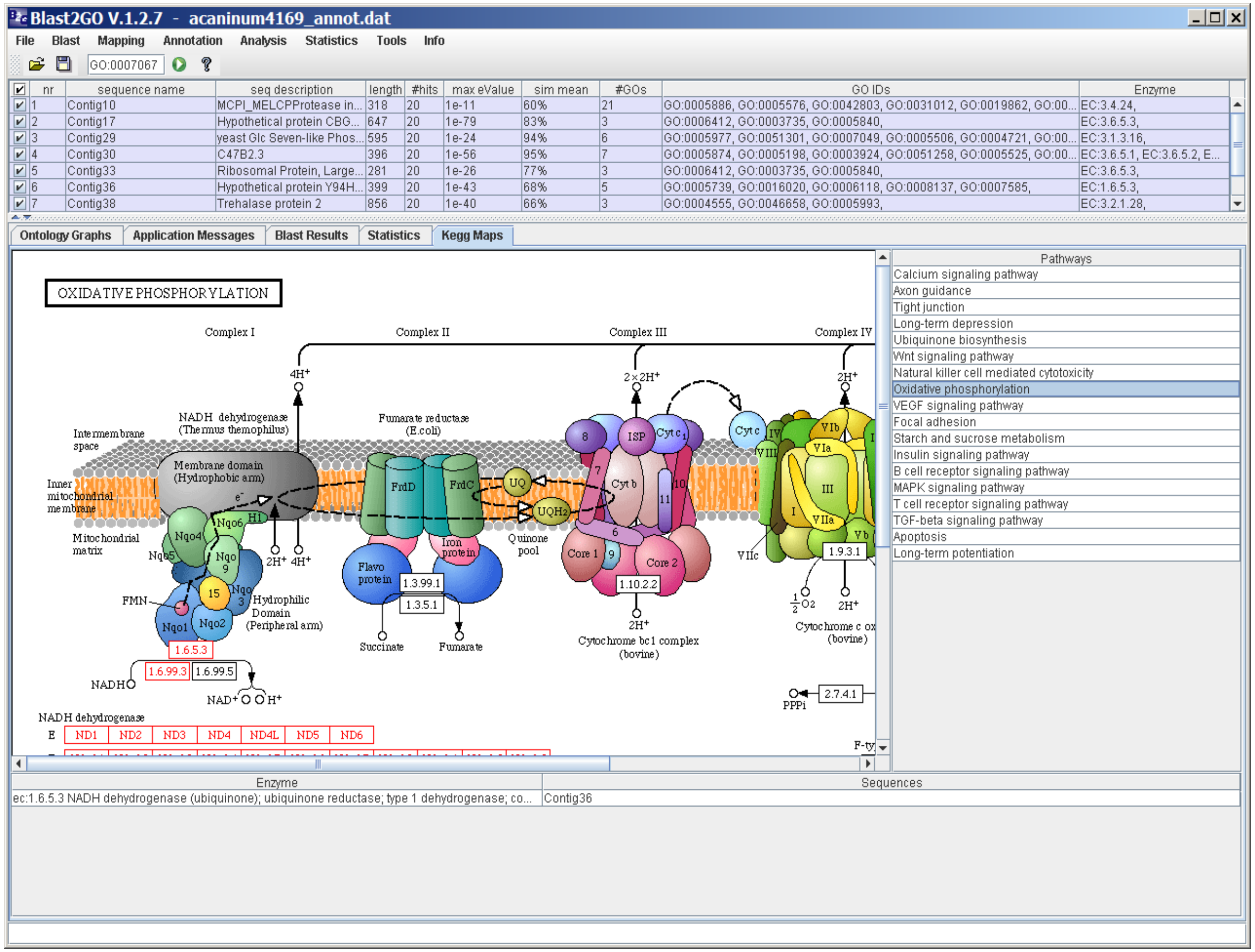
Choose GO slim: goslim_generic.obo

Different GOSlims available at Blast2GO

Enzyme annotation and Kegg Maps

GO → Enzyme Codes → KEGG maps



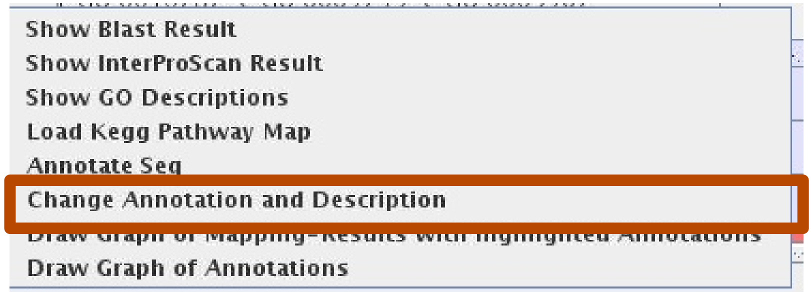


☑	nr	sequence name	seq description	length	#hits	max eValue	sim mean	#GOs	GO IDs	Enzyme
☑	1	Contig10	MCPI_MELCPProtease in...	318	20	1e-11	60%	21	GO:0005886, GO:0005576, GO:0042803, GO:0031012, GO:0019862, GO:00...	EC:3.4.24,
☑	2	Contig17	Hypothetical protein CBG...	647	20	1e-79	83%	3	GO:0006412, GO:0003735, GO:0005840,	EC:3.6.5.3,
☑	3	Contig29	yeast Glc Seven-like Phos...	595	20	1e-24	94%	6	GO:0005977, GO:0051301, GO:0007049, GO:0005506, GO:0004721, GO:00...	EC:3.1.3.16,
☑	4	Contig30	C47B2.3	396	20	1e-56	95%	7	GO:0005874, GO:0005198, GO:0003924, GO:0051258, GO:0005525, GO:00...	EC:3.6.5.1, EC:3.6.5.2, E...
☑	5	Contig33	Ribosomal Protein, Large...	281	20	1e-26	77%	3	GO:0006412, GO:0003735, GO:0005840,	EC:3.6.5.3,
☑	6	Contig36	Hypothetical protein Y94H...	399	20	1e-43	68%	5	GO:0005739, GO:0016020, GO:0006118, GO:0008137, GO:0007585,	EC:1.6.5.3,
☑	7	Contig38	Trehalase protein 2	856	20	1e-40	66%	3	GO:0004555, GO:0046658, GO:0005993,	EC:3.2.1.28,

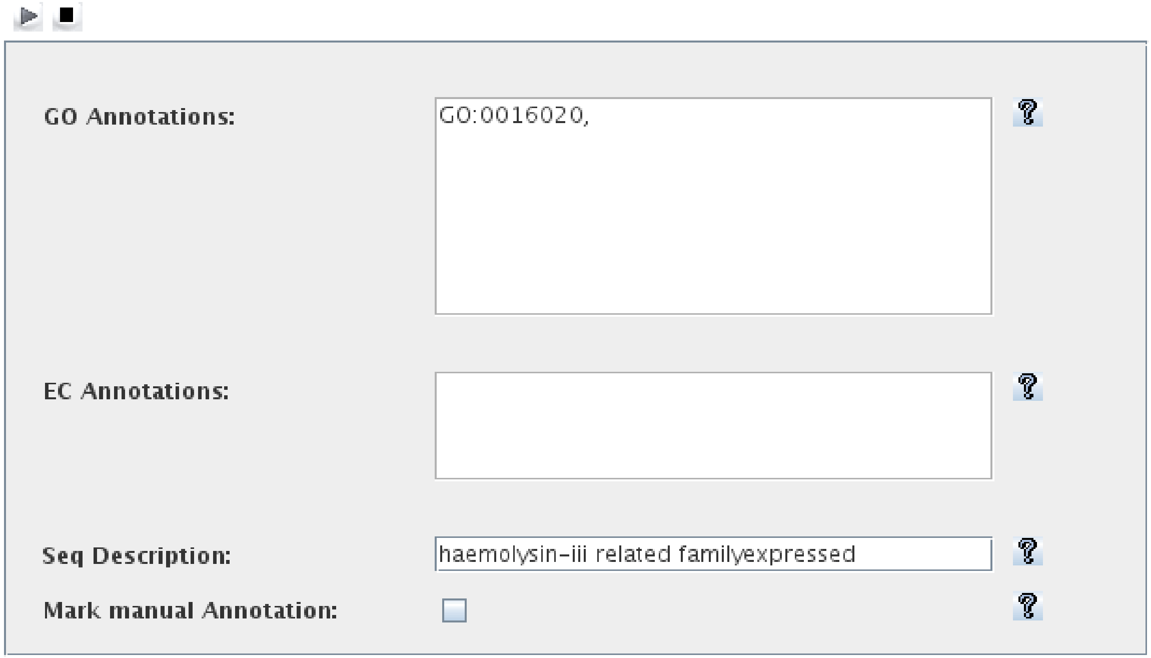
Pathways
Calcium signaling pathway
Axon guidance
Tight junction
Long-term depression
Ubiquinone biosynthesis
Wnt signaling pathway
Natural killer cell mediated cytotoxicity
Oxidative phosphorylation
VEGF signaling pathway
Focal adhesion
Starch and sucrose metabolism
Insulin signaling pathway
B cell receptor signaling pathway
MAPK signaling pathway
T cell receptor signaling pathway
TGF-beta signaling pathway
Apoptosis
Long-term potentiation

Enzyme	Sequences
ec:1.6.5.3 NADH dehydrogenase (ubiquinone); ubiquinone reductase; type 1 dehydrogenase; co...	Contig36

Additional Annotation: Manual Curation


Show Blast Result
Show InterProScan Result
Show GO Descriptions
Load Kegg Pathway Map
Annotate Seq
Change Annotation and Description
Draw Graph of Mapping-Results with highlighted Annotations
Draw Graph of Annotations

You can modify manually annotation of particular sequences



GO Annotations: GO:0016020,

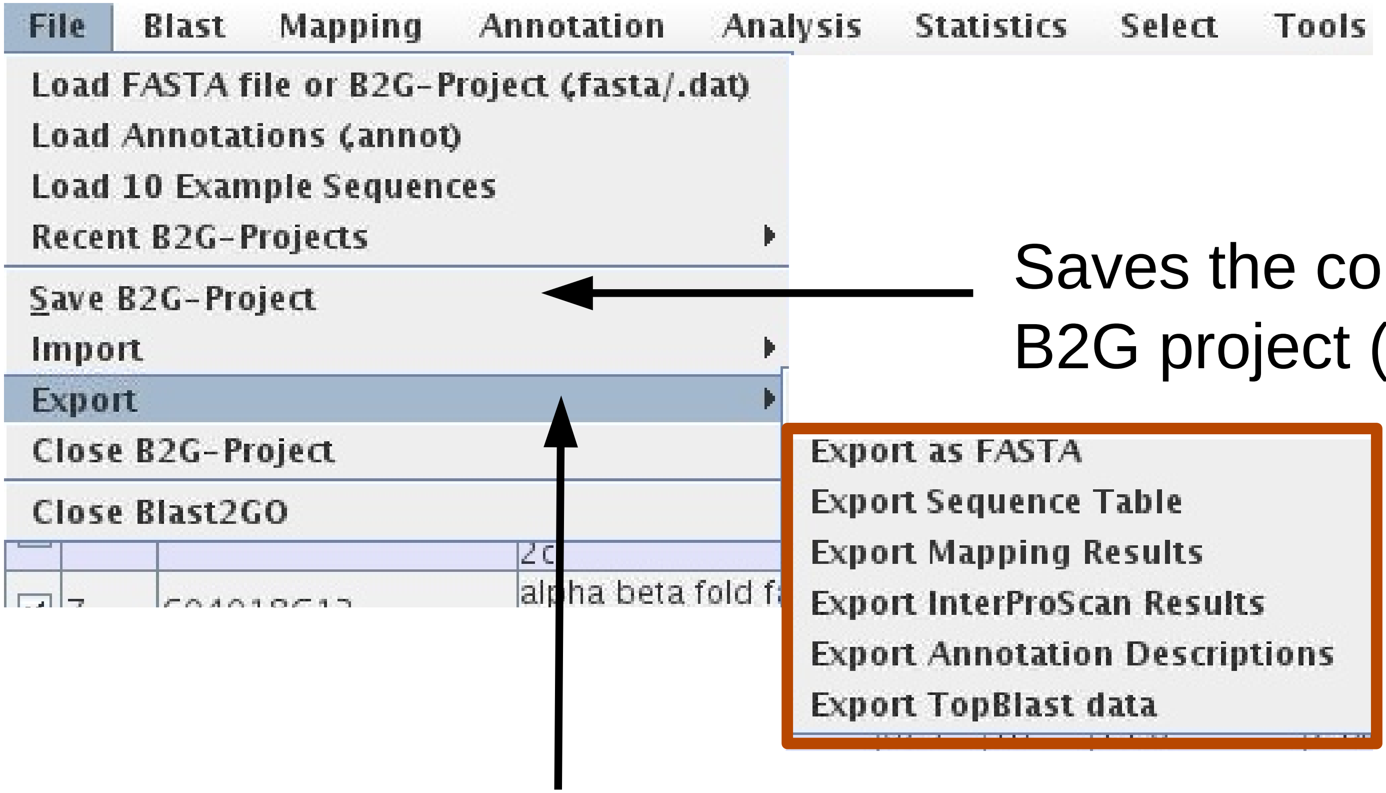
EC Annotations:

Seq Description: haemolysin-iii related familyexpressed

Mark manual Annotation: □

If you click in this box, curated sequences get purple

Export Results



Saves the complete B2G project (heavy)

Export annotation results in different formats

Export formats

Also for import!

.annot

```
C04018C10   GO:0004707     mitogen-activated protein kinase 3
C04018C10   EC:2.7.11.24
C04018A12   GO:0016798     class iv chitinase
C04018A12   GO:0000272
```

GeneSpring Format

```
C04013E10   response to water deprivation; regulation (nucleus;                    transcription factor activity;
C04013A12   translation;                                  ribosome; plastid;       structural constituent of ribosome;
C04013C12   galactose metabolic process;                  plastid;                 aldose 1-epimerase activity; carbohydrate binding;
```

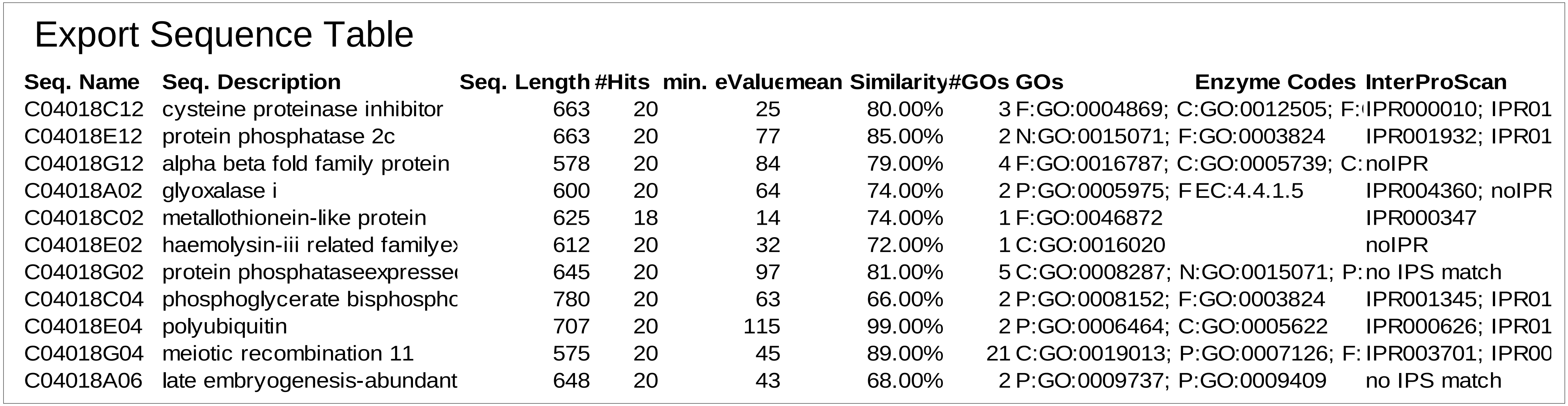
GoStat

```
C04018C10                    4707,9409,6979,10200,5524,169
C04018A12                    16798,272,44248
C04018C12                    4869,12505,8233
```

By Seq

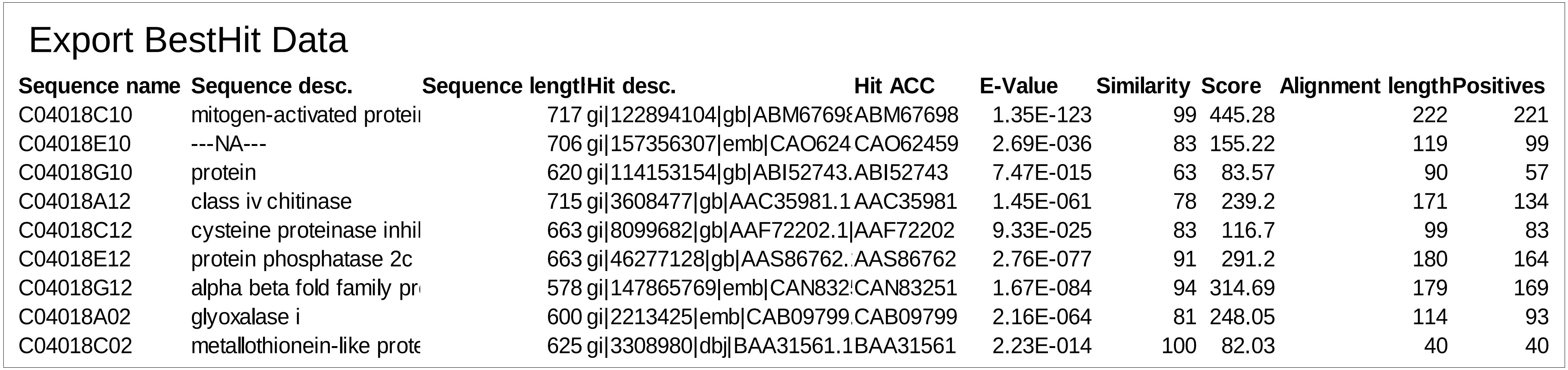
```
C04018A02   glyoxalase i                    GO:0004462   F:l
C04018C02   metallothionein-like protein    GO:0046872   F:r
C04018G02   protein phosphatase             GO:0008287   C:
```

More export formats



Export Sequence Table

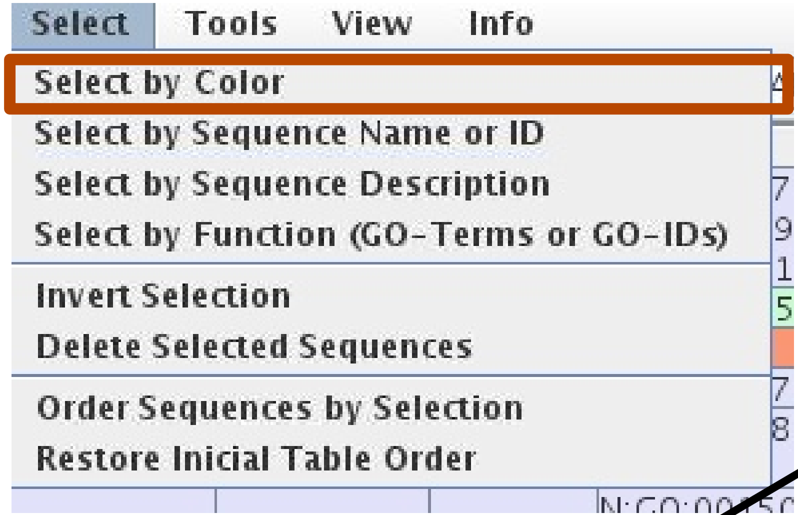
Seq. Name	Seq. Description	Seq. Length	#Hits	min. eValu	mean Similarity	#GOs	GOs	Enzyme Codes	InterProScan
C04018C12	cysteine proteinase inhibitor	663	20	25	80.00%	3	F:GO:0004869; C:GO:0012505; F:		IPR000010; IPR01
C04018E12	protein phosphatase 2c	663	20	77	85.00%	2	N:GO:0015071; F:GO:0003824		IPR001932; IPR01
C04018G12	alpha beta fold family protein	578	20	84	79.00%	4	F:GO:0016787; C:GO:0005739; C:		noIPR
C04018A02	glyoxalase i	600	20	64	74.00%	2	P:GO:0005975; F	EC:4.4.1.5	IPR004360; noIPR
C04018C02	metallothionein-like protein	625	18	14	74.00%	1	F:GO:0046872		IPR000347
C04018E02	haemolysin-iii related familye	612	20	32	72.00%	1	C:GO:0016020		noIPR
C04018G02	protein phosphataseexpresse	645	20	97	81.00%	5	C:GO:0008287; N:GO:0015071; P:		no IPS match
C04018C04	phosphoglycerate bisphospho	780	20	63	66.00%	2	P:GO:0008152; F:GO:0003824		IPR001345; IPR01
C04018E04	polyubiquitin	707	20	115	99.00%	2	P:GO:0006464; C:GO:0005622		IPR000626; IPR01
C04018G04	meiotic recombination 11	575	20	45	89.00%	21	C:GO:0019013; P:GO:0007126; F:		IPR003701; IPR00
C04018A06	late embryogenesis-abundant	648	20	43	68.00%	2	P:GO:0009737; P:GO:0009409		no IPS match



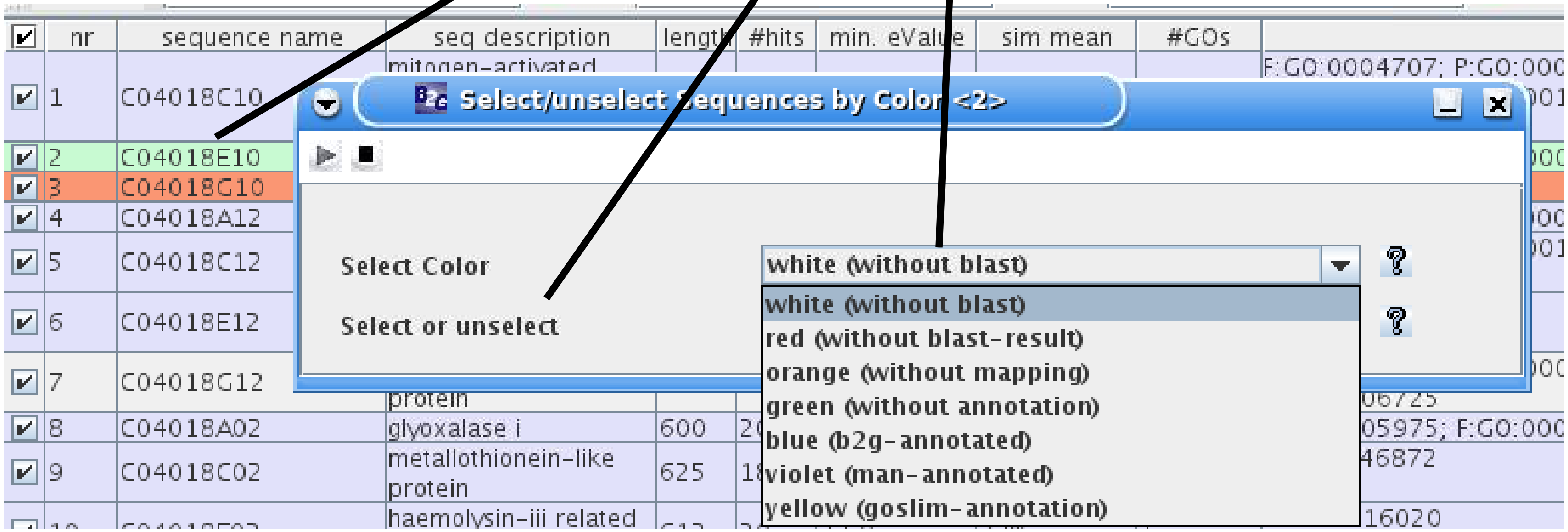
Export BestHit Data

Sequence name	Sequence desc.	Sequence lengt	Hit desc.	Hit ACC	E-Value	Similarity	Score	Alignment length	Positives
C04018C10	mitogen-activated protei	717	gi|122894104|gb|ABM6769	ABM67698	1.35E-123	99	445.28	222	221
C04018E10	---NA---	706	gi|157356307|emb|CAO624	CAO62459	2.69E-036	83	155.22	119	99
C04018G10	protein	620	gi|114153154|gb|ABI52743.	ABI52743	7.47E-015	63	83.57	90	57
C04018A12	class iv chitinase	715	gi|3608477|gb|AAC35981.1	AAC35981	1.45E-061	78	239.2	171	134
C04018C12	cysteine proteinase inhi	663	gi|8099682|gb|AAF72202.1|	AAF72202	9.33E-025	83	116.7	99	83
C04018E12	protein phosphatase 2c	663	gi|46277128|gb|AAS86762.	AAS86762	2.76E-077	91	291.2	180	164
C04018G12	alpha beta fold family pr	578	gi|147865769|emb|CAN832	CAN83251	1.67E-084	94	314.69	179	169
C04018A02	glyoxalase i	600	gi|2213425|emb|CAB09799.	CAB09799	2.16E-064	81	248.05	114	93
C04018C02	metallothionein-like prot	625	gi|3308980|dbj|BAA31561.1	BAA31561	2.23E-014	100	82.03	40	40

Sequence Selection


Select
Tools
View
Info
Select by Color
Select by Sequence Name or ID
Select by Sequence Description
Select by Function (GO-Terms or GO-IDs)
Invert Selection
Delete Selected Sequences
Order Sequences by Selection
Restore Inicial Table Order

Sequence Selection tool to obtain a selection based on annotation status



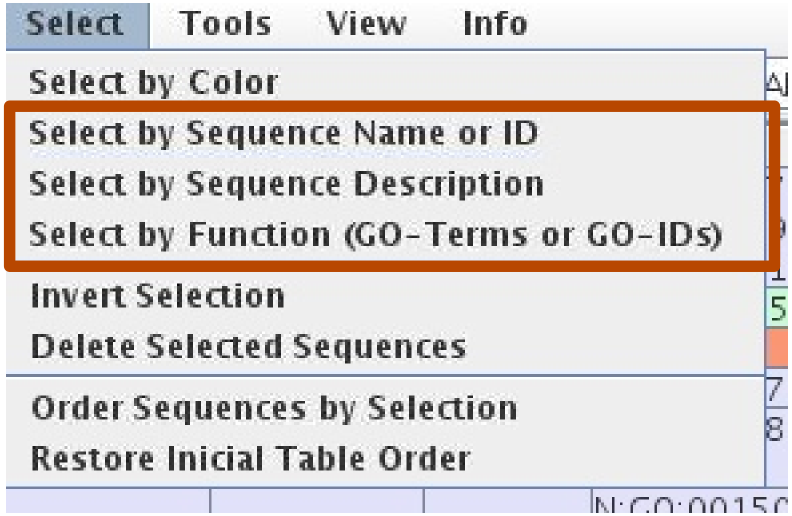
☑	nr	sequence name	seq description	length	#hits	min. eValue	sim mean	#GOs	
☑	1	C04018C10	mitogen-activated						F:GO:0004707; P:GO:000
☑	2	C04018E10							
☑	3	C04018G10							
☑	4	C04018A12							
☑	5	C04018C12							
☑	6	C04018E12							
☑	7	C04018G12	protein						06725
☑	8	C04018A02	glyoxalase i	600	2				05975; F:GO:000
☑	9	C04018C02	metallothionein-like protein	625	1				46872
	10	C04018E02	haemolysin-iii related	612					16020

Select Color: white (without blast)

Select or unselect

- white (without blast)
- red (without blast-result)
- orange (without mapping)
- green (without annotation)
- blue (b2g-annotated)
- violet (man-annotated)
- yellow (goslim-annotation)

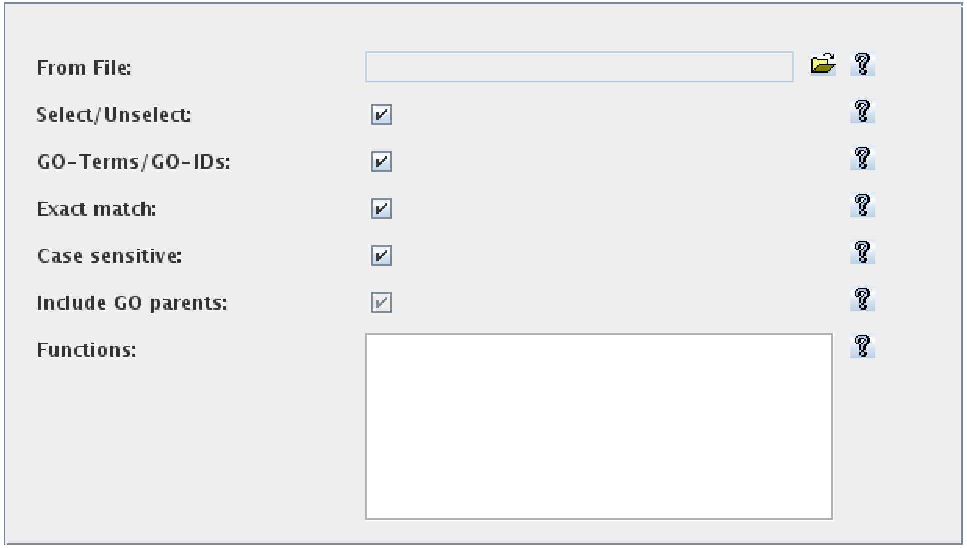
Sequence Selection



Select | Tools | View | Info

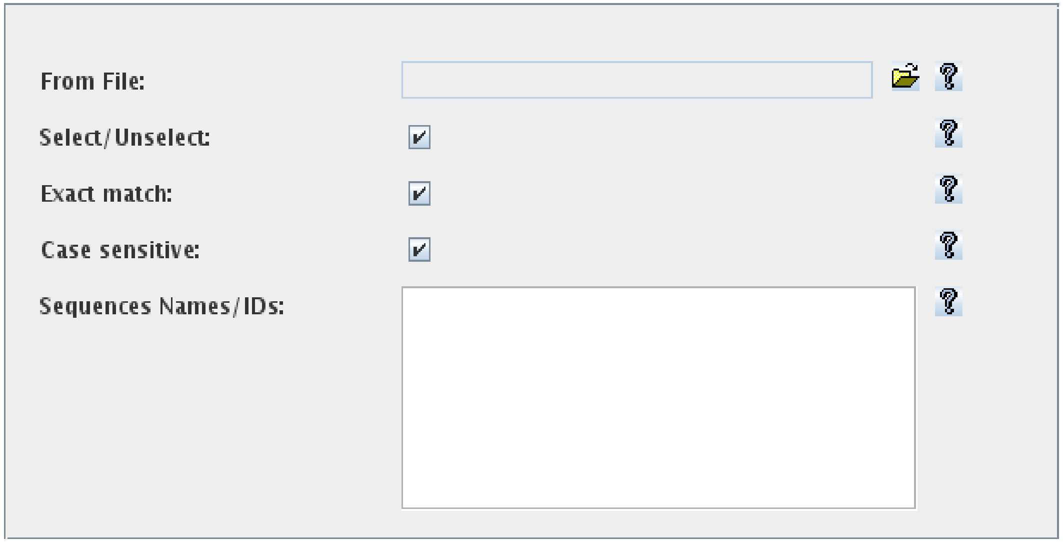
- Select by Color
- Select by Sequence Name or ID
- Select by Sequence Description
- Select by Function (GO-Terms or GO-IDs)
- Invert Selection
- Delete Selected Sequences
- Order Sequences by Selection
- Restore Inicial Table Order

By Name/Description



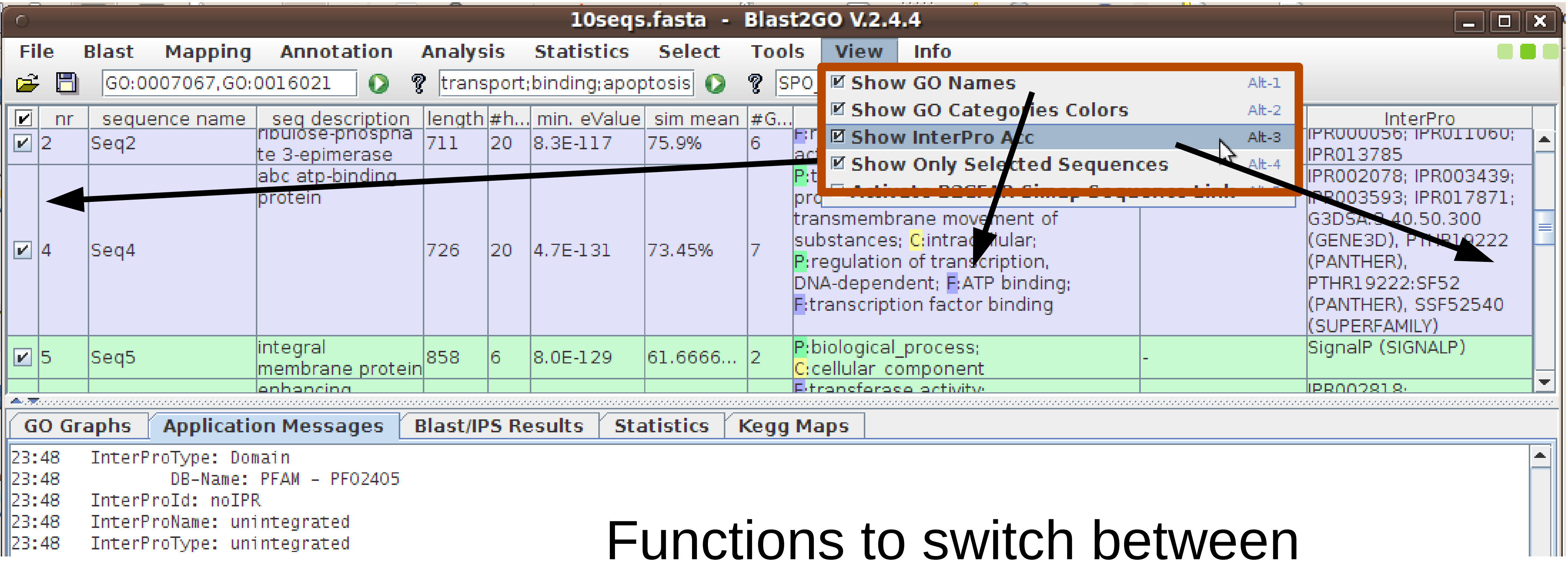
From File:	
Select/Unselect:	☑
GO-Terms/GO-IDs:	☑
Exact match:	☑
Case sensitive:	☑
Include GO parents:	☑
Functions:	

By Function



From File:	
Select/Unselect:	☑
Exact match:	☑
Case sensitive:	☑
Sequences Names/IDs:	

View Menu

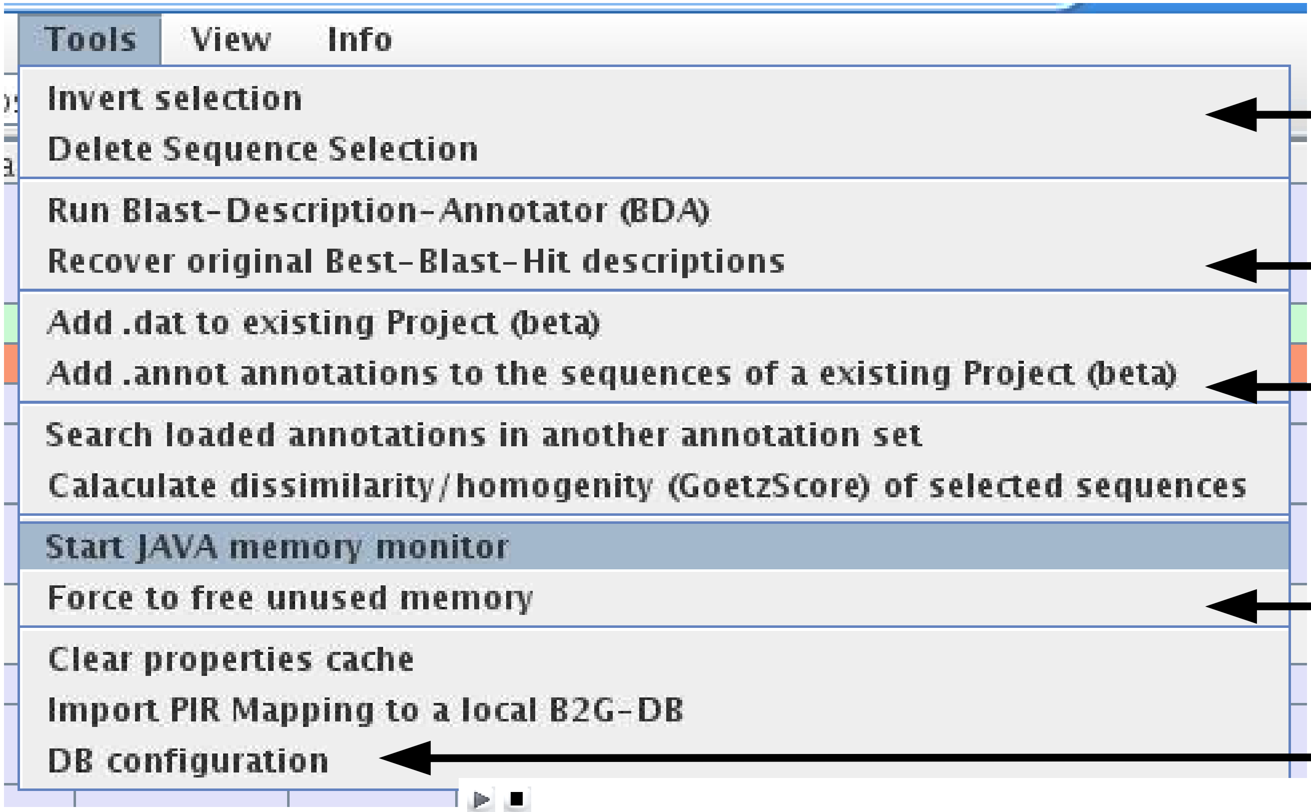

10seqs.fasta - Blast2GO V.2.4.4
File
Blast
Mapping
Annotation
Analysis
Statistics
Select
Tools
View
Info
GO:0007067,GO:0016021
transport;binding;apoptosis
SPO
☑ Show GO Names Alt-1
☑ Show GO Categories Colors Alt-2
☑ Show InterPro Acc Alt-3
☑ Show Only Selected Sequences Alt-4
□ Activate B2GFAR Single Sequence Link
GO Graphs
Application Messages
Blast/IPS Results
Statistics
Kegg Maps

☑	nr	sequence name	seq description	length	#h...	min. eValue	sim mean	#G...			InterPro
☑	2	Seq2	ribulose-phosphate 3-epimerase	711	20	8.3E-117	75.9%	6	F:r ac		IPR000056; IPR011060; IPR013785
☑	4	Seq4	abc atp-binding protein	726	20	4.7E-131	73.45%	7	P:t pro transmembrane movement of substances; C:intracellular; P:regulation of transcription, DNA-dependent; F:ATP binding; F:transcription factor binding		IPR002078; IPR003439; IPR003593; IPR017871; G3DSA:3.40.50.300 (GENE3D), PTHR19222 (PANTHER), PTHR19222:SF52 (PANTHER), SSF52540 (SUPERFAMILY)
☑	5	Seq5	integral membrane protein	858	6	8.0E-129	61.6666...	2	P:biological_process; C:cellular_component	-	SignalP (SIGNALP)
			enhancing						F:transferase activity;		IPR002818;

```
23:48   InterProType: Domain
              DB-Name: PFAM - PF02405
23:48   InterProId: noIPR
23:48   InterProName: unintegrated
23:48   InterProType: unintegrated
```

Functions to switch between displaying IDs or descriptions for GO annotation or InterPro results

Other Tools


Tools
View
Info
Invert selection
Delete Sequence Selection
Run Blast-Description-Annotator (BDA)
Recover original Best-Blast-Hit descriptions
Add .dat to existing Project (beta)
Add .annot annotations to the sequences of a existing Project (beta)
Search loaded annotations in another annotation set
Calaculate dissimilarity/homogenity (GoetzScore) of selected sequences
Start JAVA memory monitor
Force to free unused memory
Clear properties cache
Import PIR Mapping to a local B2G-DB
DB configuration

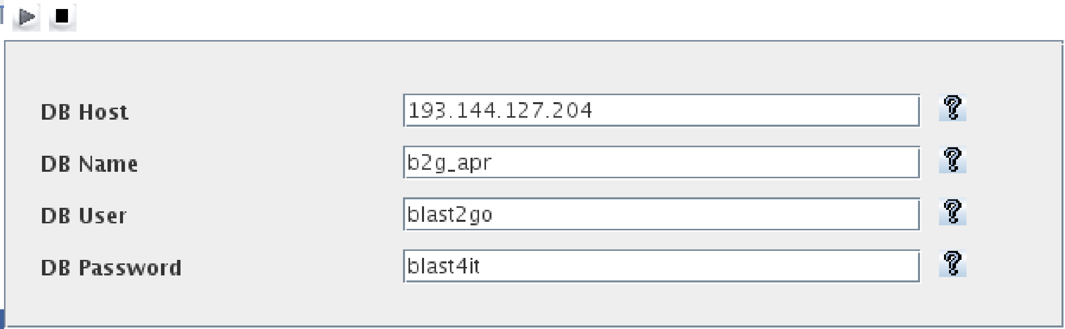
Permits to reduce the project size

Manipulation of sequence desc.

Merging .annot and .dat projects

Get more out of your memory

Check when connection problems



DB Host	193.144.127.204
DB Name	b2g_apr
DB User	blast2go
DB Password	blast4it

Outline

Concepts on Functional Annotation:

Biological Databases

Blast2GO annotation strategy

The Blast2GO annotation framework:

Annotation steps, Modulation of annotation intensity, Export/Import Functions, Sequence Selection, Additional Tools

----

Data Mining: Visualization Techniques

Combined Graph, Charts and Pies

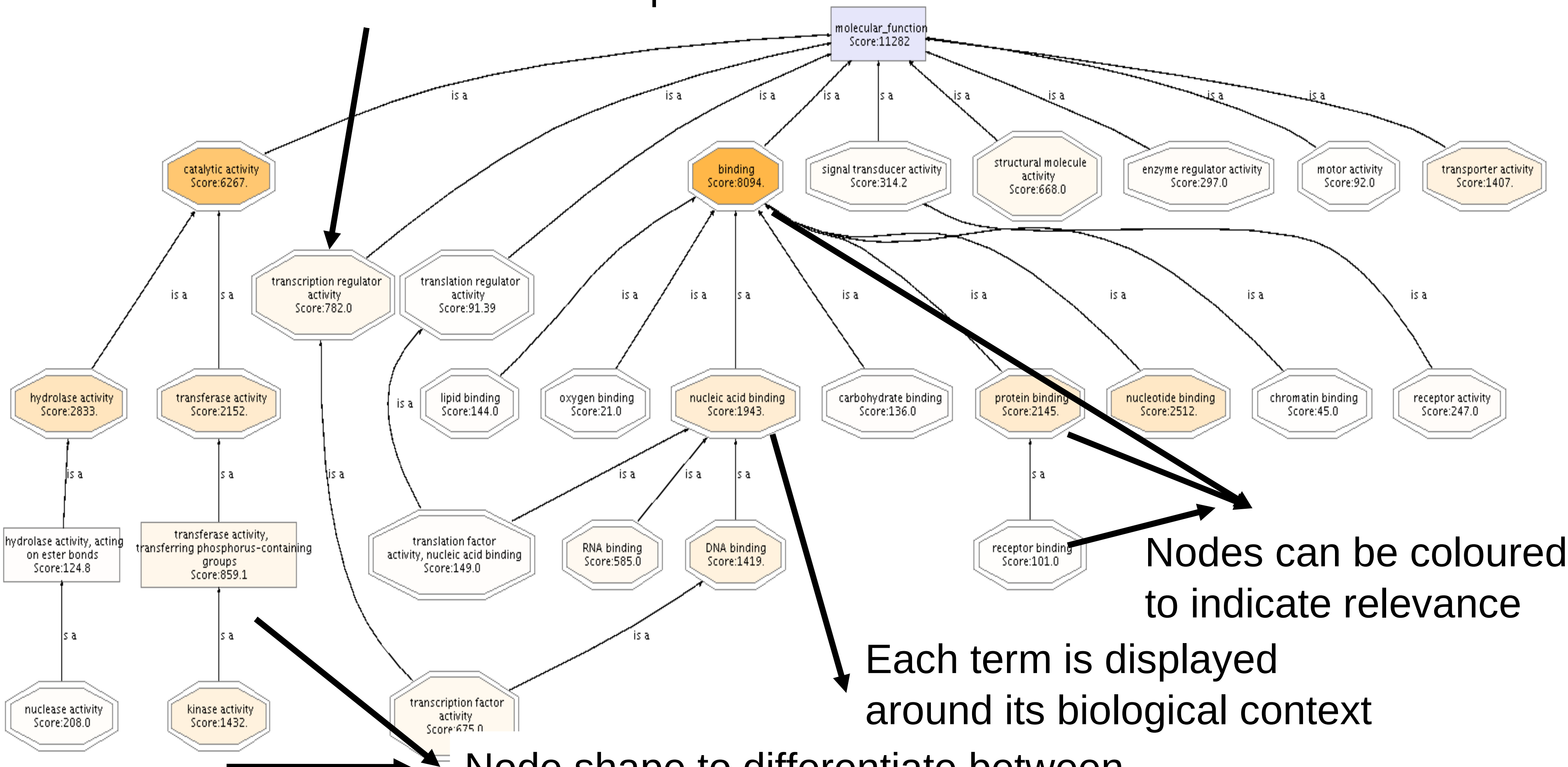
Hands on: Blast2GO annotation excercise

Data Mining on the DAG

- When working with large datasets, annotation results need to be summarized
- The DAG provides visualization of annotation data within its biological context
- In Blast2GO --> *Combined Graph* Function

Combined Graph

Each term has a number of sequences associated

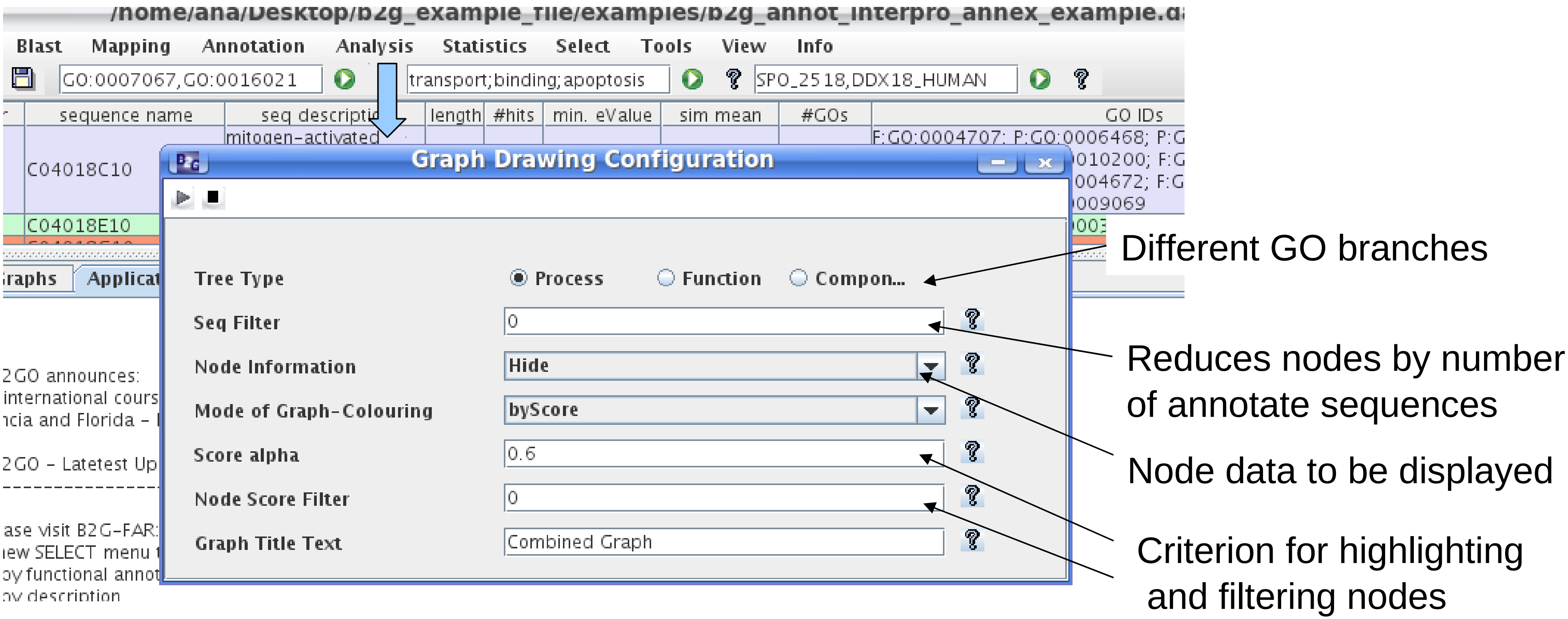

molecular_function Score:11282
is a
catalytic activity Score:6267.
binding Score:8094.
signal transducer activity Score:314.2
structural molecule activity Score:668.0
enzyme regulator activity Score:297.0
motor activity Score:92.0
transporter activity Score:1407.
transcription regulator activity Score:782.0
translation regulator activity Score:91.39
hydrolase activity Score:2833.
transferase activity Score:2152.
lipid binding Score:144.0
oxygen binding Score:21.0
nucleic acid binding Score:1943.
carbohydrate binding Score:136.0
protein binding Score:2145.
nucleotide binding Score:2512.
chromatin binding Score:45.0
receptor activity Score:247.0
hydrolase activity, acting on ester bonds Score:124.8
transferase activity, transferring phosphorus-containing groups Score:859.1
translation factor activity, nucleic acid binding Score:149.0
RNA binding Score:585.0
DNA binding Score:1419.
receptor binding Score:101.0
nuclease activity Score:208.0
kinase activity Score:1432.
transcription factor activity Score:675.0

Nodes can be coloured to indicate relevance

Each term is displayed around its biological context

Node shape to differentiate between direct and indirect annotation

Combined Graph



/home/ana/Desktop/b2g_example_file/examples/b2g_annot_interpro_annex_example.da

GO:0007067,GO:0016021

transport;binding;apoptosis

SPO_2518,DDX18_HUMAN

sequence name	seq descriptio	length	#hits	min. eValue	sim mean	#GOs	GO IDs
C04018C10	mitogen-activated						F:GO:0004707; P:GO:0006468; P:G 0010200; F:G 004672; F:G 009069
C04018E10							003

2GO announces:
international cours
ncia and Florida -

2GO - Latetest Up

ase visit B2G-FAR:
iew SELECT menu
oy functional annot
oy description

Graph Drawing Configuration

Tree Type	● Process ○ Function ○ Compon...
Seq Filter	0
Node Information	Hide
Mode of Graph-Colouring	byScore
Score alpha	0.6
Node Score Filter	0
Graph Title Text	Combined Graph

Different GO branches

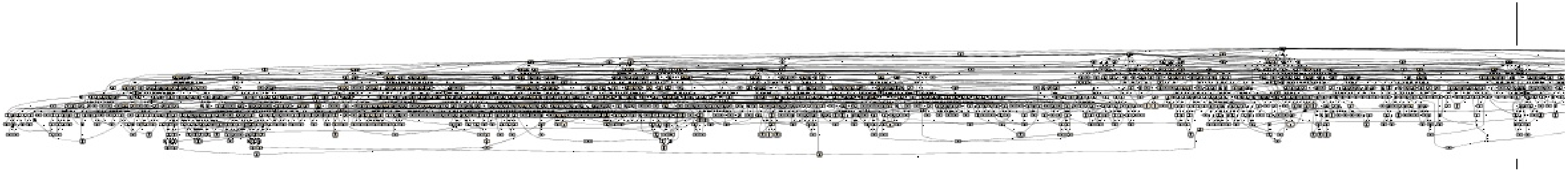
Reduces nodes by number of annotate sequences

Node data to be displayed

Criterion for highlighting and filtering nodes

Combined Graph

Let's paint the DAG of the dataset of 1000 sequences



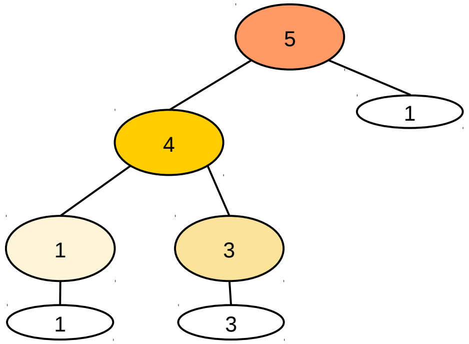
Too many nodes!!!

Need way to find relevant information

Node information content

Accumulated by node

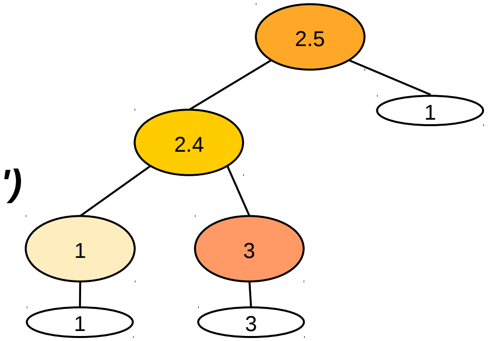
(Sequence Count)



Incomming information

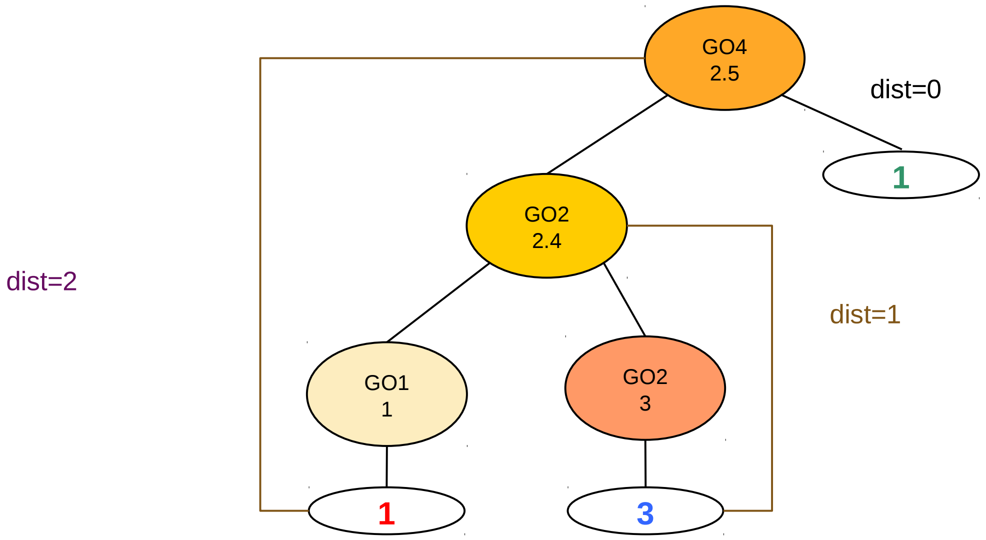
(Node Score)

$$\sum_{g \in desc(g')} seq(g) * \alpha^{\mathbf{dist\ (g,\ g')}}$$



The node score that reflects the amount of direct information at the node

Node score



$$\sum_{g \in desc(g')} seq(g) * \alpha^{dist\,(g,\, g')}$$

$$\alpha = 0.6$$

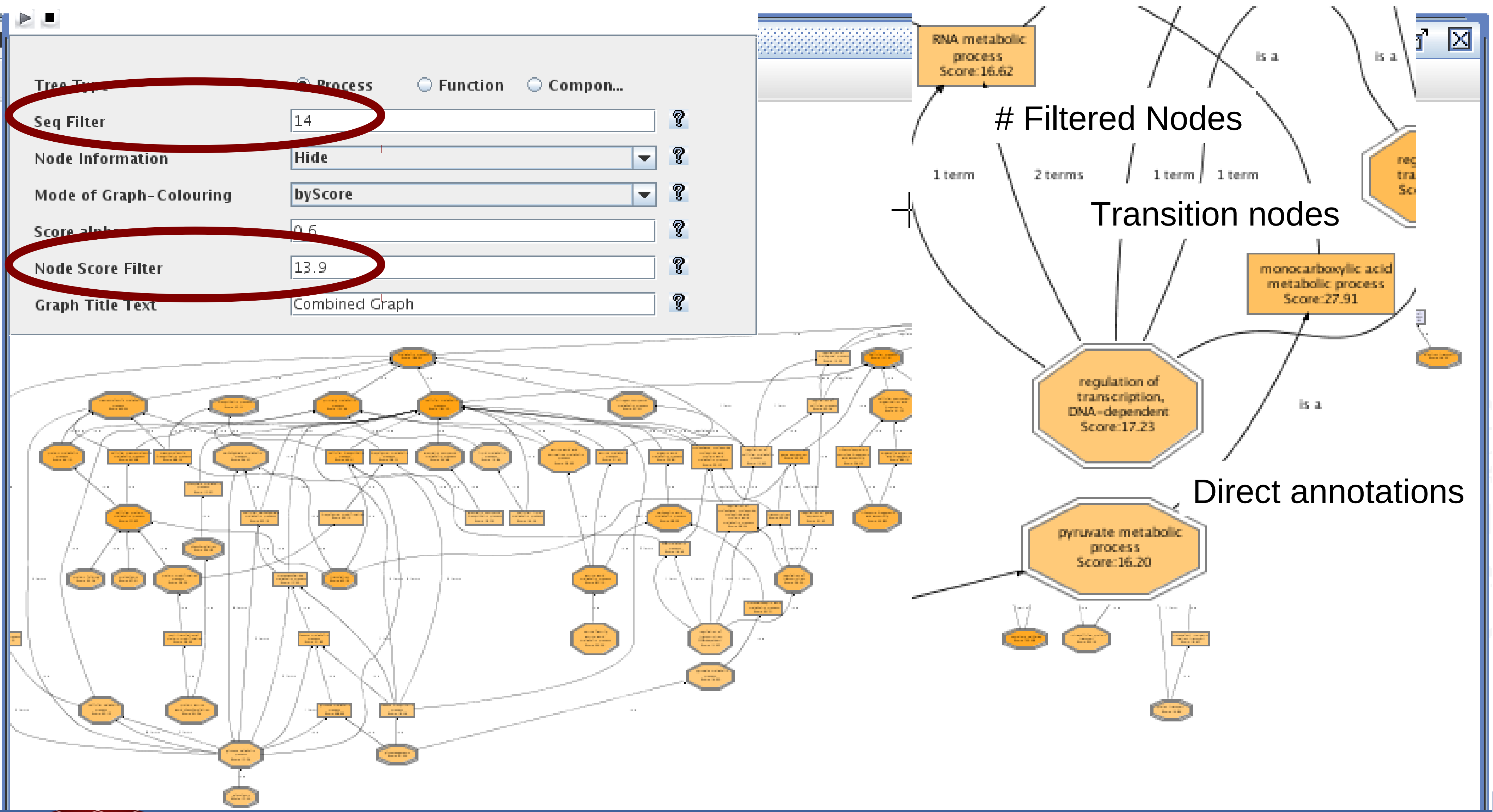
NodeScore (GO1) = 1 * 0.6^{0} = **1**

NodeScore (GO2) = **3** * 0.6^{0} = **3**

NodeScore (GO3) = **1** * 0.6^{1} + **3** * 0.6^{1} = 0.6 + 1.8 = **2.4**

NodeScore (GO4) = **1** * 0.6^{2} + **3** * 0.6^{2} + **1** * 0.6^{0} = 0.36 + 1.08 + 1 = **2.5**

Filtered Graph



Tree Type	● Process ○ Function ○ Compon...
Seq Filter	14
Node Information	Hide
Mode of Graph-Colouring	byScore
Score alpha	0.6
Node Score Filter	13.9
Graph Title Text	Combined Graph

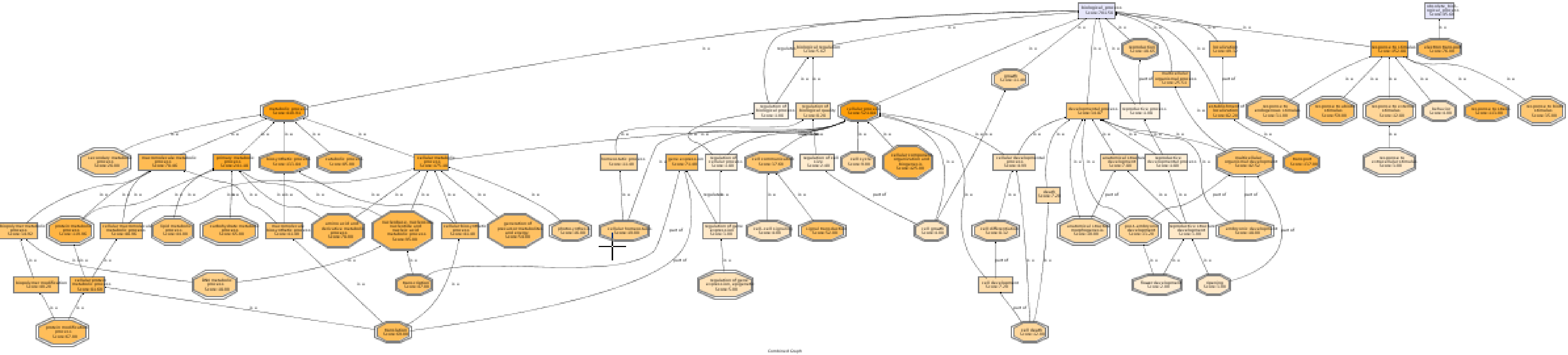
# Filtered Nodes

Transition nodes

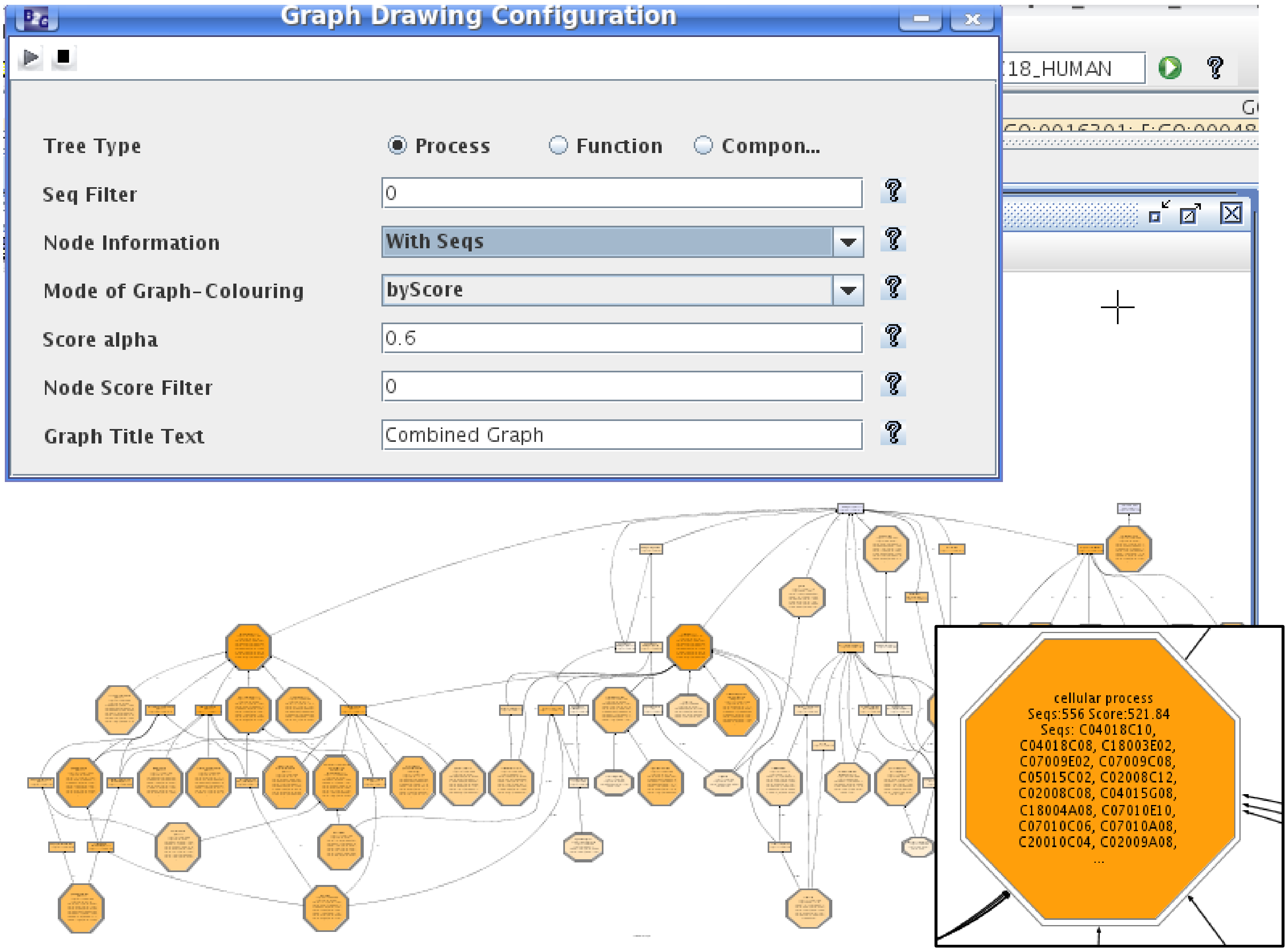
Direct annotations

RNA metabolic process Score:16.62
1 term
2 terms
1 term
1 term
is a
is a
monocarboxylic acid metabolic process Score:27.91
regulation of transcription, DNA-dependent Score:17.23
is a
pyruvate metabolic process Score:16.20

Compacting Graphs by GOSlim



Show node content



Graph Drawing Configuration

Tree Type: ● Process ○ Function ○ Compon...

Seq Filter: 0

Node Information: With Seqs

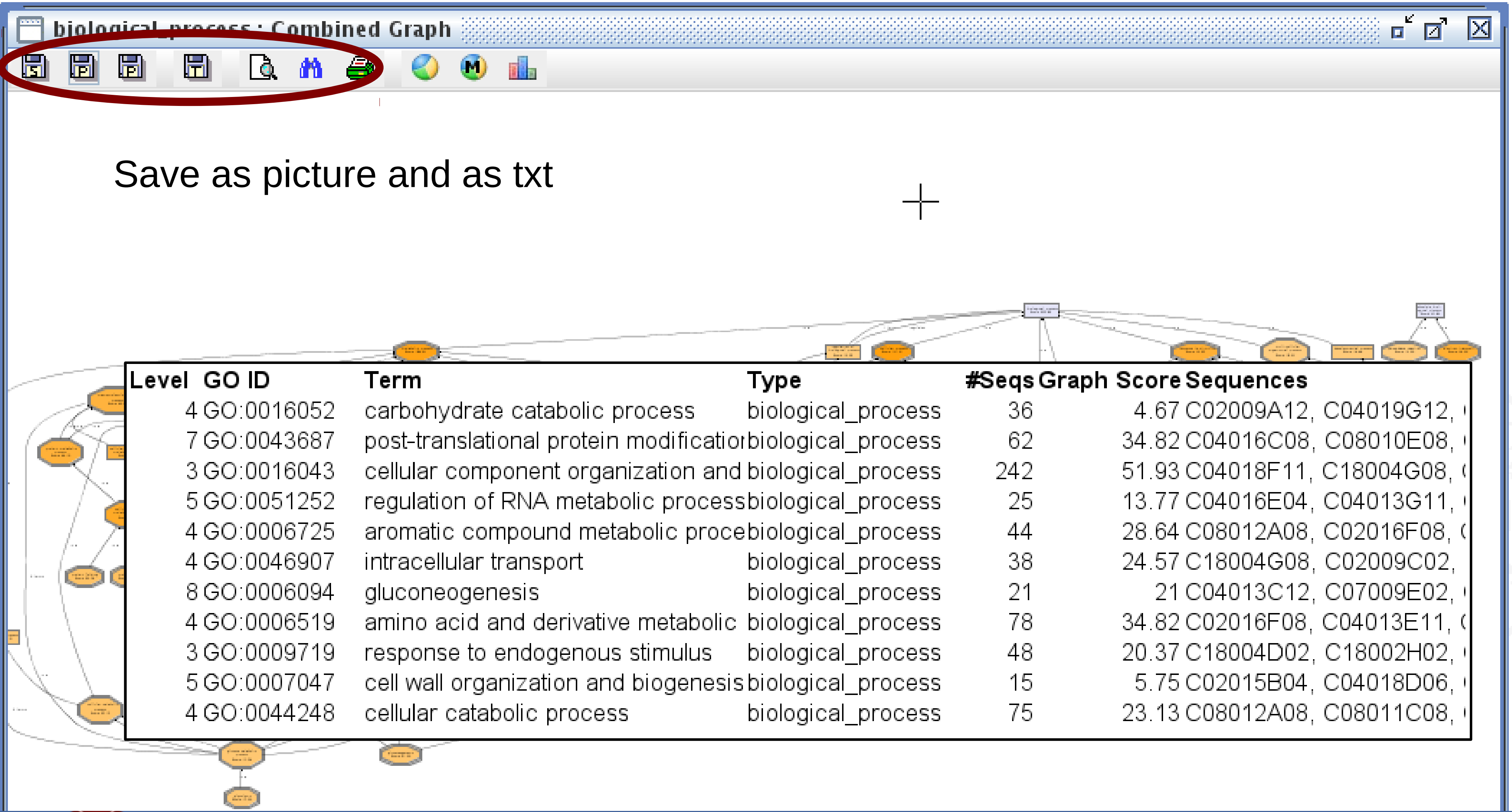
Mode of Graph-Colouring: byScore

Score alpha: 0.6

Node Score Filter: 0

Graph Title Text: Combined Graph

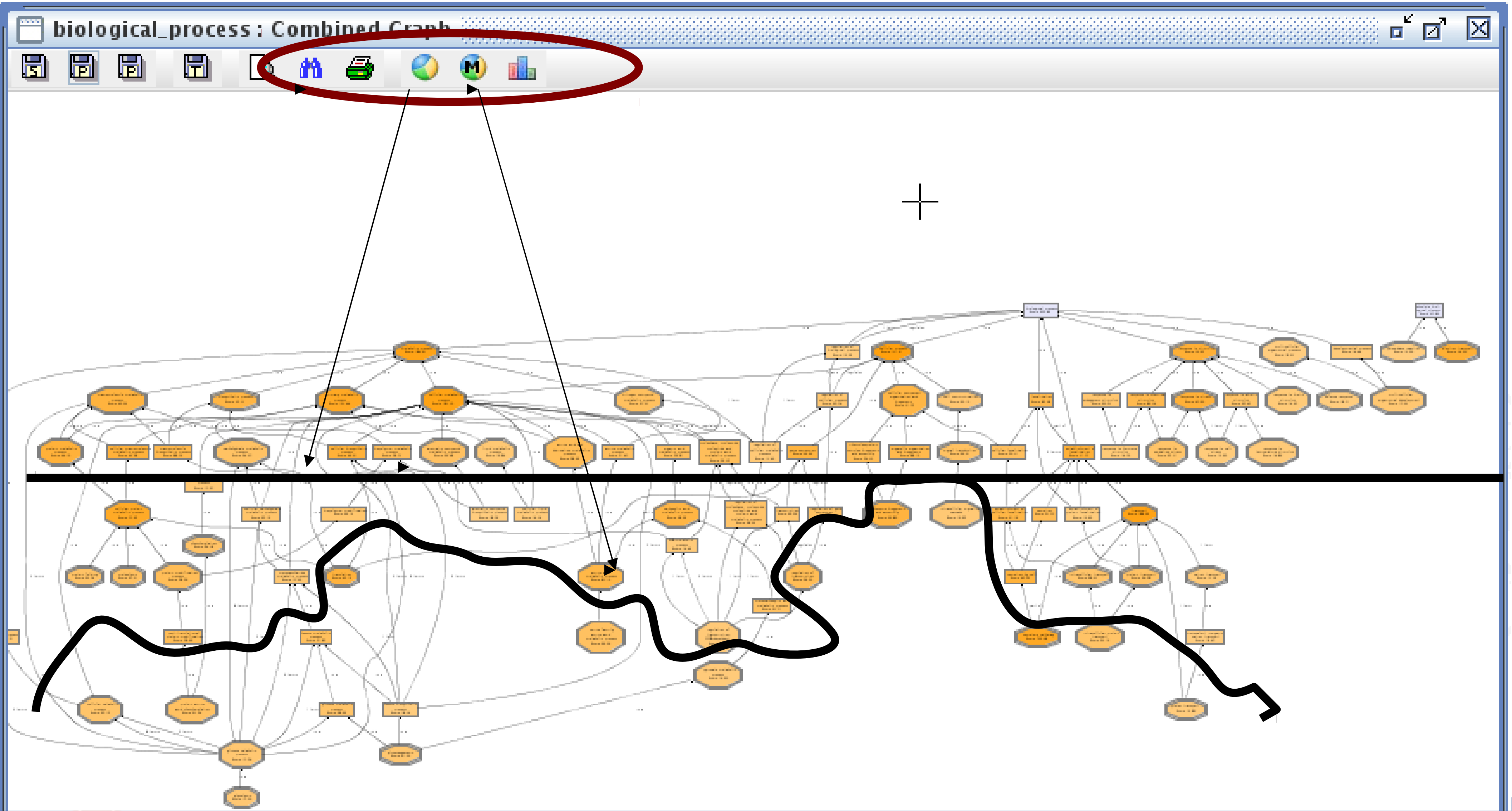
Saving Options



Save as picture and as txt

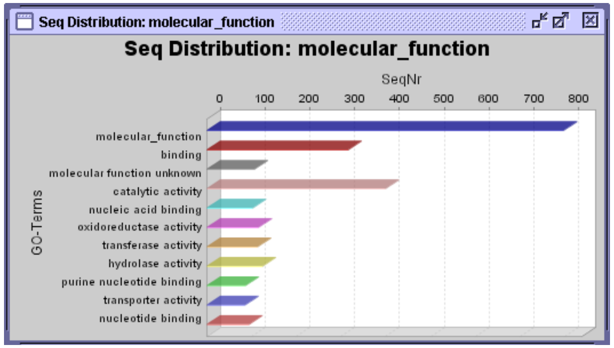
Level	GO ID	Term	Type	#Seqs	Graph Score	Sequences
4	GO:0016052	carbohydrate catabolic process	biological_process	36	4.67	C02009A12, C04019G12,
7	GO:0043687	post-translational protein modificatio	biological_process	62	34.82	C04016C08, C08010E08,
3	GO:0016043	cellular component organization and	biological_process	242	51.93	C04018F11, C18004G08,
5	GO:0051252	regulation of RNA metabolic process	biological_process	25	13.77	C04016E04, C04013G11,
4	GO:0006725	aromatic compound metabolic proce	biological_process	44	28.64	C08012A08, C02016F08,
4	GO:0046907	intracellular transport	biological_process	38	24.57	C18004G08, C02009C02,
8	GO:0006094	gluconeogenesis	biological_process	21	21	C04013C12, C07009E02,
4	GO:0006519	amino acid and derivative metabolic	biological_process	78	34.82	C02016F08, C04013E11,
3	GO:0009719	response to endogenous stimulus	biological_process	48	20.37	C18004D02, C18002H02,
5	GO:0007047	cell wall organization and biogenesis	biological_process	15	5.75	C02015B04, C04018D06,
4	GO:0044248	cellular catabolic process	biological_process	75	23.13	C08012A08, C08011C08,

Graph Charts

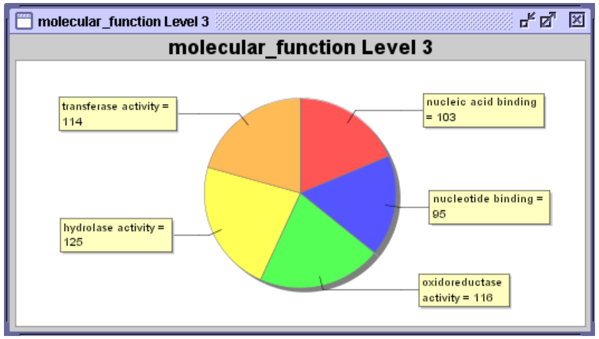


Graph Charts

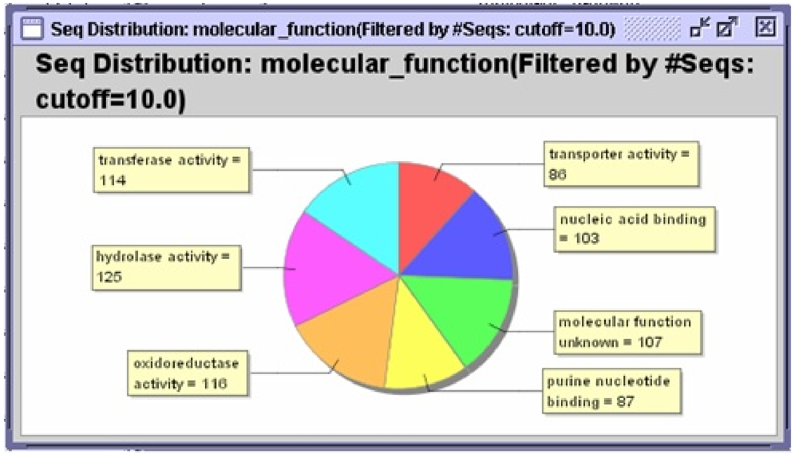
- Sequence Distribution/GO as Bar-Chart

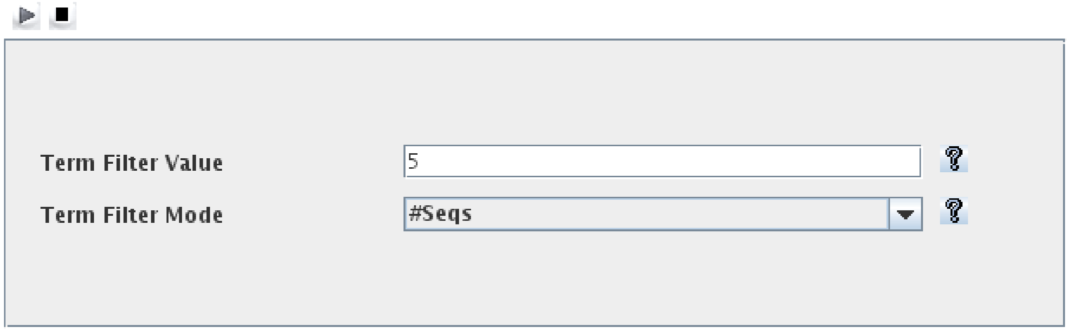


- Sequence Distribution/GO as Level-Pie (level selection)



- Sequence Distribution/GO as Multilevel-Pie (#score or #seq cutoff)

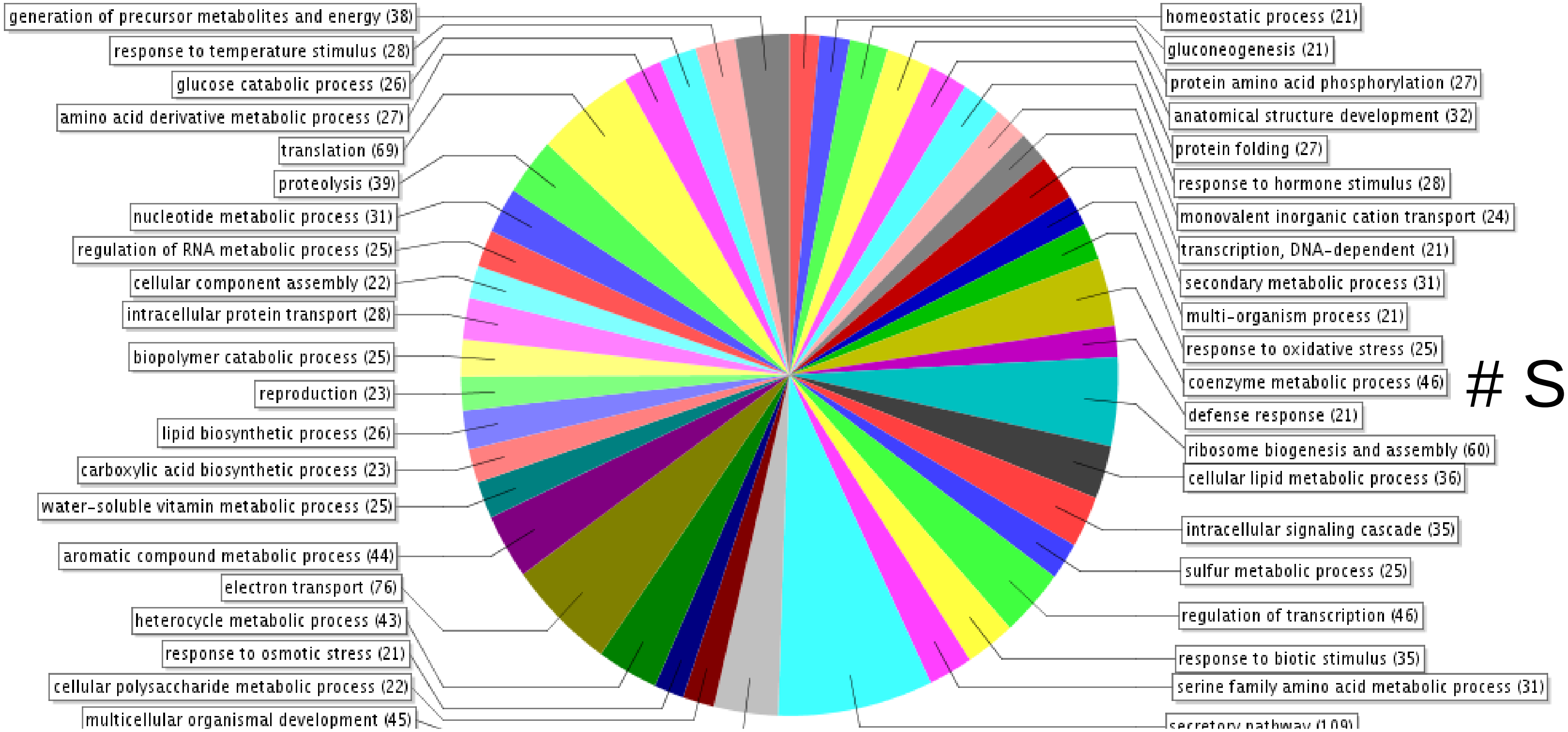




Term Filter Value 5

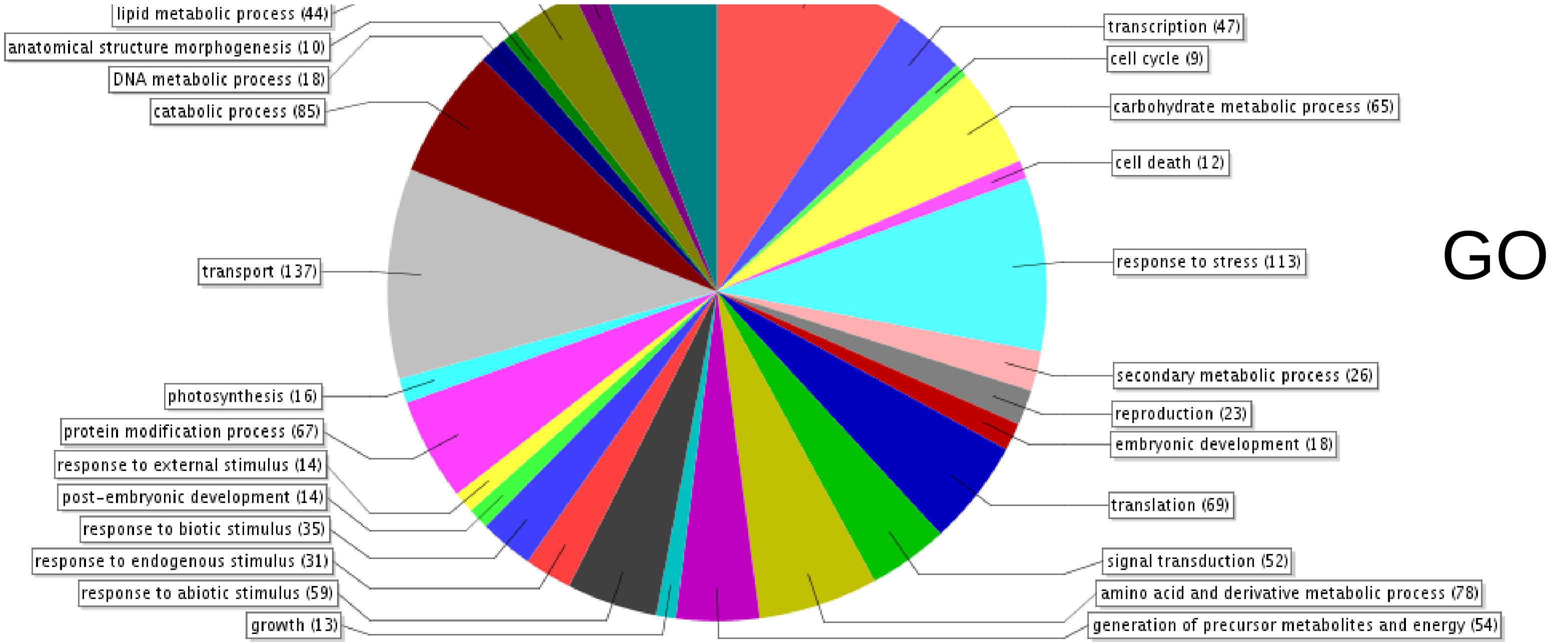
Term Filter Mode #Seqs

M-Pies



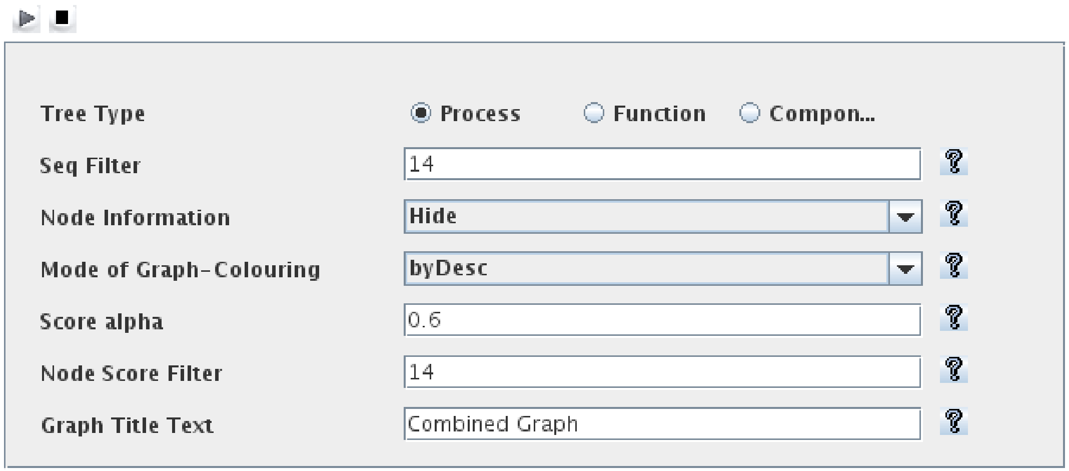
# Seq filter = 20

Handy to summarize functional content



GOSlim

Colouring yourself the DAG



Tree Type: ● Process ○ Function ○ Compon...

Seq Filter: 14

Node Information: Hide

Mode of Graph-Colouring: byDesc

Score alpha: 0.6

Node Score Filter: 14

Graph Title Text: Combined Graph

The byDesc option in the Graph-Colouring allows you to colour the DAG nodes according to an additional value

GO:0005792	GO:0005792	1.00
GO:0006412	GO:0006412	0.81
GO:0003735	GO:0003735	0.71
GO:0016705	GO:0016705	0.65
GO:0005840	GO:0005840	0.65
GO:0005506	GO:0005506	0.64
GO:0006631	GO:0006631	0.61
GO:0020037	GO..	

The “special” .annot file:
3 columns
GO name, GO ID, Value

Scale between 0 and 1 used to colour the graph

Graph Visualization

- DAGs are interesting for browsing functional annotation but can be too large
- With filtering and prunning options you can create more navegable DAGs
- Pies are good to compact information: try out levels
- GOSlim compacts to more equivaent terms than filtering the GO

HANDS ON B2G

Go to the on-line course material

Blast, map and annotate several few sequences in Blast2GO by loading the 10 test sequences (within the file menu).

Generate some singel-Seq GO graphs to review annotation. (right mouse click on sequence table)
(http://www.blast2go.org → Start → 1024MB)

Annotated 1100 Citrus-Unigenes (nt) with Blast2GO. Analyse the annotation results. Generate a Combined Graph after a GoSlim-Reduction and try to export a handy graph as PDF.