

# Functional Genomics

BME 110: CompBio Tools

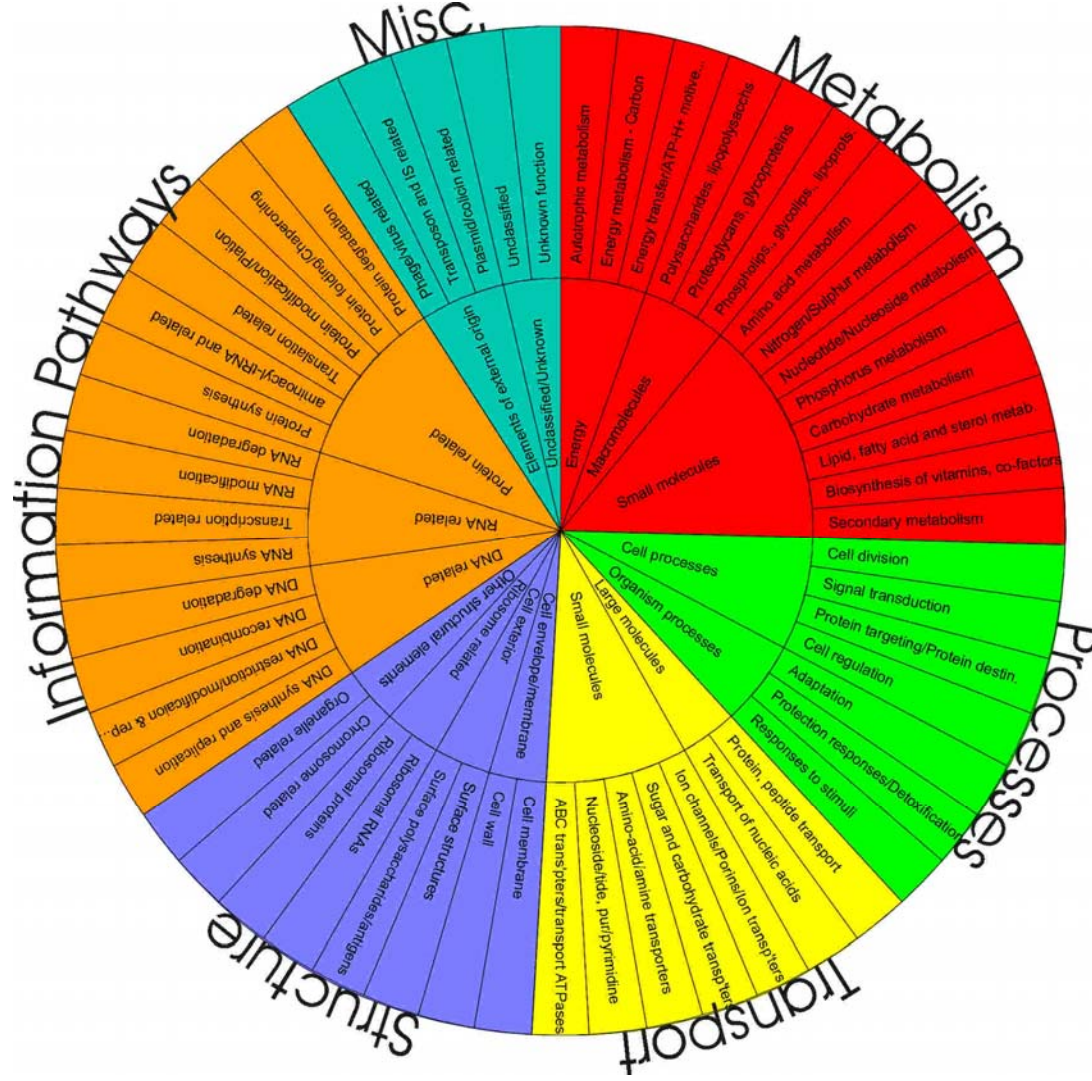
Todd Lowe

May 22, 2008

# Types of Biological Classification

- Information pathways
  - DNA, RNA and protein related
- Metabolism
- Processes
  - Signal transduction, cell division
- Transport of large and small molecules
  - Ion channels, protein and peptide transport
- Structure
  - Cell walls and membranes

# Function Wheels



# Clusters of Orthologous Genes

- Whole genome BLAST comparisons of all genomes
- “Reciprocal” best-hits form clusters of genes that are assumed to be orthologous (same function)
- Automated approach, no extensive manual curation



## Clusters of Orthologous Groups



# COG

<http://www.ncbi.nlm.nih.gov/COG/new/>

### Unicellular Clusters

Group	species
<u>A</u> rchaea	13 <a href="#">Afu</a> <a href="#">Hbs</a> <a href="#">Mac</a> <a href="#">Mth</a> <a href="#">Mja</a> <a href="#">Mka</a> <a href="#">Tac</a> <a href="#">Tvo</a> <a href="#">Pho</a> <a href="#">Pab</a> <a href="#">Pya</a> <a href="#">Sso</a> <a href="#">Ape</a>
<u>E</u> karyota	3 <a href="#">Sce</a> <a href="#">Spo</a> <a href="#">Ecu</a>
<u>B</u> acteria	10 <a href="#">Aae</a> <a href="#">Tma</a> <a href="#">Ctr</a> <a href="#">Cpn</a> <a href="#">Tpa</a> <a href="#">Bbu</a> <a href="#">Syn</a> <a href="#">Nos</a> <a href="#">Fnu</a> <a href="#">Dra</a>
<u>A</u> ctino bacteria	4 <a href="#">Cgl</a> <a href="#">Mtu</a> <a href="#">MtC</a> <a href="#">Mle</a>
<u>G</u> ramplus	12 <a href="#">Cac</a> <a href="#">Lla</a> <a href="#">Spy</a> <a href="#">Spn</a> <a href="#">Sau</a> <a href="#">Lin</a> <a href="#">Bsu</a> <a href="#">Bha</a> <a href="#">Uur</a> <a href="#">Mpu</a> <a href="#">Mpn</a> <a href="#">Mge</a>
<u>g</u> amma	11 <a href="#">Eco</a> <a href="#">EcZ</a> <a href="#">Ecs</a> <a href="#">Ype</a> <a href="#">Sty</a> <a href="#">Buc</a> <a href="#">Vch</a> <a href="#">Pae</a> <a href="#">Hin</a> <a href="#">Pmu</a> <a href="#">Xfa</a>
<u>P</u> roteo bacteria	6 <a href="#">Nme</a> <a href="#">NmA</a> <a href="#">Rso</a> <a href="#">Hpy</a> <a href="#">iHp</a> <a href="#">Cje</a>
<u>a</u> lpha	7 <a href="#">Atu</a> <a href="#">Sme</a> <a href="#">Bme</a> <a href="#">Mlo</a> <a href="#">Ccr</a> <a href="#">Rpr</a> <a href="#">Rco</a>
<b>Total</b>	<b>66</b>

### Eukaryotic Clusters

Code	Name	Abbreviation
<b>A</b>	<i>Arabidopsis thaliana</i> (thale cress)	<b>ath</b>
<b>C</b>	<i>Caenorhabditis elegans</i> (worm)	<b>cel</b>
<b>D</b>	<i>Drosophila melanogaster</i> (fruit fly)	<b>dme</b>
<b>H</b>	<i>Homo sapiens</i> (human)	<b>hsa</b>
<b>Y</b>	<i>Saccharomyces cerevisiae</i> (baker yeast)	<b>sce</b>
<b>P</b>	<i>Schizosaccharomyces pombe</i> (fission yeast)	<b>spo</b>
<b>E</b>	<i>Encephalitozoon cuniculi</i> (Microsporidia)	<b>ecu</b>

[Complete genome:](#)

[ [11620](#) ] [ [143](#) ]

[Pyrococcus abyssi GE5, complete genome](#)

[Microbial genomes](#) ▲

Sequencing center: [Genoscope](#)

Genome Info	Feature table	BLAST protein homologs	Links
Refseq: <a href="#">NC_000868</a>	<a href="#">Protein coding genes</a>	<a href="#">COGs (Clusters of Orthologous Groups)</a>	<a href="#">Refseq FTP</a>
GenBank: <a href="#">AL096836</a>	<a href="#">Structural RNAs</a>	<a href="#">3D Structure</a> (Sequences with known structure)	<a href="#">GenBank FTP</a>
Tc			<a href="#">BLAST</a>
Cc			TraceAssembly
		<a href="#">GenePlot</a> (Pairwise genome comparison)	<a href="#">CDD</a>

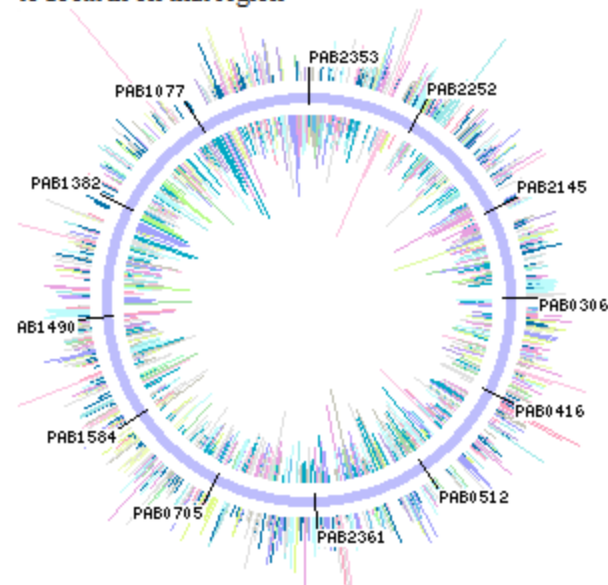
COG scheme applied to genome

Pyrococcus  
Abyssi  
(1.7 Mb)

Start from :   Search for gene

#### Protein coding genes distribution map

To see map locations of genes, click on a region in the map, to zoom in on that region

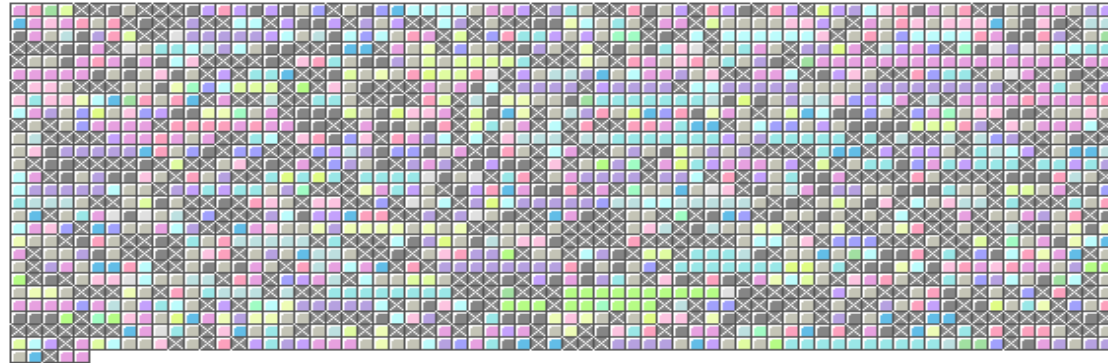


#### Gene Classification based on COG functional categories

- Translation, ribosomal structure and biogenesis
- Transcription
- DNA replication, recombination and repair
- Cell division and chromosome partitioning
- Posttranslational modification, protein turnover
- Cell envelope biogenesis, outer membrane
- Cell motility and secretion
- Inorganic ion transport and metabolism
- Signal transduction mechanisms
- Energy production and conversion
- Carbohydrate transport and metabolism
- Amino acid transport and metabolism
- Nucleotide transport and metabolism
- Coenzyme metabolism
- Lipid metabolism
- Secondary metabolites biosynthesis, transport and catabolism
- General function prediction only
- Function unknown
- No COG match

1895 proteins: distribution by COGs [functional categories](#)

1499 proteins can be found in [COGs](#) data base



COG  
classification  
by major  
functional  
categories

Code	COGs	Description
<input type="checkbox"/> J	<a href="#">151</a>	Translation
<input type="checkbox"/> A	<a href="#">1</a>	RNA processing and modification
<input type="checkbox"/> K	<a href="#">90</a>	Transcription
<input type="checkbox"/> L	<a href="#">67</a>	Replication, recombination and repair
<input type="checkbox"/> B	<a href="#">1</a>	Chromatin structure and dynamics
<input type="checkbox"/> D	<a href="#">18</a>	Cell cycle control, mitosis and meiosis
<input type="checkbox"/> Y	<a href="#">0</a>	Nuclear structure
<input type="checkbox"/> V	<a href="#">22</a>	Defense mechanisms
<input type="checkbox"/> T	<a href="#">15</a>	Signal transduction mechanisms
<input type="checkbox"/> M	<a href="#">47</a>	Cell wall/membrane biogenesis
<input type="checkbox"/> N	<a href="#">28</a>	Cell motility
<input type="checkbox"/> Z	<a href="#">0</a>	Cytoskeleton
<input type="checkbox"/> W	<a href="#">0</a>	Extracellular structures
<input type="checkbox"/> U	<a href="#">11</a>	Intracellular trafficking and secretion
<input type="checkbox"/> O	<a href="#">49</a>	Posttranslational modification, protein turnover, chaperones
<input type="checkbox"/> C	<a href="#">118</a>	Energy production and conversion
<input type="checkbox"/> G	<a href="#">84</a>	Carbohydrate transport and metabolism
<input type="checkbox"/> E	<a href="#">131</a>	Amino acid transport and metabolism



## KEGG: Kyoto Encyclopedia of Genes and Genomes

A grand challenge in the post-genomic era is a complete computer representation of the cell and the organism, which will enable computational prediction of higher-level complexity of cellular processes and organism behaviors from genomic information. Towards this end we have been developing a bioinformatics resource named KEGG, Kyoto Encyclopedia of Genes and Genomes, as part of the research projects in the Kanehisa Laboratory of Kyoto University Bioinformatics Center.

### ● KEGG Table of Contents ●

Entry to the KEGG web service

KEGG Release 34.0, April 2005 (plus daily updates)  
[Release notes](#)

---

#### Introduction

User manuals  
References

#### Standards

Data formats  
API

#### Links

Related databases

#### Distribution

Disclaimer  
FTP access

---

[Feedback](#)

[GenomeNet](#)

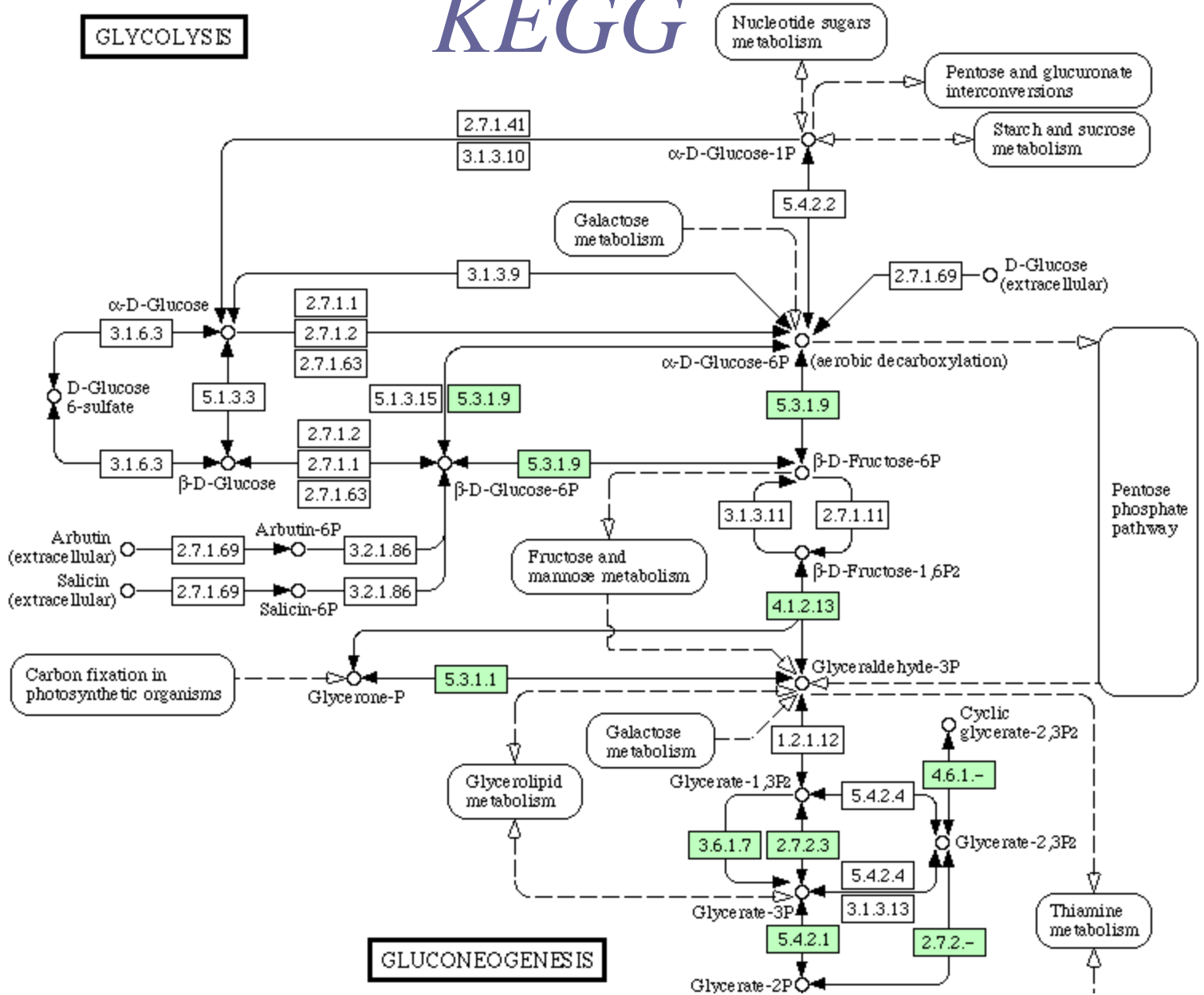
---

Best Pathway Database:

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

**GLYCOLYSIS**

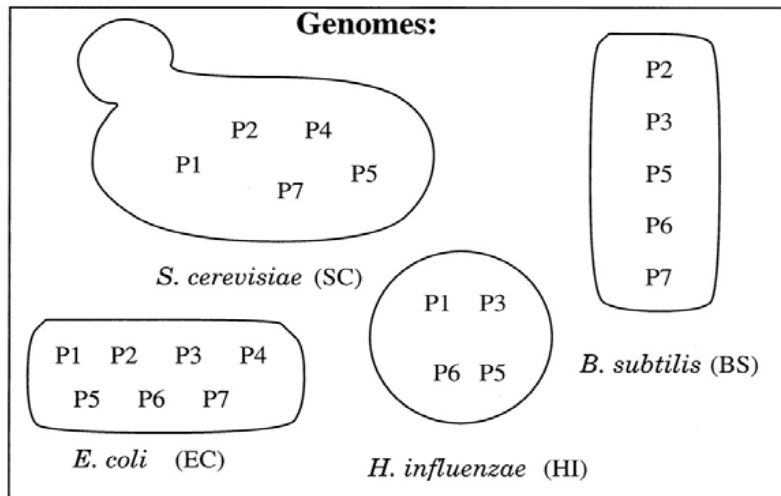
# KEGG



# Where do we get function assignments?

- Experimental methods
  - Microarray analyses
  - Other large-scale functional screens (i.e. global disruptions, molecular interactions)
  - Biochemical & enzymatic analyses
  - Traditional genetics
- Computational methods
  - Sequence / motif / domain comparisons
  - Phylogenetic tree methods
  - Rosetta Stone methods

# Phylogenetic profiles



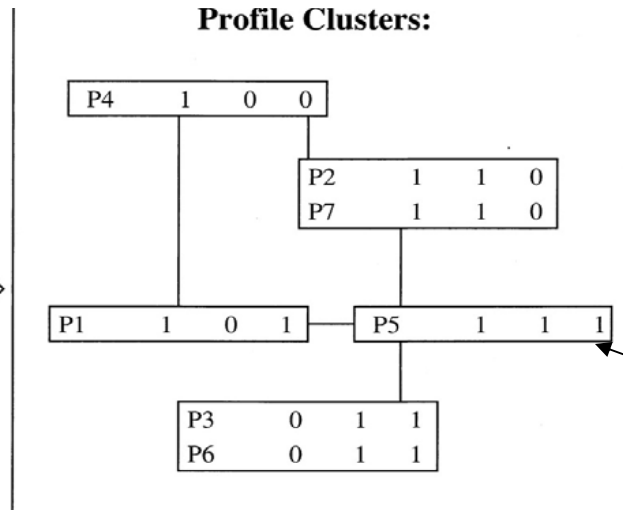
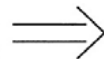
Pellegrini M et al., "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles." PNAS (1999) 96(8):4285-8

Phylogeny Tree -> Multiple Alignment -> Profile



**Phylogenetic Profile:**

	EC	SC	BS	HI
P1	1	0	1	1
P2	1	1	0	0
P3	0	1	1	1
P4	1	0	0	0
P5	1	1	1	1
P6	0	1	1	1
P7	1	1	0	0



Cluster by presence in genomes



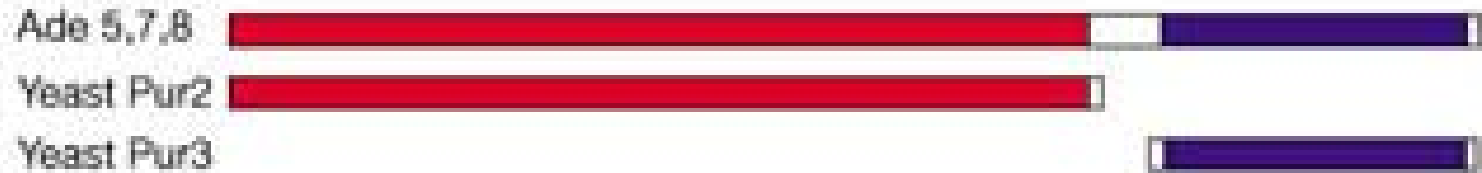
**Conclusion:** P2 and P7 are functionally linked, P3 and P6 are functionally linked

# Rosetta Stone method

## General concept



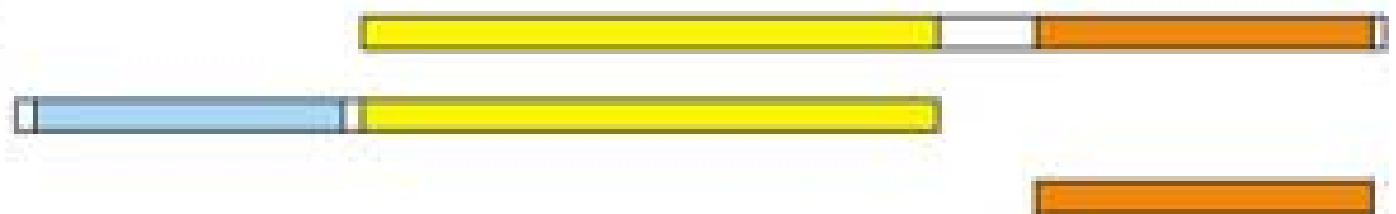
## *C. elegans*



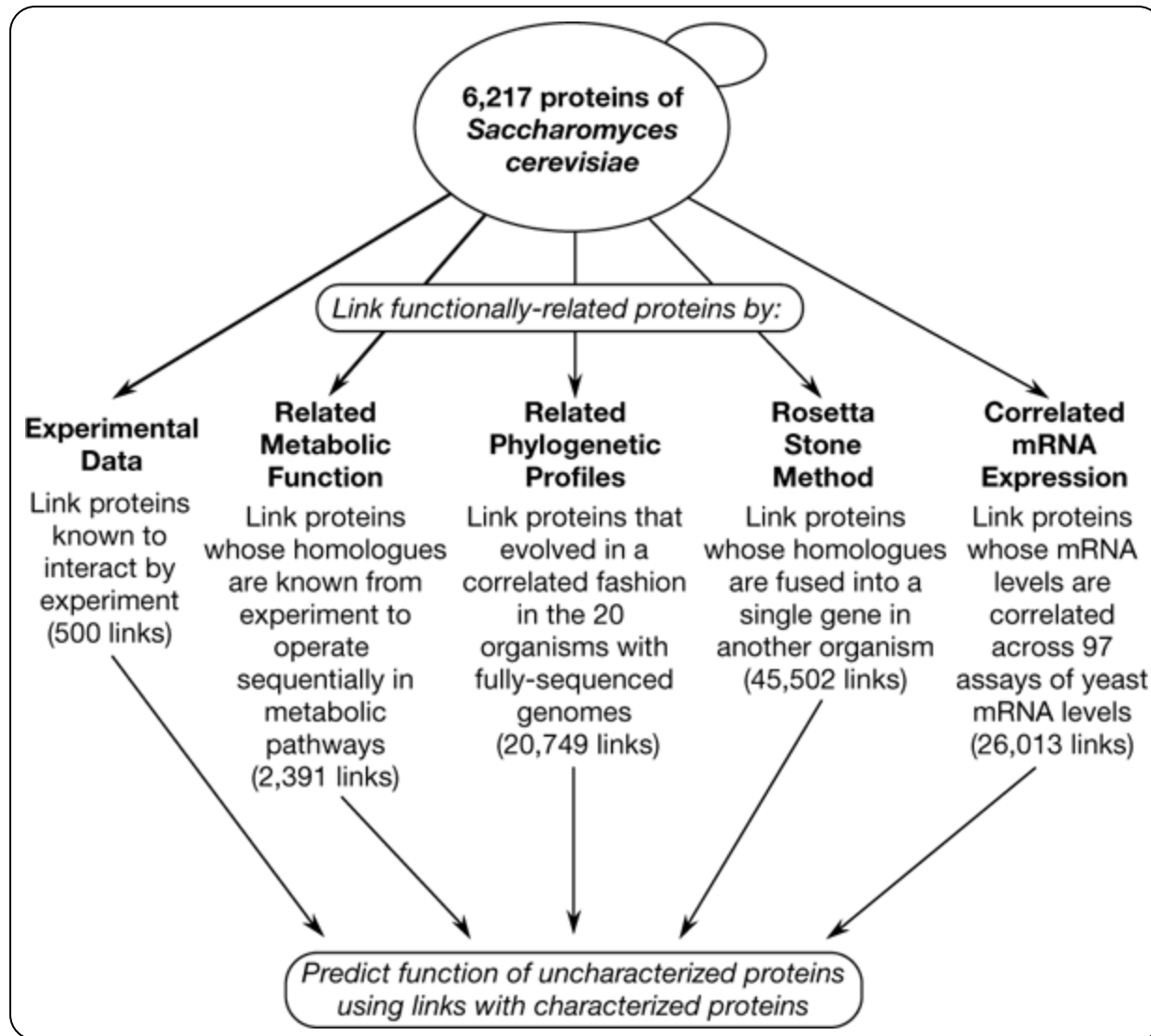
## *E. coli* TrpC

## Yeast TrpG

## Yeast TrpF



# More methods...



Marcotte EM, *et al.*, Nature (1999) 402:83-86

Enright AJ, *et al.*, Nature (1999) 404:86-90

# The Gene Ontologies (GO)

## Molecular Function Ontology

- the tasks performed by individual gene products; examples are *carbohydrate binding* and *ATPase activity*

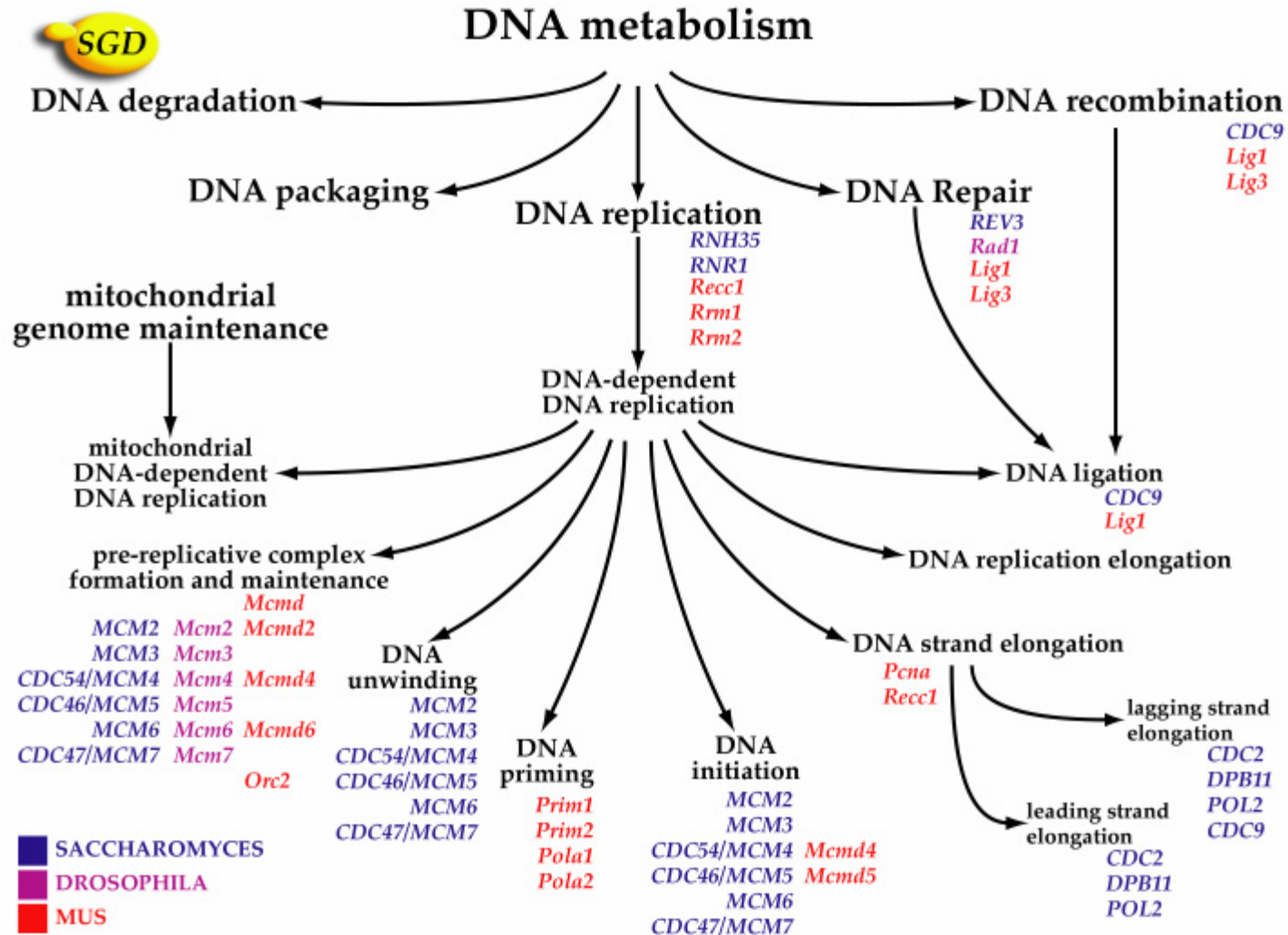
## Biological Process Ontology

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

## Cellular Component Ontology

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

# Gene Ontology - Process



# Gene Ontology – Linking Homologs Between Species

- Major Founding Members of Consortium
  - Drosophila (fruit fly) - FlyBase
  - Saccharomyces Genome Database (SGD)
  - Mouse Genome Database (MGD)

<http://www.geneontology.org/>

## Mappings of External Classification Systems to GO

These files contain concepts from systems external to GO e.g. Enzyme Commission numbers, SWISS-PROT keywords and TIGR roles, indexed to equivalent GO terms. The mappings are typically made manually, details can be found in the file header. The files are of the format:

```
external system identifier: external system term name/id > GO:GO term name ; GO:id
```

Database	Index File	Source	Date of last update
UniProt Knowledgebase	<a href="#">spkw2go</a>	Evelyn Camon ( <i>Note: spkw2go used to be called swp2go, all files remain the same.</i> )	Monthly
COG Functional Categories	<a href="#">cog2go</a>	Michael Ashburner and Jane Lomax	June 2004
Enzyme Commission	<a href="#">ec2go</a>	Michael Ashburner, GO Editorial Office	Monthly
EGAD	<a href="#">egad2go</a>	Michael Ashburner	October 2000
GenProtEC	<a href="#">genprotec2go</a>	Heather Butler and Michael Ashburner	December 2000
TIGR Role	<a href="#">tigr2go</a>	Michael Ashburner	January 2004
TIGR Families	<a href="#">tigrfams2go</a>	TIGR Staff	September 2004
InterPro	<a href="#">interpro2go</a>	Nicola Mulder	Monthly
MIPS Funct	<a href="#">mips2go</a>	Michael Ashburner and Midori Harris	August 2002
MetaCyc Pathways and Reactions	<a href="#">metacyc2go</a>	Michael Ashburner, Midori Harris and Amelia Ireland	Daily
MultiFun Classifications	<a href="#">multifun2go</a>	Michael Ashburner, Jane Lomax and Margrethe Hauge Serres	October 2003
Pfam Domains	<a href="#">pfam2go</a>	Generated electronically from InterPro2go by Daniel Barrell, Original by Nicola Mulder.	Monthly
Prodom Domains	<a href="#">prodom2go</a>	Generated electronically from InterPro2go by Daniel Barrell, Original by Nicola Mulder.	Monthly
Prints Domains	<a href="#">prints2go</a>	Generated electronically from InterPro2go by Daniel Barrell, Original by Nicola Mulder.	Monthly
ProSite Domains	<a href="#">prosite2go</a>	Generated electronically from InterPro2go by Daniel Barrell, Original by Nicola Mulder.	Monthly
Reactome biological processes	<a href="#">reactome2go</a>	Lisa Matthews and Amelia Ireland	Daily
Smart Domains	<a href="#">smart2go</a>	Generated electronically from InterPro2go by Daniel barrell, Original by	Monthly