

*Annotations des génomes
Analyse du transcriptome
Analyse des voies métaboliques*

Jacques van Helden

Jacques.van-Helden@univ-amu.fr

Aix-Marseille Université, France

Technological Advances for Genomics and Clinics
(TAGC, INSERM Unit U1090)

<http://jacques.van-helden.perso.luminy.univ-amu.fr/>

Annotation des génomes

- Rappel des prérequis
 - Structure des gènes et des génomes
 - Alignements
 - Homologie
- Annotation des génomes
 - Organisation des génomes
 - Bases de données génomiques
 - Localisation des gènes
 - Annotation de fonction par similarité de séquences
 - Coupable par association
- L'analyse du transcriptome
 - Détection de gènes exprimés différentiellement
 - Clustering des gènes
- Annotation métabolique
 - Bases de données métaboliques
 - Projection métabolique
- Enrichissement fonctionnel
- Réseaux d'interaction

Statistiques pour l'analyse du transcriptome

- Rappels/prérequis
 - Test de comparaison de moyenne
 - Choix d'un test en fonction des hypothèses de travail (Student, Welch, Wilcoxon)
 - Population, échantillon, échantillonnage
 - Interprétation d'une p-valeur
- Sources de variations en analyse du transcriptome
- Interprétation de la p-valeur
 - Ce qu'elle veut dire et ne veut pas dire
 - Significativité versus effet: volcano plots
 - Corrections de tests multiples
- Contrôles des modèles
 - Genèse de jeux de contrôle: données artificielles, permutations des valeurs, ...
 - Distributions de p-valeurs
 - Courbes de ROC
 - Evaluation de la robustesse par rééchantillonnage
- Clustering
- Enrichissement fonctionnel

Quelques rappels

Structure d'un gène eucaryote

- Dessin au tableau – photo à insérer ici

Structure d'un gène eucaryote

- **Exon does not mean coding !!!**
 - 3' UTR, 5' UTR
 - There are non-coding genes (tRNAs, rRNAs, lncRNAs, ...), which may be spliced.
- The only valid definition of exon / intro relates to the splicing mechanism !!!

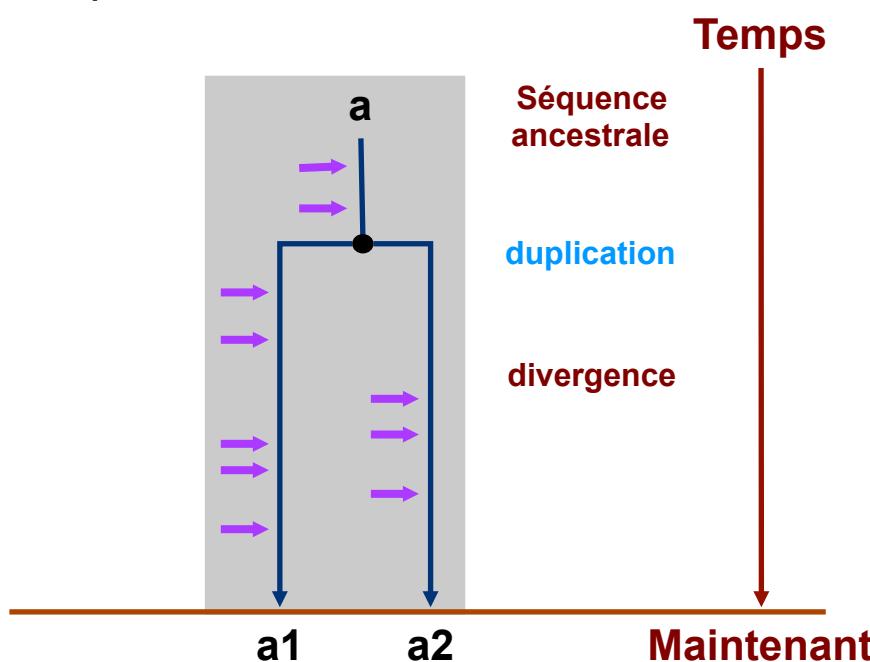
<https://en.wikipedia.org/wiki/Exon>



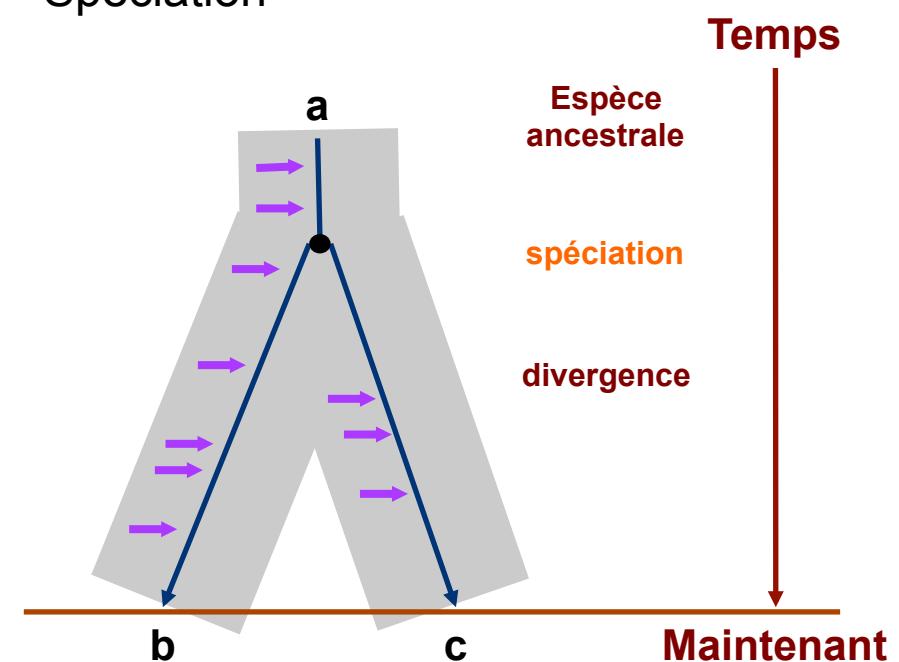
Scénarios évolutifs

- Nous disposons de deux séquences, et nous supposons qu'elles divergent d'un ancêtre commun.
- La divergence peut résulter
 - d'une **duplication** (dédoublement d'un segment d'ADN menant à la formation de plusieurs copies dans le même génome)
 - ou d'une **spéciation** (formation d'espèces séparées à partir d'une espèce unique).
- Les **flèches violettes** indiquent les mutations (substitutions, délétions, insertions) qui s'accumulent au sein d'une séquence particulière au cours de son histoire évolutive. Ces mutations sont à l'origine de la diversification des séquences, des structures et des fonctions.

Duplication

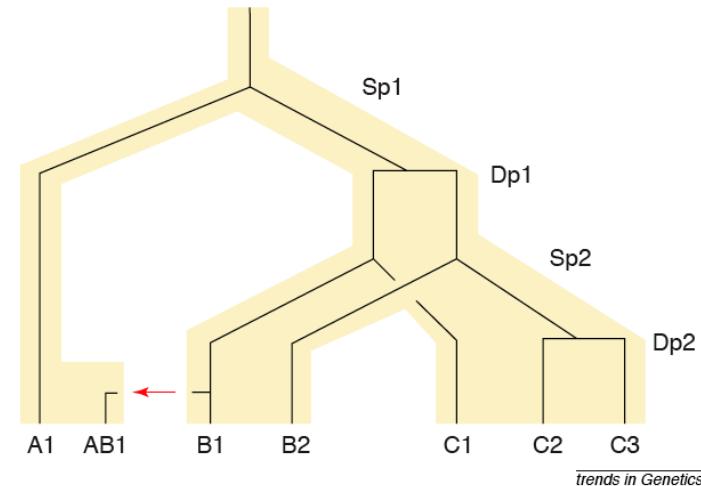


Spéciation



Représentation détaillée des événements de spéciation / duplication

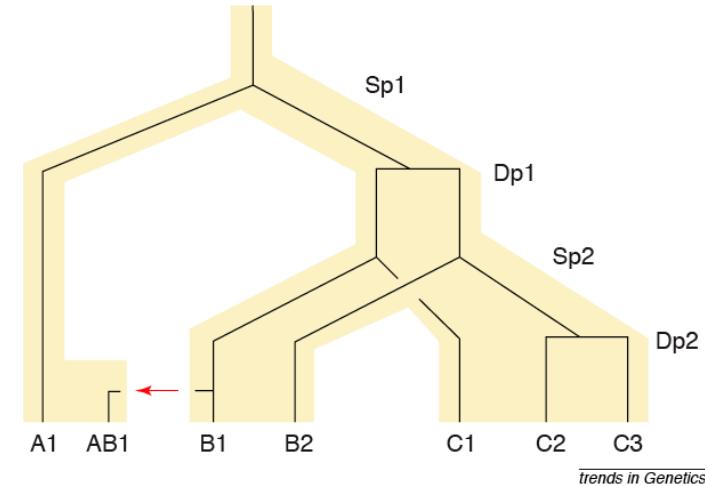
- La figure de droite combine deux niveaux de représentation
 - Les lignes fines représentent les relations évolutives entre molécules (arbre des molécules).
 - Les ombrages épais représentent l'arbre des espèces.
- Les **spéciations (Sp)** sont représentées par des branchements triangulaires sur l'arbre des espèces
 - En cas de spéciation, la molécule ancestrale se retrouve dans chacune des espèces dérivées.
- Les **duplications (Dp)** sont représentées par des branchements rectangulaires.
 - En cas de duplication, on retrouve au sein de la même espèce deux copies de la séquence ancestrale.



The idealized evolution of a gene (lines) is shown from a common ancestor in an ancestral population (the gray background), descending to three populations labelled A, B and C. There are two speciation events (Sp1 and Sp2), each occurring at the junctions shown as an upside down Y. There are also two gene-duplication events (Dp1 and Dp2), depicted by a horizontal bar. Two genes whose common ancestor resides at a horizontal bar junction (speciation) are orthologous. Two genes whose common ancestor resides at a horizontal bar junction (gene duplications) are paralogous. Thus, C2 and C3 are paralogous to each other but are orthologous to B2. Both are paralogous to B1 but orthologous to A1. The red arrow denotes the transfer of the B1 gene from species B to species A. As a result, the AB1 gene is xenologous to all six other genes. All three subtype relationships are reflexive, that is, $A1 \Rightarrow B1$ implies $B1 \Rightarrow A1$ where \Rightarrow should be read, for example, as 'is orthologous to.' However, the relationships are not transitive. Thus, $C2 \Rightarrow A1 \Rightarrow C3$ might be true, but it is not necessarily therefore true that $C2 \Rightarrow C3$, as indeed it is not in the figure if \Rightarrow is read as 'is orthologous to.' A different non-transitivity occurs for 'is paralogous to' with $B2 \Rightarrow C1 \Rightarrow C2$.

Définitions des concepts d'après Fitch (2000)

- L'article de Fitch (2000) définit les concepts suivants.
 - Fitch, W. M. (2000). Homology a personal view on some of the problems. Trends Genet 16, 227-31.
- **Homologie**
 - Owen (1843). « le même organe sous toutes ses variétés de forme et de fonction ».
 - Fitch (2000). L'homologie est la relation entre toute paire de caractères qui descendent, généralement avec divergence, d'un caractère ancestral commun.
 - Note: "caractère" peut se référer à un trait phénotypique, un site d'une séquence, à un gène entier, ...
 - Application moléculaire: deux gènes sont homologues s'ils divergent d'un gène ancestral commun.
- **Analogie:** relation entre deux caractères qui se sont développés de façon convergente à partir d'ancêtres non-apparentés.
- **Cénancêtre:** l'ancêtre commun le plus récent pour les groupes taxonomiques considérés.
- **Orthologie:** relation entre deux caractères homologues dont l'ancêtre commun se trouve chez le cénancêtre des taxa à partir desquels les séquences ont été obtenues.
- **Paralogie:** relation entre deux caractères émanant d'une duplication de gène pour ce caractère.
- **Xénologie:** relation entre deux caractères dont l'histoire, depuis leur dernier ancêtre commun, inclut un transfert entre espèces (horizontal) du matériel génétique pour au moins l'un de ces caractères.



The idealized evolution of a gene (lines) is shown from a common ancestor in an ancestral population (the gray background), descending to three populations labelled A, B and C. There are two speciation events (Sp1 and Sp2), each occurring at the junctions shown as an upside down Y. There are also two gene-duplication events (Dp1 and Dp2), depicted by a horizontal bar. Two genes whose common ancestor resides at a Y junction (speciation) are orthologous. Two genes whose common ancestor resides at a horizontal bar junction (gene duplications) are paralogous. Thus, C2 and C3 are paralogous to each other but are orthologous to B2. Both are paralogous to B1 but orthologous to A1. The red arrow denotes the transfer of the B1 gene from species B to species A. As a result, the AB1 gene is xenologous to all six other genes. All three subtype relationships are reflexive, that is, $A1 \Rightarrow B1$ implies $B1 \Rightarrow A1$ where \Rightarrow should be read, for example, as 'is orthologous to.' However, the relationships are not transitive. Thus, $C2 \Rightarrow A1 \Rightarrow C3$ might be true, but it is not necessarily therefore true that $C2 \Rightarrow C3$, as indeed it is not in the figure if \Rightarrow is read as 'is orthologous to.' A different non-transitivity occurs for 'is paralogous to' with $B2 \Rightarrow C1 \Rightarrow C2$.

Analogie

Homologie

Paralogie

Xénologie ou non
(xénologues issus de paralogues)

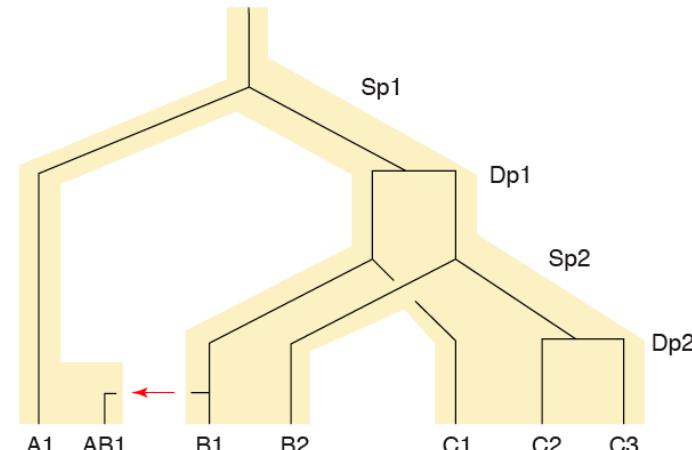
Orthologie

Xénologie ou non
(xénologues issus d'orthologues)

Exercice

- Sur base des définitions de **Zvelebil & Baum's** (ci-dessous), qualifiez la relation entre chaque paire de gènes dans le schéma de Fitch (ci-contre).

- P paralogie
- O orthologie
- X xenologie
- A analogie



Trends in Genetics

	A1	AB1	B1	B2	C1	C2	C3
A1							
AB1							
B1							
B2							
C1							
C2							
C3							

- Paire d'**orthologues**: paire de gènes dont le dernier ancêtre commun précède immédiatement un événement de spéciation (ex: a_1 and a_2).
- Paire de **paralogues**: paire de gènes dont le dernier ancêtre commun précède immédiatement une duplication génique (ex: b_2 and $b_{2'}$).

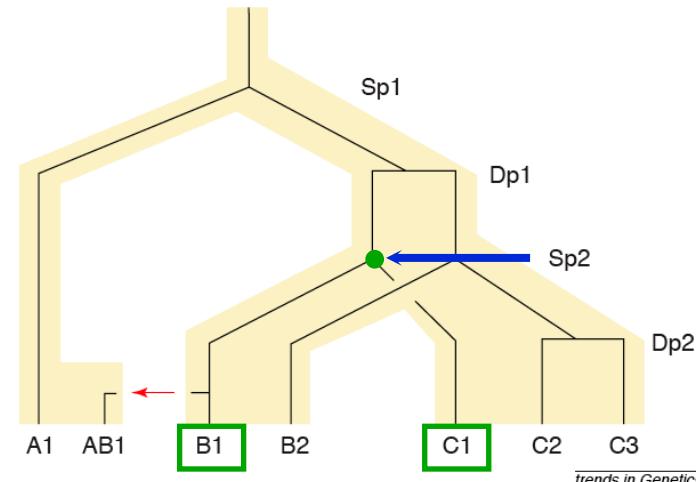
Source: Zvelebil & Baum, 2000

Exercice

- **Exemple: B1 versus C1**

- Les deux séquences (B1 and C1) proviennent respectivement des taxa B and C.
- Le cénancêtre (**flèche bleue**) est le taxon qui précède le second événement de spéciation (Sp2).
- Le gène ancestral commun (**point vert**) coïncide avec le cénancêtre.

- -> B1 et C1 sont orthologues



	A1	AB1	B1	B2	C1	C2	C3
A1							
AB1							
B1							
B2							
C1							
C2							
C3							

- **Paire d'orthologues:** paire de gènes dont le dernier ancêtre commun précède immédiatement un événement de spéciation (ex: a_1 and a_2).
- **Paire de paralogues:** paire de gènes dont le dernier ancêtre commun précède immédiatement une duplication génique (ex: b_2 and $b_{2'}$).

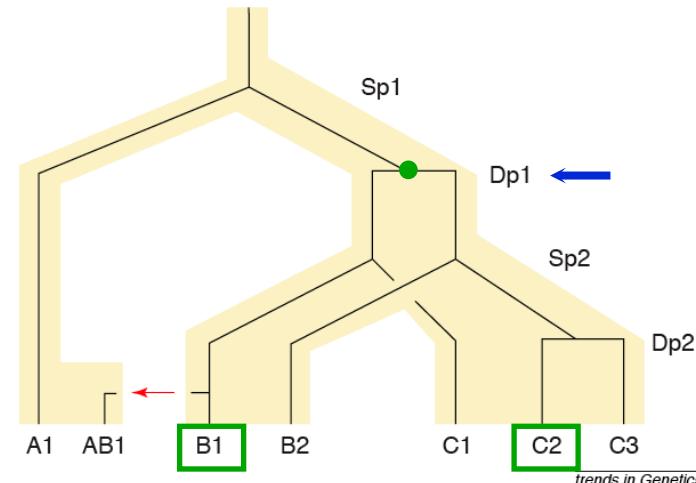
Source: Zvelebil & Baum, 2000

Exercice

- **Exemple: B1 versus C2**

- Les deux séquences (B1 and C2) proviennent respectivement des taxa B and C.
- Le dernier gène ancestral commun (**point vert**) est celui qui précède immédiatement la duplication Dp1.
- Cet ancêtre commun est bien antérieur à la spéciation qui a séparé les espèces B et C (**flèche bleue**).

- -> B1 et C2 sont paralogues



	A1	AB1	B1	B2	C1	C2	C3
A1							
AB1							
B1							
B2							
C1			O				
C2				P			
C3							

- **Paire d'orthologues:** paire de gènes dont le dernier ancêtre commun précède immédiatement un événement de spéciation (ex: a_1 and a_2).
- **Paire de paralogues:** paire de gènes dont le dernier ancêtre commun précède immédiatement une duplication génique (ex: b_2 and b_2').

Source: Zvelebil & Baum, 2000

Bases de données biomoléculaires

Exemples de bases de données biomoléculaires

- Séquence et structure des macromolécules
 - Séquences protéiques ([UniProt](#))
 - Séquences nucléotidiques (EMBL / [ENA](#), [Genbank](#), [DDBJ](#))
 - Structures tridimensionnelles des protéines ([PDB](#))
 - Motifs structurels ([CATH](#))
 - Motifs dans les séquences ([PROSITE](#), [PRODOM](#))
- Génomes
 - Bases de données génériques ([Ensembl](#), [UCSC](#), [Integr8](#), [NCBI genome](#), ...)
 - Bases de données spécifiques d'un organisme (SGD, FlyBase, AceDB, PlasmoDB, ...)
- Fonctions moléculaires
 - Fonctions enzymatiques, catalyses ([Exasy](#), [LIGAND/KEGG](#), [BRENDA](#))
 - Régulation transcriptionnelle ([JASPAR](#), [TRANSFAC](#), [RegulonDB](#), ...)
- Processus biologiques
 - Voies métaboliques ([MetaCyc](#), [KEGG pathways](#), [Biocatalysis/biodegradation](#))
 - Interactions protéine-protéine ([DIP](#), BIND, [MINT](#))
 - Transduction de signal (Transpath)
 - ...

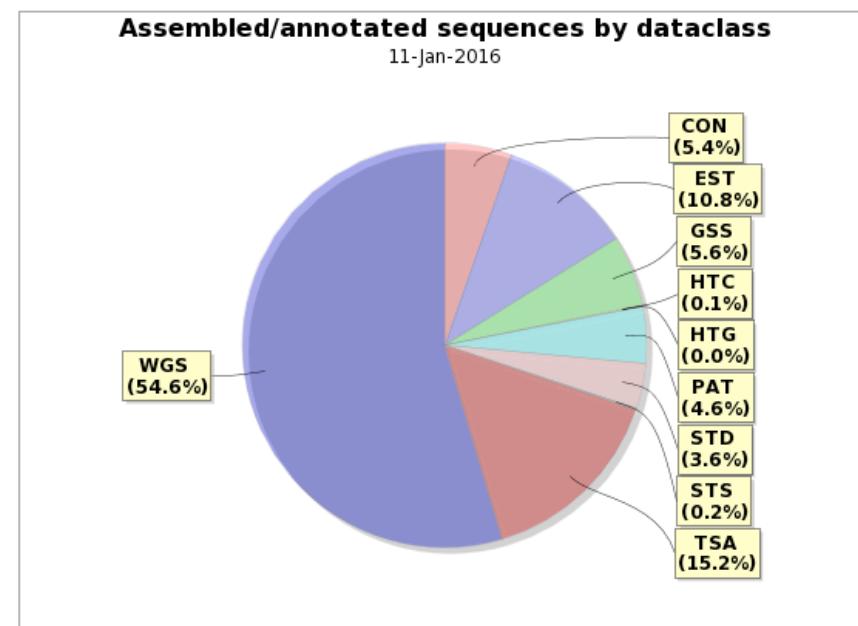
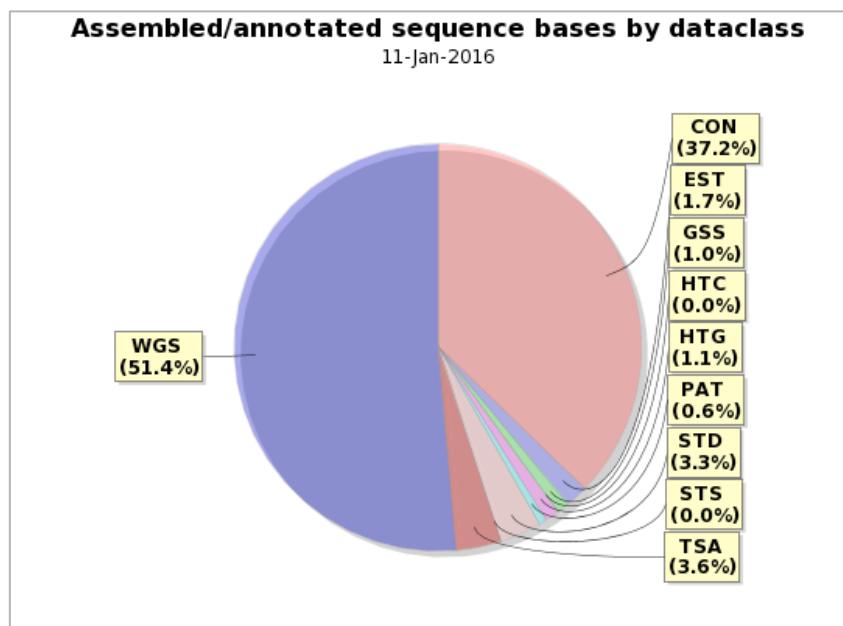
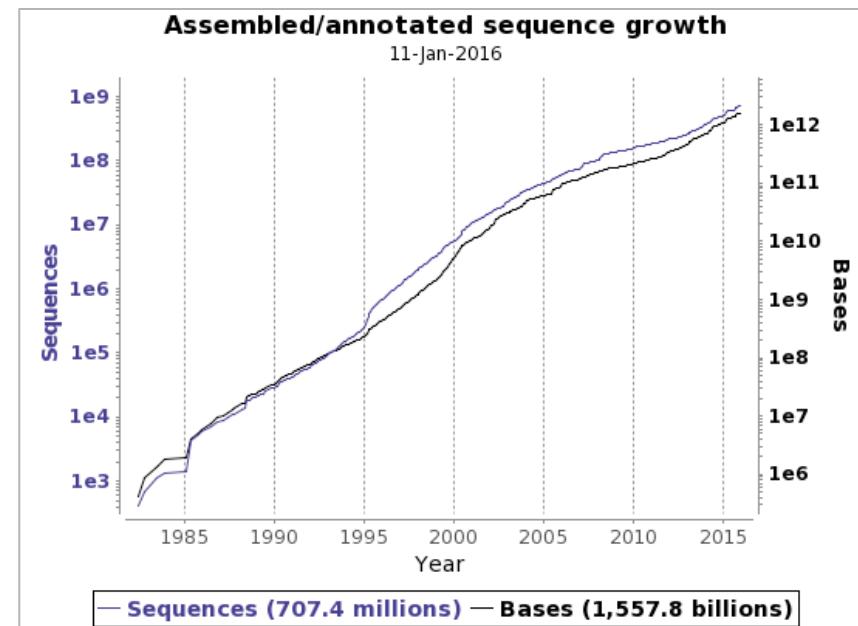
Bases de données de bases de données

- Il existe des centaines de bases de données spécialisées pour la biologie moléculaire et la biochimie. Ce nombre augmente chaque année.
- Pour s'y retrouver, la revue Nucleic Acids Research consacre chaque année son numéro de janvier à une revue des bases de données existantes, et maintient un catalogue des bases de données:
 - <http://www.oxfordjournals.org/nar/database/c/>
- Plusieurs centaines de bases de données sont disponibles.



Annotation des génomes

- Depuis 1985 on observe une croissance exponentielle des séquences soumises aux bases de données.
- Actuellement, la majorité des séquences proviennent de projets de séquençage génomique (WGS=Whole Genome Sequencing).

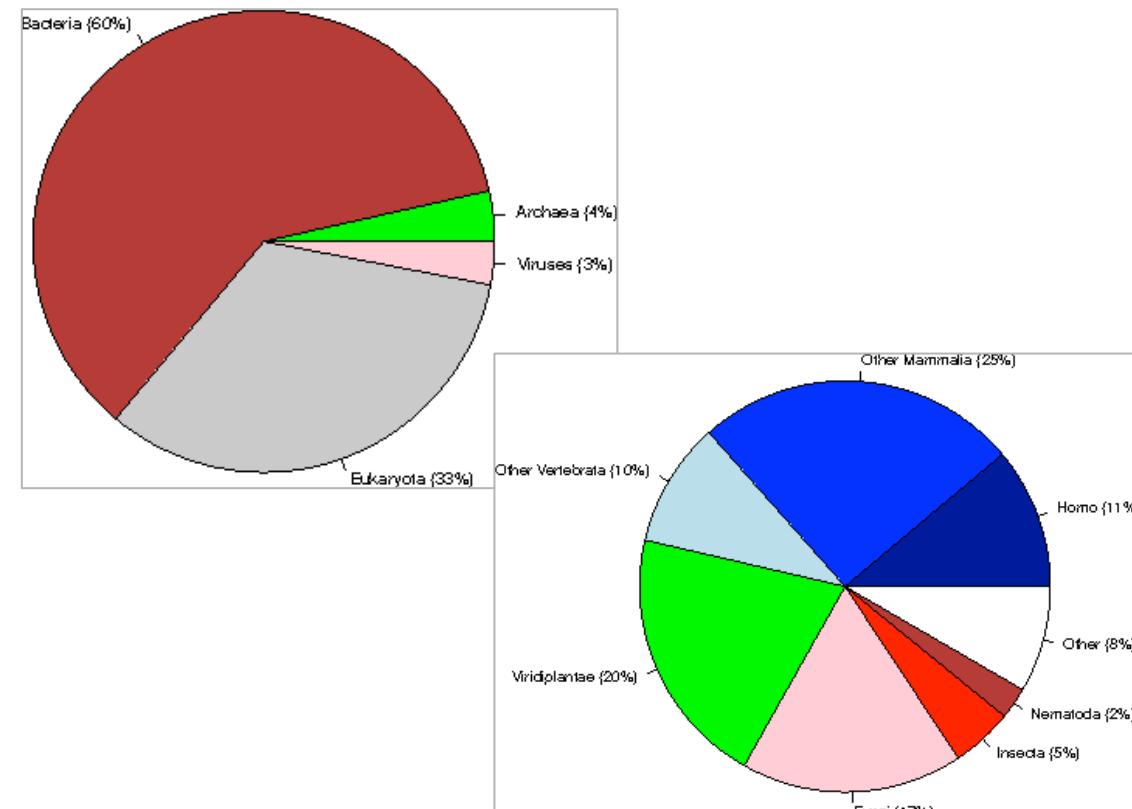
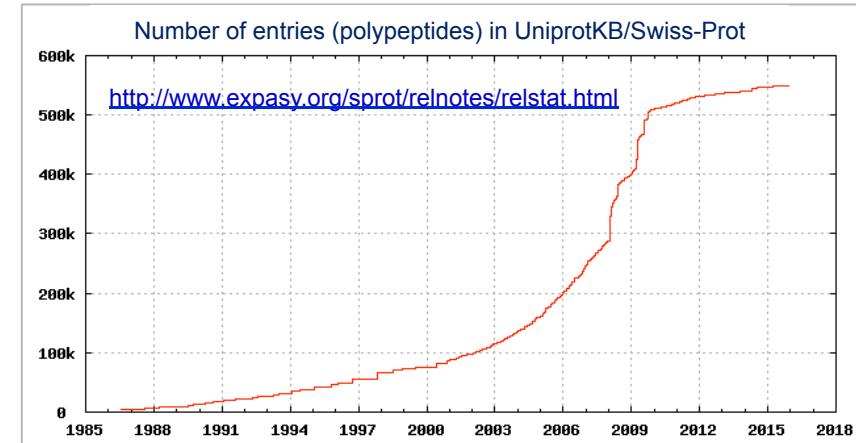


UniProt - the Universal Protein Resource

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



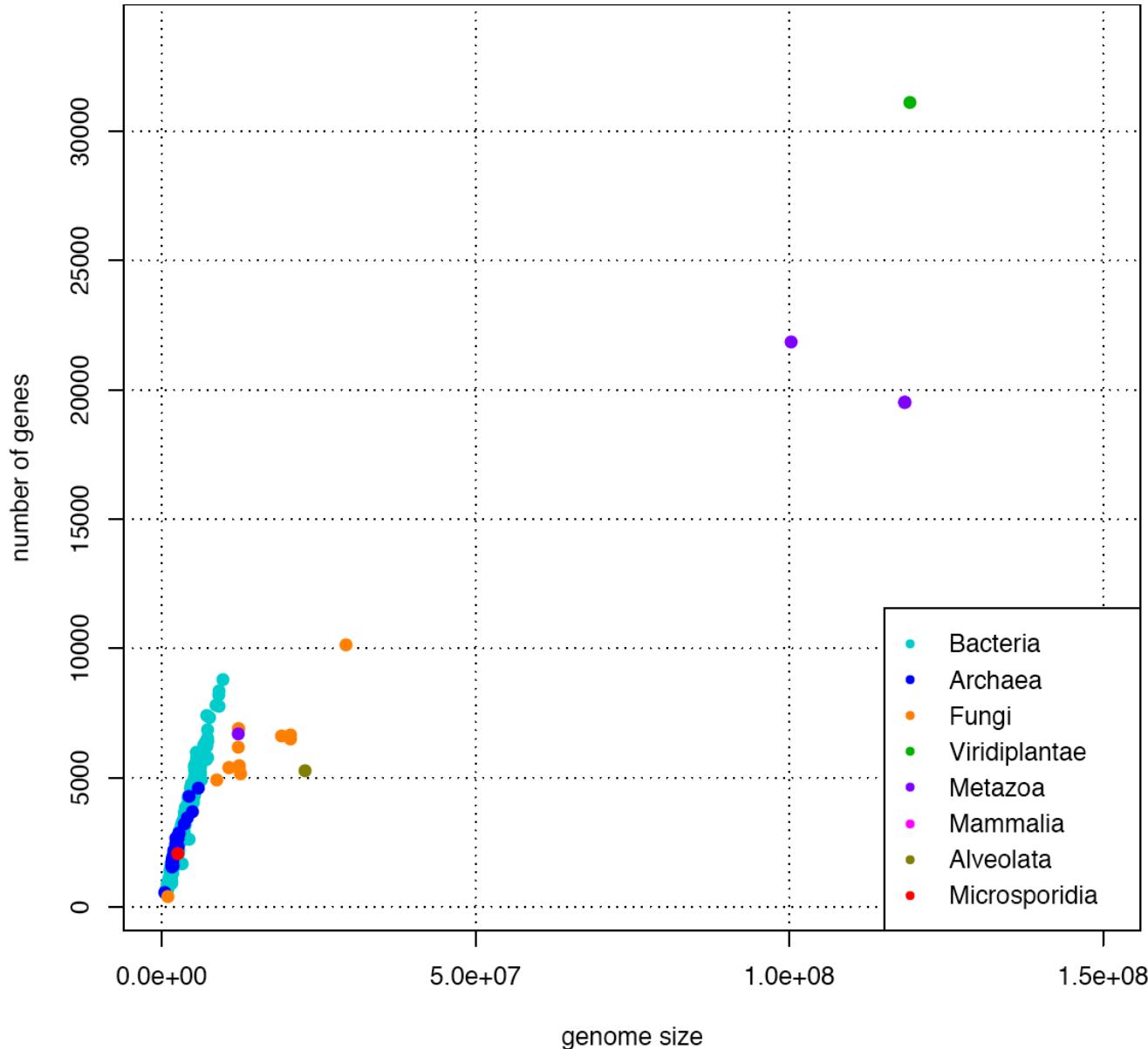
- Contenu Uniprot (11 Janvier 2016)
 - Unreviewed (TrEMBL)
 - **55.270.679 protéines**
 - Traduction et annotation automatique de toutes les séquences codantes d'EMBL
 - Section Swiss-Prot d'UniProtKB (**« reviewed »**):
 - **550.116 protéines**
 - annotation par des experts
 - Contenu informationnel important.
 - Nombreuses références à la littérature scientifique.
 - Bonne fiabilité des informations.
 - La majorité des annotations de séquences protéiques sont donc faites automatiquement, sans être vérifiées par un être humain !!!
- Swissprot
 - La base de données de protéines la plus complète au monde.
 - Une énorme équipe: >100 annotateurs + développeurs d'outils.
 - Annotation par experts, spécialistes des différents types de protéines.
- References
 - Bairoch et al. The SWISS-PROT protein sequence data bank. Nucleic Acids Res (1991) vol. 19 Suppl pp. 2247-9
 - The UniProt Consortium. The Universal Protein Resource (UniProt) 2009. Nucleic Acids Res (2008). Database Issue.



Some milestones

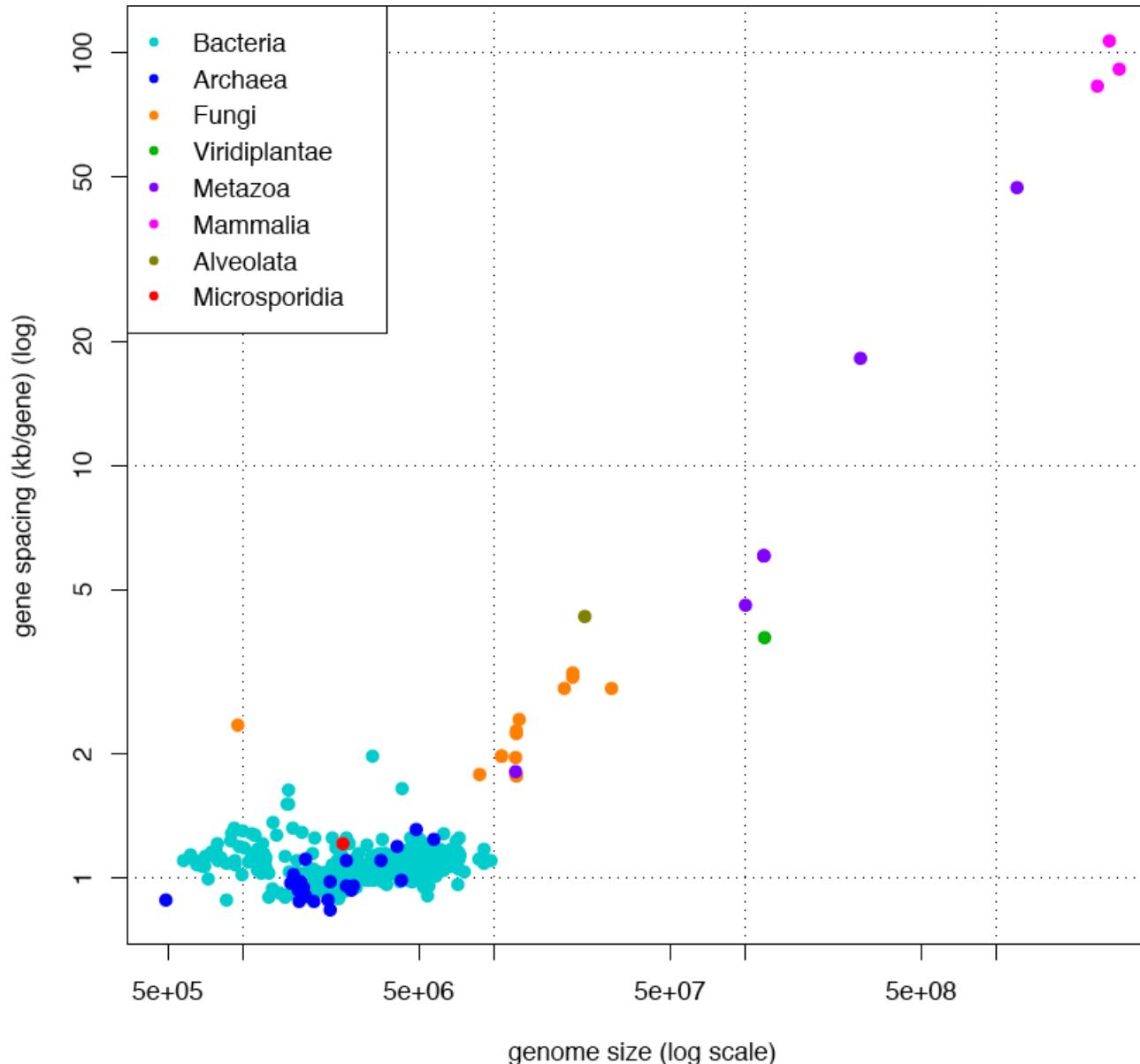
Species name	Common name	Publication year	Genome size Mb	Gene number	Mean intergenic distance Kb	Fraction of coding sequences %	Non-coding fraction %	Repetitive elements %	Transcribed fraction %	Remarks
Bactéries										
<i>Mycoplasma genitalium</i>	<i>Mycoplasma</i>	1995	0.6	481	1.2	90	10			Small genome (intracellular)
<i>Haemophilus influenzae</i>		1995	1.8	1 717	1.0	86	14			First bacterial genome sequenced
<i>Escherichia coli</i>	<i>Enterobacteria</i>	1997	4.6	4 289	1.1	87	13			
Levures										
<i>Saccharomyces cerevisiae</i>	<i>Baker's yeast</i>	1996	12	6 286	1.9	72	28			First eukaryote genome
Animaux										
<i>Caenorhabditis elegans</i>	<i>Nematod worm</i>	1998	97	19 000	5	27	73			First metazoan genome
<i>Drosophila melanogaster</i>	<i>Fruit fly</i>	2000	165	16 000	10	15	85			
<i>Ciona intestinalis</i>			174	14 180	12					
<i>Danio rerio</i>	<i>Zebrafish</i>		1 527	18 957	81					
<i>Xenopus laevis</i>	<i>Amphibian</i>		1 511	18 023	84					
<i>Gallus gallus</i>	<i>Chicken</i>		2 961	16 736	177					
<i>Ornithorynchus anatinus</i>	<i>Ornithorhynchus</i>		1 918	17 951	107					
<i>Mus musculus</i>	<i>Mouse</i>	2002	3 421	23 493	146					
<i>Pan troglodytes</i>	<i>Chimp</i>		2 929	20 829	141					
<i>Homo sapiens</i>	<i>Human</i>	2001	3 200	21 528	149	2	98	46	28	Draft version in 2001
1000 génomes humains		> 2008								Project announced Jan 2008
Plantes										
<i>Arabidopsis thaliana</i>		2001	120	27 000	4	30	70			First plant genome sequenced
<i>Oryza sativa</i>	<i>Rice</i>		390	37 544	10					
<i>Zea mays</i>	<i>Maize</i>		2 500	50 000	50		50			Nb of gene is an approximation
<i>Triticum aestivum</i>	<i>Wheat</i>		16 000							Hexaploid genome
<i>Lilium</i>			120 000							
<i>Psilotum nudum</i>			250 000							

Genes and genome size



- In prokaryotes, the number of genes increases linearly with genome size
- In eukaryotes, this is not the case: the genome size increases faster than the number of genes

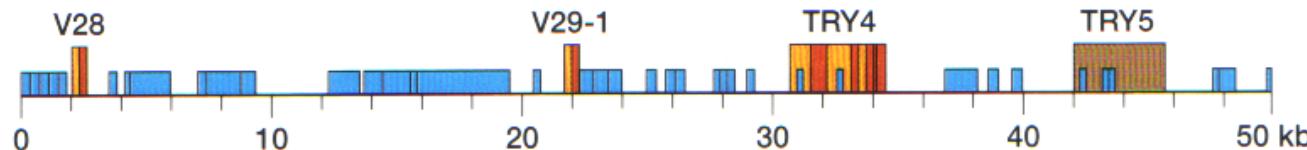
Gene spacing



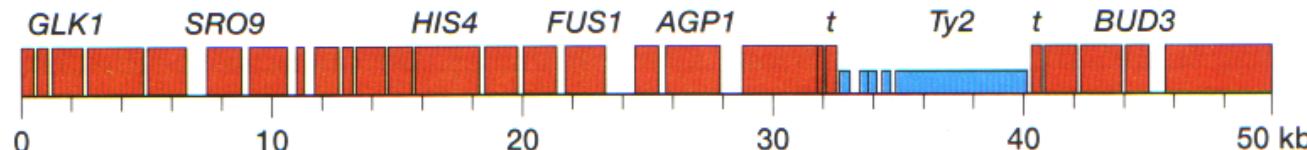
- Gene spacing increases considerably with the complexity off the organisms.
- Note: the X axis si logarithmic, not the Y axis -> the increase seems grossly exponential.

Gene organization

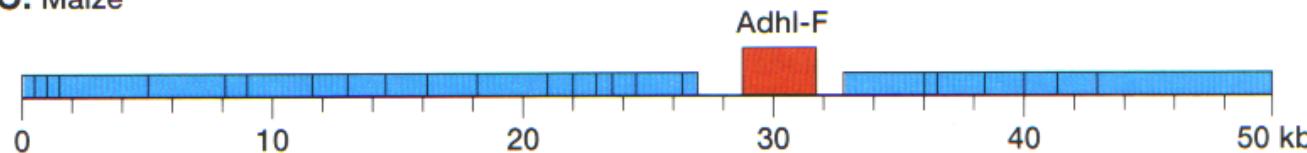
A. Human



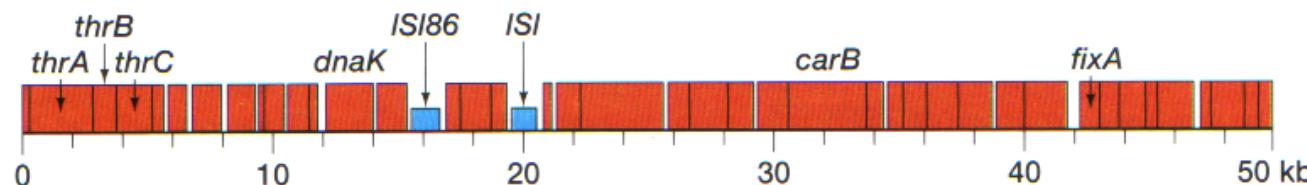
B. *Saccharomyces cerevisiae*



C. Maize



D. *Escherichia coli*



KEY

Gene Intron Human pseudogene Genome-wide repeat t tRNA gene

Source: Mount (2000)

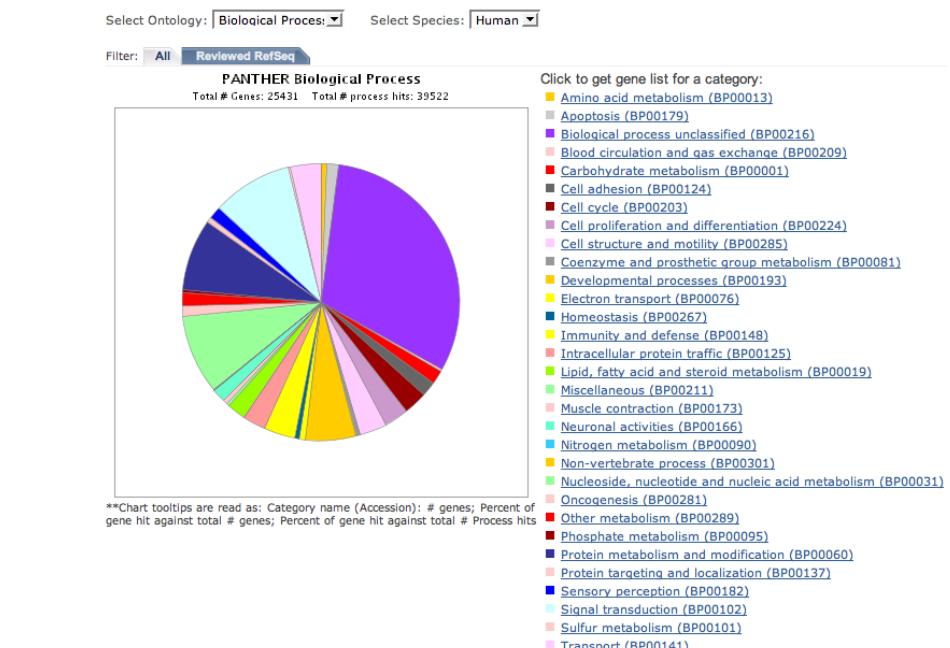
Approches d'annotation de la fonction des gènes

- Expérimentation: phénotypes de perte et gain de fonction, action d'inhibiteurs, caractérisation biochimique des protéines, ...
- Annotation par similarité de séquence.
- Approches "coupable par association"
 - Appartenance au même opéron
 - Fusions de gènes
 - Profils phylogénétiques
- Analyse du transcriptome
 - Groupes de gènes co-exprimés
- Analyses de réseaux
 - Interactions protéines-protéines
 - Interactions fonctionnelles / génétiques

Gene function

- After having localized genes on the sequence, we have to predict their function.
- Some genes have already been characterized before the genome project, but these are generally a minority of those found in the genome.
- For the majority of the genes, one tries to predict function on the basis of similarities between the sequence of the newly sequenced gene and some previously known genes (**function assignation by sequence similarity**).
- Example: yeast genome (1996): there are still 2500 genes (39%) whose function is completely unknown. However
 - Yeast is among the best known model organisms (genetics, molecular biology).
 - The full genome is available since 1996.
- When the first draft of the Human genome has been published, 60% of the predicted genes were of unknown function.

```
>PHO4_SPBC428.03C : THIAMINE-REPRESSIBLE ACID PHOSPHATASE PRECURSOR
: Q01682;Q9UU70;
Length = 463 Score = 161 bits (408), Expect = 1e-40
Identities = 138/473 (29%), Positives = 223/473 (46%), Gaps = 47/473 (9%)
Query: 9 ILAASLVNAGTIPLGKLSIDDKIGTQTEIFPFLGGSPYYSPGDYGISRDLPESCEMKQ 68
+LAAS+V+AG S + + LG Y+ P G + PESC +KQ
Sbjct: 10 LLAASIVHAGK-----SQFEAFENEFYFKDHLGTTISVYHE-PYFNGPTTSFPESCAIKQ 62
Query: 69 VQMVRGHGERYPT-----VSKAKSIMTTWYKLSNYTGQFSGALSFLNDDYEFFIRDTK 121
V ++ RHG R PT VS A+ I KL N G S+ + F T
Sbjct: 63 VHLLQRHGSRNPTGDDTATDVSSAQYIDIFQNKLLN--GSIPVNFSYPENPLYFVKHWTP 120
Query: 122 NLEMETTLANSVNVLNPYTGEMNAKRHARDFLAQYGYMVENQTSFAVFTSNRCHDTAQ 181
++ E S + G + R +Y Y + + + + T+ R D+A+
Sbjct: 121 VIKAENADQLSSS-----GRIELFDLGRQVFERY-YELFDTDVYDINTAAQERVVDSAE 173
Query: 182 YFIDGL-GDKFN--ISLOTISEAESAGANTLSAHHSCPawanDDVVNDDILKK----YDTK 233
+F G+ GD + + E +SAGAN+L+ ++SCP ++D+ D+ + +
Sbjct: 174 WFSYGMFGDDDMQNKTNFIVLPEDDSAGANSLAMYYSCPVYEDNNIDENTTEAAHTSWRN 233
Query: 234 YLSGIAKRLNKE-NKGLNLNTSSDANTFFAWCAYEINARGYSDCNICNIFTKDELVRFSYQD 292
+L IA RLNK + G NLT SD + + C YEI R SD C++FT E + F Y D
Sbjct: 234 FLKPIANRLNKYFDSGYNLTVDVRSLYICVYEINARGYSDCNICNIFTKDELVRFSYQD 293
Query: 293 LETSYQTGPYGVYDVRSLYYICVYEINARGYSDCNICNIFTKDELVRFSYQD 350
L+ Y GP + ++G N L++ + D+KV+L+FTHD+ I+ +G
Sbjct: 294 LDYAYWGGPASEWASTLGGAYVNNLANNLRKGVNNSADRKVFLAFTHDQSIIIPVEAALGF 353
Query: 351 IDDKNNLTAEH-VPFMENTF---HRSWYVPOQARVYTFKFOCS-NDTYVRYVINDAVVP 404
D +T EH +P +N F S +VP + TE F CS N YVR+++N V P
Sbjct: 354 FPD---ITPEHPLPTDKNIFTYSLKTSFVFPFAGNLITEFLCSDNKYYVRHLVNQQVYP 410
Query: 405 IETCSTGPGFS---CEINDFYDYAEKRVAGTDFLKVCNVSSVSNSTELTFFW 453
+ C GP + CE++ + + + + + ++ + N ++ST +T ++
Sbjct: 411 LTDCGYGPGSGASDGLCELSAYLNSSVRVNSTSNGIANFNSQCQAHSTNVTVYY 463
```



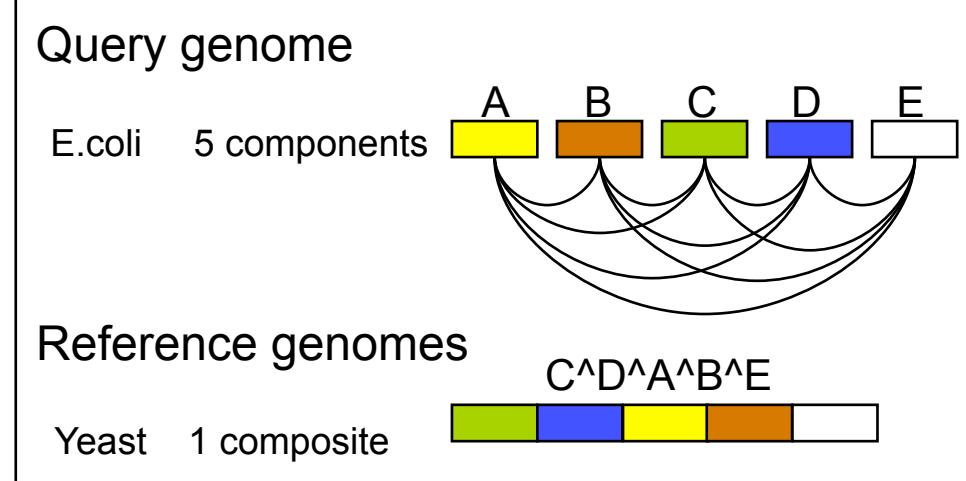
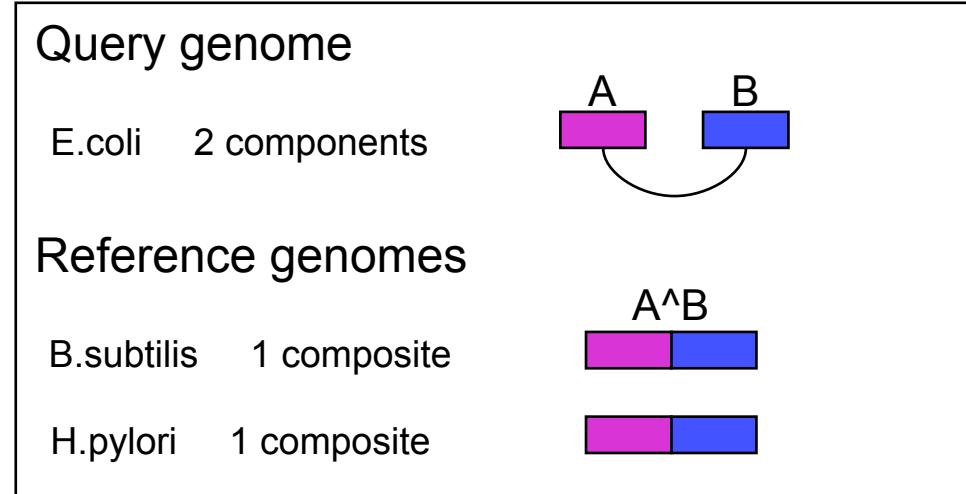
Phylogenetic profiles

- For each gene of the query genome (e.g. *E.coli*), orthologs are searched in all the sequenced genomes
- Each gene is characterized by a profile of presence/absence in all the sequenced genomes
- Groups of genes having similar phylogenetic profiles are likely to be functionally related

Gene	<i>A.aeolicus</i>	<i>C.muridarum</i>	<i>C.pneumoniae.AR39</i>	<i>Nostoc.sp</i>	<i>Synechocystis.PCC6803</i>	<i>B.halodurans</i>	<i>B.subtilis</i>	<i>C.acetobutylicum</i>	<i>C.glutamicum</i>	<i>C.perfringens</i>	<i>L.innocua</i>	<i>L.lactis</i>	<i>M.genitalium</i>	<i>M.leprae</i>	<i>M.pneumoniae</i>	<i>M.pulmonis</i>	<i>S.aureus.MW2</i>	<i>S.coelicolor</i>	<i>S.pneumoniae.R6</i>	<i>S.pyogenes</i>	<i>T.tengcongensis</i>	<i>U.urealyticum</i>	<i>F.nucleatum</i>	<i>A.tumefaciens.C58</i>	<i>B.aphidicola.Sg</i>	<i>B.melitensis</i>	<i>C.crescens</i>	<i>C.jejuni</i>	<i>H.influenzae</i>	<i>H.pylori.26695</i>	<i>M.loji</i>	<i>N.meningitidis.MC58</i>	<i>P.aeruginosa</i>	<i>R.conorii</i>	<i>S.melin<i>n*</i></i>		
16127995	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16127996	1	0	0	1	1	1	1	1	1	1	1	1	0	0	1	0	0	1	1	1	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1
16127997	1	0	0	1	1	1	1	1	1	0	1	1	0	1	0	0	0	1	1	1	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1
16127998	0	0	0	0	1	0	0	1	1	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1
16127999	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16128000	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	1	0	0	0	1	0	0	0	1	0	0	1	1	1	1	1	1	0
16128001	0	1	1	0	0	1	1	0	1	1	0	0	0	0	0	1	1	1	1	1	0	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1
16128002	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0
16128003	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	1	1	0	0	0	0	0	0	1	0
16128004	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16128005	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16128006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16128007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16128008	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
16128009	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
16128010	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Gene fusion analysis

- It is quite frequent to observe that two genes of a given organism are fused into a single gene in another organism.
- Fusions between more than 2 genes are occasionally observed.
- Fused genes are likely to be functionally related.



References

- Marcotte, et al. (1999). Science 285(5428), 751-3.
Marcotte, et al. (1999). Nature 402(6757), 83-6.
Enright, et al. (1999). Nature 402(6757), 86-90.

Bases de données biomoléculaires

La base de données “Gene ontology” (GO)

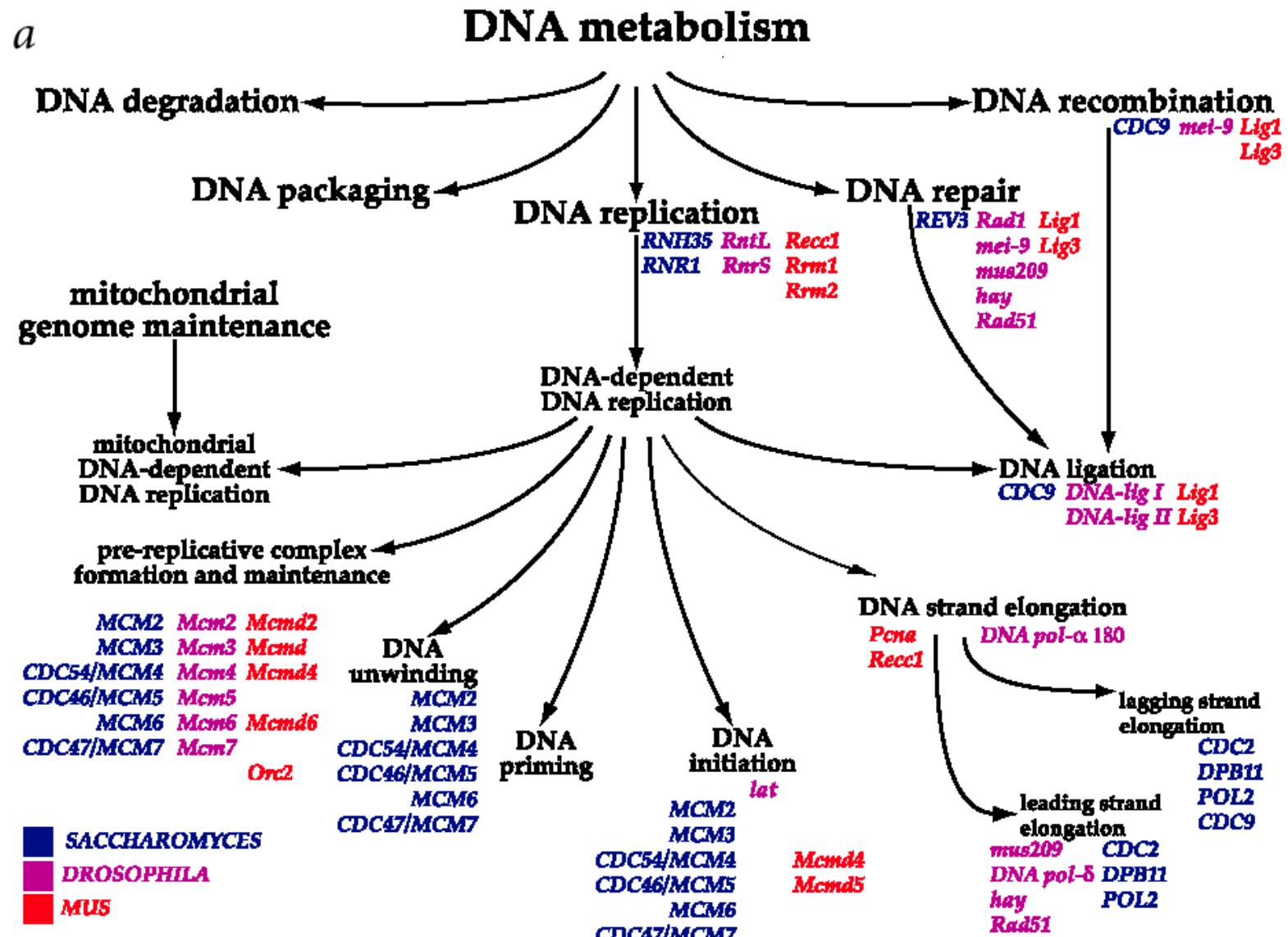
Ontologie – définition générale

- Ontologie: partie de la métaphysique qui s'intéresse à l'être en tant qu'être, indépendamment de ses déterminations particulières
- *Le Petit Robert - dictionnaire alphabétique et analogique de la langue française. 1993*

Les "bio-ontologies"

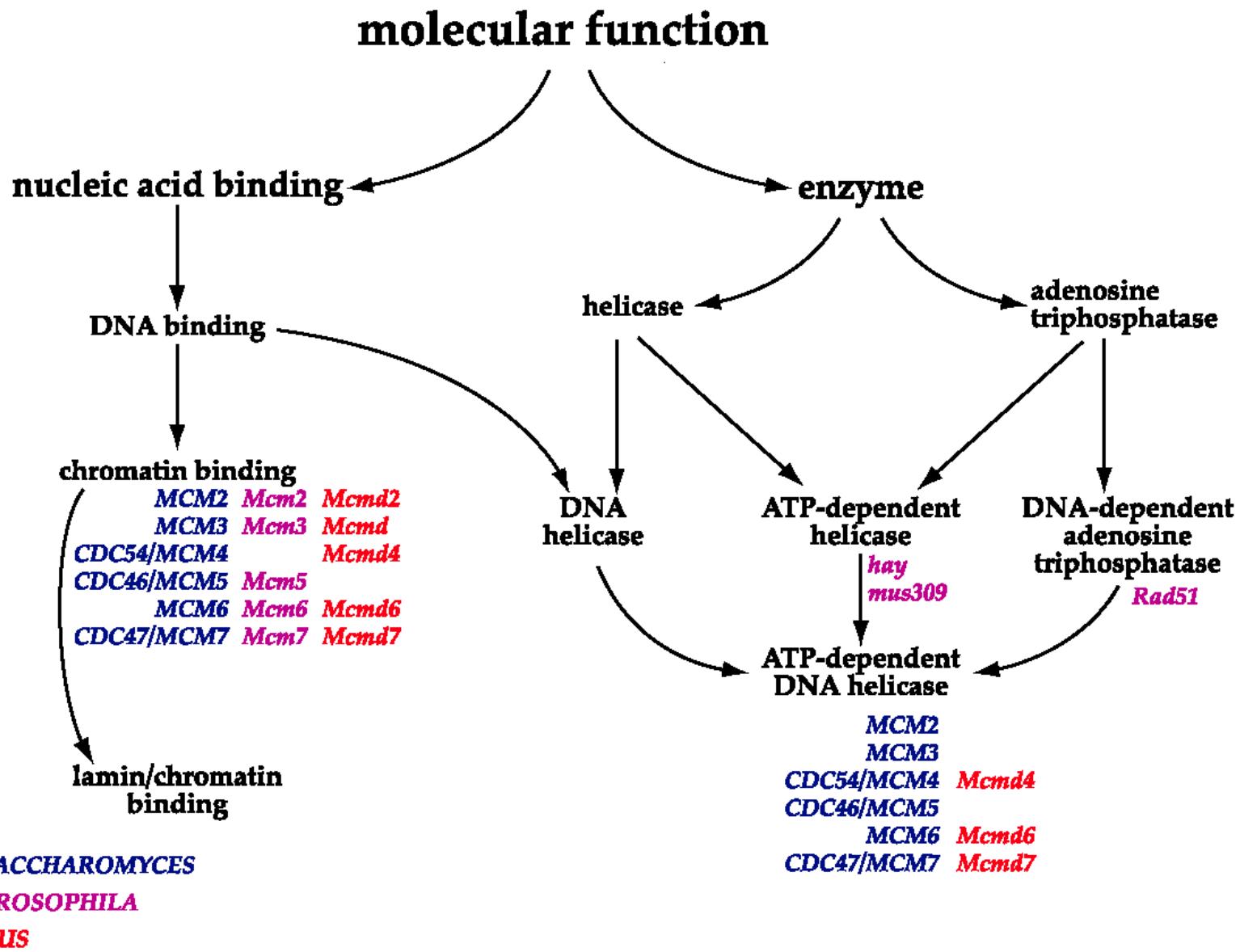
- Les bio-ontologies ne constituent pas une « ontologie » au sens philosophique du terme, elles se rapportent à un sens dérivé en informatique: classification des concepts liés à un champ disciplinaire.
- Les bio-ontologies visent à répondre au problème d'inconsistance entre annotations.
- Pour y répondre, on met en place
 - Un vocabulaire contrôlé
 - On utilise toujours le même mot pour désigner le même concept.
 - Les listes de synonymes permettent d'établir les correspondances.
 - Classification hiérarchique entre les termes de ce vocabulaire contrôlé.
- La « Gene ontology » établit une classification des gènes et protéines selon trois critères complémentaires:
 - Fonction moléculaire (ex: aspartokinase, transporteur de glucose, ...).
 - Processus biologique (ex: biosynthèse de la méthionine, réPLICATION, ...).
 - Composante cellulaire (ex: membrane mitochondriale, noyau, ...).

Gene ontology: exemples de processus biologiques

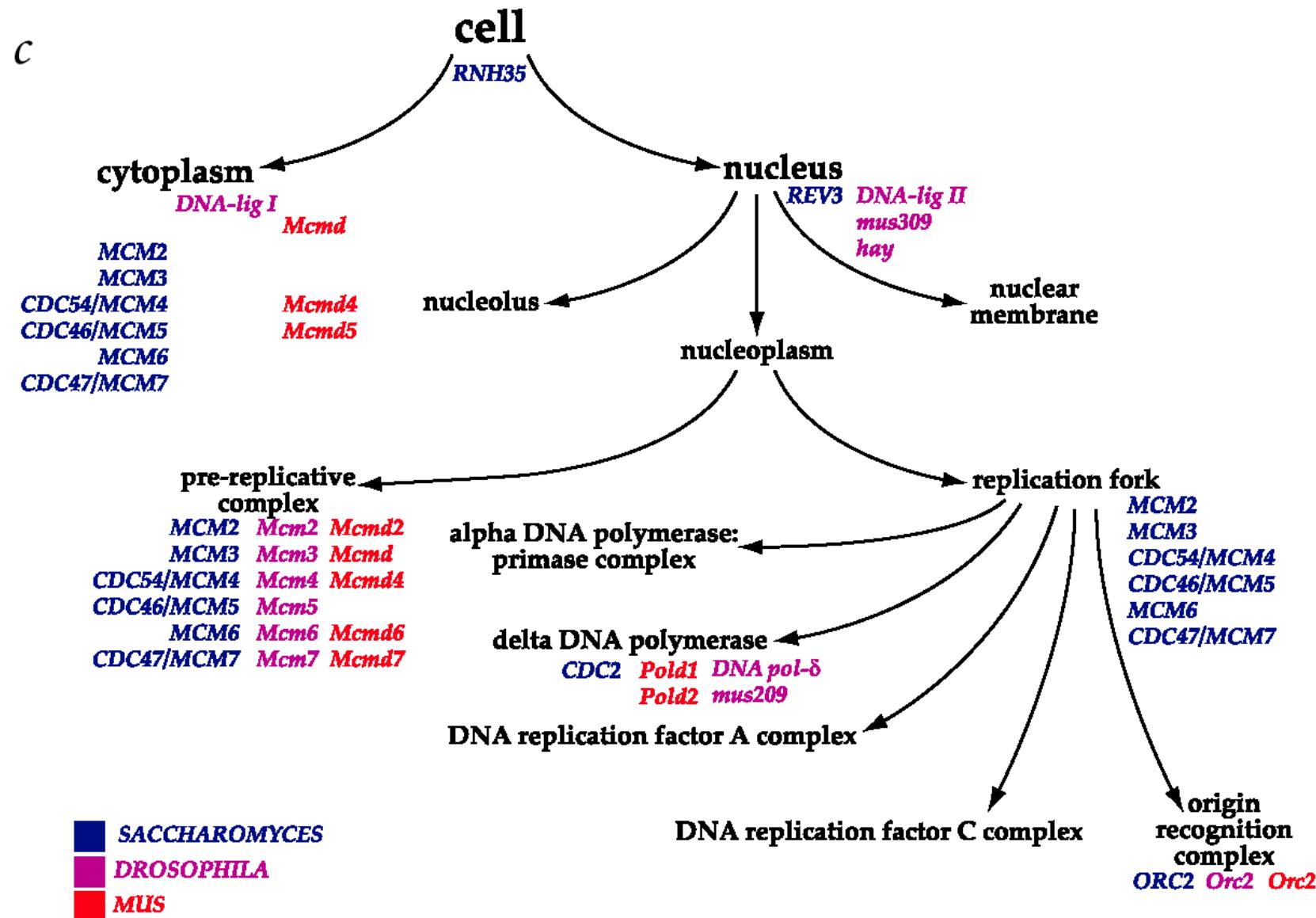


Gene ontology: exemples de fonctions moléculaires

b

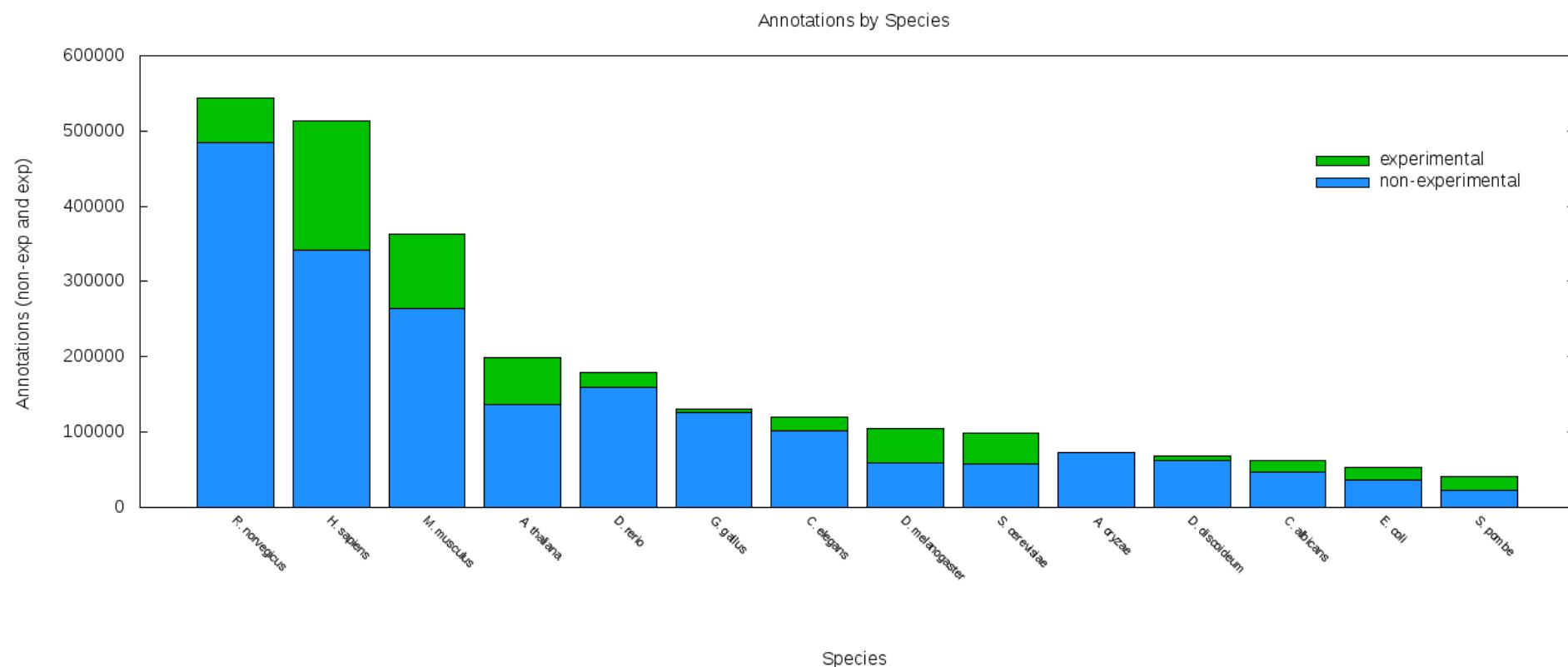


Gene ontology: exemples de composantes cellulaires



Gene ontology – some statistics

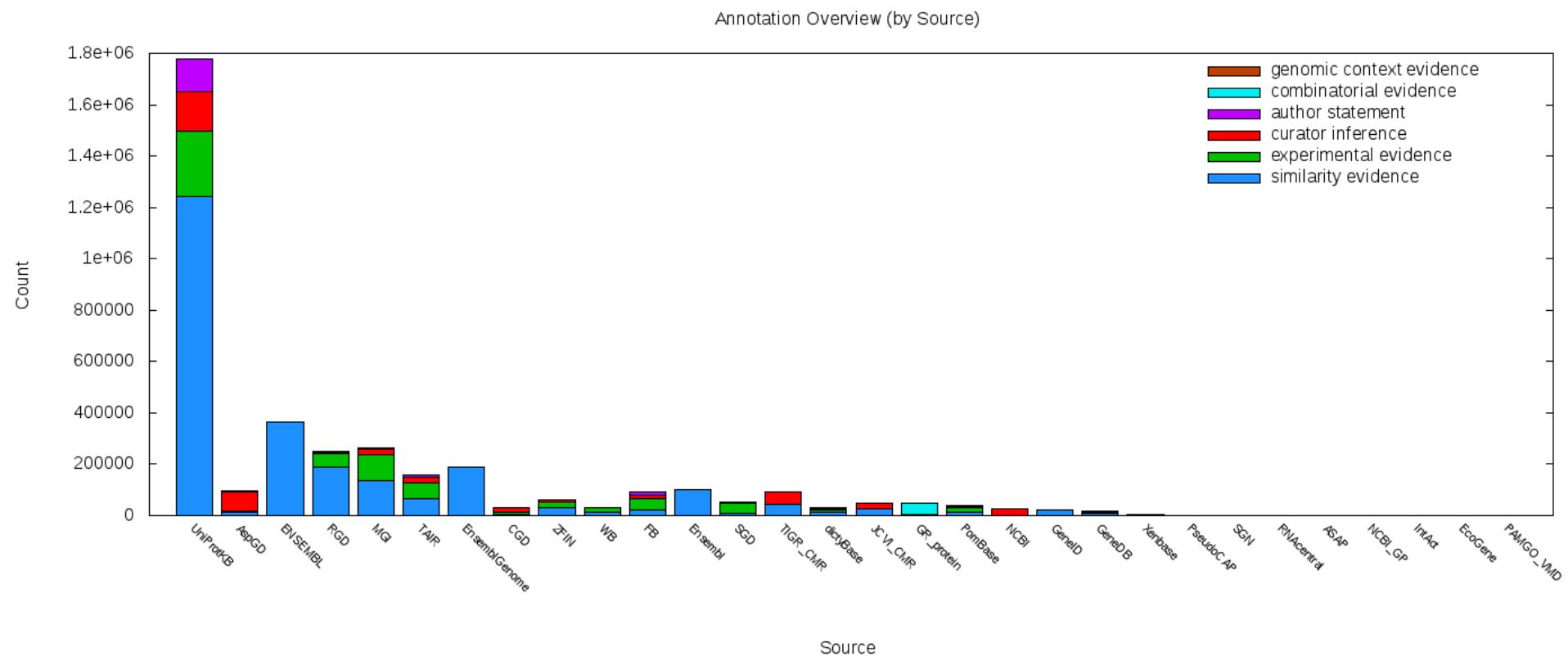
- Even for model organisms, the majority of annotations relies on "non-experimental" indications.



- <http://geneontology.org/page/current-go-statistics>

Gene ontology – some statistics

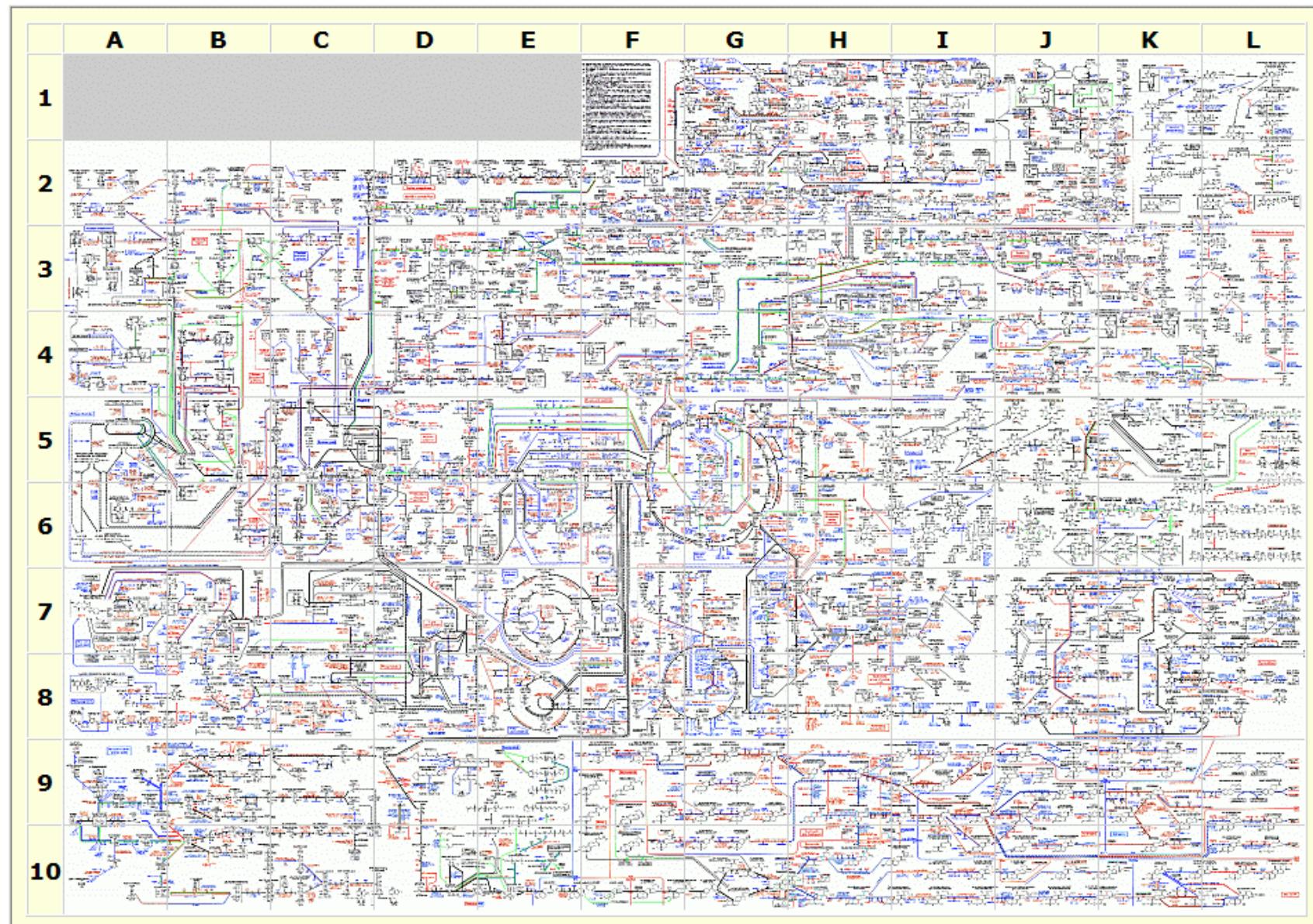
- Annotation by sequence similarity remains the main approach to annotate genes.



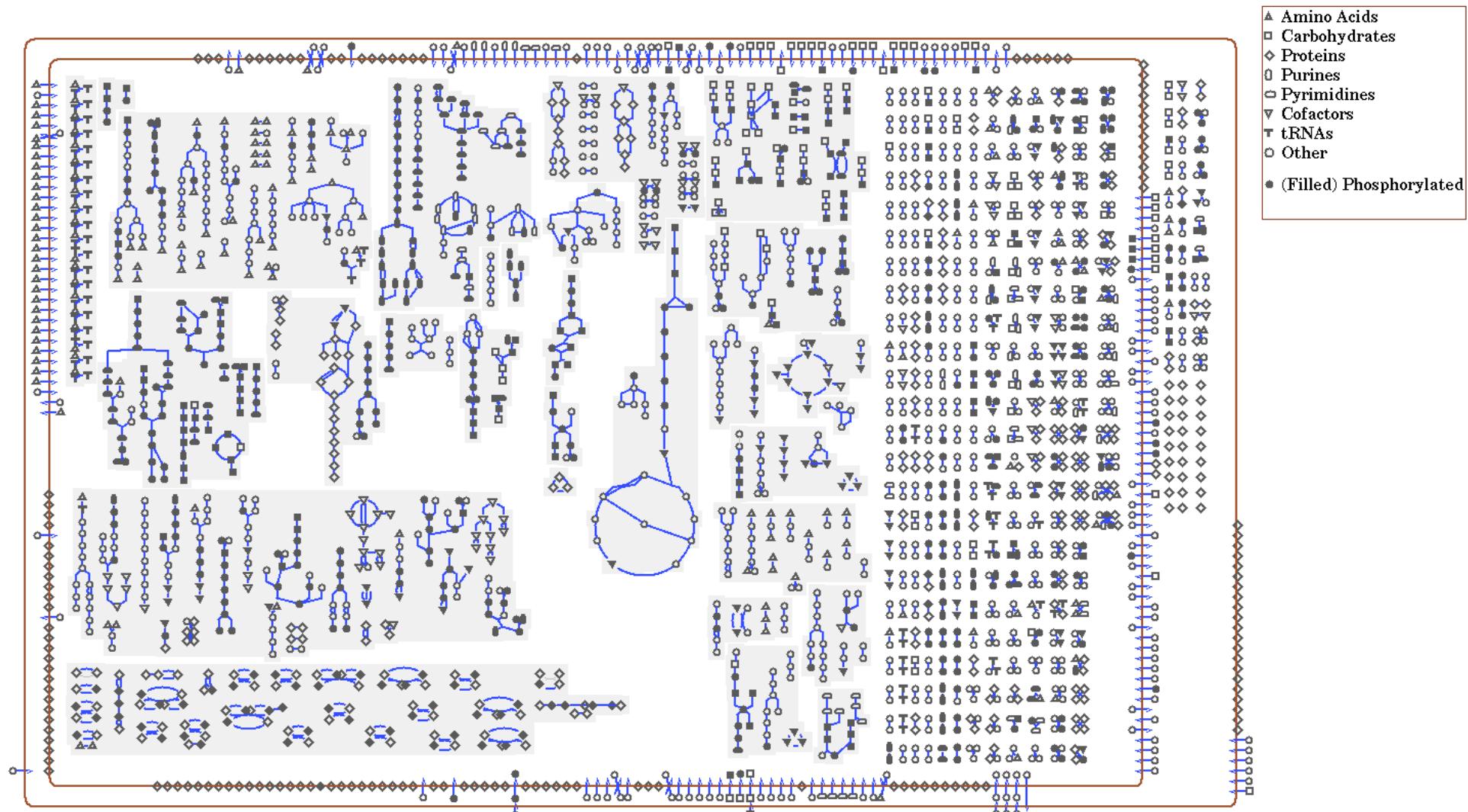
- <http://geneontology.org/page/current-go-statistics>

Annotation des voies métaboliques

Boehringer-Mannheim Metabolic Wall Chart



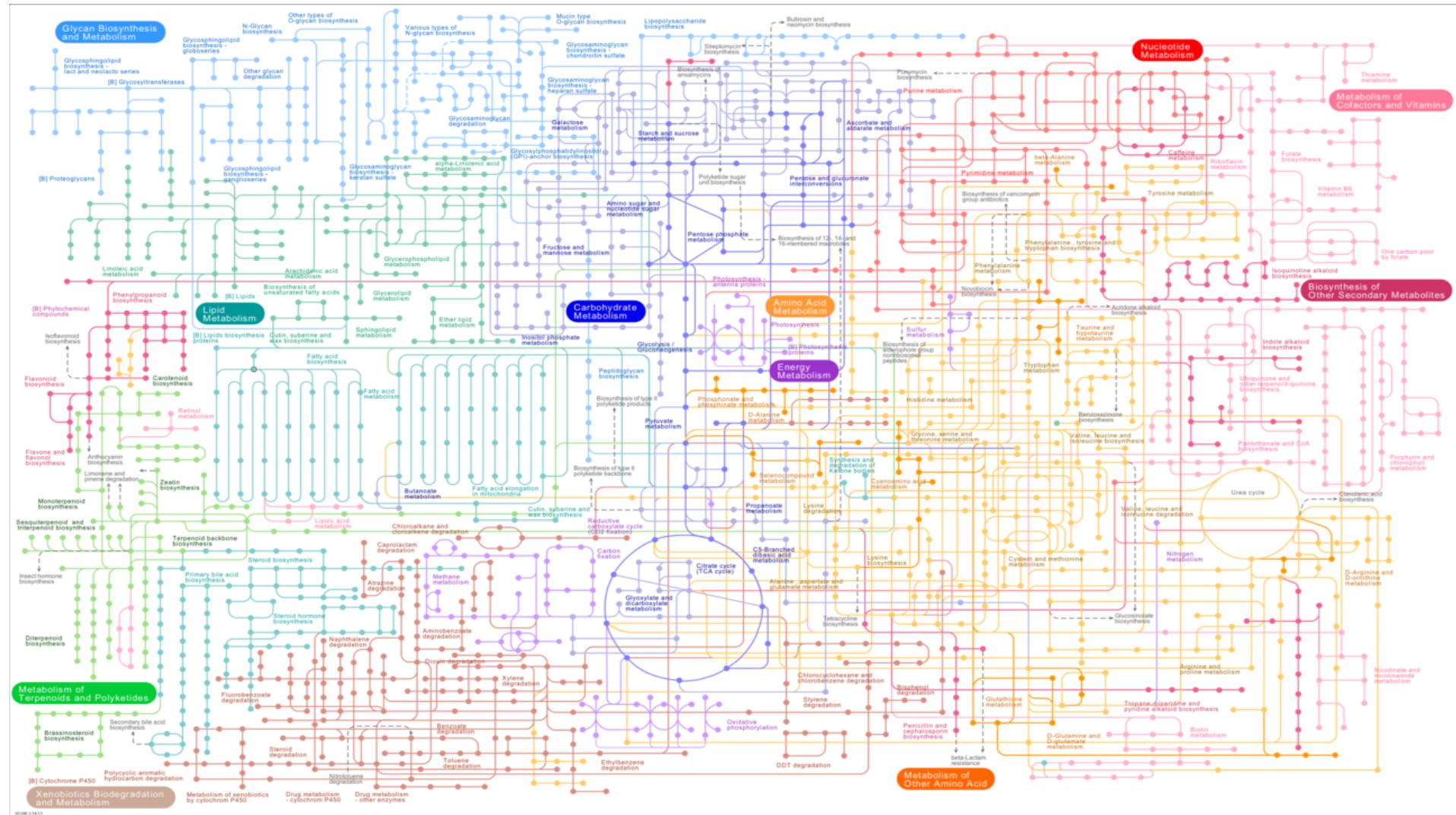
EcoCyc metabolic chart



<http://biocyc.org/ECOLI/new-image?type=OVERVIEW>

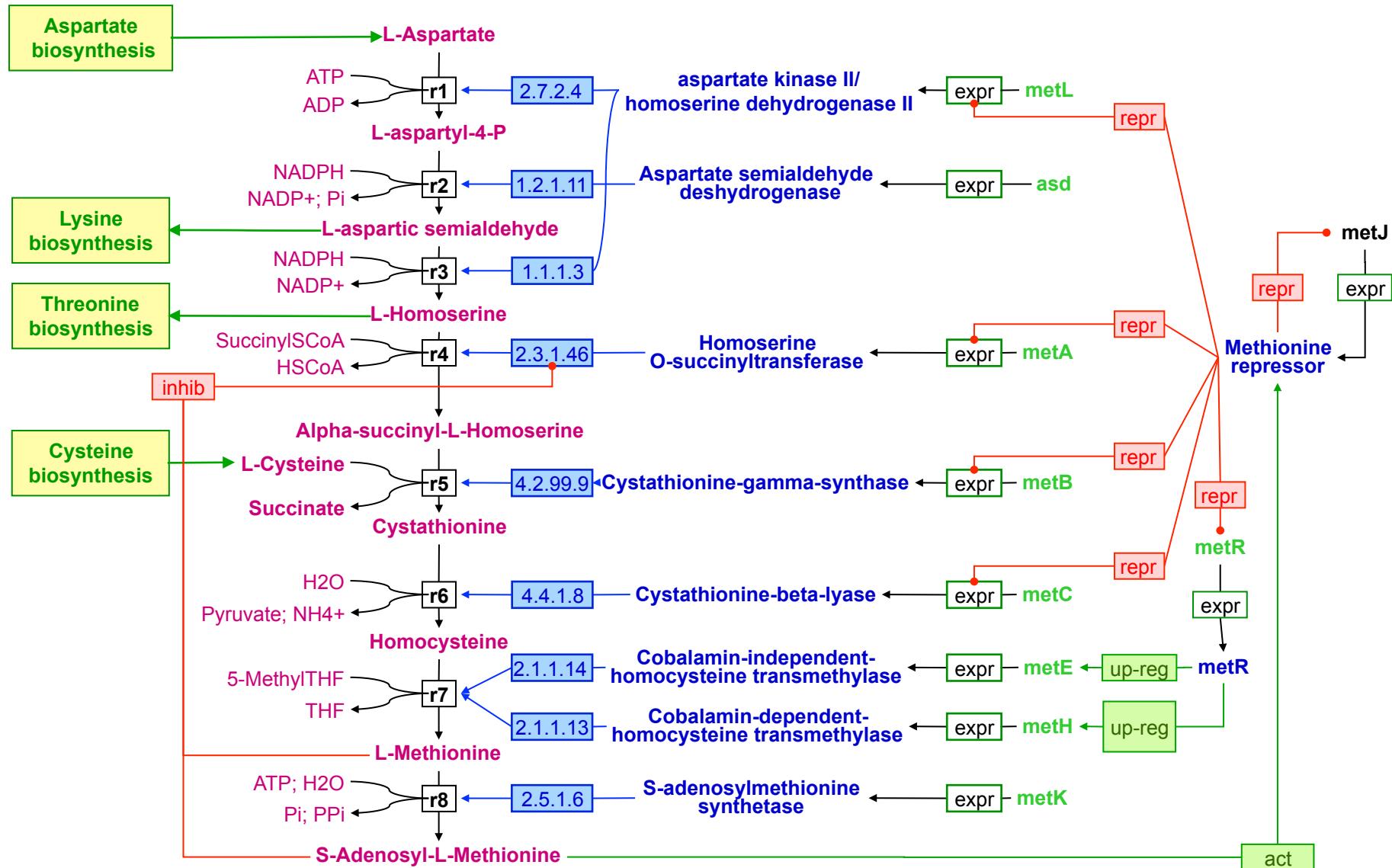
KEGG - Kyoto Encyclopaedia of Genes and Genomes

- La “carte globale” donne une vue d’ensemble de la complexité du métabolisme. Chaque point représente une molécule, chaque ligne une réaction métabolique.

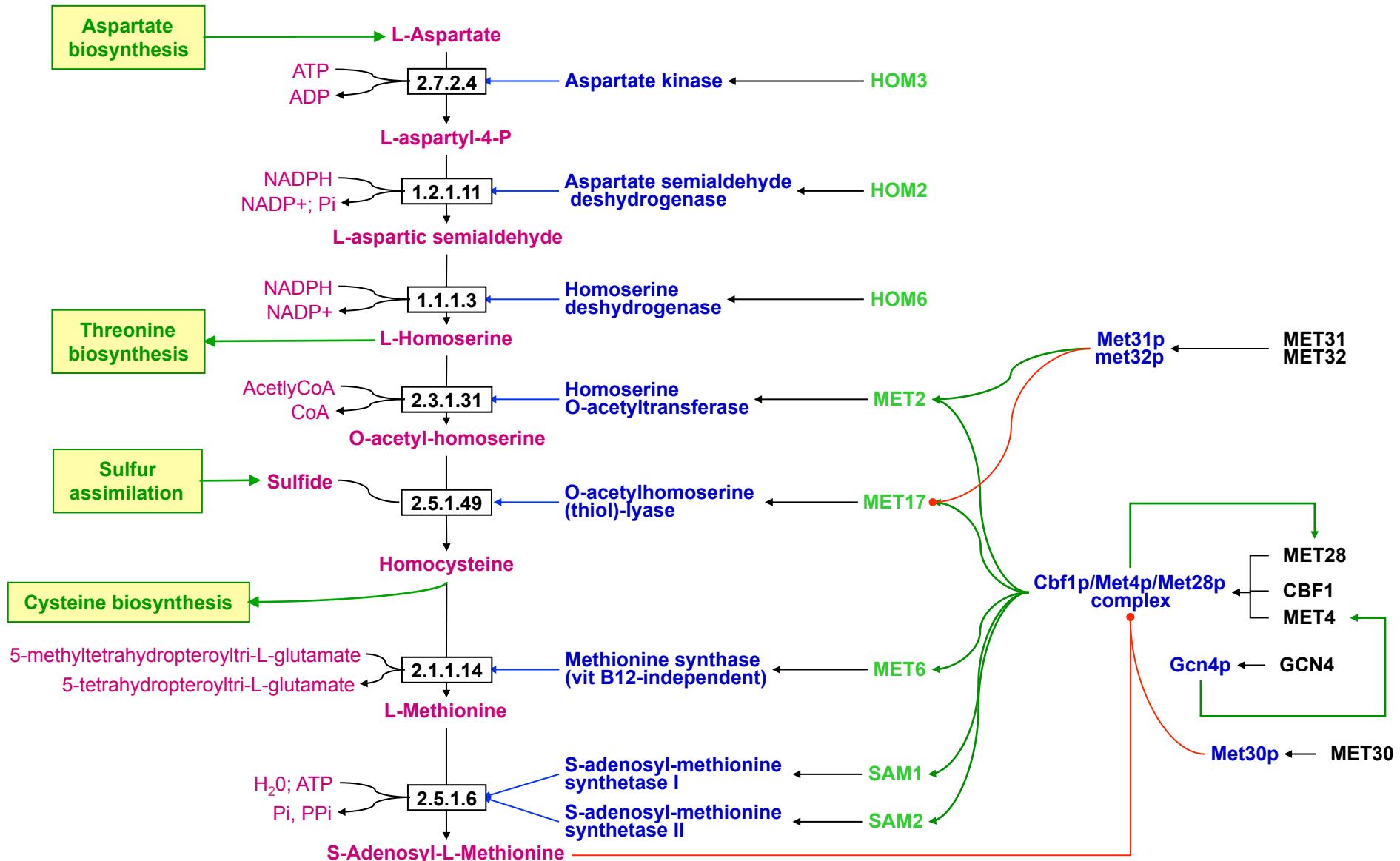


- KEGG global pathway map: http://www.genome.jp/kegg-bin/show_pathway?map01100

Methionine Biosynthesis in *E.coli*

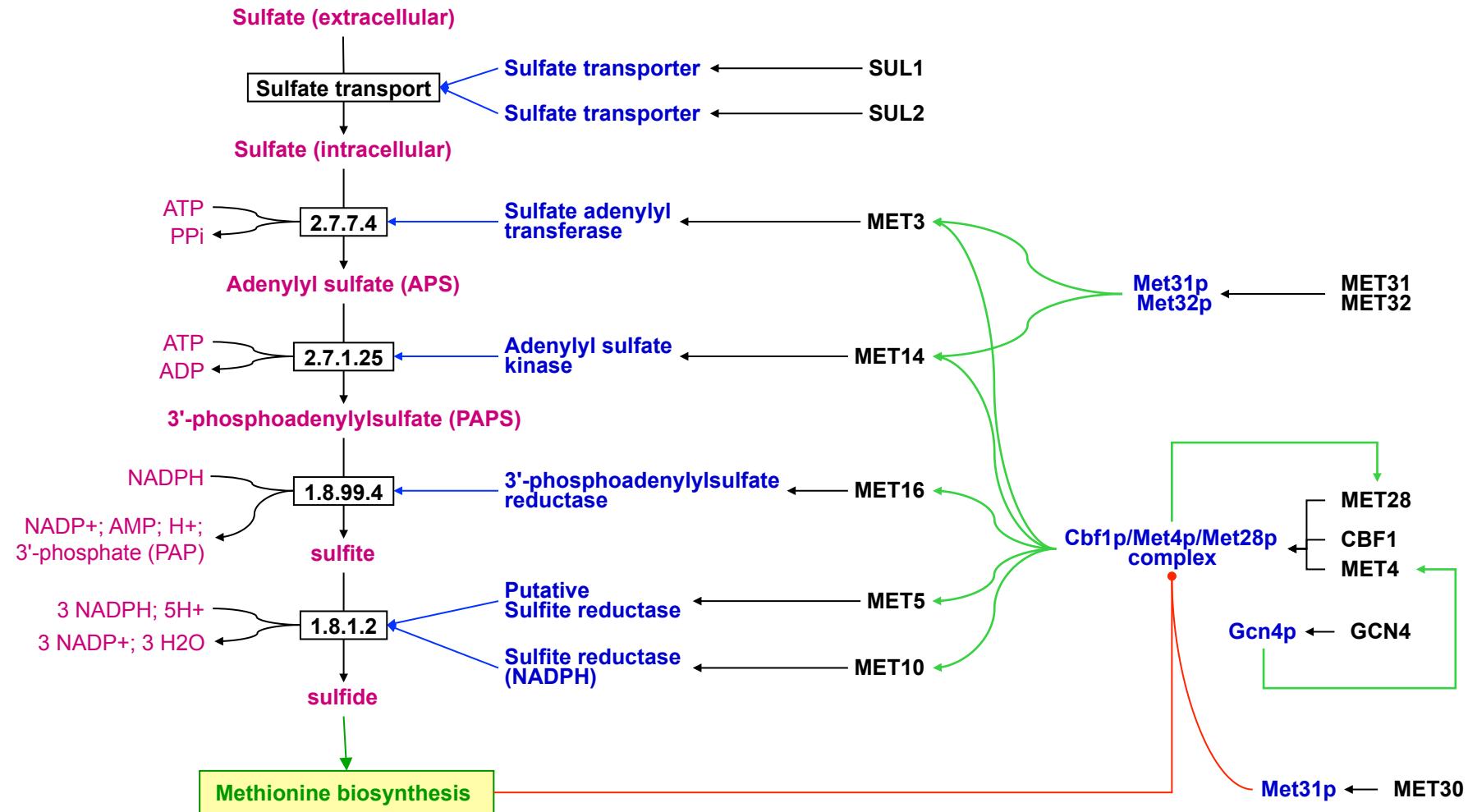


Methionine Biosynthesis in *S.cerevisiae*



■ Data provided by Georges Cohen, 1998

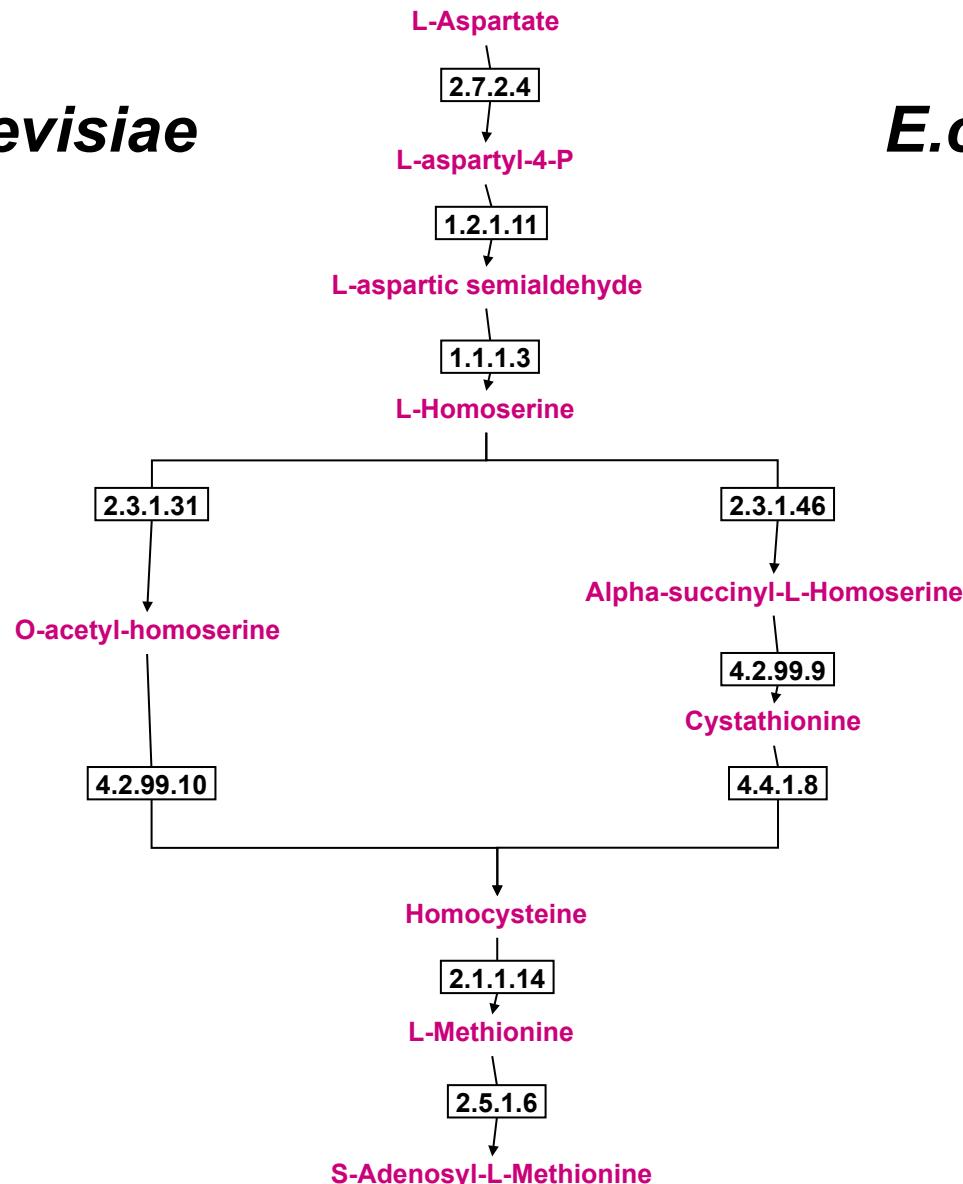
Sulfur Assimilation in yeast

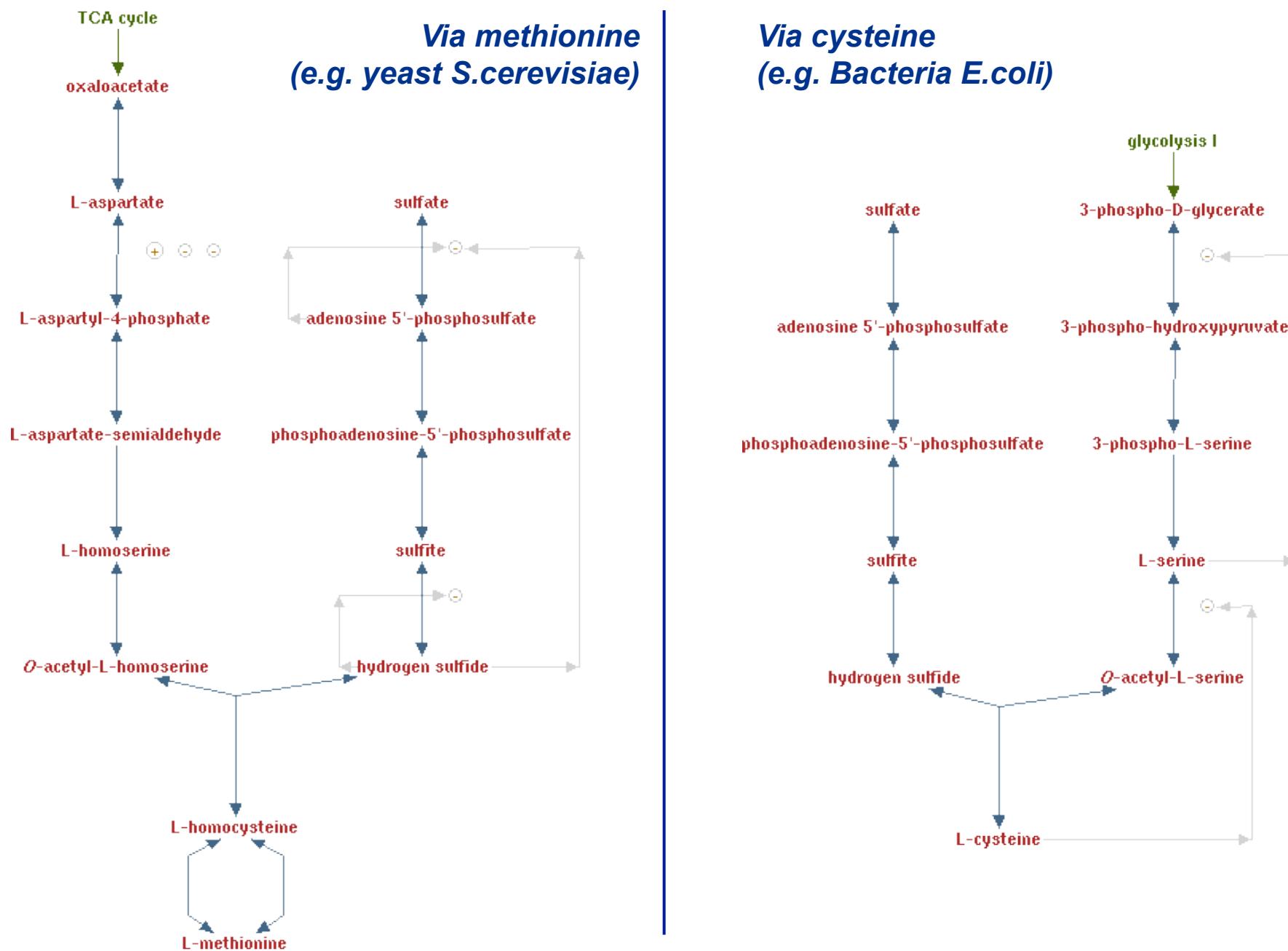


Alternative methionine pathways

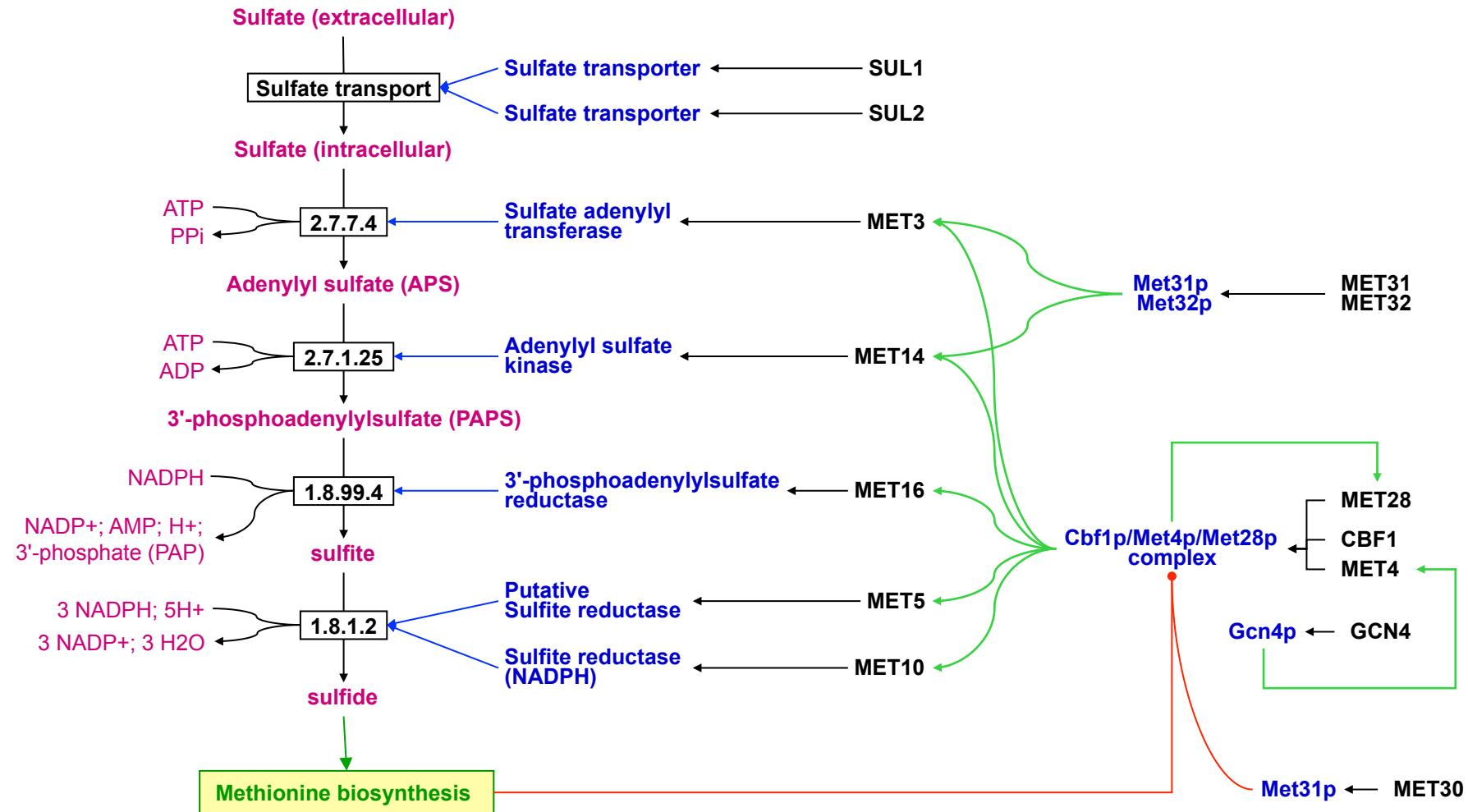
- In yeast, Sulfur is incorporated in amino acids in the methionine biosynthetic pathway, and then transferred from methionine to cysteine.
- In E.coli, sulfur is incorporated in amino acids through the cysteine biosynthetic pathway, and then transferred to methionine.

S.cerevisiae

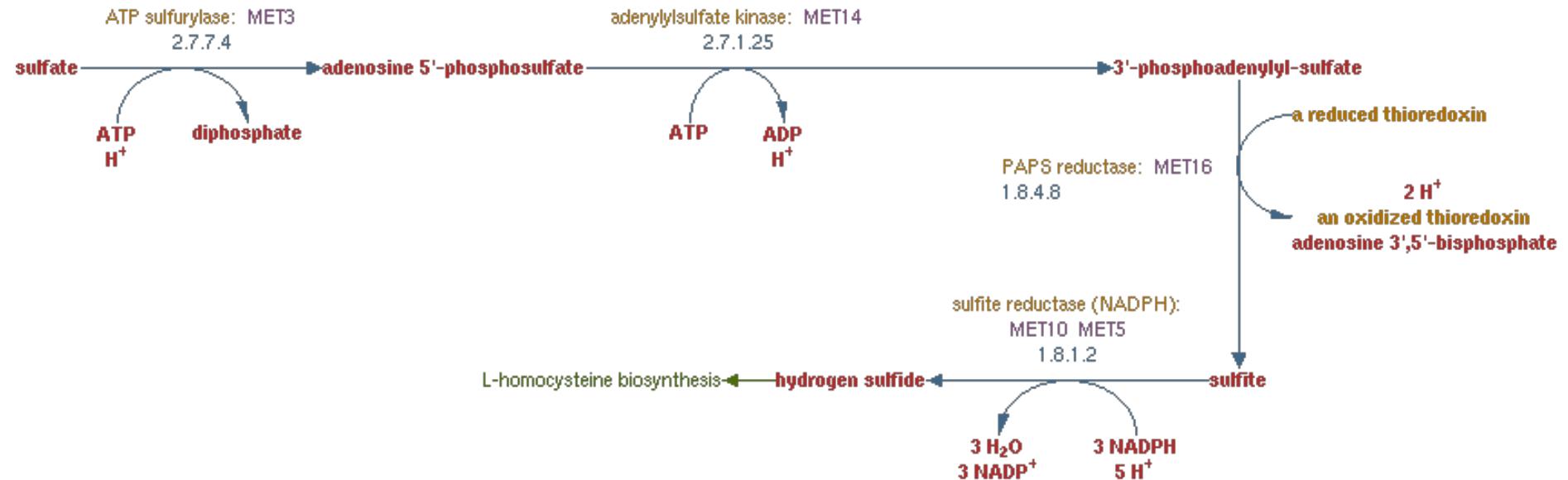




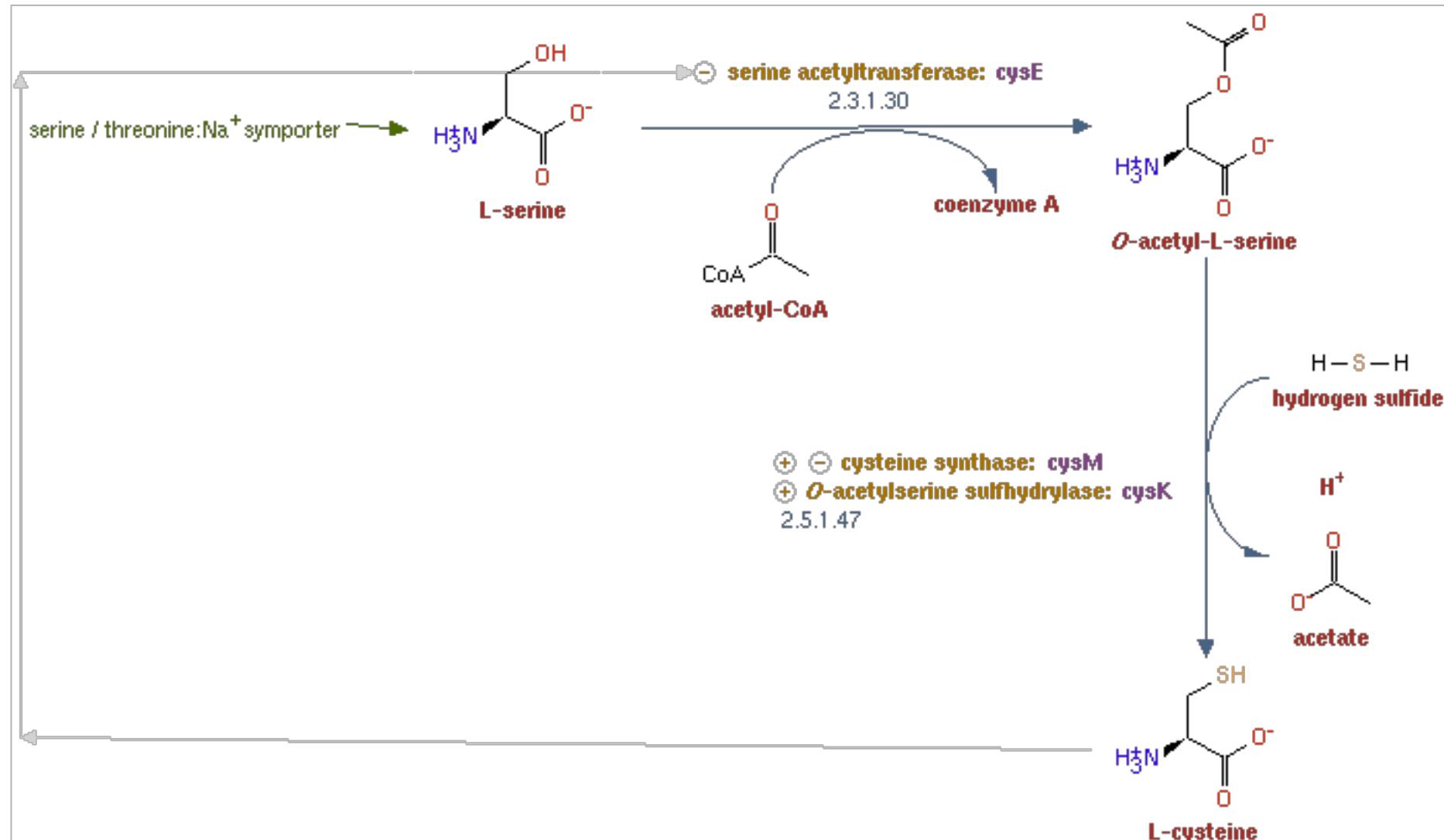
Sulfate reduction in yeast



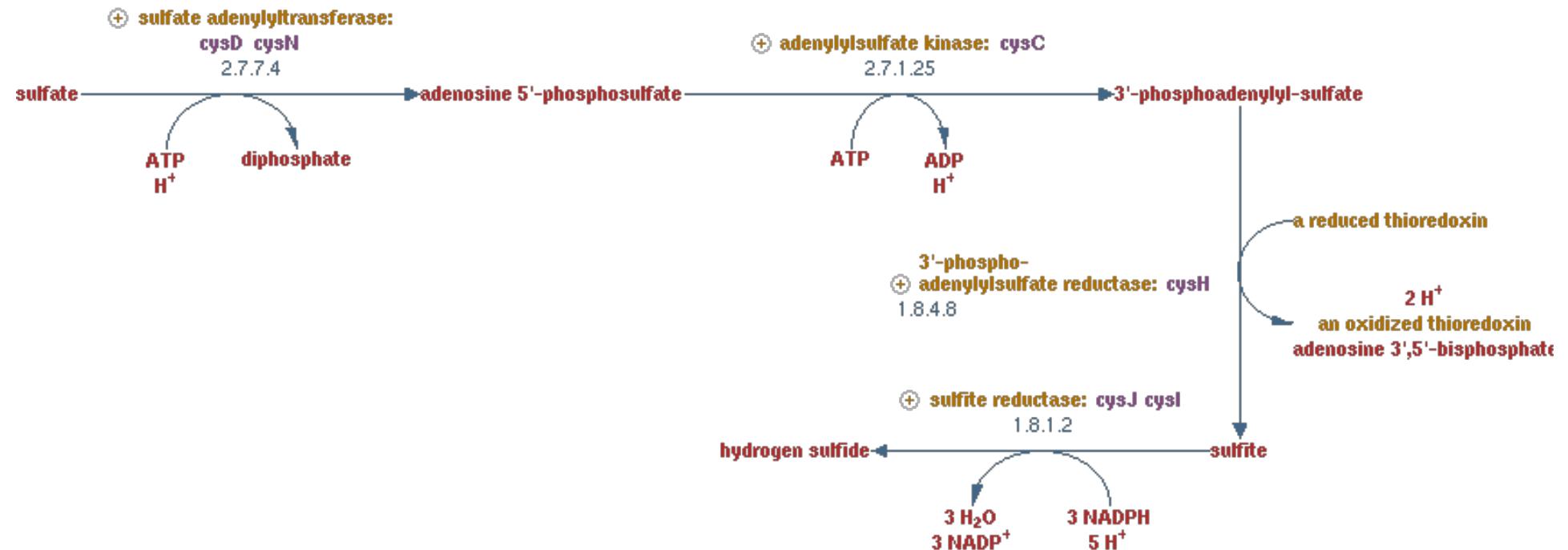
MetaCyc Saccharomyces cerevisiae Pathway: sulfate reduction I (assimilatory)



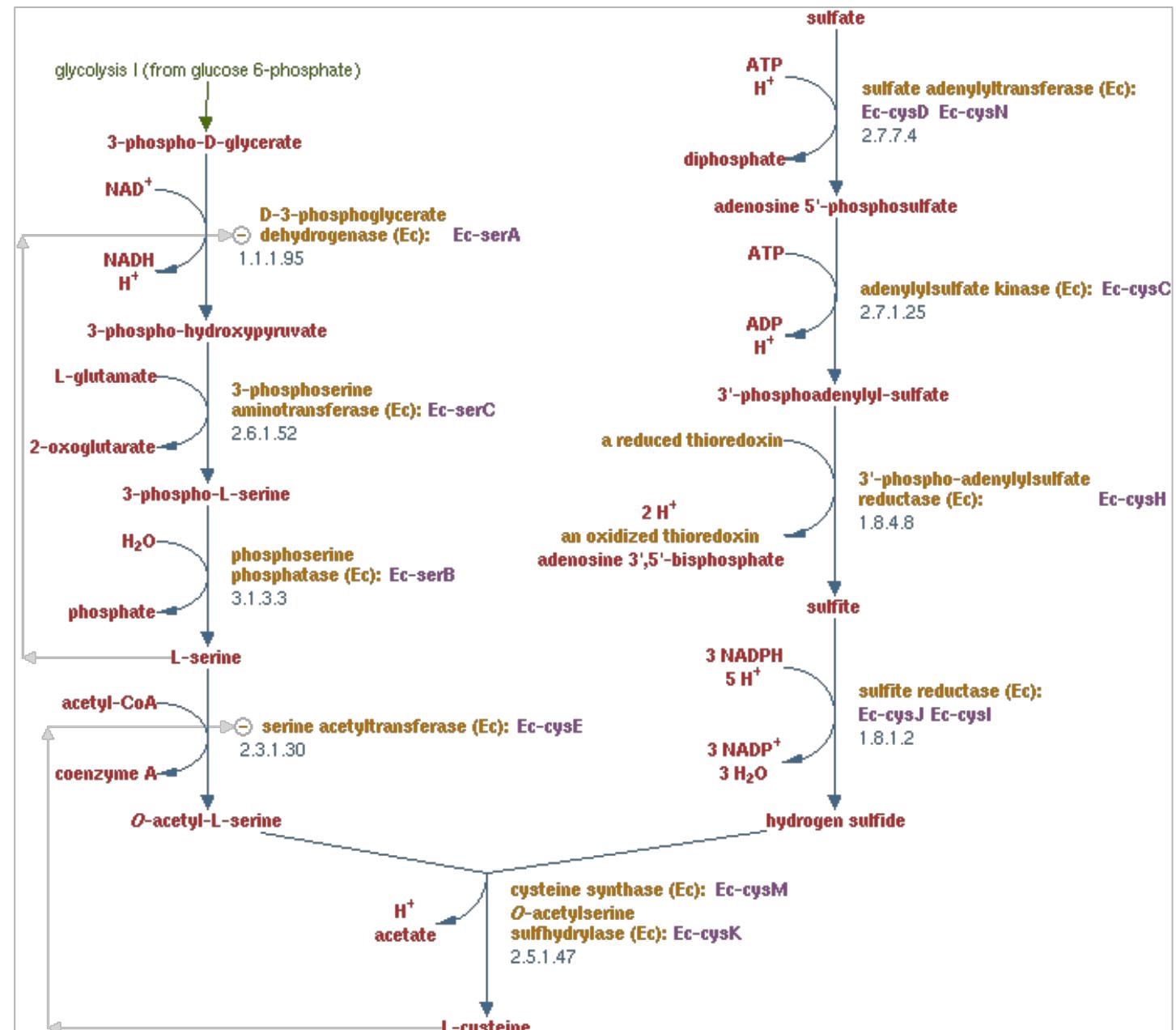
EcoCyc – Cysteine biosynthesis I



BioCyc Escherichia coli K-12 MG1655 - sulfate reduction I (assimilatory)

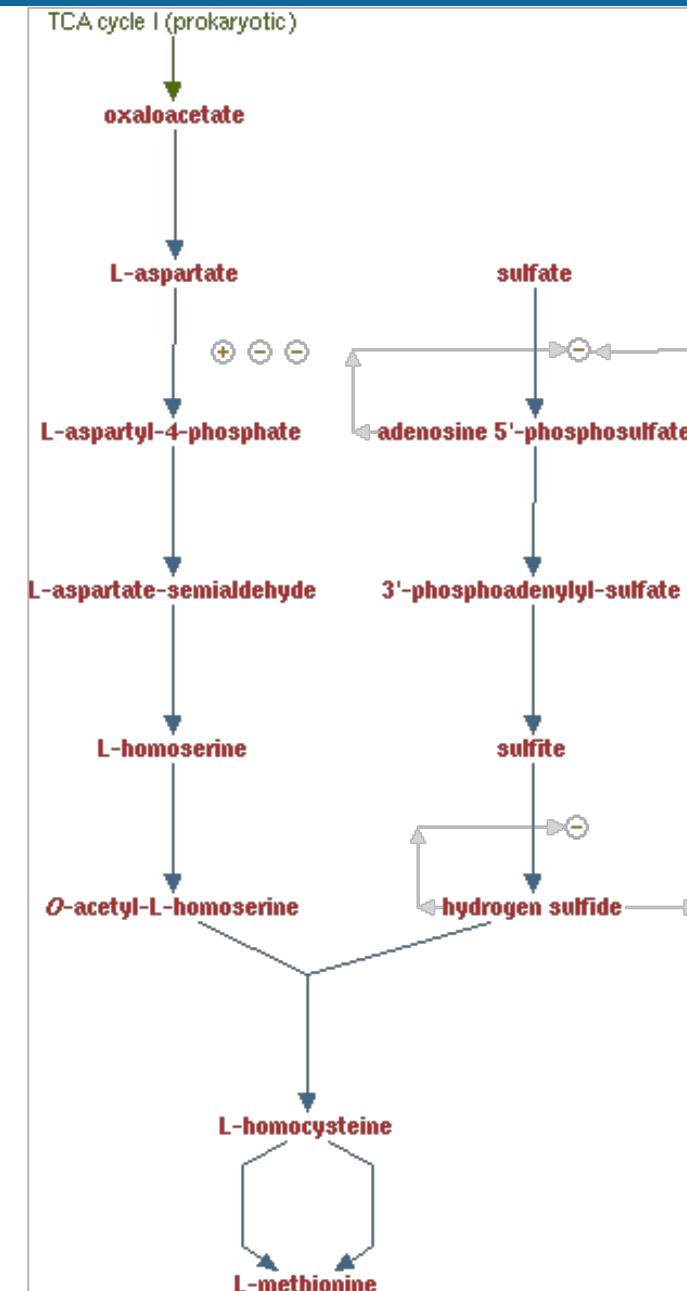


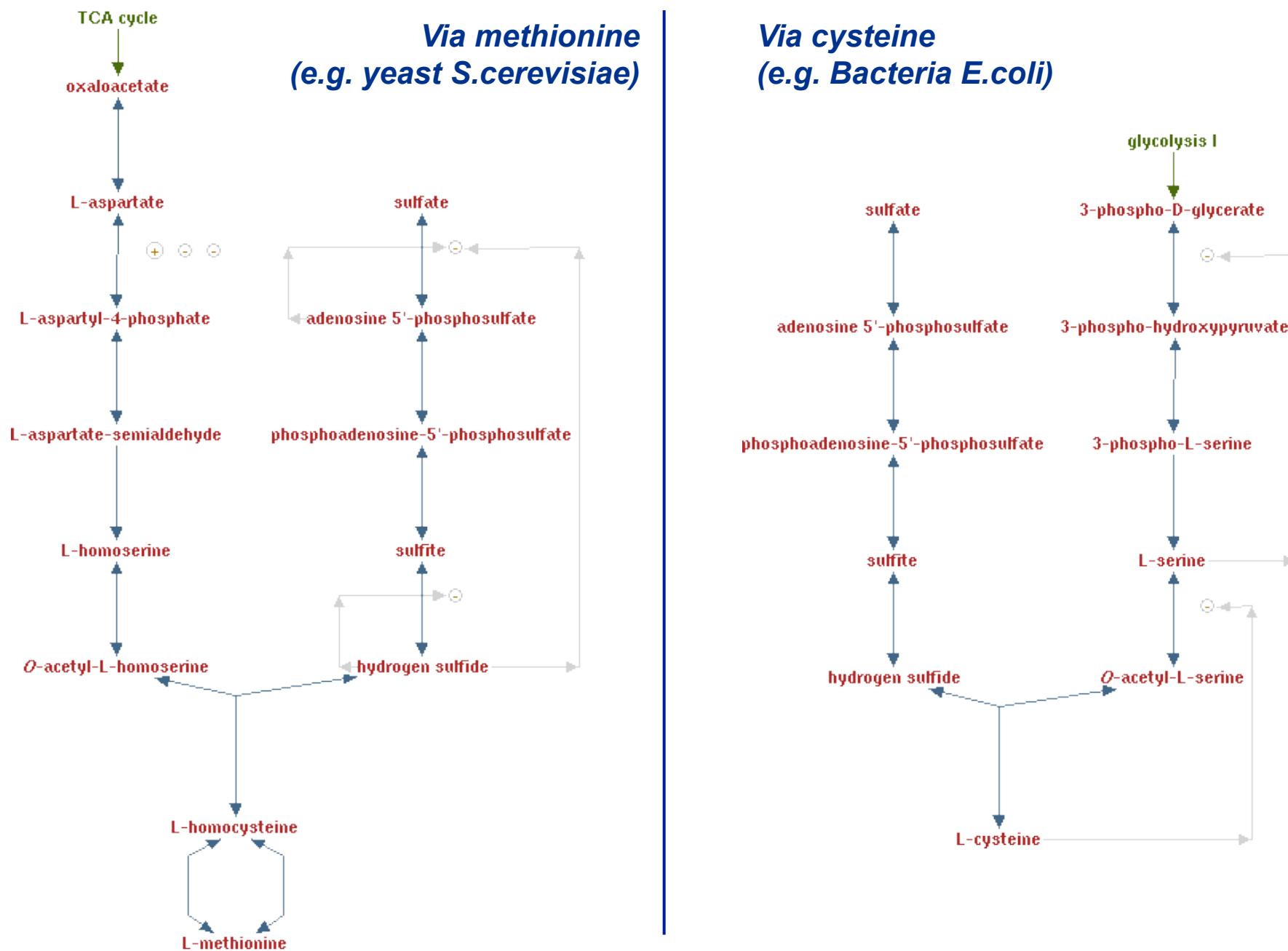
EcoCyc - Superpathway of sulfate assimilation and cysteine biosynthesis



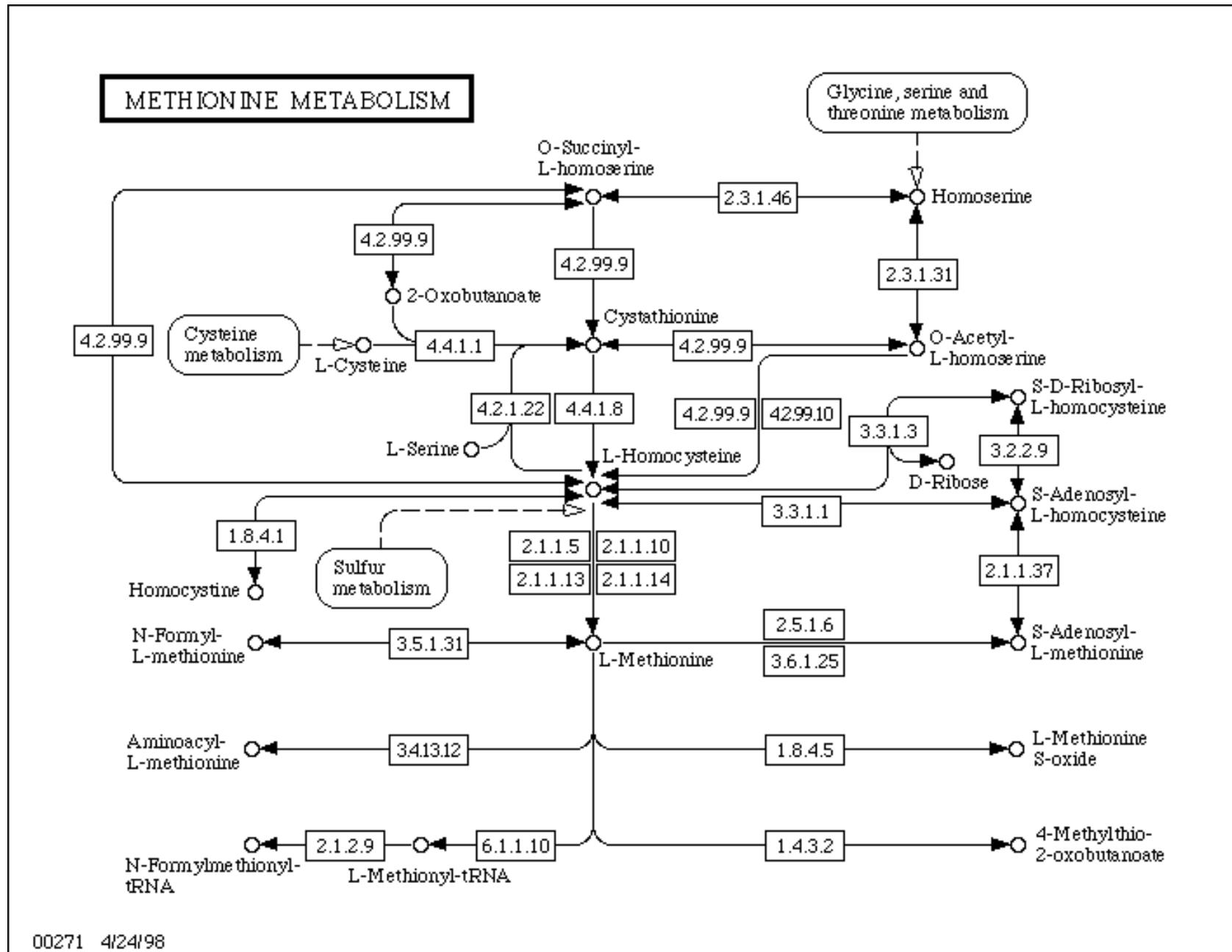
MetaCyc - Superpathway of methionine biosynthesis by sulfhydrylation

- Chez les levures (notamment), l'incorporation du soufre fait partie de la voie de biosynthèse de la méthionine. Le soufre est ensuite transféré de la méthionine à la cystéine.



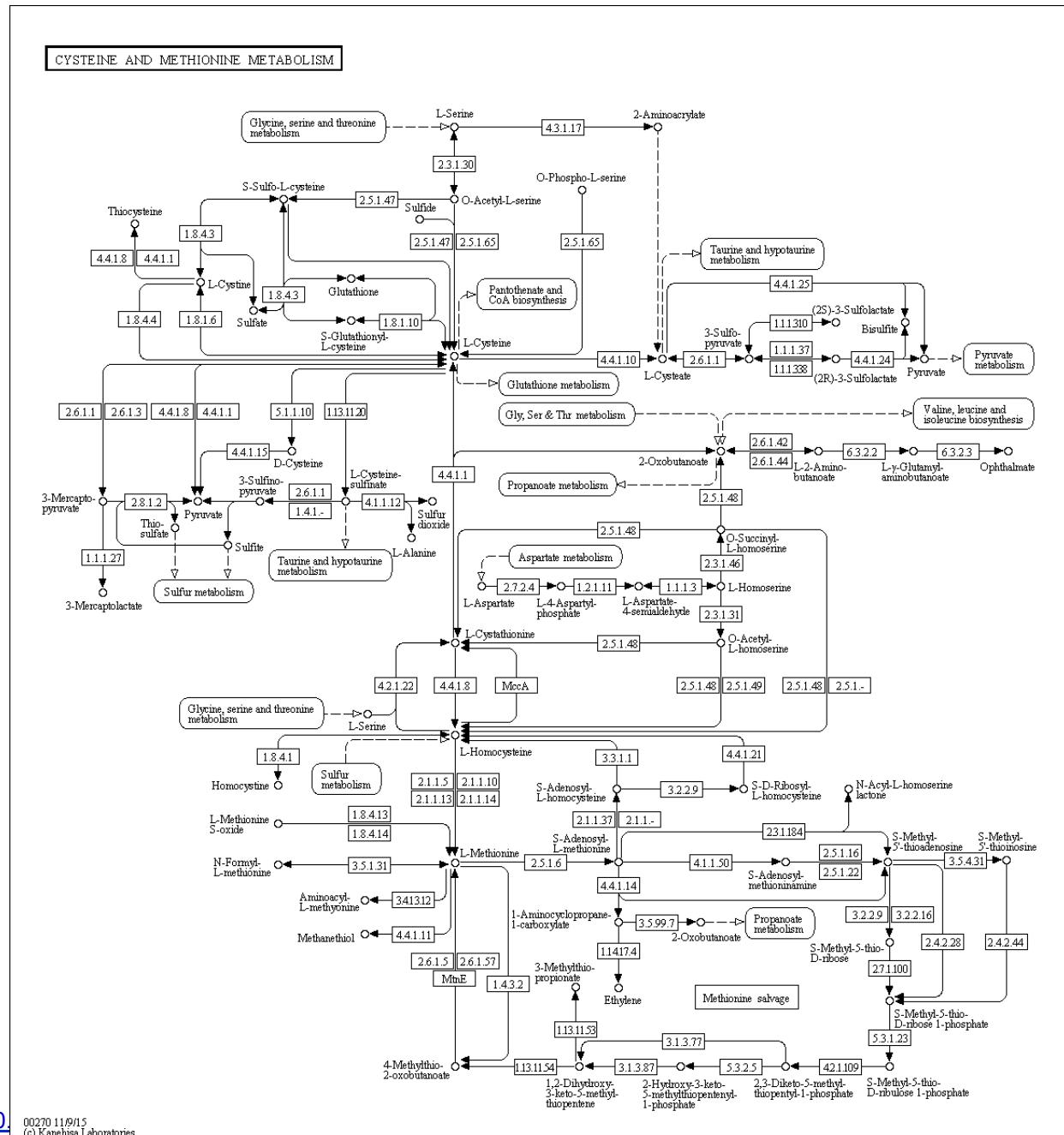


KEGG “reference” pathway - Methionine metabolism (1998)



KEGG “reference” map - Cysteine and methionine metabolism (Jan 2016)

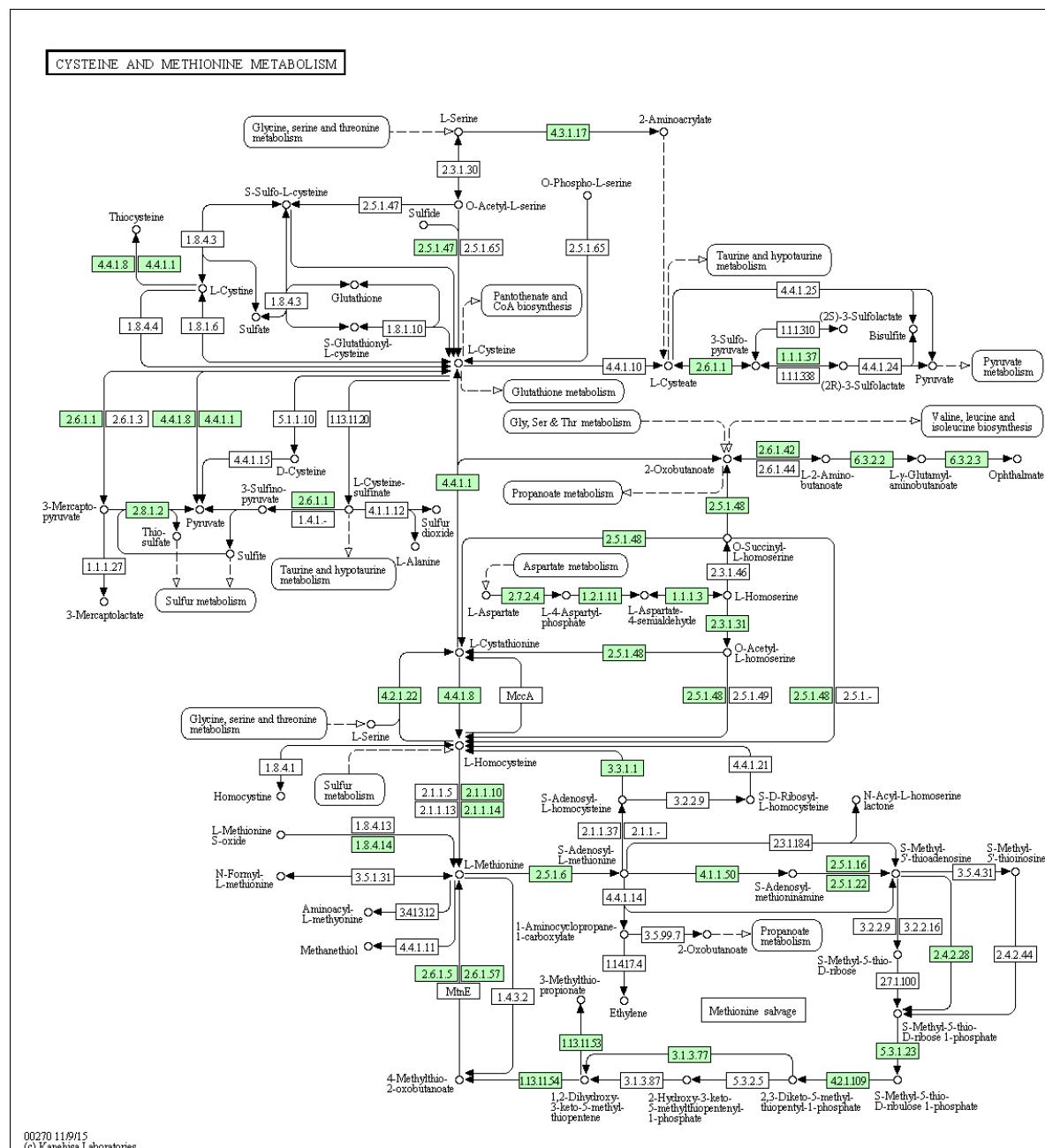
- In principle, merging methionine and cysteine should highlight the relationship between the two sulfur-containing amino acids.
- Questions:
 - Where is L-Cysteine ?
 - Where is L-Methionine ?



■ <http://www.genome.jp/kegg/pathway/map/map00270>

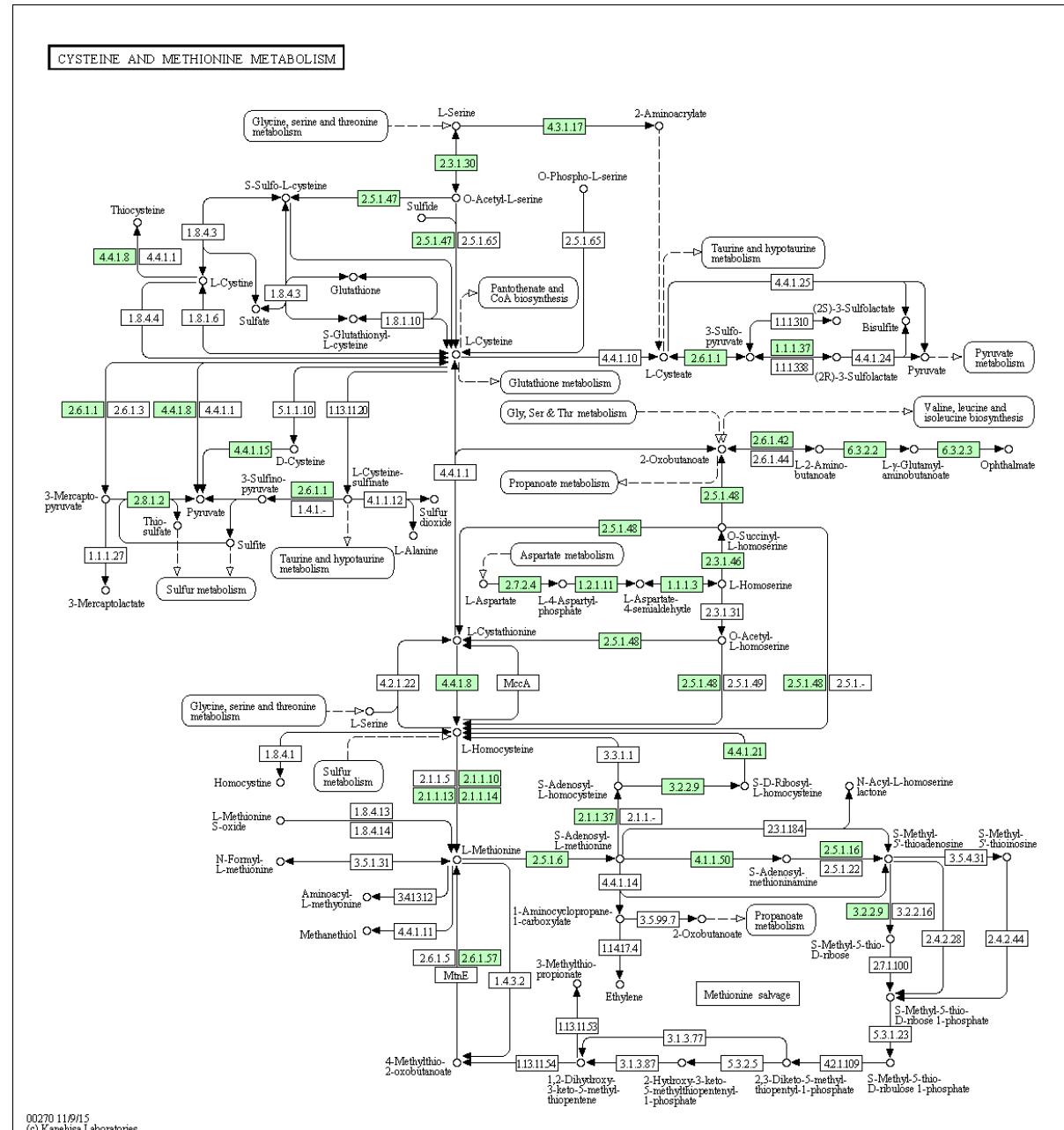
KEGG - Cysteine and methionine metabolism (2016) – *S.cerevisiae*

- KEGG cysteine and methionine pathway.
- *Saccharomyces cerevisiae*.
- Question
 - How is sulfur incorporated into amino acids in this Fungus ?



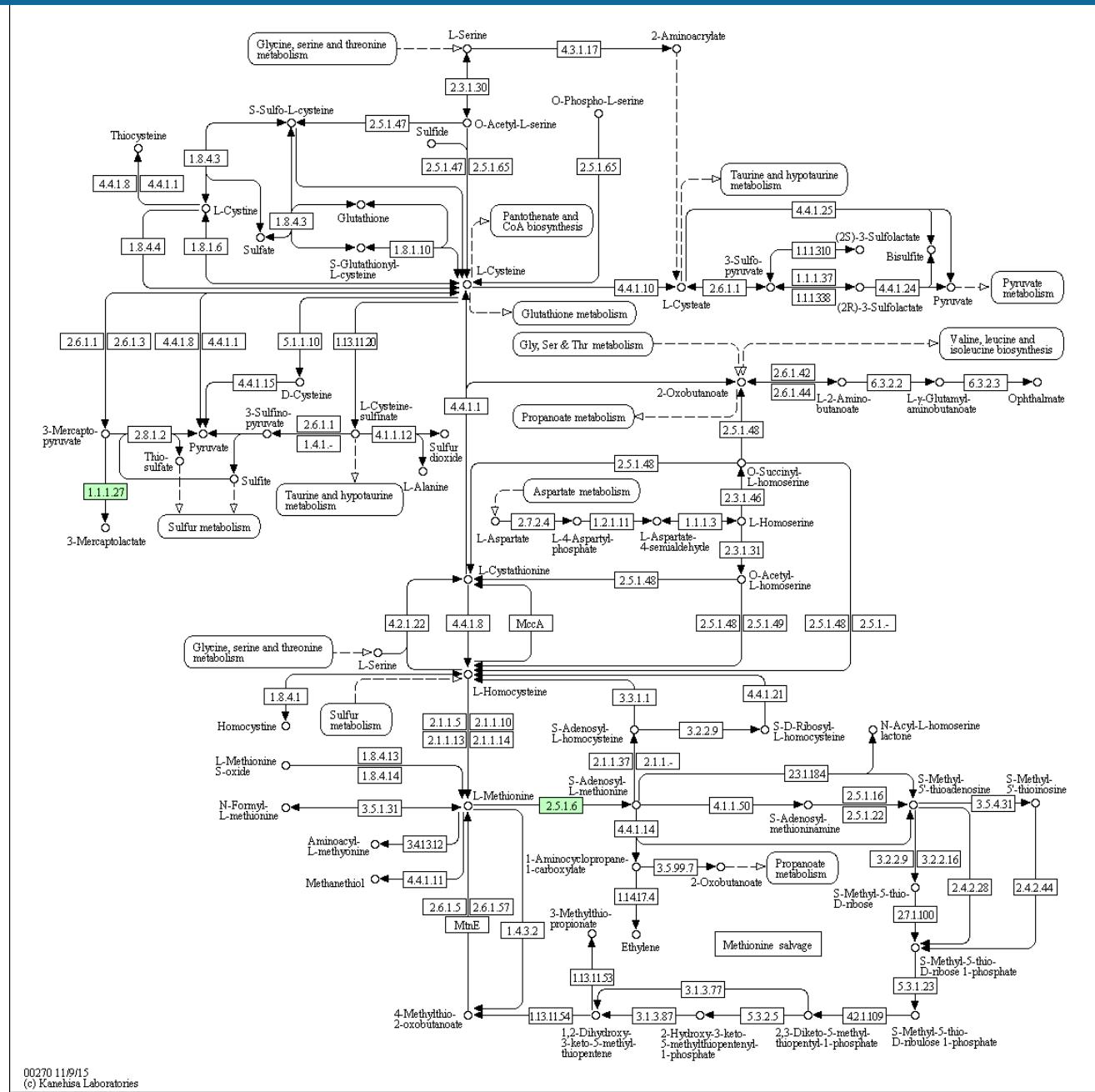
KEGG - Cysteine and methionine metabolism (2016) – E.coli

- KEGG cysteine and methionine pathway.
- *Escherichia coli* K12.
- Question
 - How is sulfur incorporated into amino acids in this in this Bacteria?



KEGG - Cysteine and methionine metabolism (2016) – *M.genitalium*

- *Mycoplasma genitalium*
 - Very small genome (500 genes).
 - Intra-cellular parasite.
 - Parasitism allowed to loose many pathways.
 - Relies on host for the corresponding compounds.
- For genome annotation, pathway impoverishment is indicative of the metabolic conditions.



00270 11/9/15
 (c) Kanehisa Laboratories

Résumé

- Les voies métaboliques connues ont été caractérisées à partir d'une poignée d'organismes modèles.
- Selon les organismes, la même molécule peut être construite par des voies alternatives.
- Les voies métaboliques sont régulées à plusieurs niveaux: transcription, activité enzymatique, transport.
- Les bases de données offrent des perspectives complémentaires sur le métabolisme.
 - EcoCyc: annotations détaillées pour un organisme modèle (*Escherichia coli* K12).
 - BioCyc: annotations détaillées pour quelques organismes de référence.
 - MetaCyc: **modèles métaboliques** élaborés à partir de quelques organismes modèles (projection).
 - KEGG: **cartes de référence** rassemblant diverses voies alternatives pour un métabolisme donné.
- Pour l'énorme majorité des organismes actuellement séquencés, on ne sait quasiment rien du métabolisme. Les annotations métaboliques reposent majoritairement sur la projection des enzymes identifiées dans le génome sur les cartes de référence.

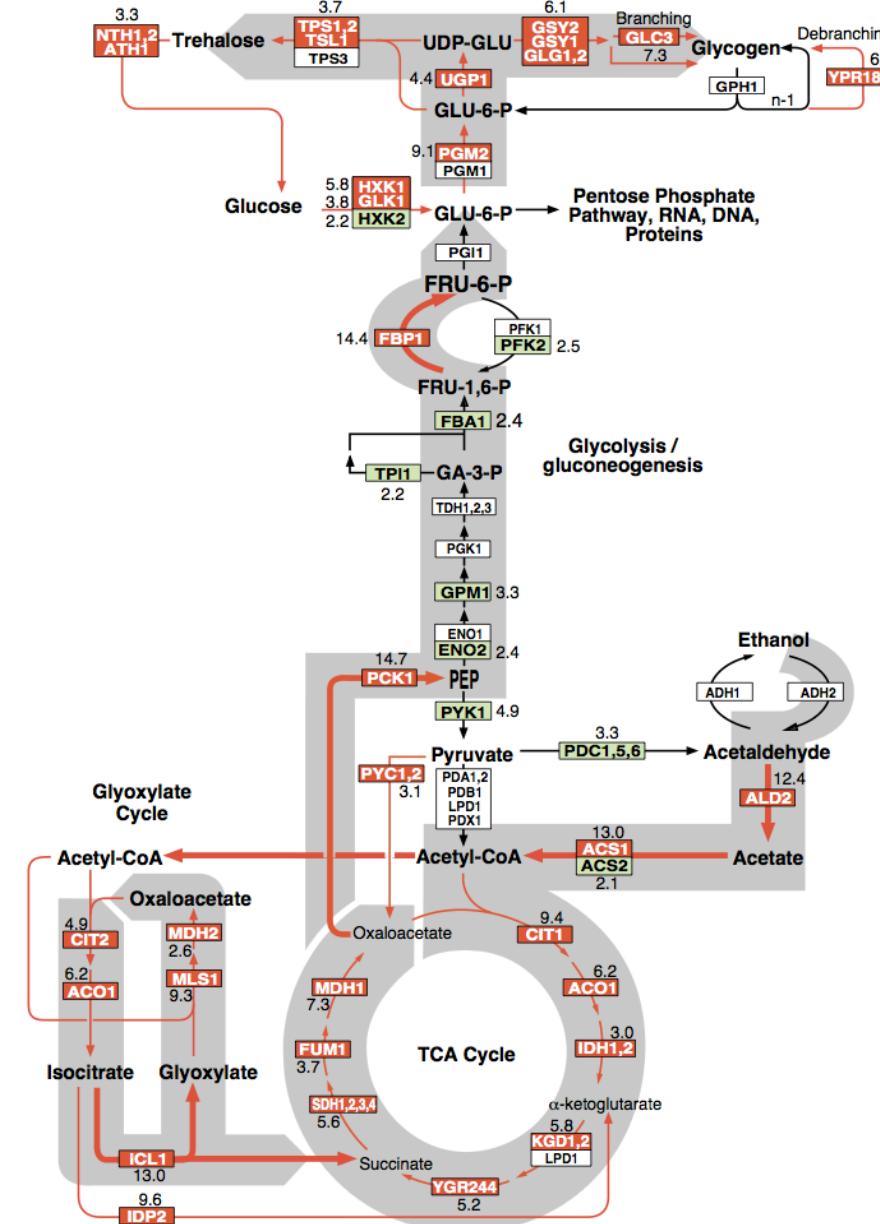
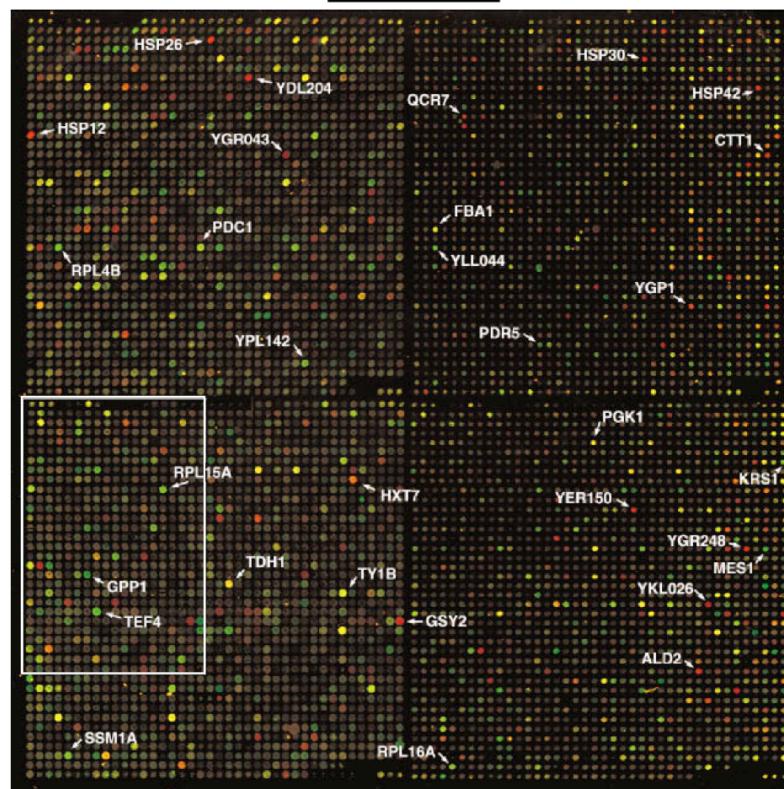
Transcriptome et métabolisme

The diauxic shift revisited by DeRisi, Iyer and Brown (1997)

Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale

Joseph L. DeRisi, Vishwanath R. Iyer, Patrick O. Brown*

DNA microarrays containing virtually every gene of *Saccharomyces cerevisiae* were used to carry out a comprehensive investigation of the temporal program of gene expression accompanying the metabolic shift from fermentation to respiration. The expression profiles observed for genes with known metabolic functions pointed to features of the metabolic reprogramming that occur during the diauxic shift, and the expression patterns of many previously uncharacterized genes provided clues to their possible functions. The same DNA microarrays were also used to identify genes whose expression was affected by deletion of the transcriptional co-repressor *TUP1* or overexpression of the transcriptional activator *YAP1*. These results demonstrate the feasibility and utility of this approach to genome-wide exploration of gene expression patterns.



- DeRisi, J. L., Iyer, V. R. & Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680-6.

Distinct temporal patterns of induction or repression

- The diauxic shift experiment reveals ***clusters of co-expressed genes***, which are induced or repressed at specific time points of the experiment.
- These temporal patterns are consistent with the known function of these genes.
- The experiment also uncovers genes of unknown function that are co-expressed with genes of known function.
- Co-expression might be a hint about possible involvement in a same biological process.

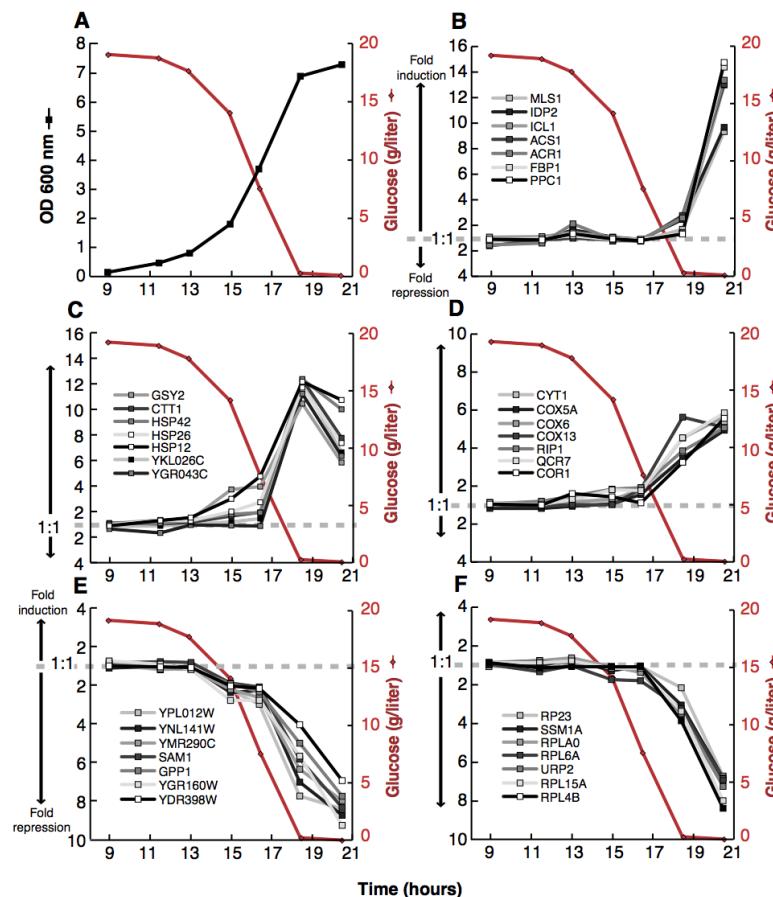


Fig. 5. Distinct temporal patterns of induction or repression help to group genes that share regulatory properties. **(A)** Temporal profile of the cell density, as measured by OD at 600 nm and glucose concentration in the media. **(B)** Seven genes exhibited a strong induction (greater than ninefold) only at the last timepoint (20.5 hours). With the exception of *IDP2*, each of these genes has a CSRE UAS. There were no additional genes observed to match this profile. **(C)** Seven members of a class of genes marked by early induction with a peak in mRNA levels at 18.5 hours. Each of these genes contain STRE motif repeats in their upstream promoter regions. **(D)** Cytochrome c oxidase and ubiquinol cytochrome c reductase genes. Marked by an induction coincident with the diauxic shift, each of these genes contains a consensus binding motif for the HAP2,3,4 protein complex. At least 17 genes shared a similar expression profile. **(E)** *SAM1*, *GPP1*, and several genes of unknown function are repressed before the diauxic shift, and continue to be repressed upon entry into stationary phase. **(F)** Ribosomal protein genes comprise a large class of genes that are repressed upon depletion of glucose. Each of the genes profiled here contains one or more RAP1-binding motifs upstream of its promoter. RAP1 is a transcriptional regulator of most ribosomal proteins.