JGB71E - Bioinformatique appliquée

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

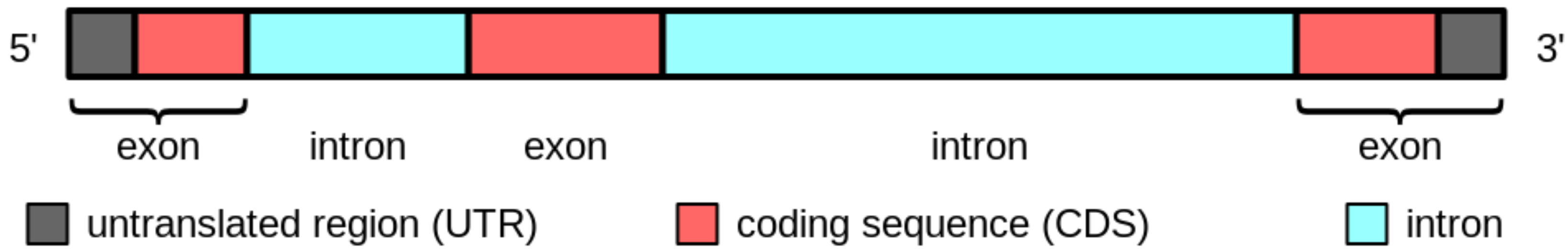
Structure d'un gène eucaryote

- 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 !!!

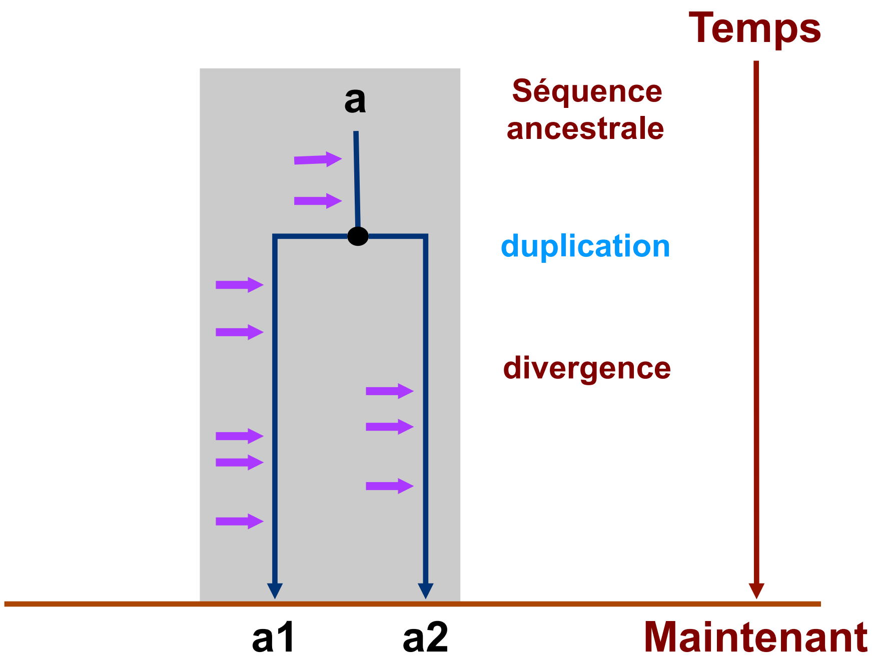
https://en.wikipedia.org/wiki/Exon



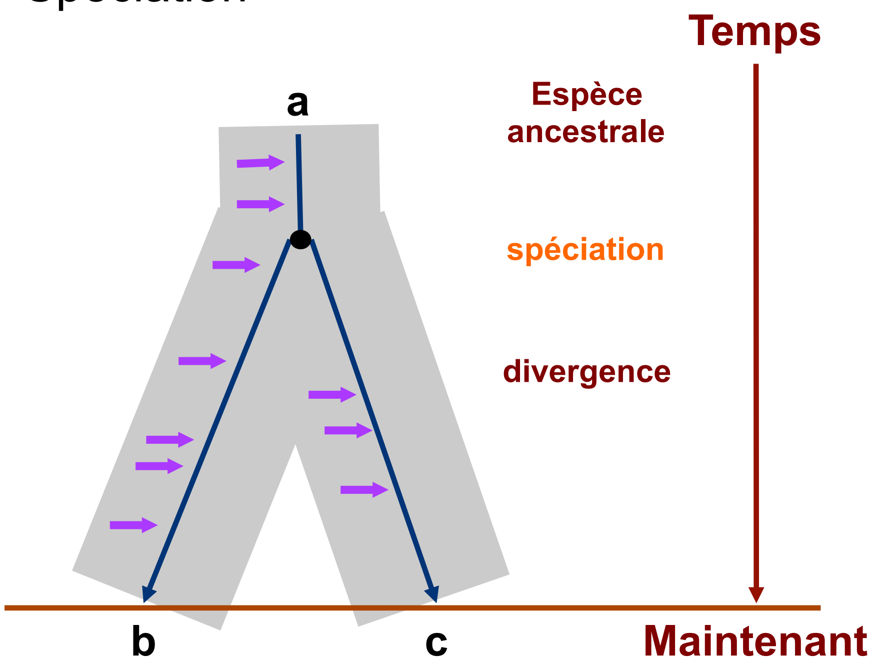
Scénarios évolutifs

- 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

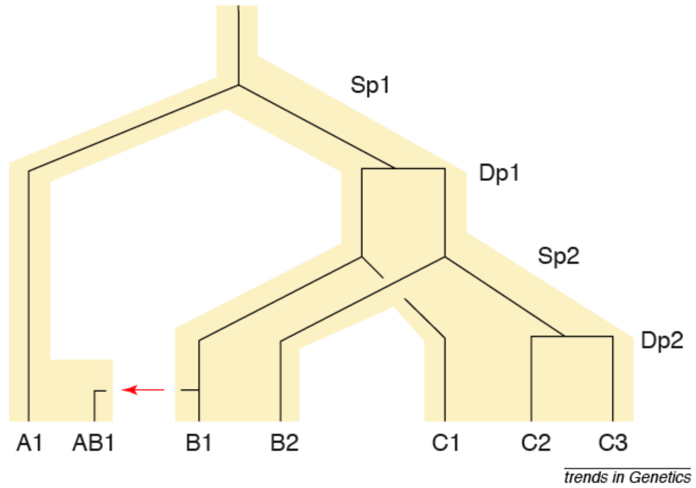


Spéciation



Représentation détaillée des événements de spéciation / duplication

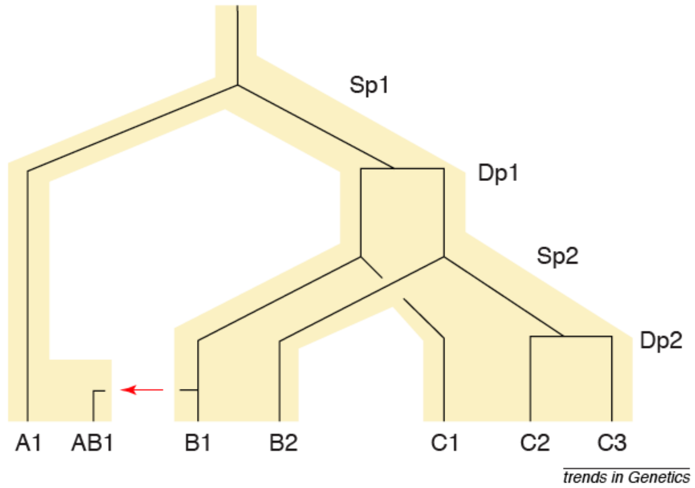
- La figure de droite combine deux niveaux de représentation
  - Les lignes noires fines représentent les relations évolutives entre molécules (arbre des molécules).
  - Les ombrages épais représentent l'arbre des espèces.
- Les ***spéciations (Sp)*** sont représentées par des branchements triangulaires 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 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=>B1 implies B1=>A1 where => should be read, for example, as 'is orthologous to.' However, the relationships are not transitive. Thus, C2=>A1=>C3 might be true, but it is not necessarily therefore true that C2=>C3, as indeed it is not in the figure if => is read as 'is orthologous to.' A different non-transitivity occurs for 'is paralogous to' with B2=>C1=>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 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.

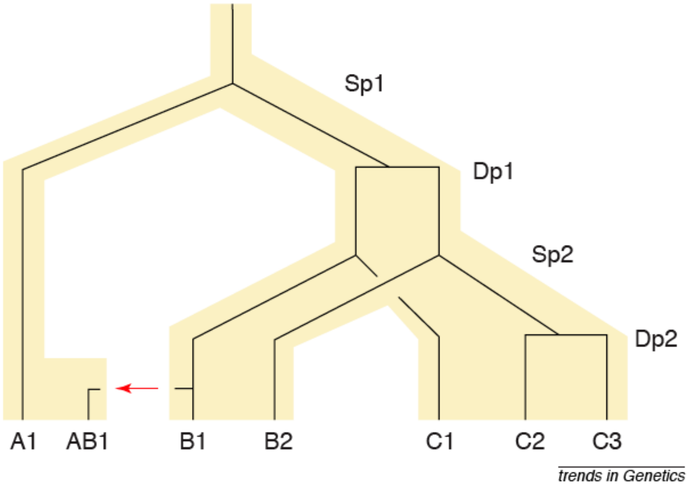


The idealized evolution of a gene (lines) is shown from a common ancestor in an ancestral population (the gray background), descending to three populations labelled A, B and C. There are two speciation events (Sp1 and Sp2), each occurring at the junctions shown as an upside down Y. There are also two 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=>B1 implies B1=>A1 where => should be read, for example, as 'is orthologous to.' However, the relationships are not transitive. Thus, C2=>A1=>C3 might be true, but it is not necessarily therefore true that C2=>C3, as indeed it is not in the figure if => is read as 'is orthologous to.' A different non-transitivity occurs for 'is paralogous to' with B2=>C1=>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



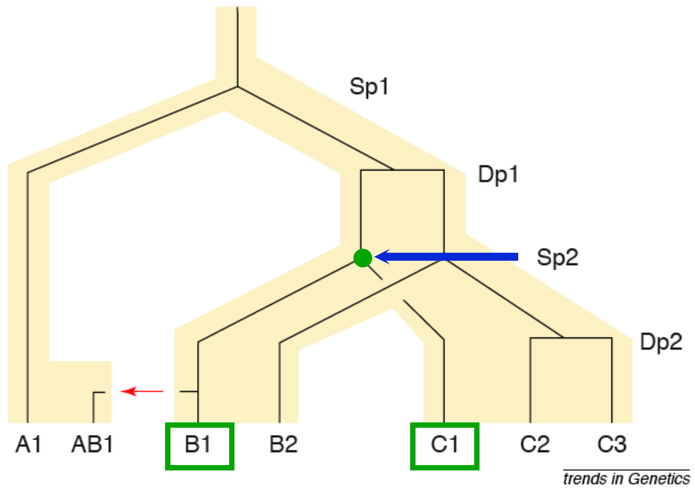
	A1	AB1	B1	B2	C1	C2	C3
A1							
AB1							
B1							
B2							
C1							
C2							
C3							

- 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



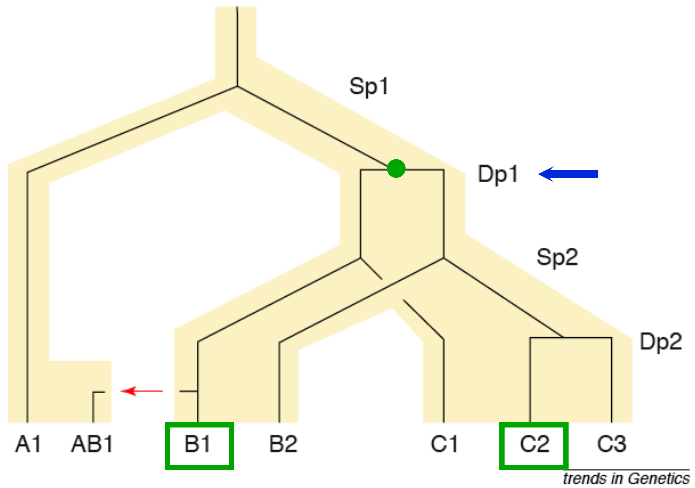
	A1	AB1	B1	B2	C1	C2	C3
A1							
AB1							
B1							
B2							
C1			O				
C2							
C3							

- 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



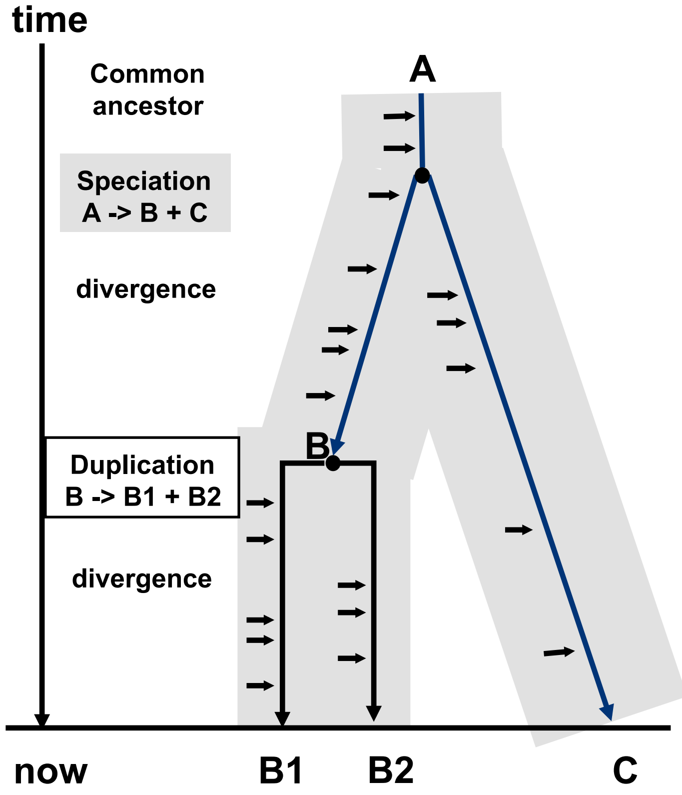
	A1	AB1	B1	B2	C1	C2	C3
A1							
AB1							
B1							
B2							
C1			O				
C2			P				
C3							

- 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

Non-transitivity of the orthology relationship

- In the figure
  - B and C are orthologs, because their last common ancestor lies just before the speciation
    A -> B + C
  - B1 and B2 are paralogs because the first event that follows their last common ancestor (B) is the duplication
    B -> B1 + B2
- Beware ! These definitions are often misunderstood, even in some textbooks. Contrarily to a strong belief, orthology can be a 1 to N relationship.
  - *B1 and C are orthologs*, because the first event after their last common ancestor (A) was the speciation A -> B + C
  - *B2 and C are orthologs* because the first event after their last common ancestor (A) was the speciation A -> B + C
- The orthology relationship is ***reciprocal*** but ***not transitive***.
  - C <-[orthologous]-> B1
  - C <-[orthologous]-> B2
  - B1 <-[paralogous]-> B2



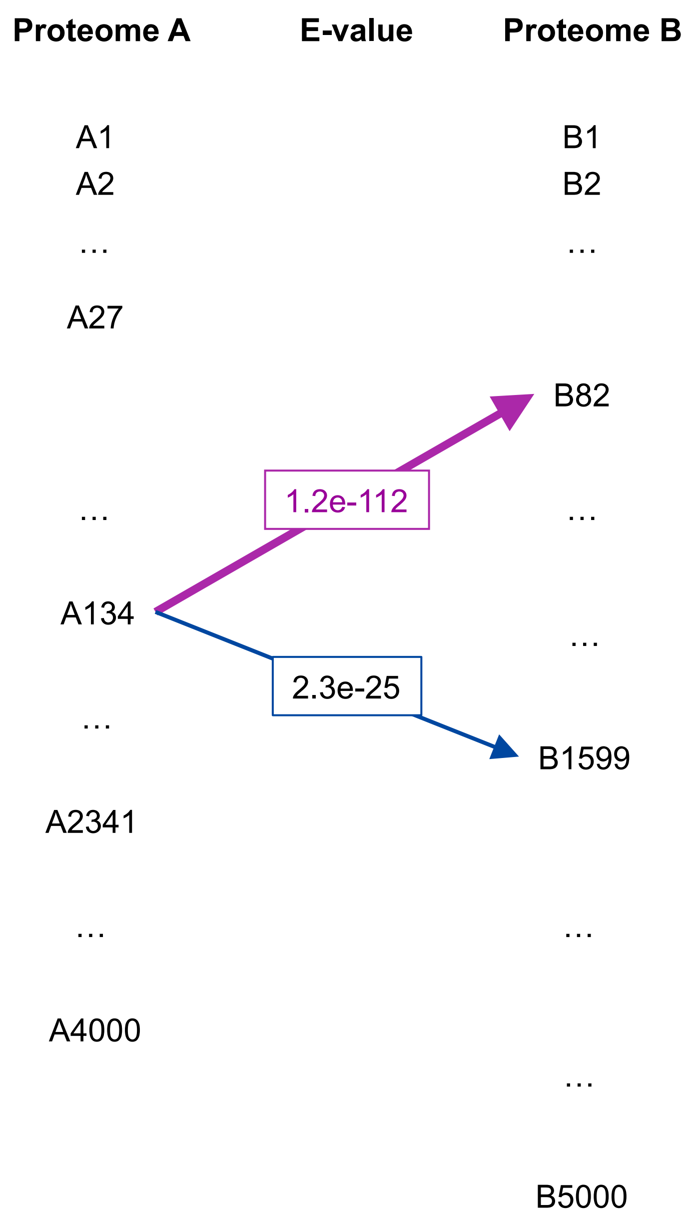
Orthologs are sequences whose **last common ancestor** occurred immediately before a **speciation** event.
Paralogs are sequences whose **last common ancestor** occurred immediately before a **duplication** event.
(Fitch, 1970; Zvelebil & Baum, 2000)

Inferring orthology / paralogy by phylogenetic inference

- To assess whether a pair of homologous genes are orthologs or paralogs, the most suitable method is to reconcile molecular and species trees.
  - In Ensembl and EnsemblGenomes, orthology/paralogy is inferred by phylogenetic tree reconciliation.
  - However, this may become complex: When the number of species increases, computing time increases quadratically or worse.
  - In 2014, EnsemblGenomes contains >10,000 Bacteria, but the orthology/paralogy is established for 123 of them only.

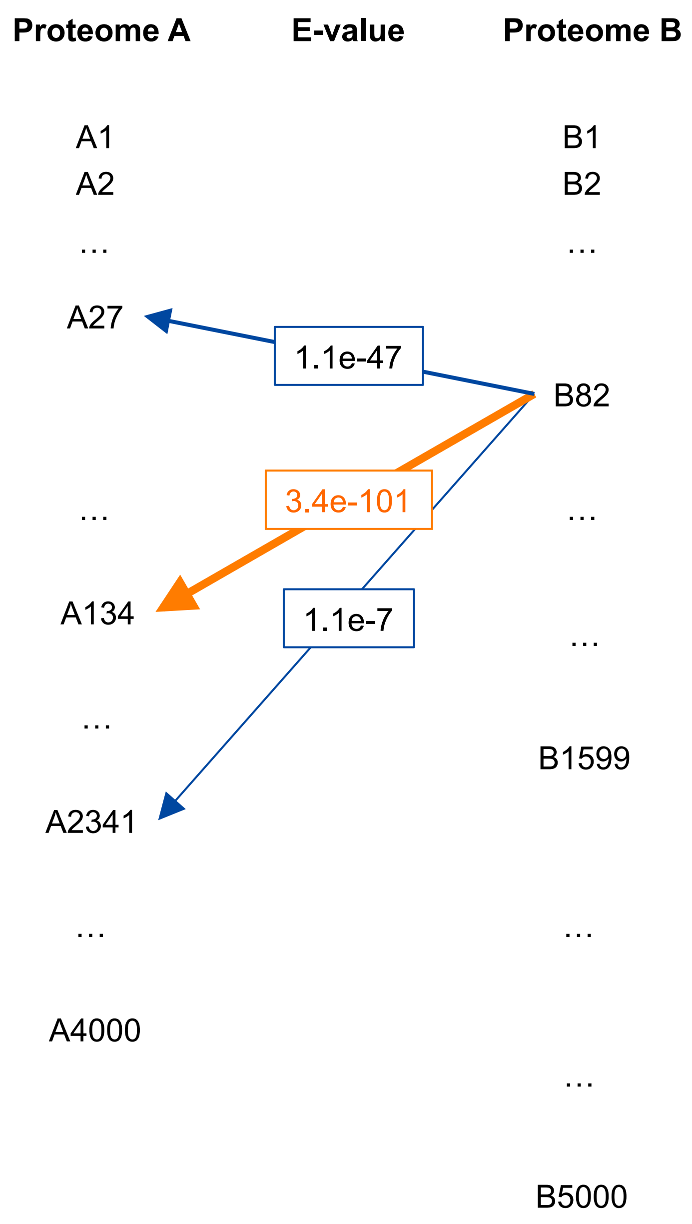
Inferring orthology / paralogy by reciprocal best hits

- Fallback approach: use ***heuristics*** that approximate the solution.
  - The most commonly used method: ***bidirectional best hits (BBH)***, also called ***reciprocal best hits (RBH)***.
- Let us assume
  - Genome A contains 4000 protein-coding genes.
  - Genome B contains 5000 protein-coding genes
- Procedure
  - BLAST each protein of proteome A (query) against each protein of proteome B (database).
  - For each protein, identify ***best hit from A in B***.
  - Note: the best hit is the hit with the ***lowest E-value***.



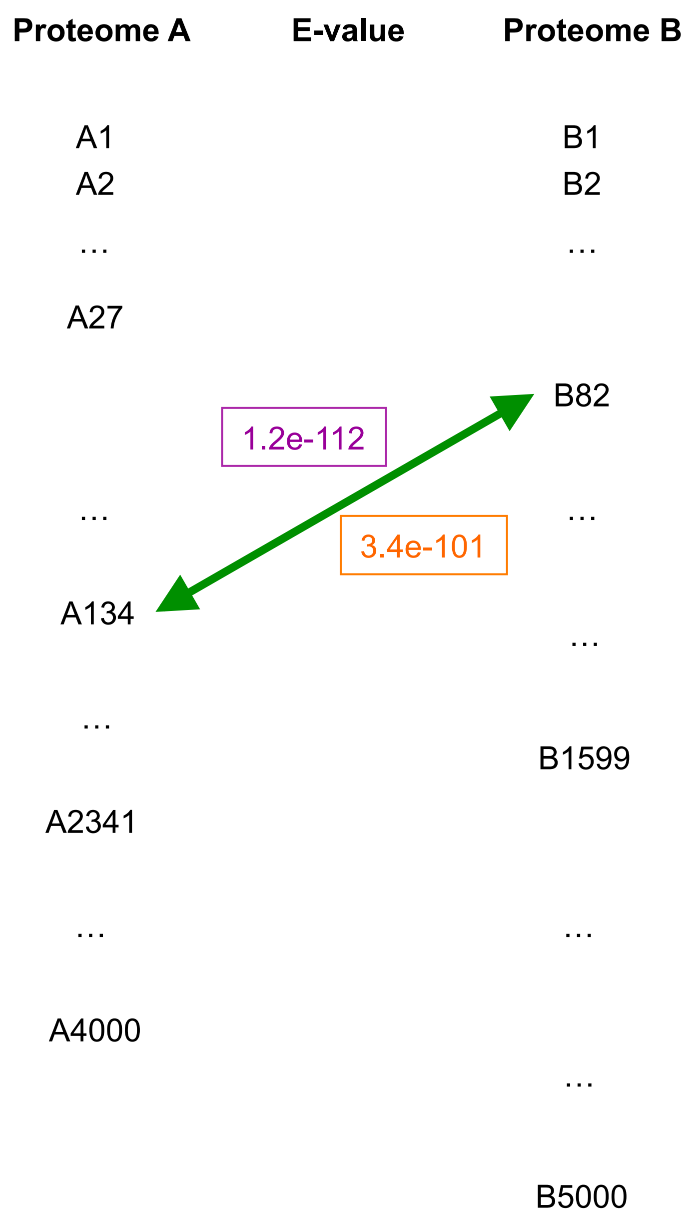
Inferring orthology / paralogy by reciprocal best hits

- Fallback approach: use ***heuristics*** that approximate the solution.
  - The most commonly used method: ***bidirectional best hits (BBH)***, also called ***reciprocal best hits (RBH)***.
- Let us assume
  - Genome A contains 4000 protein-coding genes.
  - Genome B contains 5000 protein-coding genes
- Procedure
  - BLAST each protein of proteome A (query) against each protein of proteome B (database).
  - For each protein, identify best hit from A in B.
  - BLAST each protein of proteome B (query) against each protein of proteome A (database).
  - For each protein, identify ***best hit from B in A***.
  - Note: the best hit is the hit with the ***lowest E-value***.



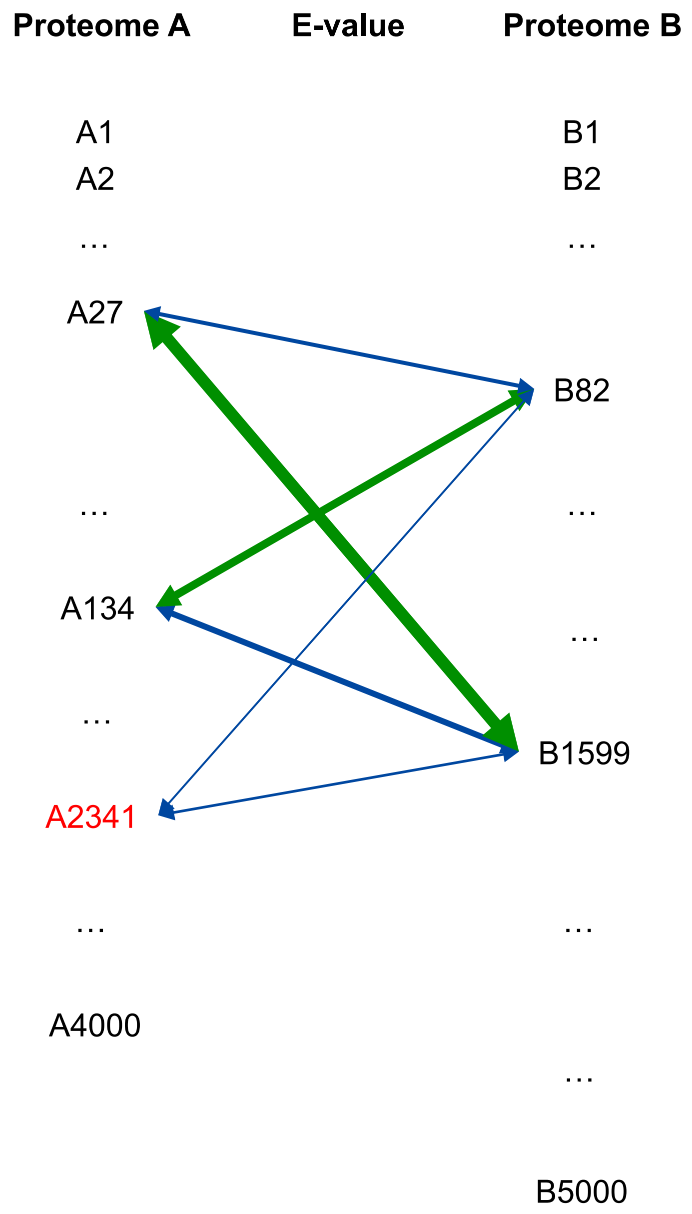
Inferring orthology / paralogy by reciprocal best hits

- Fallback approach: use ***heuristics*** that approximate the solution.
  - The most commonly used method: ***bidirectional best hits (BBH)***, also called ***reciprocal best hits (RBH)***.
- Let us assume
  - Genome A contains 4000 protein-coding genes.
  - Genome B contains 5000 protein-coding genes
- Procedure
  - BLAST each protein of proteome A (query) against each protein of proteome B (database).
  - For each protein, identify ***best hit from A in B***.
  - BLAST each protein of proteome B (query) against each protein of proteome A (database).
  - For each protein, identify ***best hit from B in A***.
  - Identify ***bidirectional best hits***.
  - Note: scores may differ depending on the BLAST direction.
- Advantages
  - Scales up with large number of species.
- Limitations
  - May miss a large number of true orthologies.
  - Intrinsic conceptual flaw: BBH is by definition a 1-to-1 relationship, whereas true orthology is n-to-n.



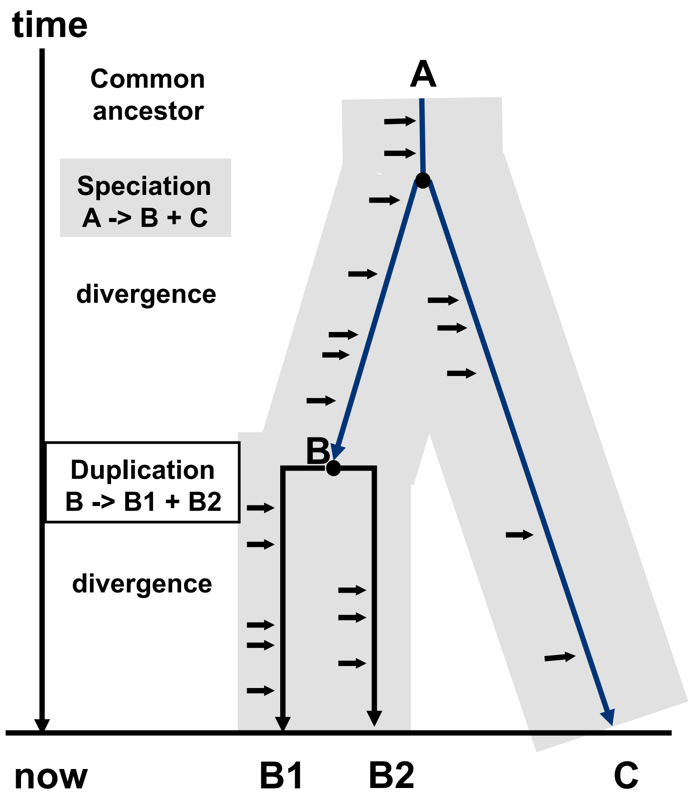
Inferring orthology / paralogy by reciprocal best hits

- For some proteins, there may be no reciprocal best hit.
- In this figure, arrow widths are proportional to the significance of the hit (lower E-values are thicker).
- Bidirectional best hits
  - For A27, the best hit is B1599.
  - For B1599, the best hit is A27.
  - A27 and B1599 are thus BBH.
  - Same reasoning for A134 and B82.
- Protein without BBH
  - For A2341, the best hit is B1599.
  - But for B1599, the best hit is A27.
  - There is thus no BBH for A2341.



Conceptual problem with the RBH/BBH approach

- Let us come back to the schematic example:
  - B and C are orthologs, because their last common ancestor lies just before the speciation
    A -> B + C
  - B1 and B2 are paralogs because the first event that follows their last common ancestor (B) is the duplication
    B -> B1 + B2
- Beware ! These definitions are often misunderstood, even in some textbooks. Contrarily to a strong belief, orthology can be a 1 to N relationship.
  - *B1 and C are orthologs*, because the first event after their last common ancestor (A) was the speciation A -> B + C
  - *B2 and C are orthologs* because the first event after their last common ancestor (A) was the speciation A -> B + C
- The orthology relationship is ***reciprocal*** but ***not transitive***.
  - C <-[orthologous]-> B1
  - C <-[orthologous]-> B2
  - B1 <-[paralogous]-> B2
- Consequences
  - The strategy to **search reciprocal best hits** (**RBH**) is thus a simplification that misses many true orthologs (it is essentially justified by pragmatic reasons).
  - The commonly used ***concept "clusters of orthologous genes (COG)" is thus an aberration***.



Orthologs are sequences whose **last common ancestor** occurred immediately before a **speciation** event.

Paralogs are sequences whose **last common ancestor** occurred immediately before a **duplication** event.
(Fitch, 1970; Zvelebil & Baum, 2000)

Limitations of the BBH approach to infer orthology

- Concepts
  - Best hit (BH)
  - Reciprocal (RBH) or bidirectional (BBH) best hit.
- Problem 1: ***non-reciprocity of the BH relationship***, which may result from various effects
  - Multidomain proteins -> non-transitivity of the homology relationship
    - Detection: no paralogy
  - Paralogs in one genome corresponding to the same ortholog in the other genome
  - Non-symmetry of the BLAST result (can be circumvented by using dynamical programming, e.g. Smith-Waterman)
- Problem 2: ***unequivocal but fake reciprocal best hit***
  - Duplication followed by a deletion
  - Two paralogs can be BBH, but the true orthologs are not present anymore in the genome (due to duplication).
  - Ex: Hox genes
- Conceptual problem: ***intrinsically unable to treat multi-orthology relationships***
  - Ex: Fitch figure: B2 is orhtolog to both C2 and C3, but only one of these will be its Best Hit.
- Conclusion: the analysis of BBH is intrinsically unable to reveal the true orthology relationships

How to circumvent the weaknesses of RBH ?

- Solutions to the problems with RBH
  - ***Domain analysis***: analyse the location of the hits in the alignments
    - Resolves the problems of gene fusion (two different fragments of a protein in genome A correspond to 2 distinct proteins of genome B)
  - Analysis of the evolutionary history : full phylogenetic inference + reconciliation of the sequence tree and the species tree
    - Resolves the cases of multiple orthology relationships (n to n)
    - Does not resolve the problems of differential deletions after regional duplications
  - Solving the problem of regional duplications followed by differential deletion
    - Analysis of synteny: neighbourhood relationships between genes across genomes
    - Analysis of pseudo-genes: allows to infer the presence of a putative gene in the common ancestor
    - This is OK when the duplication affects a regions sufficiently large to encompass multiple genes.
- These solutions require a case-by-case analysis -> this is not what you will find in the large-scale databases.

- Resources:
  - EnsEMBL database
  - SPRING database

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 (Expasy, 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)
  - …

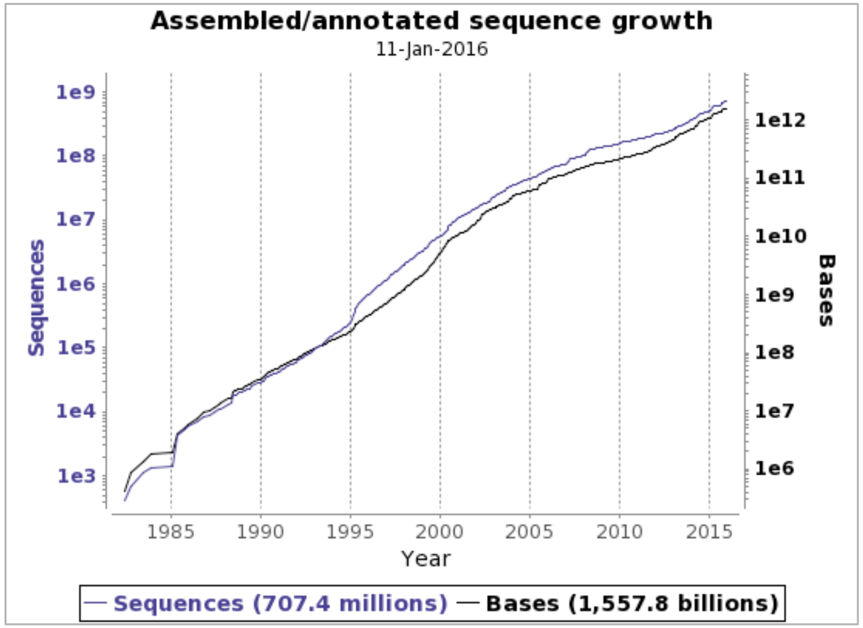
Bases de données de bases de données

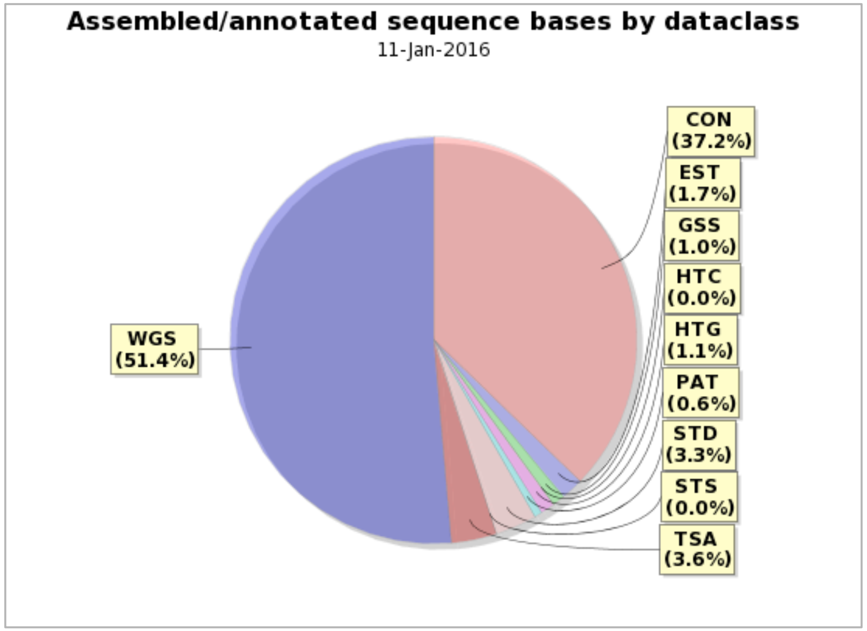
- Il existe des centaines de bases de données spécialisées pour la biologie moléculaire et la biochimie. Ce nombre augmente chaque année.
- Pour s'y retrouver, la revue Nucleic Acids Research consacre chaque année son numéro de janvier à une revue des bases de données existantes, et maintient un catalogue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

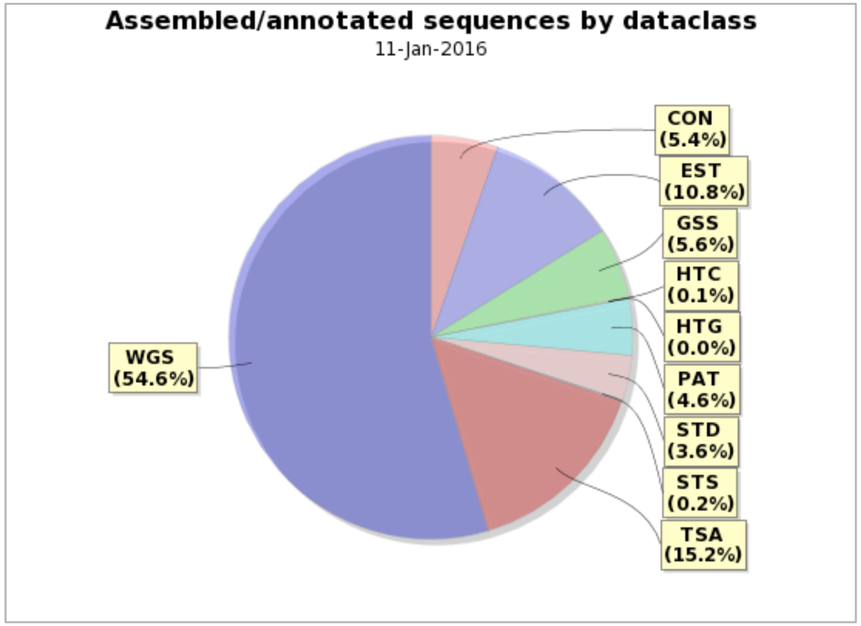
- Fernandez-Suarez and Galperin, 2013, NAR D1-D7

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).



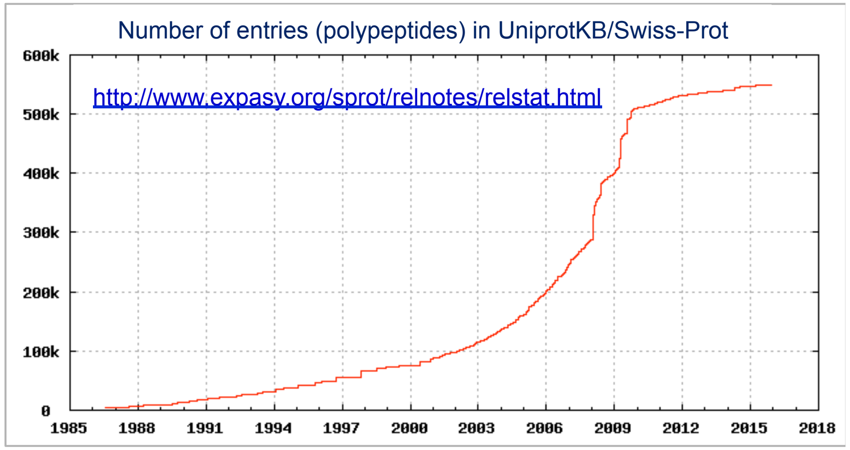


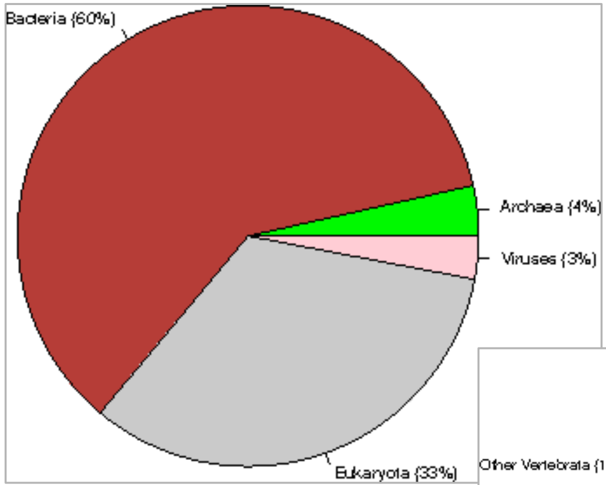


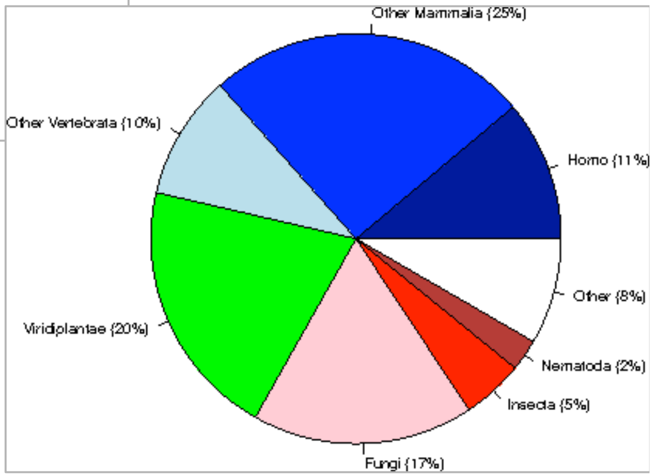
- http://www.ebi.ac.uk/ena/about/statistics

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.



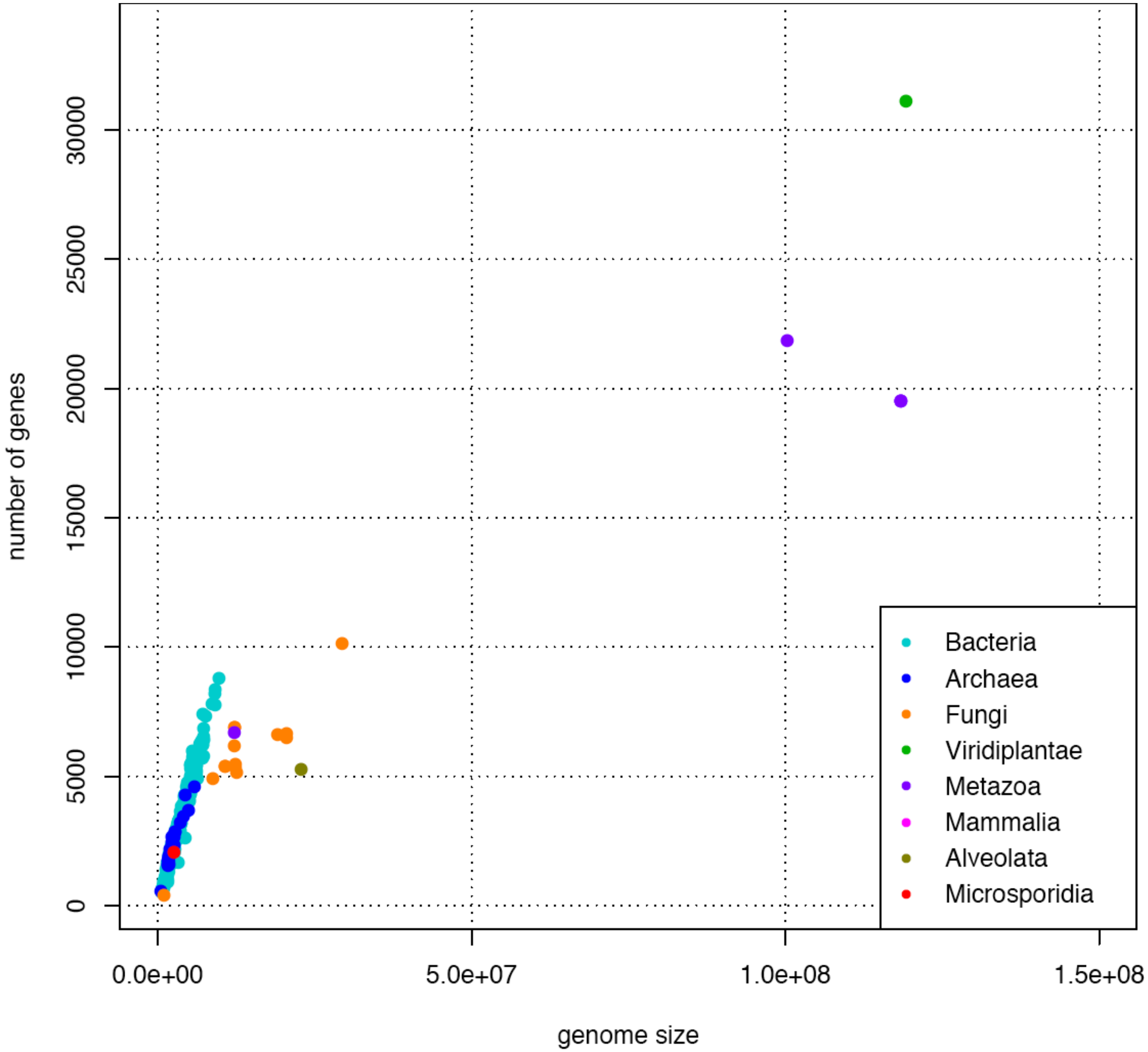




Some milestones

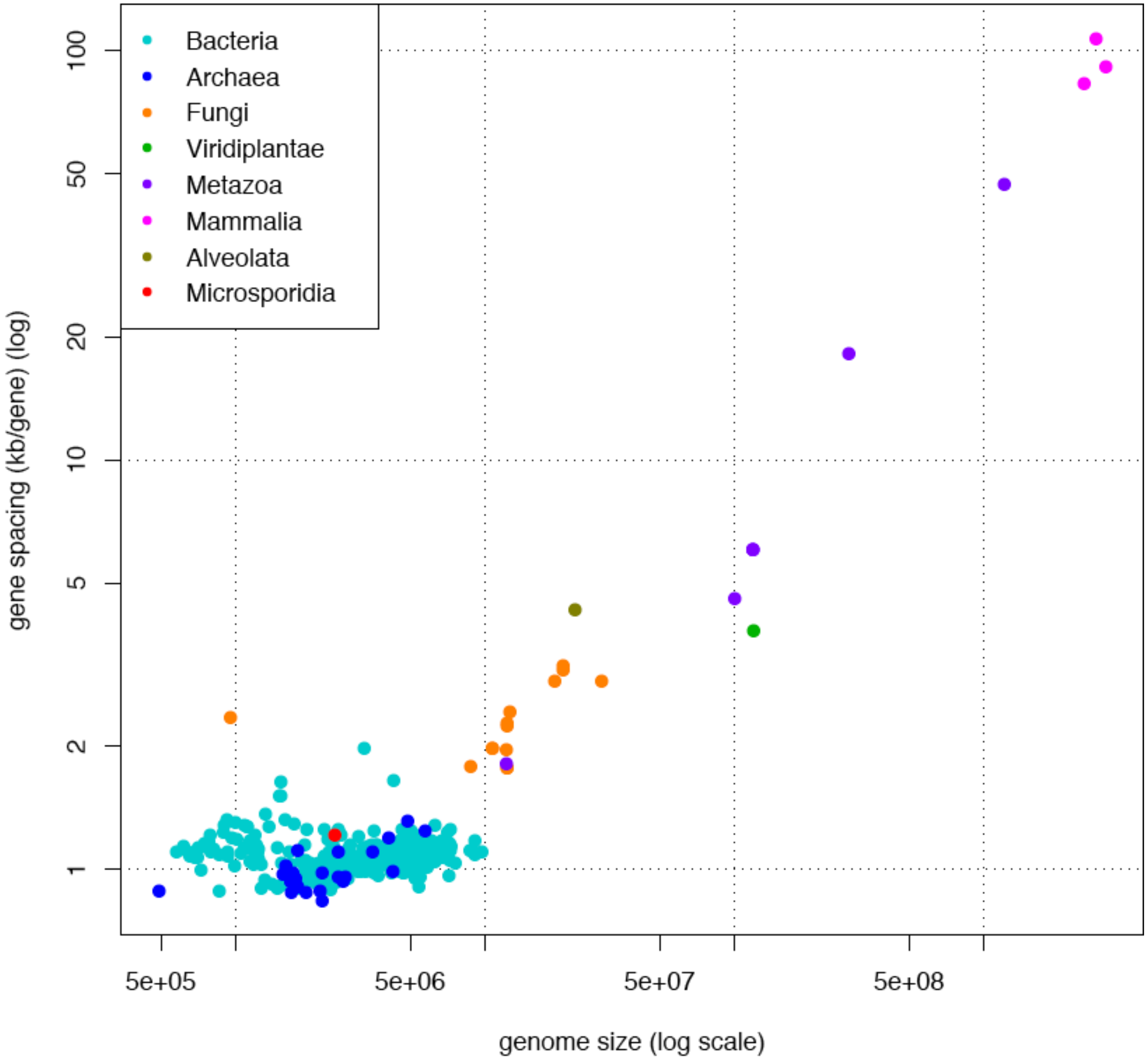
Speices name	Common name	Publication year	Genome size Mb	Gene number	Mean intergenic distance Kb	Fraction of coding sequences %	Non-coding fraction %	Repetitive elements %	Transcribed fraction %	Remarks
Bactérie										
Mycoplasma genitalium	*Mycoplasma*	1995	0.6	481	1.2	90	10			Small genome (intracellular)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

Genes and genome size



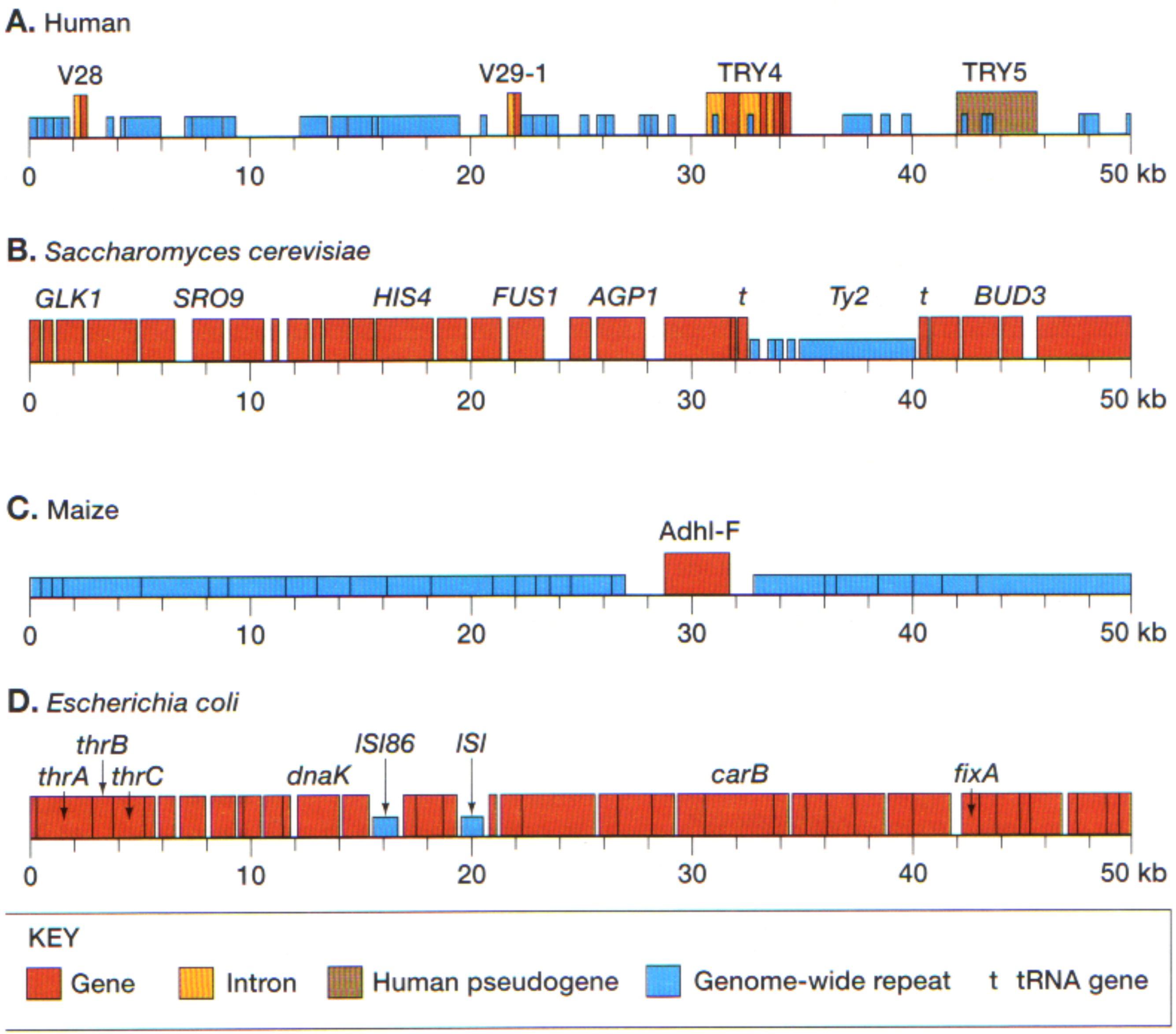
- In prokaryotes, the number of genes increases linearly with genome size
- In eukaryotes, this is not the case: the genome size increases faster than the number of genes

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



Source: Mount (2000)

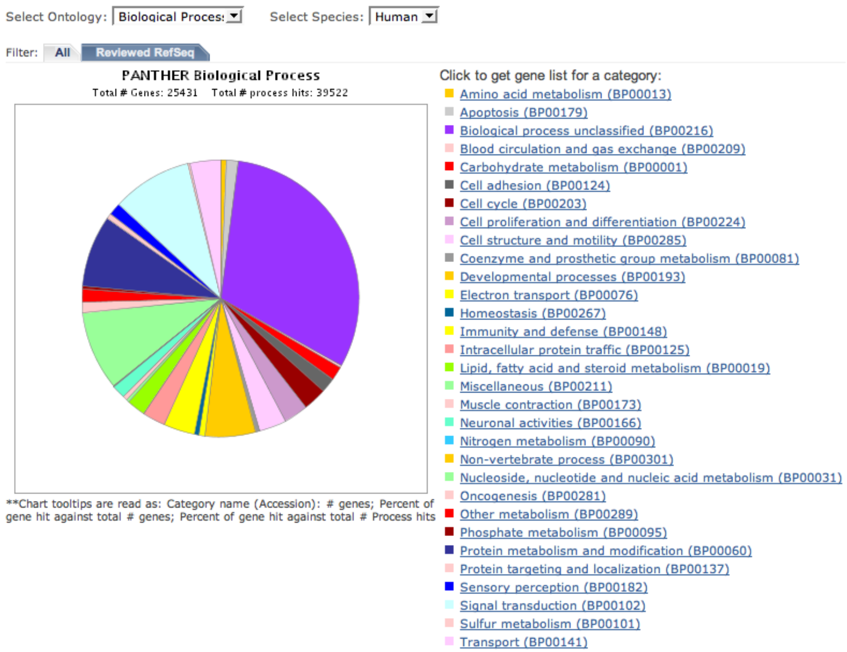
Approches d'annotation de la fonction des gènes

- 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 unknwown 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   ILAASLVNAGTIPLGKLSDIDKIGTQTEIFPFLGGSGPYYSFPGDYGISRDLPESCEMKQ 68
           +LAAS+V+AG       S  +    +      LG    Y+  P   G +   PESC +KQ
Sbjct: 10  LLAASIVHAGK------SQFEAFENEFYFKDHLGTISVYHE-PYFNGPTTSFPESCAIKQ 62
Query: 69  VQMVGRHGERYPT-------VSKAKSIMTTWYKLSNYTGQFSGALSFLNDDYEFFIRDTK 121
           V ++ RHG R PT       VS A+ I     KL N  G      S+  +   F    T
Sbjct: 63  VHLLQRHGSRNPTGDDTATDVSSAQYIDIFQNKLLN--GSIPVNFSYPENPLYFVKHWTP 120
Query: 122 NLEMETTLANSVNVLNPYTGEMNAKRHARDFLAQYGYMVENQTSFAVFTSNSNRCHDTAQ 181
            ++ E     S +      G +      R    +Y Y + +   + + T+   R  D+A+
Sbjct: 121 VIKAENADQLSSS------GRIELFDLGRQVFERY-YELFDTDVYDINTAAQERVVDSAE 173
Query: 182 YFIDGL-GDKFN--ISLQTISEAESAGANTLSAHHSCPAWDDDVNDDILKK-----YDTK 233
           +F  G+ GD      +   + E +SAGAN+L+ ++SCP ++D+  D+   +     +
Sbjct: 174 WFSYGMFGDDMQNKTNFIVLPEDDSAGANSLAMYYSCPVYEDNNIDENTTEAAHTSWRNV 233
Query: 234 YLSGIAKRLNKE-NKGLNLTSSDANTFFAWCAYEINARGYSDICNIFTKDELVRFSYGQD 292
           +L  IA RLNK  + G NLT SD  + +  C YEI  R  SD C++FT  E + F Y  D
Sbjct: 234 FLKPIANRLNKYFDSGYNLTVSDVRSLYYICVYEIALRDNSDFCSLFTPSEFLNFEYDSD 293
Query: 293 LETYYQTGPGYDVVRSVGANLFNASVKLLKE--SEVQDQKVWLSFTHDTDILNYLTTIGI 350
           L+  Y GP   +   ++G    N     L++  +   D+KV+L+FTHD+ I+     +G
Sbjct: 294 LDYAYWGGPASEWASTLGGAYVNNLANNLRKGVNNASDRKVFLAFTHDSQIIPVEAALGF 353
Query: 351 IDDKNNLTAEH-VPFMENTF----HRSWYVPQGARVYTEKFQCS-NDTYVRYVINDAVVP 404
             D   +T EH +P  +N F      S +VP    + TE F CS N  YVR+++N  V P
Sbjct: 354 FPD---ITPEHPLPTDKNIFTYSLKTSSFVPFAGNLITELFLCSDNKYYVRHLVNQQVYP 410
Query: 405 IETCSTGPGFS----CEINDFYDYAEKRVAGTDFLKVCNVSSVSNSTELTFFW 453
           +  C  GP  +    CE++ + + + +  + ++ +   N    ++ST +T ++
Sbjct: 411 LTDCGYGPSGASDGLCELSAYLNSSVRVNSTSNGIANFNSQCQAHSTNVTVYY 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	A.aeolicus	C.muridarum	C.pneumoniae.AR39	Nostoc.sp	Synechocystis.PCC6803	B.halodurans	B.subtilis	C.acetobutylicum	C.glutamicum	C.perfringens	L.innocua	L.lactis	M.genitalium	M.leprae	M.pneumoniae	M.pulmonis	S.aureus.MW2	S.coelicolor
16127995	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	1	0	0	1	1
16127997	1	0	0	1	1	1	1	1	1	0	1	1	0	1	0	0	1	1
16127998	0	0	0	0	1	0	0	1	1	0	0	1	0	1	0	0	0	0
16127999	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
16128001	0	1	1	0	0	1	1	0	1	1	0	0	0	0	0	0	1	1
16128002	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
16128003	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1
16128004	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
16128006	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
16128008	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
16128009	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

Gene	S.pneumoniae.R6	S.pyogenes	T.tengcongensis	U.urealyticum	F.nucleatum	A.tumefaciens.C58	B.aphidicola.Sg	B.melitensis	C.crescentus	C.jejuni	H.influenzae	H.pylori.26695	M.loti	N.meningitidis.MC58	P.aeruginosa	R.conorii	R.solanacearum	S.melil
16127995	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16127996	1	0	1	0	0	1	1	1	1	1	1	1	1	1	1	0	1	1
16127997	1	0	1	0	0	0	1	0	0	1	1	1	0	0	0	0	0	0
16127998	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	1	1
16127999	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16128000	1	1	0	0	1	0	0	0	1	0	1	0	0	1	1	1	1	0
16128001	1	1	0	0	1	0	0	1	0	1	1	1	0	1	1	0	0	1
16128002	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0
16128003	0	0	0	0	0	0	0	1	0	1	1	1	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
16128005	0	0	0	0	0	0	0	0	0	0	0	1	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
16128007	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	1	1
16128009	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

Pellegrini et al. (1999). Proc Natl Acad Sci U S A 96(8), 4285-8.

Phylogenetic profiles reveal groups of functionally related genes

- In 1999, based on the 16 genomes available at that time, Pellegrini et al. propose a method relying on ***phylogenetic profiles***
  - For each gene of a reference genome, detect all orthologs in a set of other genomes (phylogenetic profiles of gene occurrence)
  - Detect groups of co-occurring genes : similar profiles of presence / absence across genomes
- This method can now be applied to several thousands of genomes. Its power increases with the number of genomes.

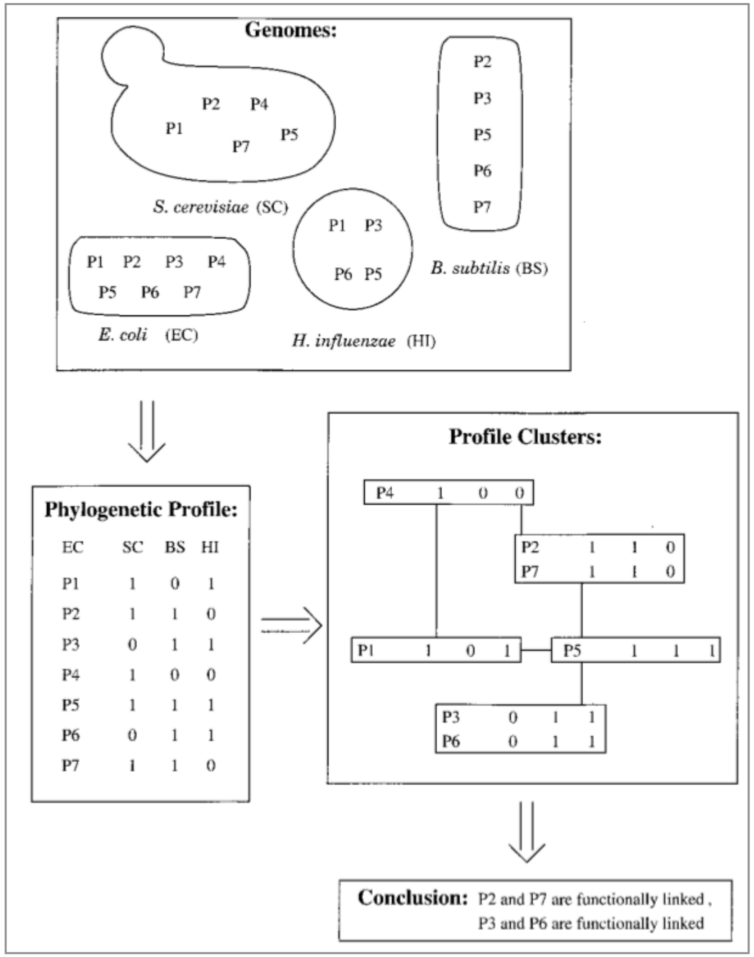


Table 1. Phylogenetic profiles link protein with similar keywords

Keyword	No. proteins	No. neighbors in keyword group	No. neighbors in random group
Ribosome	60	197	27
Transcription	36	17	10
tRNA synthase and ligase	26	11	5
Membrane proteins*	25	89	5
Flagellar	21	89	3
Iron, ferric, and ferritin	19	31	2
Galactose metabolism	18	31	2
Molybdoterin and Molybdenum, and molybdoterin	12	6	1
Hypothetical†	1,084	108,226	8,440

Proteins grouped on the basis of similar keywords in SwissProt have more similar phylogenetic profiles than random proteins. Column 2 gives the number of nonhomologous proteins in the keyword group. Column 3 gives the number of protein pairs in the keyword group with profiles that differ by less than 3 bits. These pairs are called neighbors. Column 4 lists the number of neighbors found on average for a random group of proteins of the same size as the keyword group.

*Only membrane proteins without uniformly zero phylogenetic profiles were included.

†Unlike the other rows of the table, the hypothetical proteins do contain homologous pairs.

- Pellegrini et al. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA (1999) vol. 96 (8) pp. 4285-8

Phylogenetic profiles reveal groups of functionally related genes

- The phylogenetic profile table indicates the presence/absence of genes (one row per gene) in a set of genomes (one column per cross-species comparison).
- Reference organism: Escherichia coli K12 MG1665
- Query genomes
  - Selected 154 Bacteria among 2065 (1 species for each group at depth 5 of the taxonomic tree, to avoid redundant genomes).
  - Reference genome contains 4322 CDS.
- Ortholog identification: BLAST BBH
  - Max expect: 1e-10
  - Min identity: 30%
  - Min length: 50
  - At least one non-E.coli ortholog (BBH) found for 1994 genes.

Reference organism: Escherichia coli K12 MG1665
Query genomes
154 Bacteria (among 2065) Selected 1 species for each group at depth 5 of the taxonomic tree, to avoid redundant genomes.

Stats on organisms per reference gene

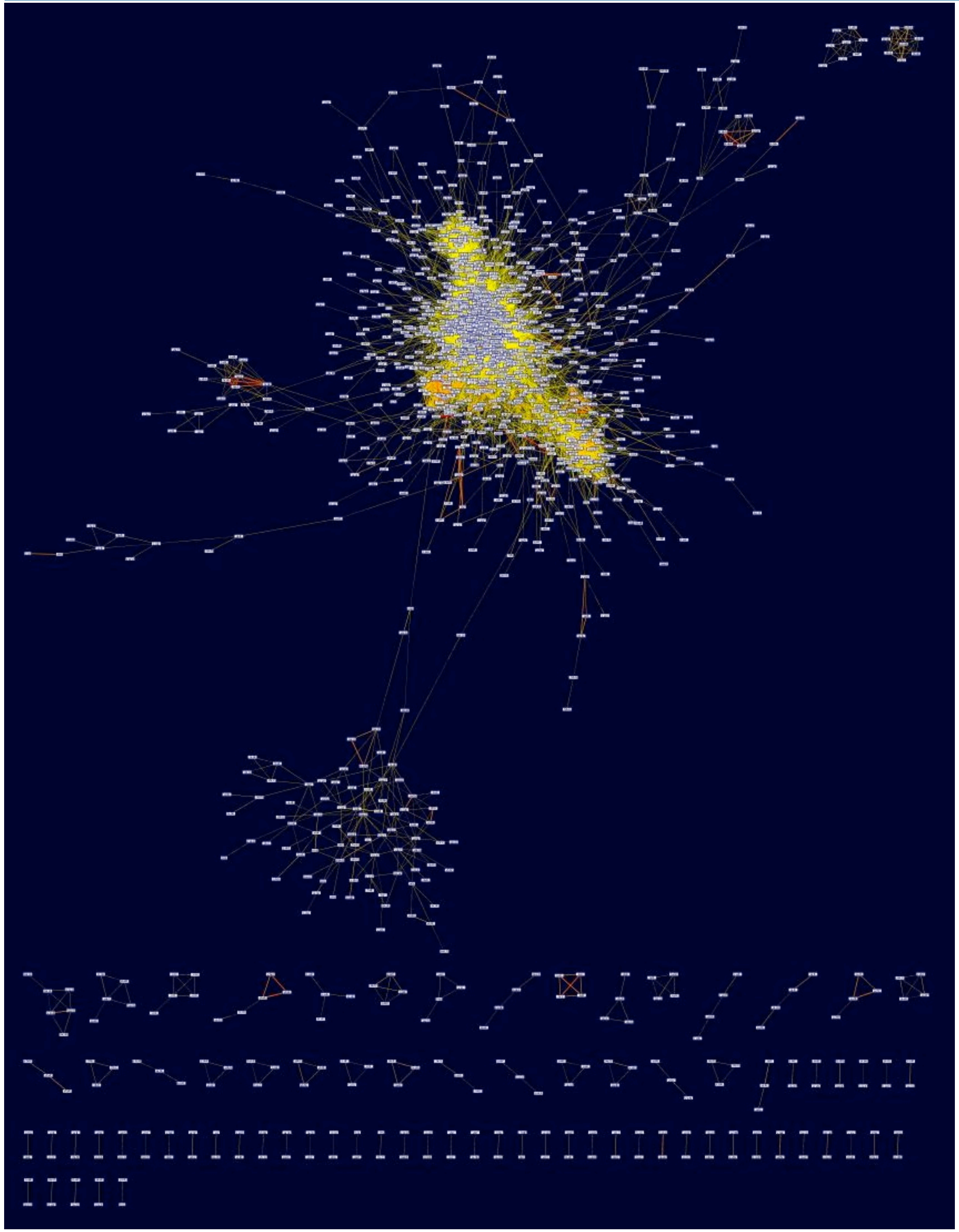
min	1
max	123
mean	23.3
median	10.0

Stats on orthologs per organism

min	73
max	3968
mean	758.2
median	677

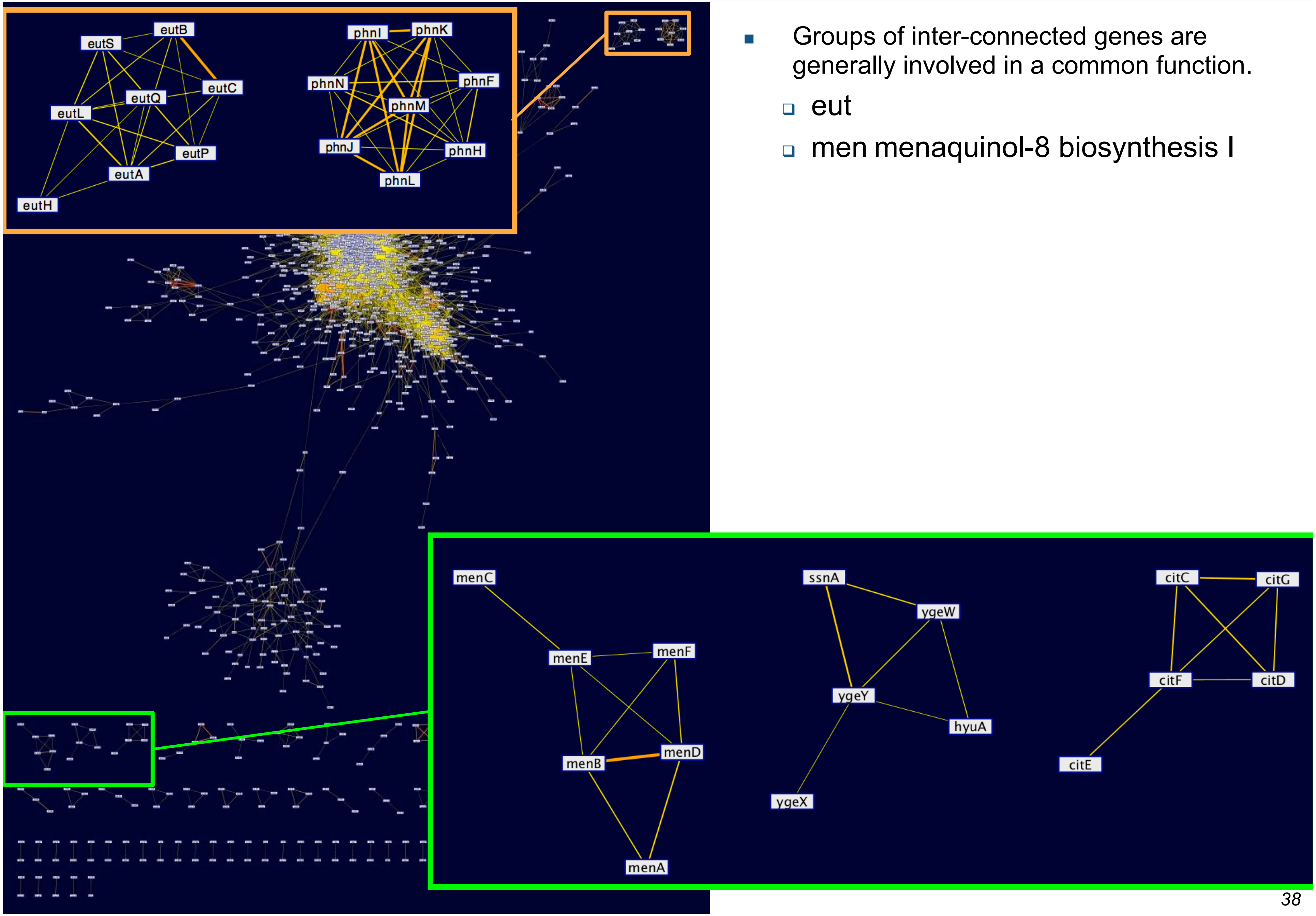
Gene name	Gene ID	Number of organisms per reference gene	Escherichia_coli_K_12_substr__MG1655_uid57779	Acidobacterium_capsulatum_ATCC_51196_uid59127	Candidatus_Chloracidobacterium_thermophilum_B_uid73587	Candidatus_Koribacter_versatilis_Ellin345_uid58479	Candidatus_Solibacter_usitatus_Ellin6076_uid58139	Rubrobacter_xylanophilus_DSM_9941_uid58057	Conexibacter_woesei_DSM_14684_uid43467	Aquifex_aeolicus_VF5_uid57765	Desulfurobacterium_thermolithotrophum_DSM_11699_uid63405	Persephonella_marina_EX_H1_uid58119	Rhodothermus_marinus_DSM_4252_uid41729	Salinibacter_ruber_uid47323	Chitinophaga_pinensis_DSM_2588_uid59113	Chlorobaculum_parvum_NCIB_8327_uid59185	Dehalogenimonas_lykanthroporepellens_BL_DC_9_uid48131	Herpetosiphon_aurantiacus_ATCC_23779_uid58599	Sphaerobacter_thermophilus_DSM_20745_uid41997	Desulfurispirillum_indicum_S5_uid45897	Acaryochloris_marina_MBIC11017_uid58167	Microcystis_aeruginosa_NIES_843_uid59101	Synechococcus_CC9311_uid58123	Thermosynechococcus_elongatus_BP_1_uid57907	Gloeobacter_violaceus_PCC_7421_uid58011	Trichodesmium_erythraeum_IMS101_uid57925	Prochlorococcus_marinus_AS9601_uid58307	Deinococcus_deserti_VCD115_uid58615	Marinithermus_hydrothermalis_DSM_14884_uid65783
Number of orthologs per organism			3968	782	673	892	964	709	793	575	540	619	765	677	849	671	488	770	726	774	799	713	628	640	770	687	489	783	684
thrA	NP_414543.1	84	1	0	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
thrB	NP_414544.1	23	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	1	1	1	1	1	1	0	1	0
thrC	NP_414545.1	46	1	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
yaaX	NP_414546.1	1	1	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
yaaA	NP_414547.1	22	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
yaaJ	NP_414548.1	41	1	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1
talB	NP_414549.1	37	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	1	1	0
mog	NP_414550.1	46	1	0	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1
yaaH	NP_414551.1	17	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
yaaW	NP_414552.1	1	1	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
yaaI	NP_414554.1	1	1	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
dnaK	NP_414555.1	115	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	0	1	0	1	1	1	1	1	1
dnaJ	NP_414556.1	116	1	1	1	1	0	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
mokC	NP_414559.1	1	1	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
nhaA	NP_414560.1	36	1	0	1	1	0	0	1	0	0	0	1	0	1	1	1	0	1	0	0	0	0	0	0	0	0	1	0
nhaR	NP_414561.1	13	1	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
rpsT	NP_414564.1	86	1	1	0	0	1	1	0	0	0	1	1	0	0	0	1	0	0	1	1	0	1	1	1	1	1	0	1
yaaY	NP_414565.1	1	1	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
ribF	NP_414566.1	112	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
ileS	NP_414567.1	93	1	1	1	1	1	1	0	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0	0
lspA	NP_414568.1	74	1	0	1	1	1	0	1	1	1	1	0	0	0	0	1	1	1	1	0	1	0	1	1	1	1	0	0
fkpB	NP_414569.1	19	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
ispH	NP_414570.1	72	1	1	0	1	1	0	1	1	1	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	1	1
rihC	NP_414571.1	7	1	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
dapB	NP_414572.1	85	1	1	0	1	0	1	0	1	1	1	1	1	1	1	0	0	0	1	0	0	0	1	1	0	0	0	0
carA	NP_414573.1	108	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
carB	NP_414574.1	105	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

Co-occurrence network extracted from phylogenetic profiles



- Co-occurrence network extracted from phylogenetic profiles
  - Reference organism: Escherichia coli K12 MG1665
  - Query genomes
    - 154 Bacteria (among 2065)
    - Selected 1 species for each group at depth 5 of the taxonomic tree, to avoid redundant genomes.
  - Similarity metrics: hypergeometric significance
- Resulting network
  - 1433 nodes (genes)
  - 20728 edges
- For a discussion about network inference parameters, see
  - Ferrer et al. A systematic study of genome context methods: calibration, normalization and combination. BMC Bioinformatics 2010 11:493 (2010) vol. 11 (1) pp. 493

Co-occurrence network extracted from phylogenetic profiles

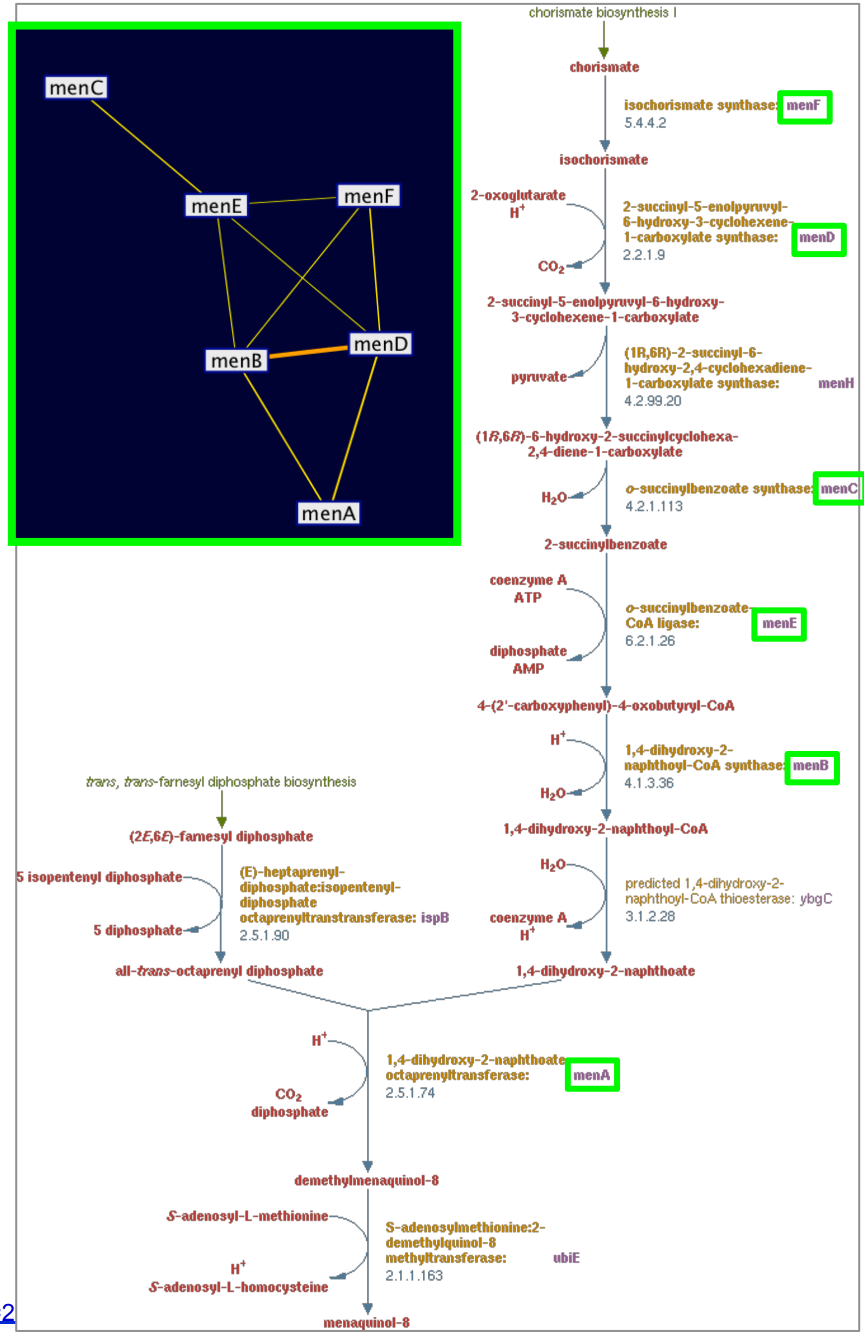


- Groups of inter-connected genes are generally involved in a common function.
  - eut
  - men menaquinol-8 biosynthesis I

Clusters of co-occurring genes reveal pathways

- Phylogenetic profiles reveal a group of co-occurring genes whose name starts by “men”.
- Phylogenetic profiles revealed the associations between these genes without any indication of their function.
- However these genes can be mapped onto annotated pathways to indicate their respective roles.
- The 6 genes of the “men” cluster code for enzymes that catalyze 6 among 10 reactions of the superpathway “menaquinol-8 biosynthesis I”.

- Without any prior information, the footprint-discovery approach thus revealed the functional relationship between these 6 genes, on the simple base of their co-occurrence across genomes.



- http://ecocyc.org/ECOLI/NEW-IMAGE?type=PATHWAY&object=PWY-5838&detail-level=2

Gene fusions / fissions

- In 1999, two groups propose a method to predict functional interactions between genes based on cross-genome identification of gene fusions.
  - Marcotte et al., Science, 1999
  - Enright et al., Nature, 1999

Protein interaction maps for complete genomes based on gene fusion events

Anton J. Enright, Ioannis Iliopoulos, Nikos C. Kyrpides*
& Christos A. Ouzounis

Computational Genomics Group, Research Programme, The European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge CB10 1SD, UK
** Integrated Genomics Inc., 2201 West Campbell Park Drive, Chicago, Illinois 60612, USA*

A large-scale effort to measure, detect and analyse protein–protein interactions using experimental methods is under way[1,2]. These include biochemistry such as co-immunoprecipitation or crosslinking, molecular biology such as the two-hybrid system or phage display, and genetics such as unlinked noncomplementing mutant detection[3]. Using the two-hybrid system[4], an international effort to analyse the complete yeast genome is in progress[5]. Evidently, all these approaches are tedious, labour intensive and inaccurate[6]. From a computational perspective, the question is how can we predict that two proteins interact from structure or sequence alone. Here we present a method that identifies gene-fusion events in complete genomes, solely based on sequence comparison. Because there must be selective pressure for certain genes to be fused over the course of evolution, we are able to predict functional associations of proteins. We show that 215 genes or proteins in the complete genomes of *Escherichia coli*,

Detecting Protein Function and Protein-Protein Interactions from Genome Sequences

Edward M. Marcotte, Matteo Pellegrini, Ho-Leung Ng, Danny W. Rice, Todd O. Yeates, David Eisenberg*

A computational method is proposed for inferring protein interactions from genome sequences on the basis of the observation that some pairs of interacting proteins have homologs in another organism fused into a single protein chain. Searching sequences from many genomes revealed 6809 such putative protein-protein interactions in *Escherichia coli* and 45,502 in yeast. Many members of these pairs were confirmed as functionally related; computational filtering further enriches for interactions. Some proteins have links to several other proteins; these coupled links appear to represent functional interactions such as complexes or pathways. Experimentally confirmed interacting pairs are documented in a Database of Interacting Proteins.

- Marcotte et al. Detecting protein function and protein-protein interactions from genome sequences. Science (1999) vol. 285 (5428) pp. 751-3
- Enright et al. Protein interaction maps for complete genomes based on gene fusion events. Nature (1999) vol. 402 (6757) pp. 86-90

Inferring groups of functionally related genes from gene fusions

- Marcotte et al. further show that groups of fused genes (Fig B, D) are functionally linked.
- They show two examples of gene groups coding for the enzymes of specific metabolic pathways (A, C).
- This opens the perspective to guess the function of unknown genes based on their fusion with genes of known function (method called “***guilty by association***”).

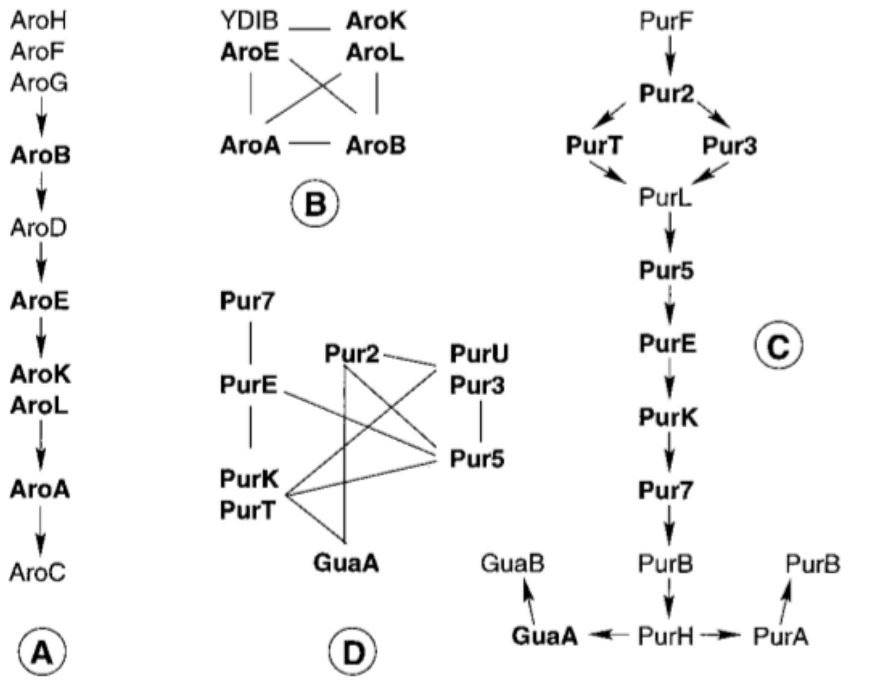
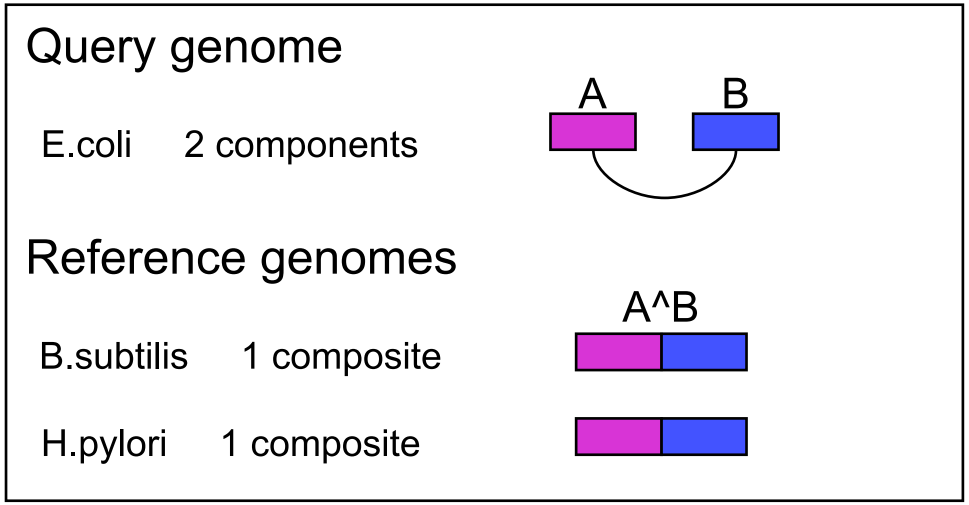


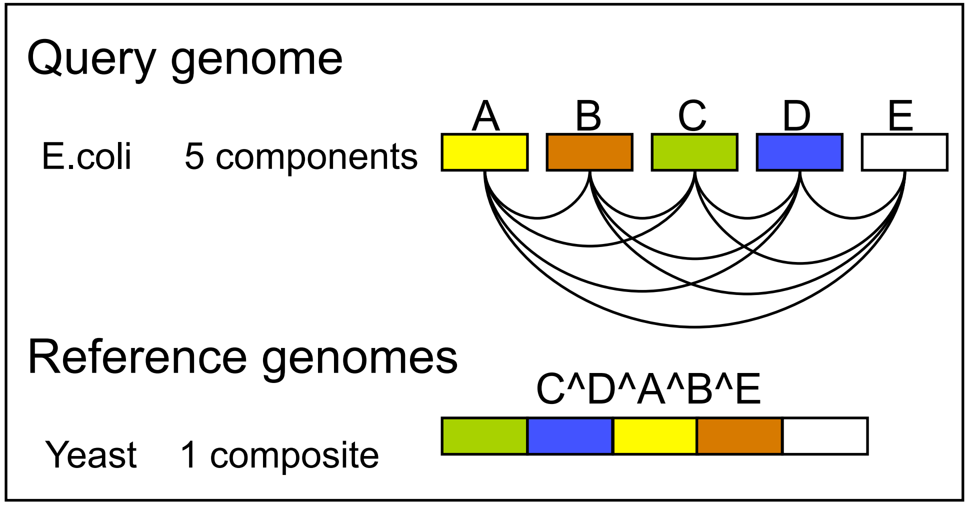
Fig. 2. Reconstruction of two metabolic pathways in *E. coli*, with only interactions predicted by the domain fusion method. Pathways A and C are the known pathways for biosynthesis of shikimate and purine, respectively; they are ordered by the traditional method of successive action of the enzymes on the known metabolites. Pathways B and D are constructed from the proteins in pathways A and C with connections predicted by the domain fusion method. In both cases, more than half of the proteins in the biochemical pathway are predicted by the domain fusion method to interact with other proteins of the pathway. It is possible that these groupings represent multiprotein complexes. Enzymes stacked together (for example, AroK and AroL) are homologs.

- Marcotte et al. Detecting protein function and protein-protein interactions from genome sequences. Science (1999) vol. 285 (5428) pp. 751-3

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.

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*

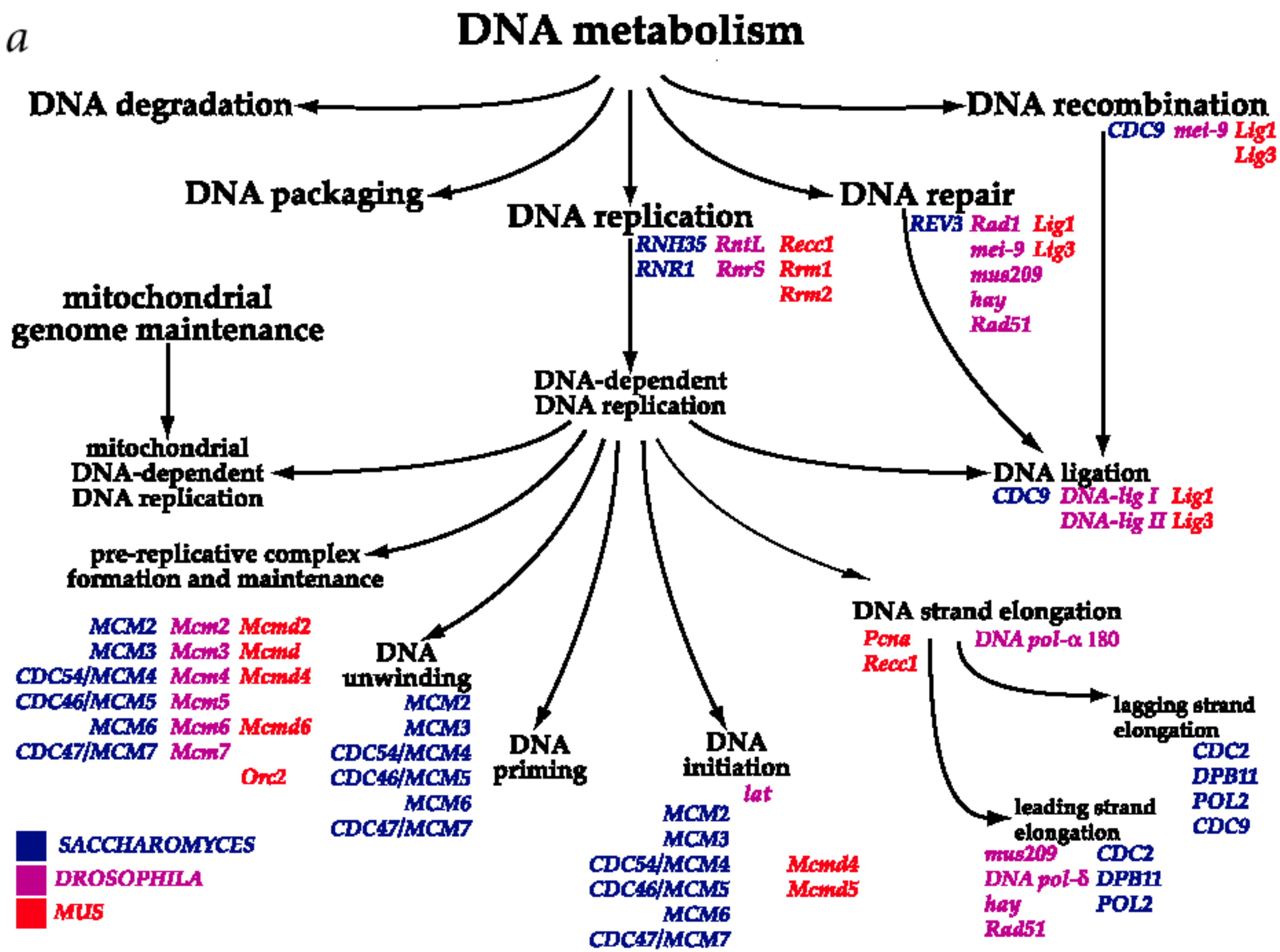
Les"bio-ontologies"

- 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.

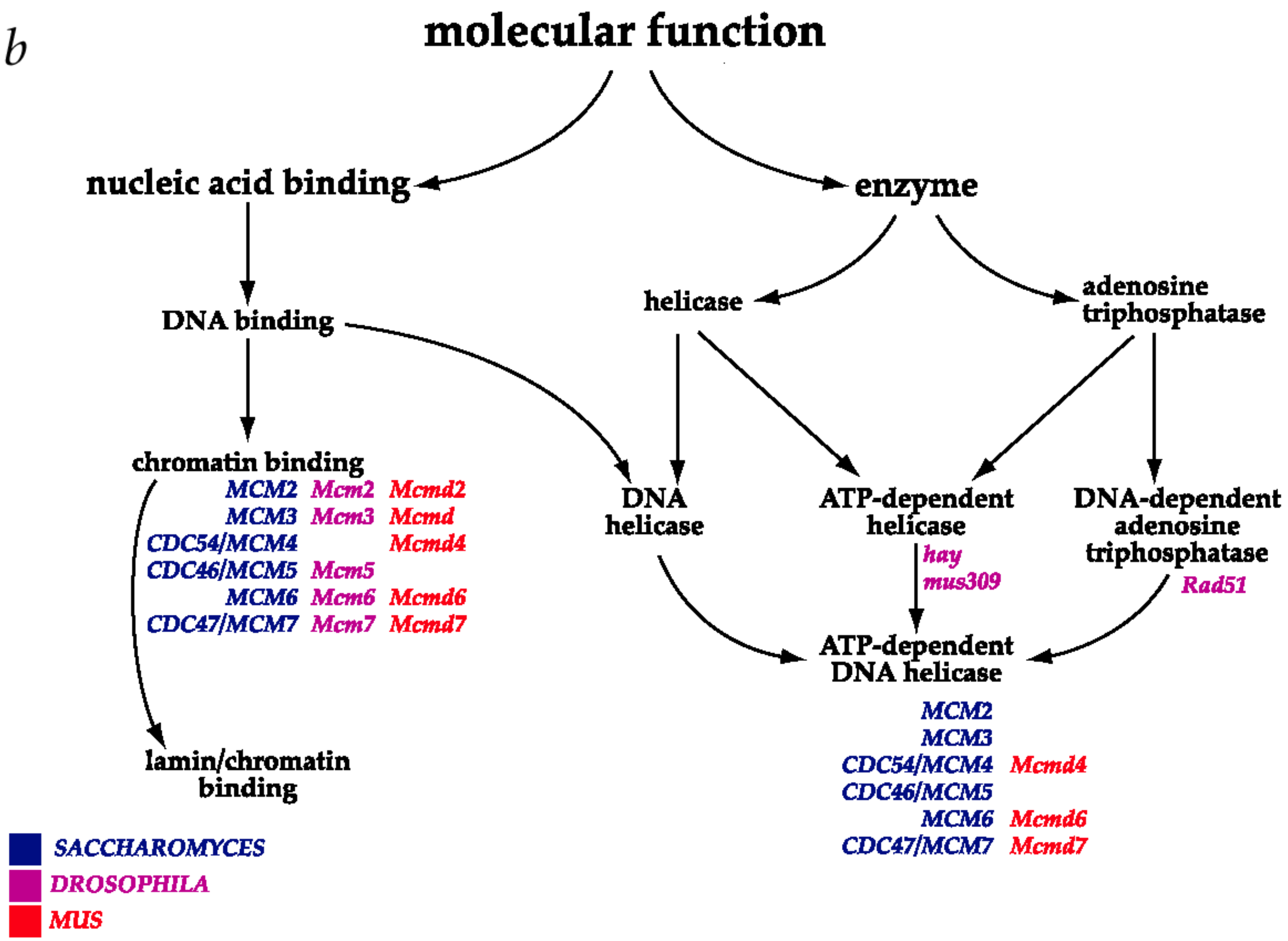
- Les bio-ontologies visent à répondre au problème d'inconsistance entre annotations.
- Pour y répondre, on met en place
  - Un vocabulaire contrôlé
    - On utilise toujours le même mot pour désigner le même concept.
    - Les listes de synonymes permettent d'établir les correspondances.
  - Classification 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, …).

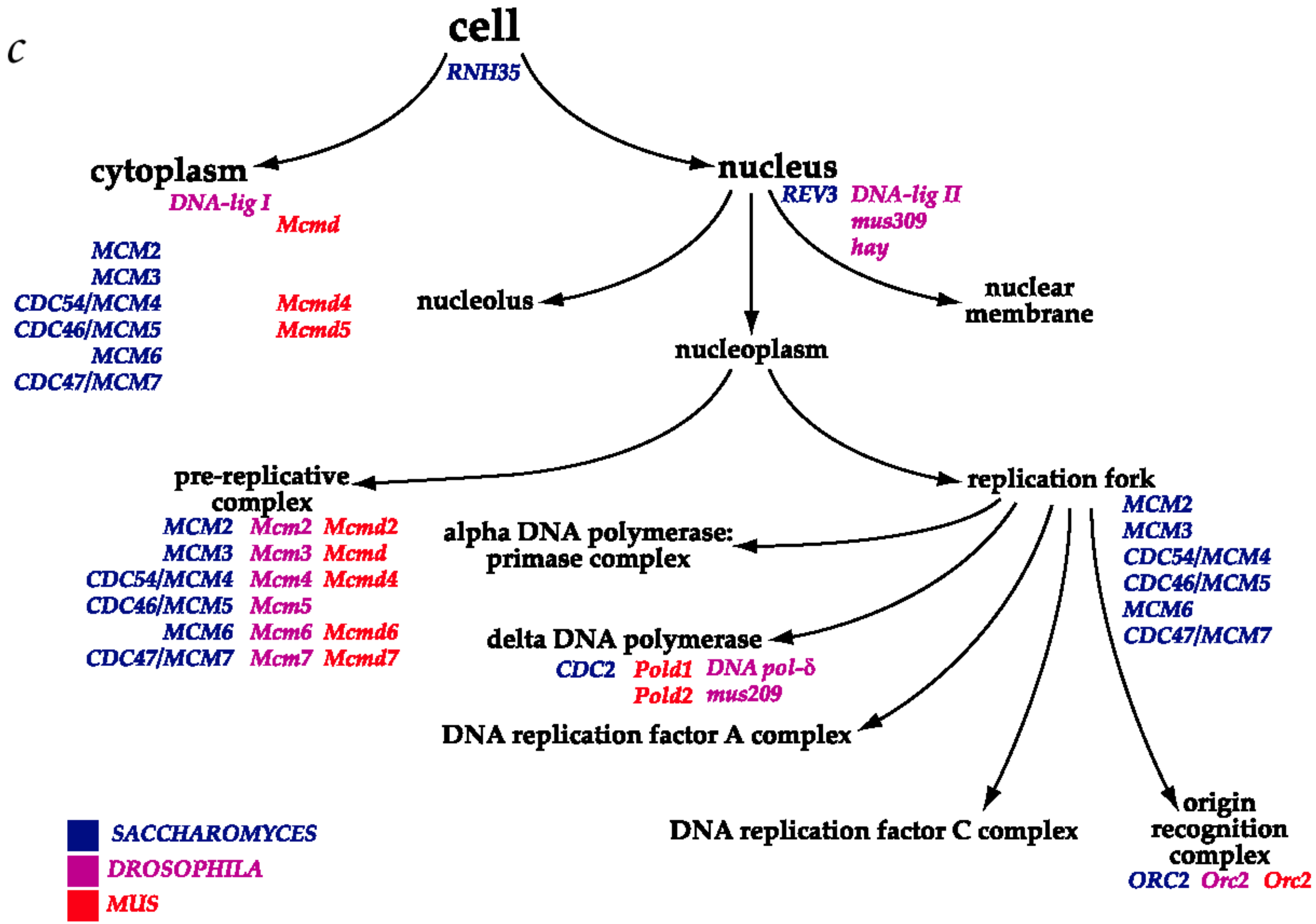
Gene ontology: exemples de processus biologiques



Gene ontology: exemples de fonctions moléculaires

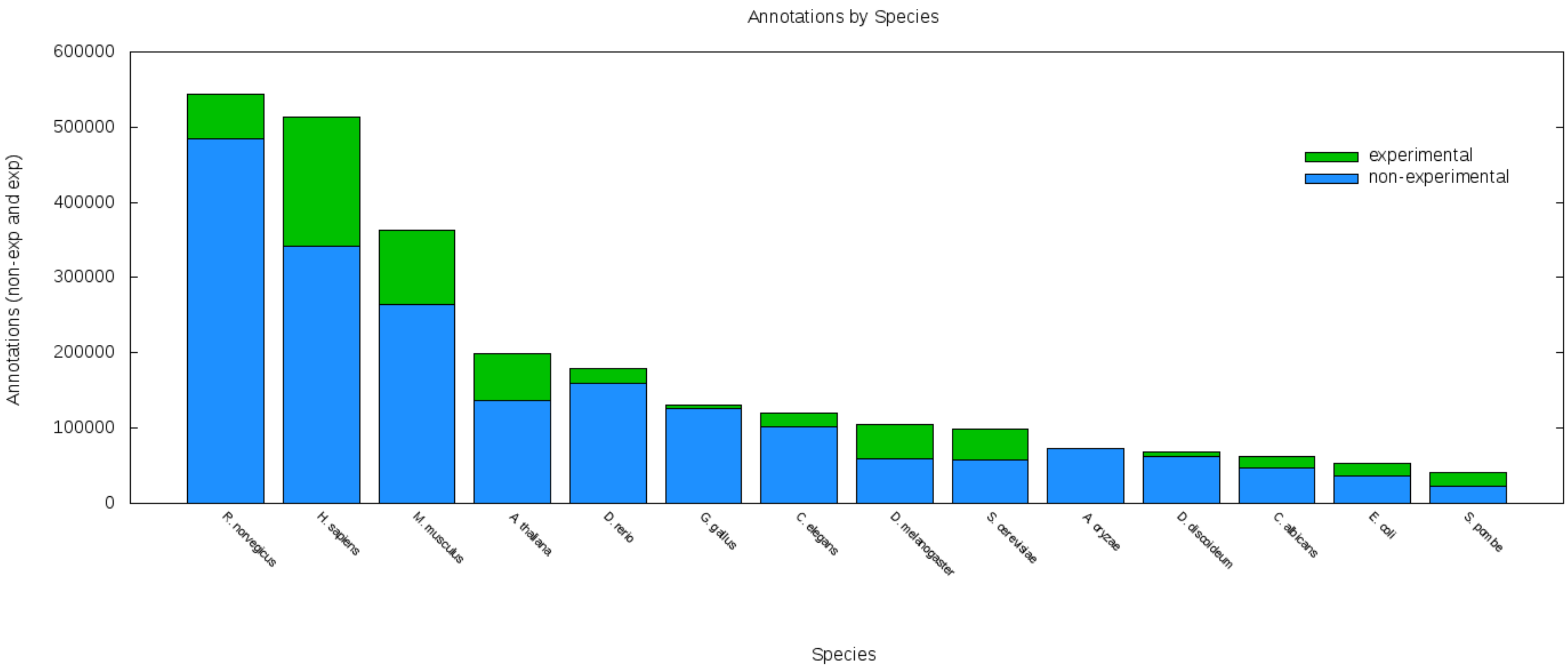


Gene ontology: exemples de composantes cellulaires



Gene ontology – some statistics

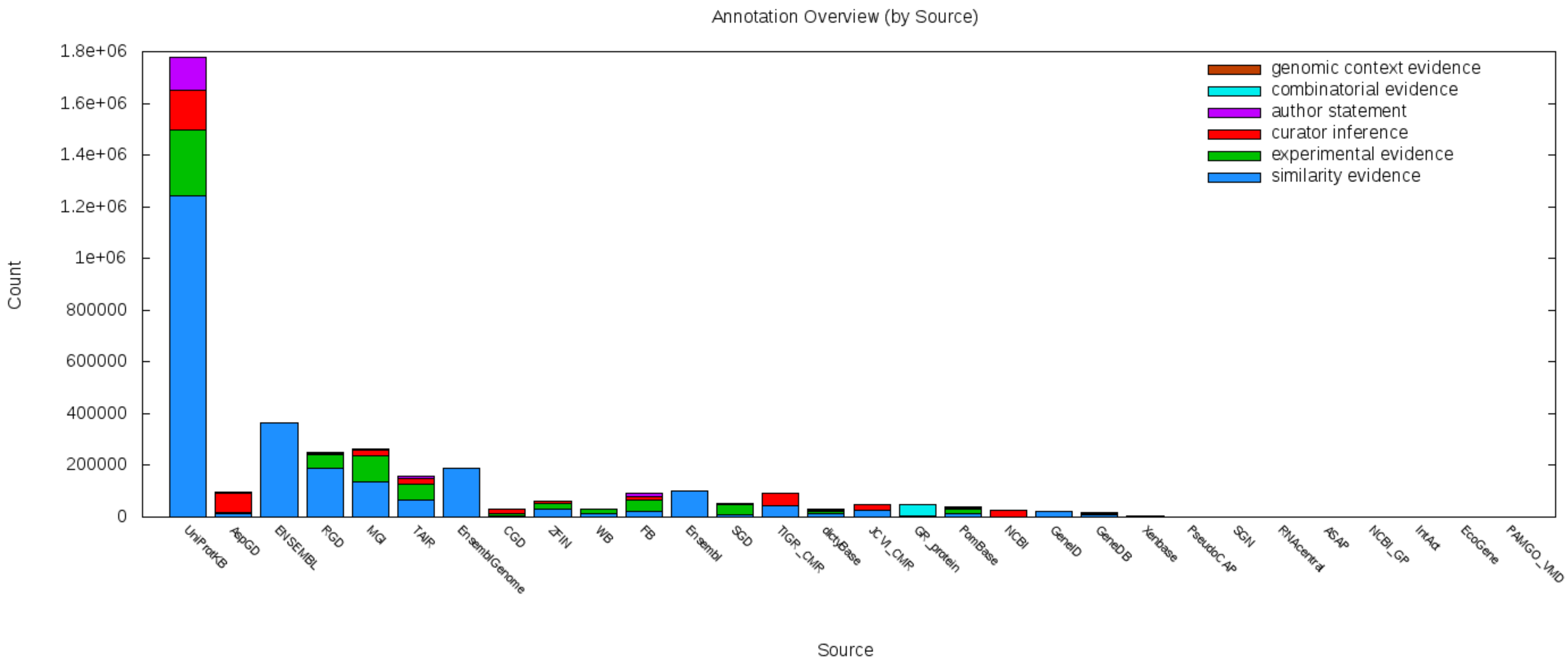
- Even for model organisms, the majority of annotations relies on "non-experimental" indications.



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

Gene ontology – some statistics

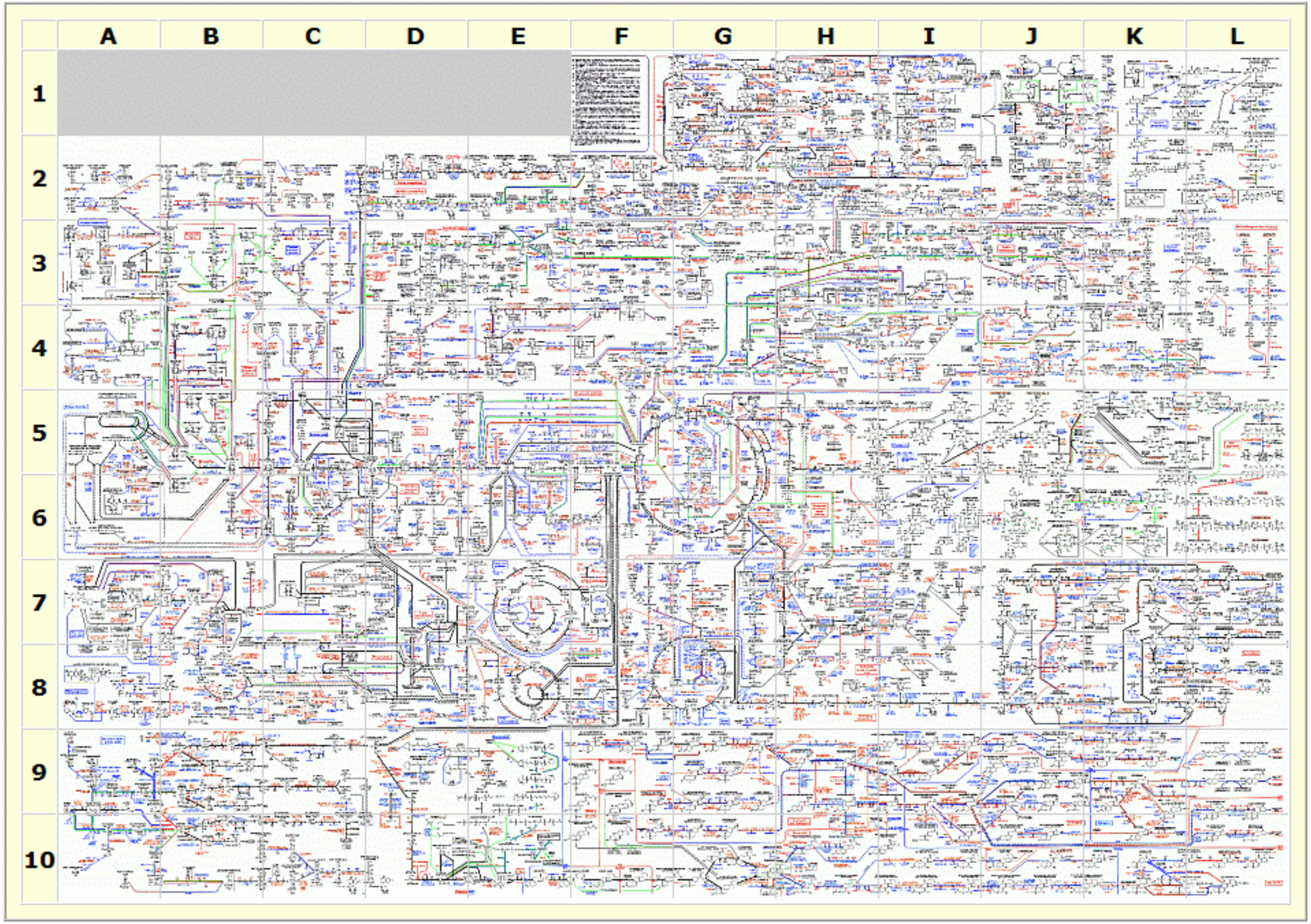
- Annotation by sequence similarity remains the main approach to annotate genes.



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

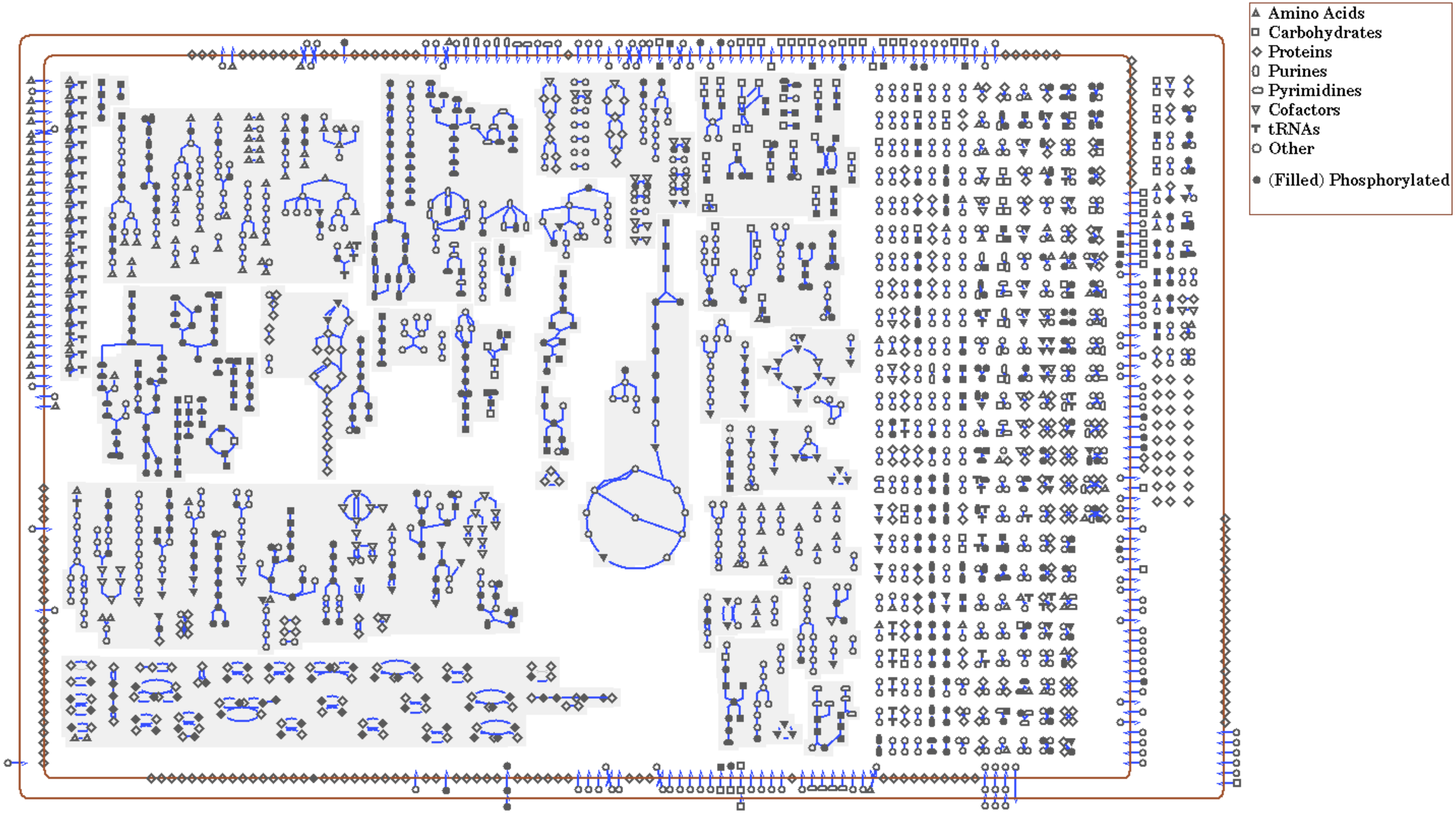
Annotation des voies métaboliques

Boerhinger-Mannheim Metabolic Wall Chart



http://www.expasy.ch/cgi-bin/show_thumbnails.pl

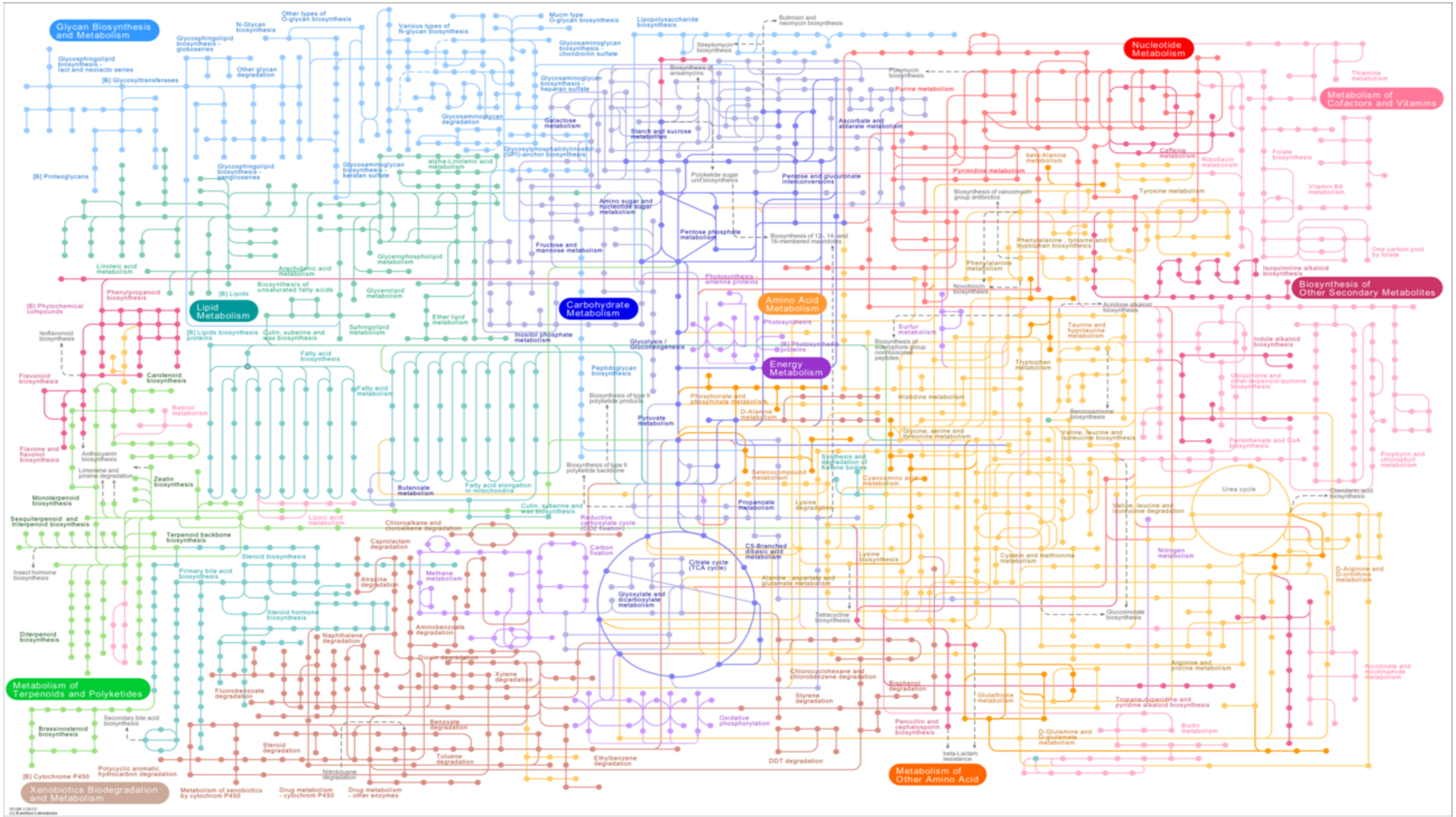
EcoCyc metabolic chart



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

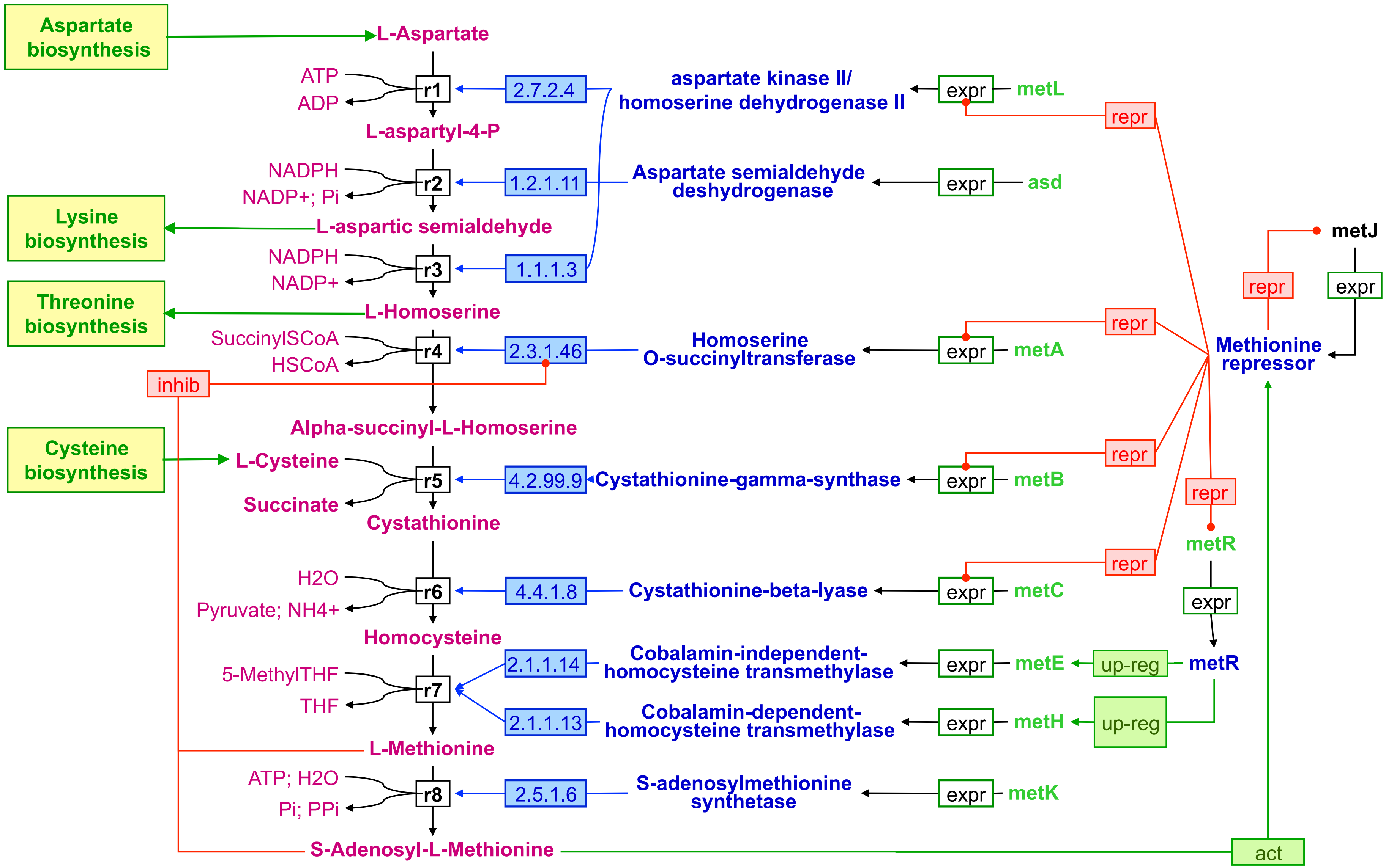
KEGG - Kyoto Encycplopaedia of Genes and Genomes

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



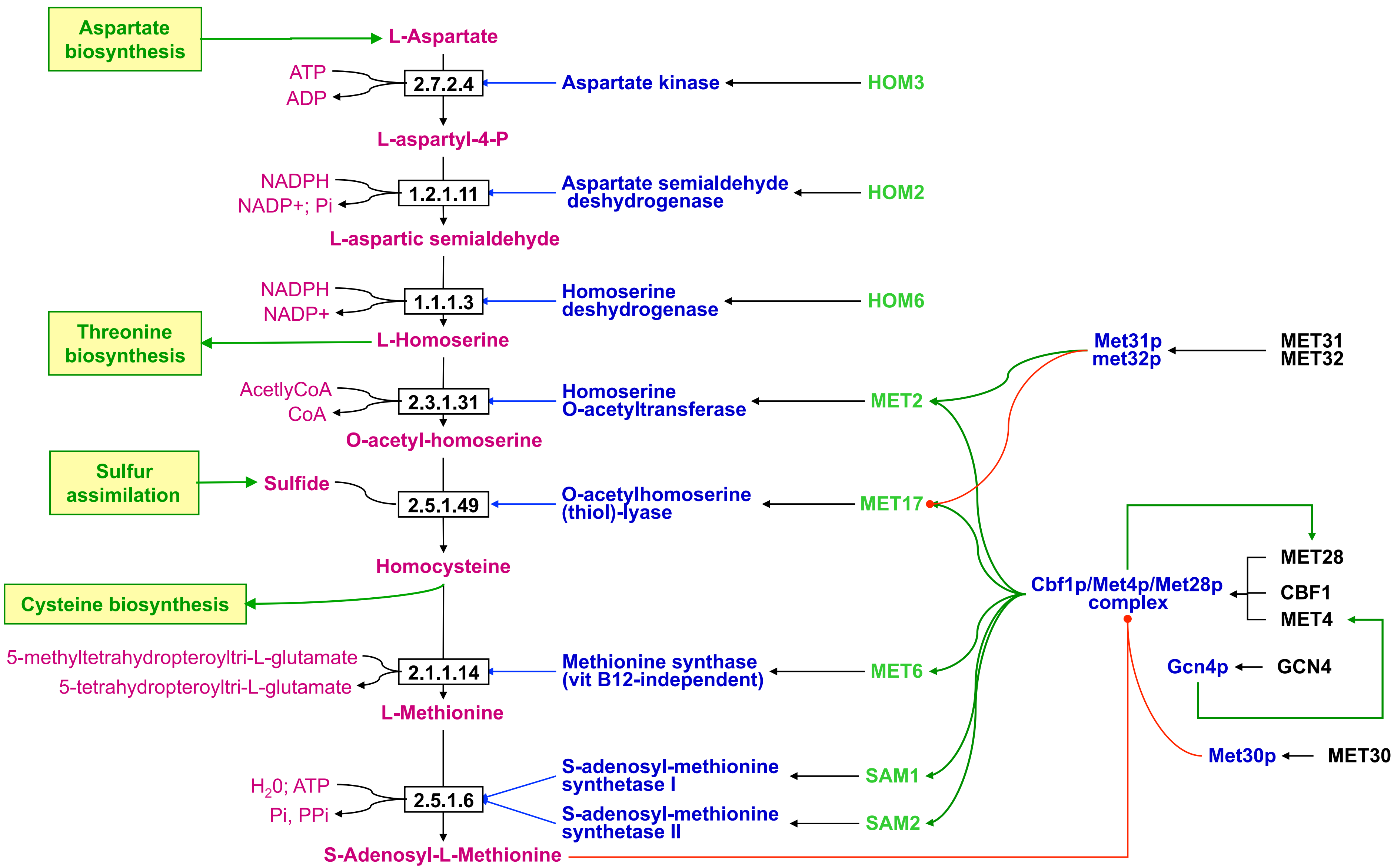
- KEGG global pathway map: http://www.genome.jp/kegg-bin/show_pathway?map01100

Methionine Biosynthesis in E.coli



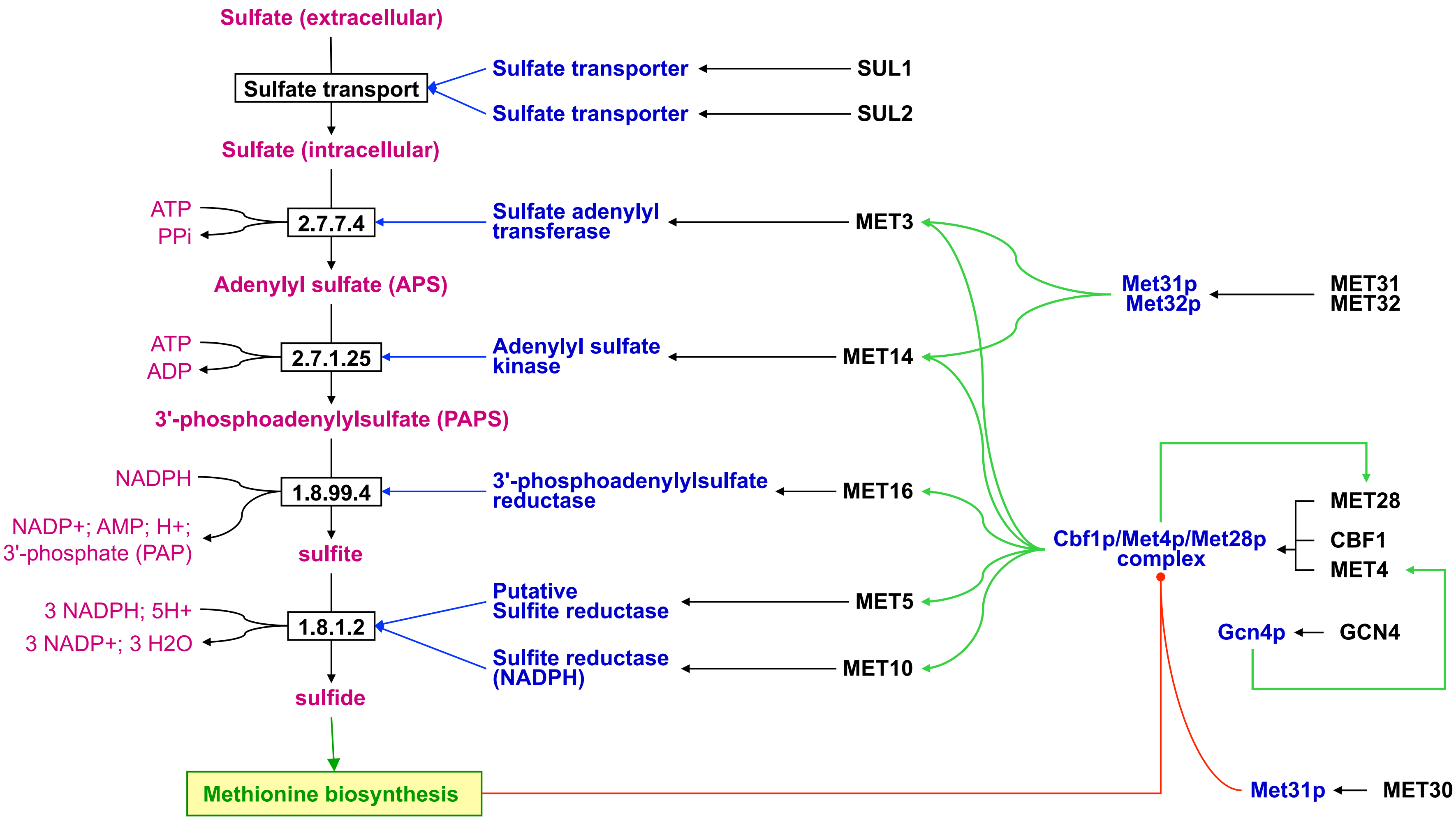
- Data provided by Georges Cohen, 1998

Methionine Biosynthesis in S.cerevisiae



- Data provided by Georges Cohen, 1998

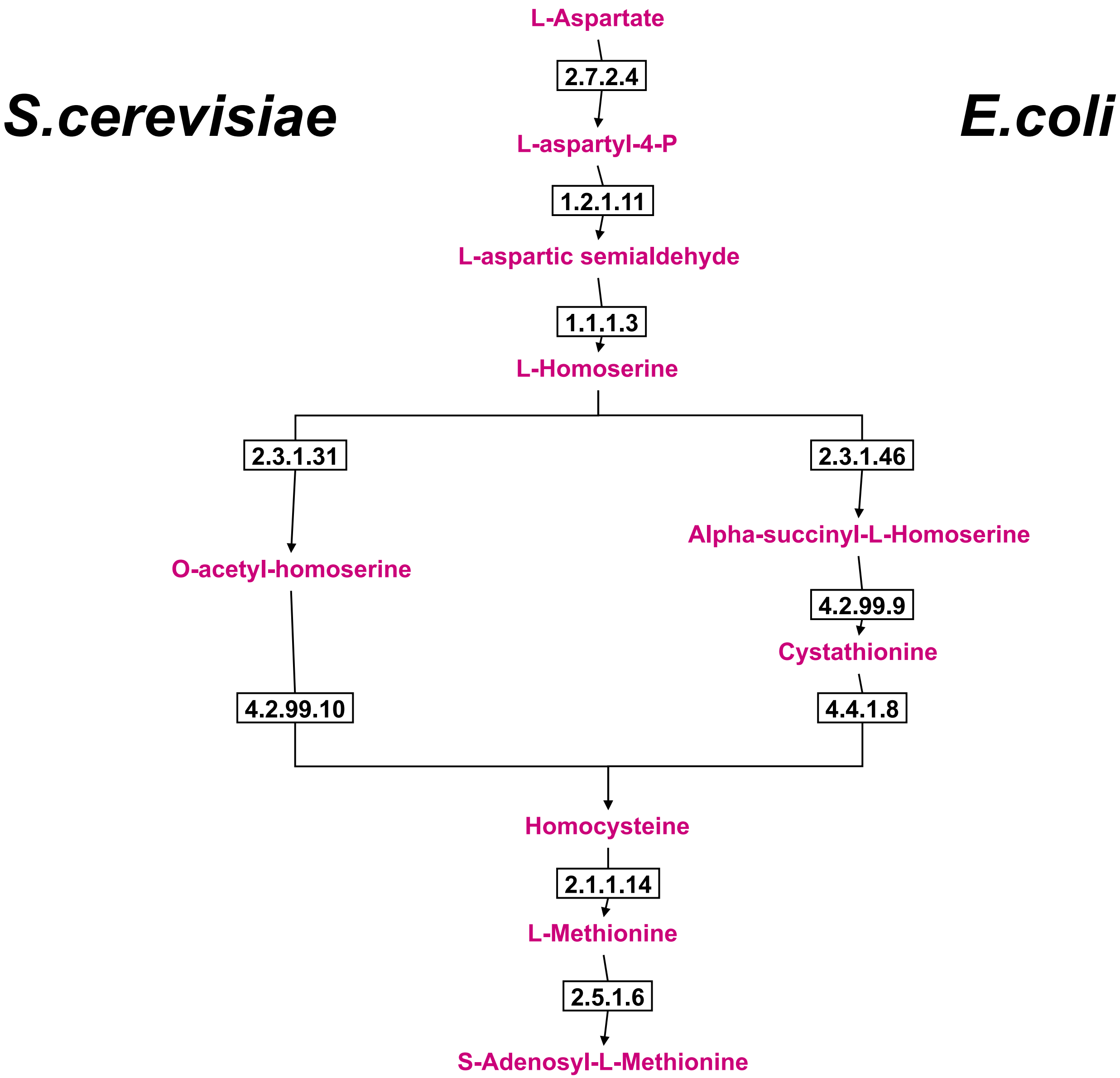
Sulfur Assimilation in yeast



- Data provided by Georges Cohen, 1998

Alternative methionine pathways

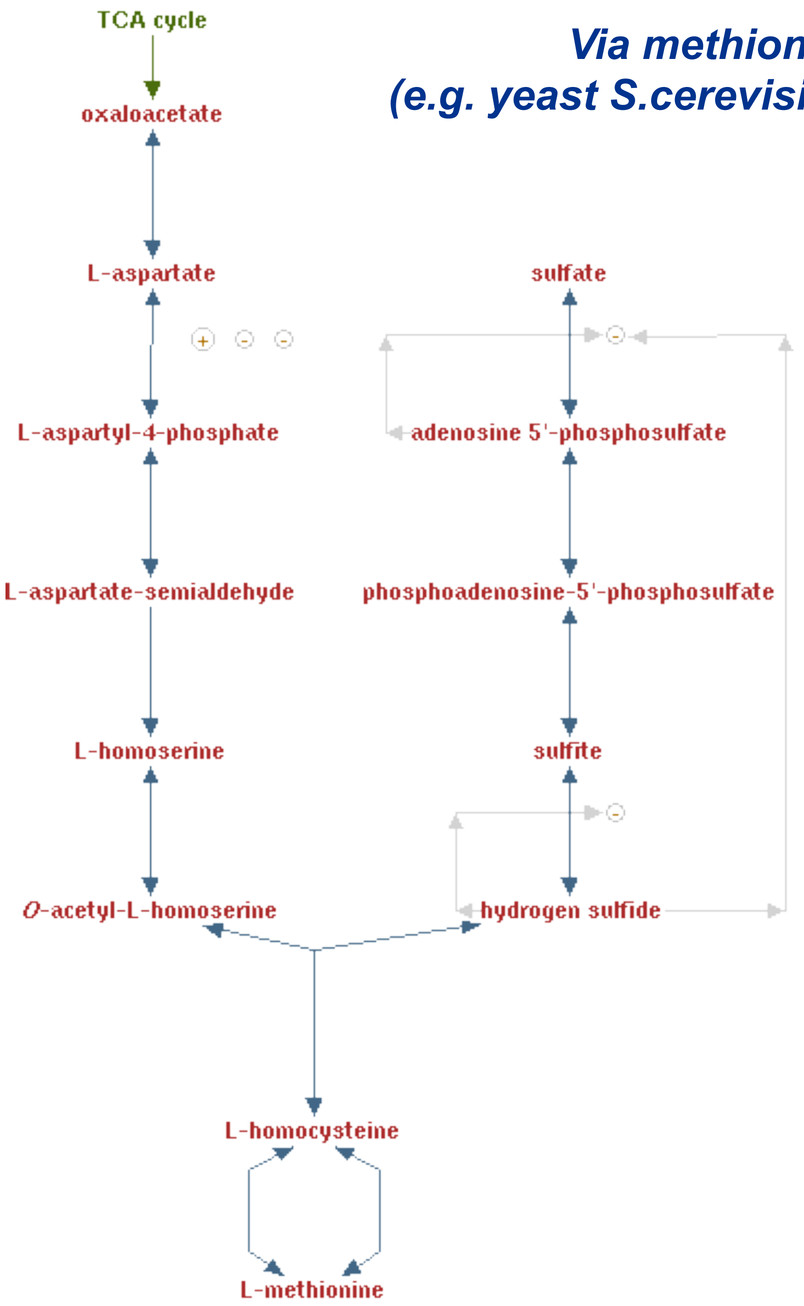
- In yeast, Sulfur is incorporated in amino acids in the methionine biosynthetic pathway, and then transferred from methionine to cysteine.
- In E.coli, sulfur is incorporated in amino acids through the cysteine biosynthetic pathway, and then transferred to methionine.



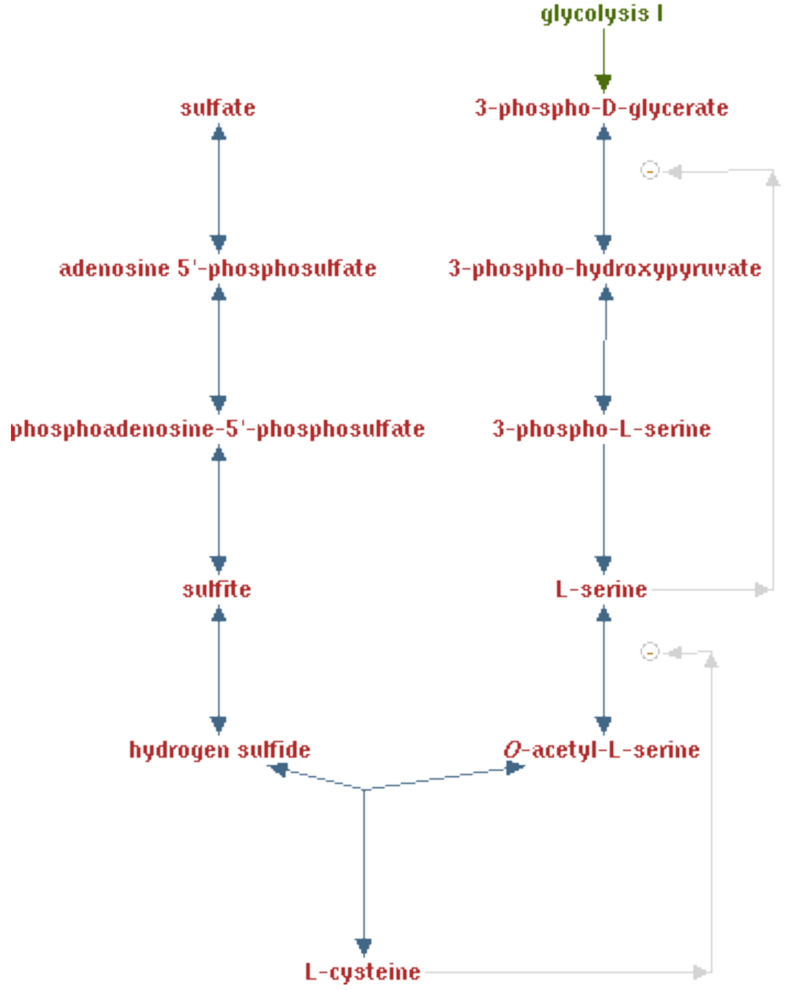
- Data provided by Georges Cohen, 1998

MetaCyc – Sulfur incoroporation in amino-acids

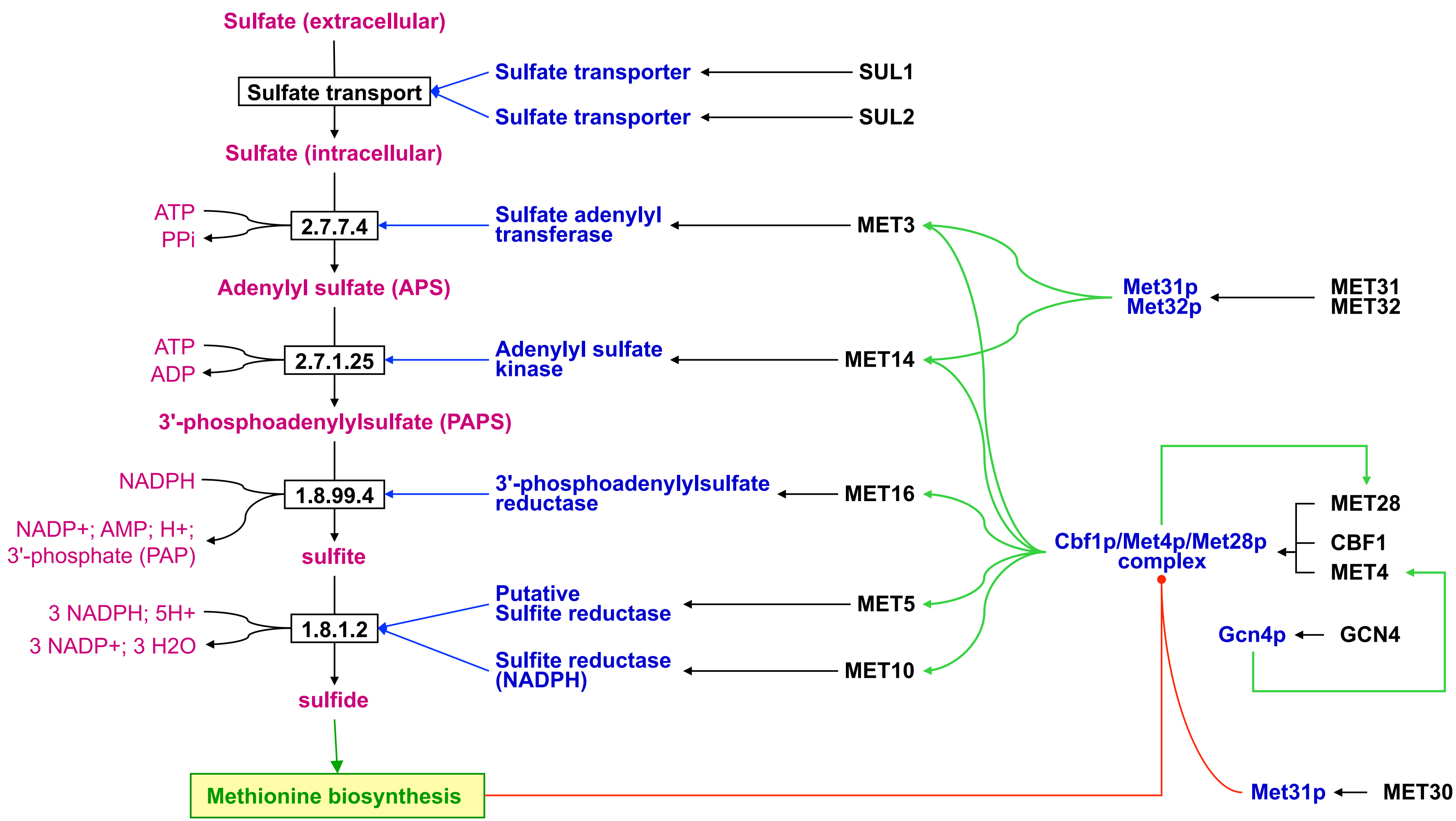
Via methionine (e.g. yeast S.cerevisiae)



Via cysteine (e.g. Bacteria E.coli)

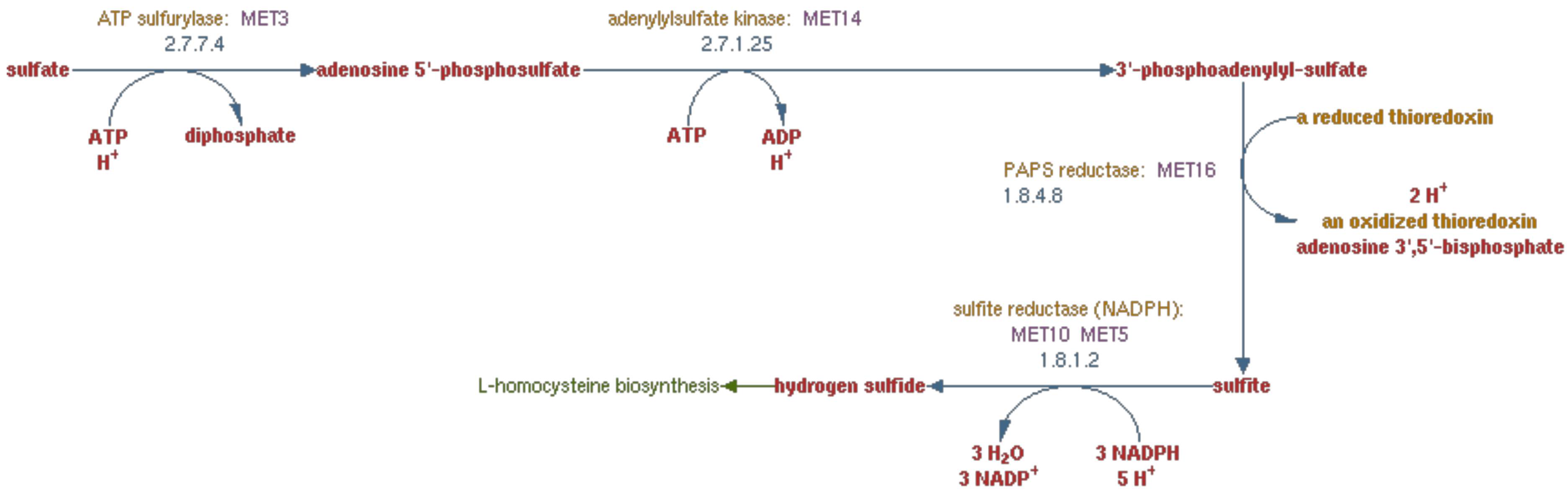


Sulfate reduction in yeast



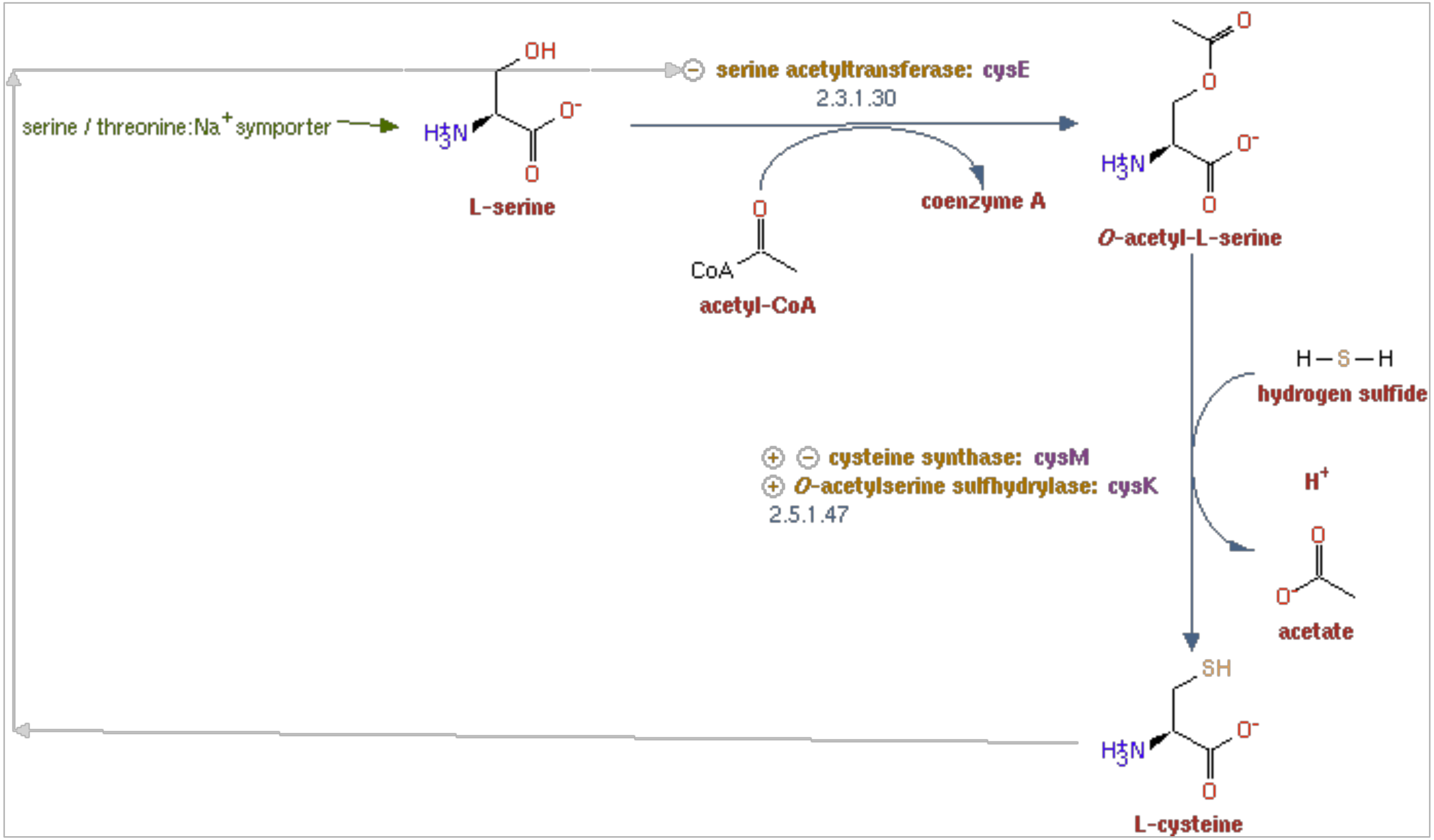
- Data provided by Georges Cohen, 1998

MetaCyc Saccharomyces cerevisiae Pathway: sulfate reduction I (assimilatory)



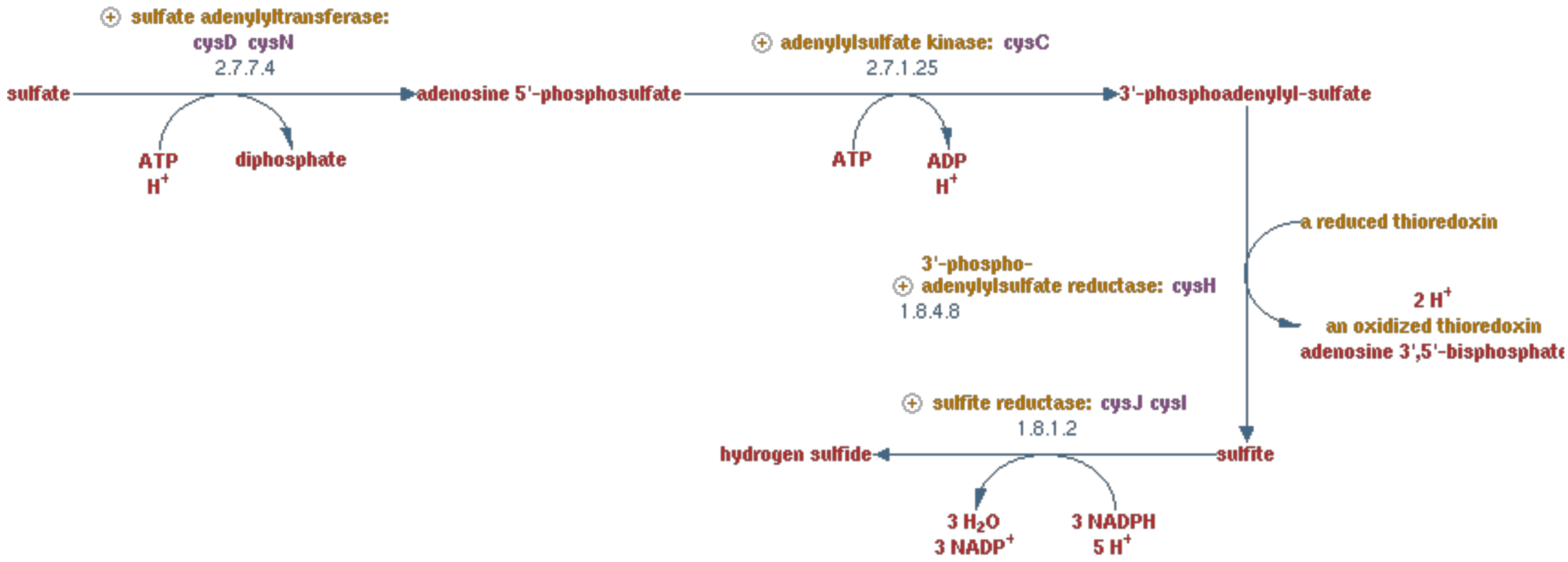
- http://biocyc.org/YEAST/NEW-IMAGE?type=PATHWAY&object=SO4ASSIM-PWY&detail-level=2

EcoCyc – Cysteine biosynthesis I



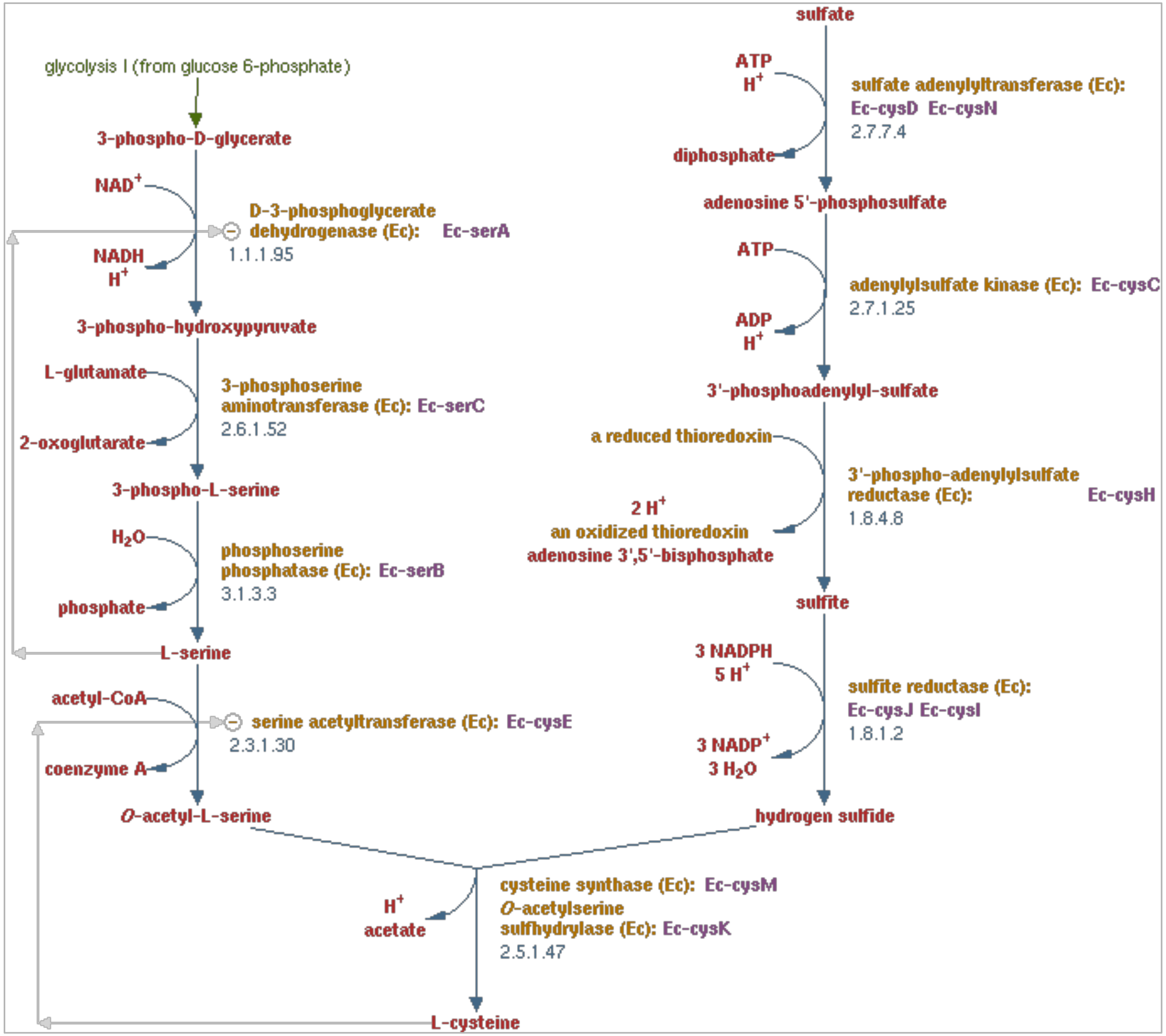
- http://ecocyc.org/ECOLI/NEW-IMAGE?type=PATHWAY&object=CYSTSYN-PWY&detail-level=4

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



- http://biocyc.org/ECOLI/NEW-IMAGE?type=PATHWAY&object=SO4ASSIM-PWY

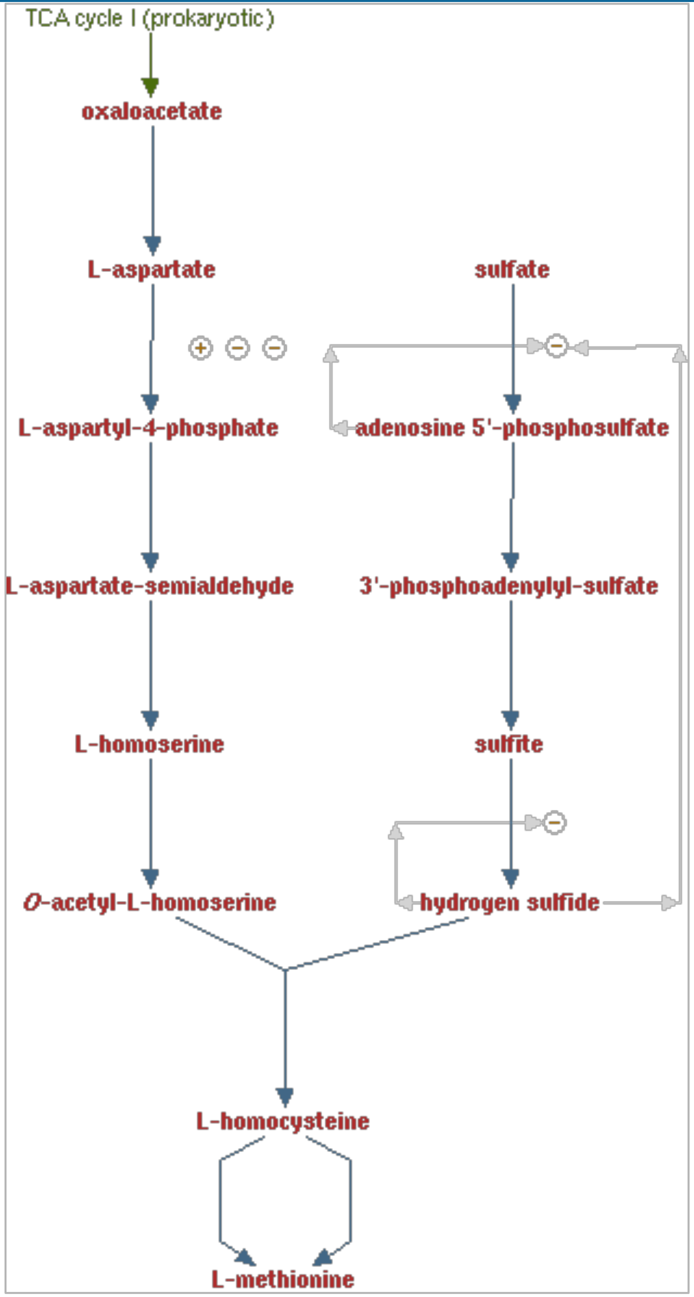
EcoCyc - Superpathway of sulfate assimilation and cysteine biosynthesis



- http://biocyc.com/META/NEW-IMAGE?type=PATHWAY&object=SULFATE-CYS-PWY&detail-level=2

MetaCyc - Superpathway of methionine biosynthesis by sulfhydrylation

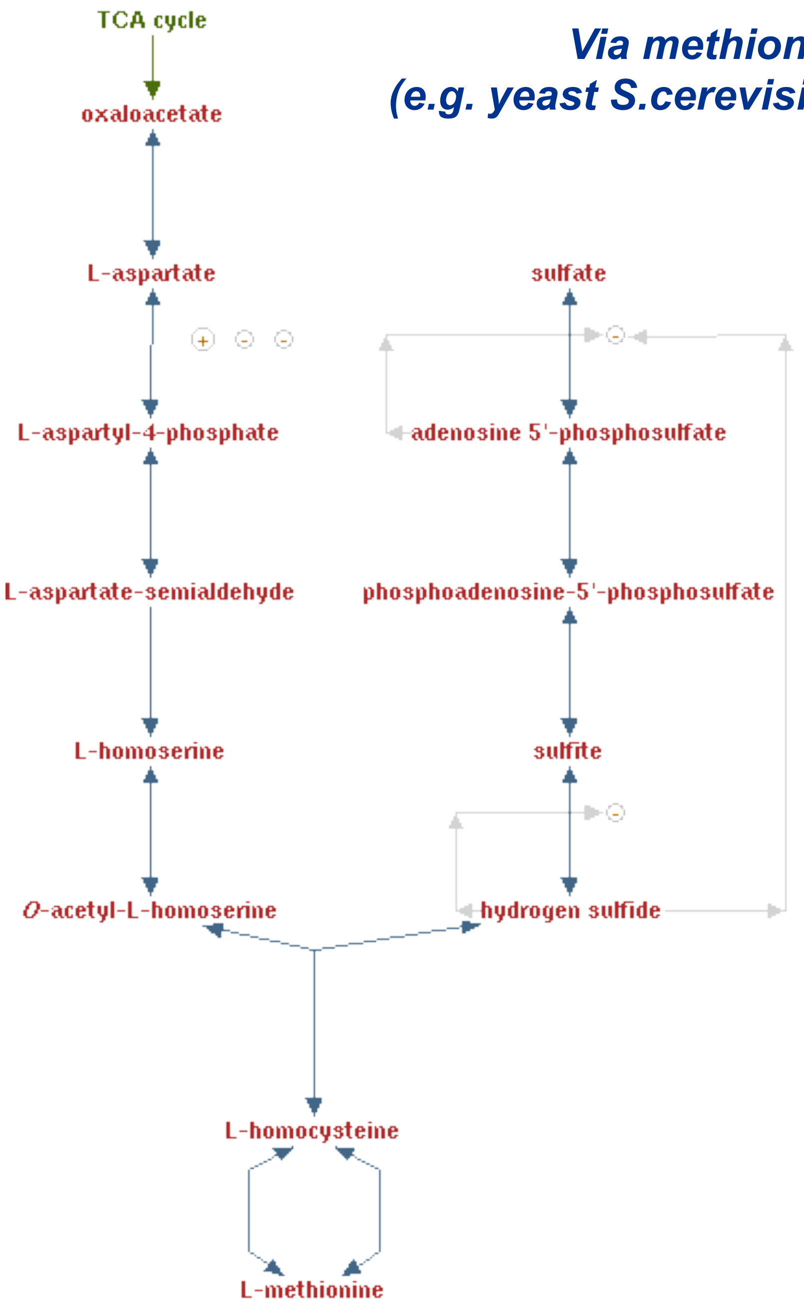
- Chez les levures (notamment), l'incorporation du soufre fait partie de la voie de biosynthèse de la méthionine. Le soufre est ensuite transféré de la méthionine à la cystéine.



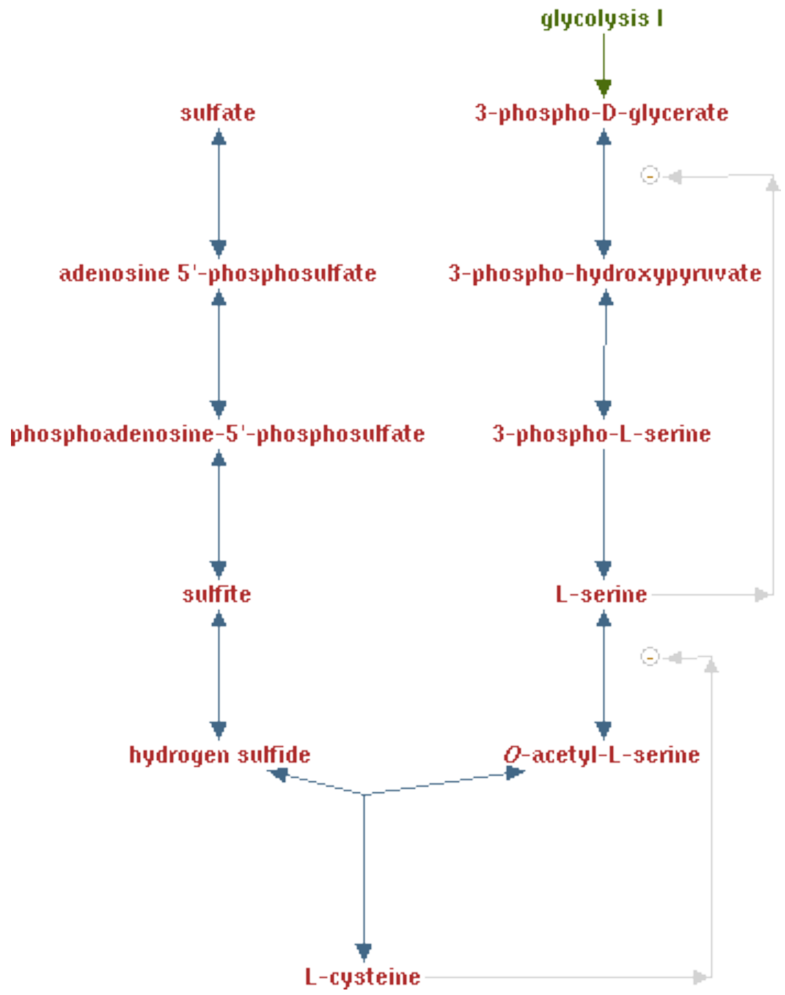
- http://biocyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-5345&detail-level=1

MetaCyc – Sulfur incoroporation in amino-acids

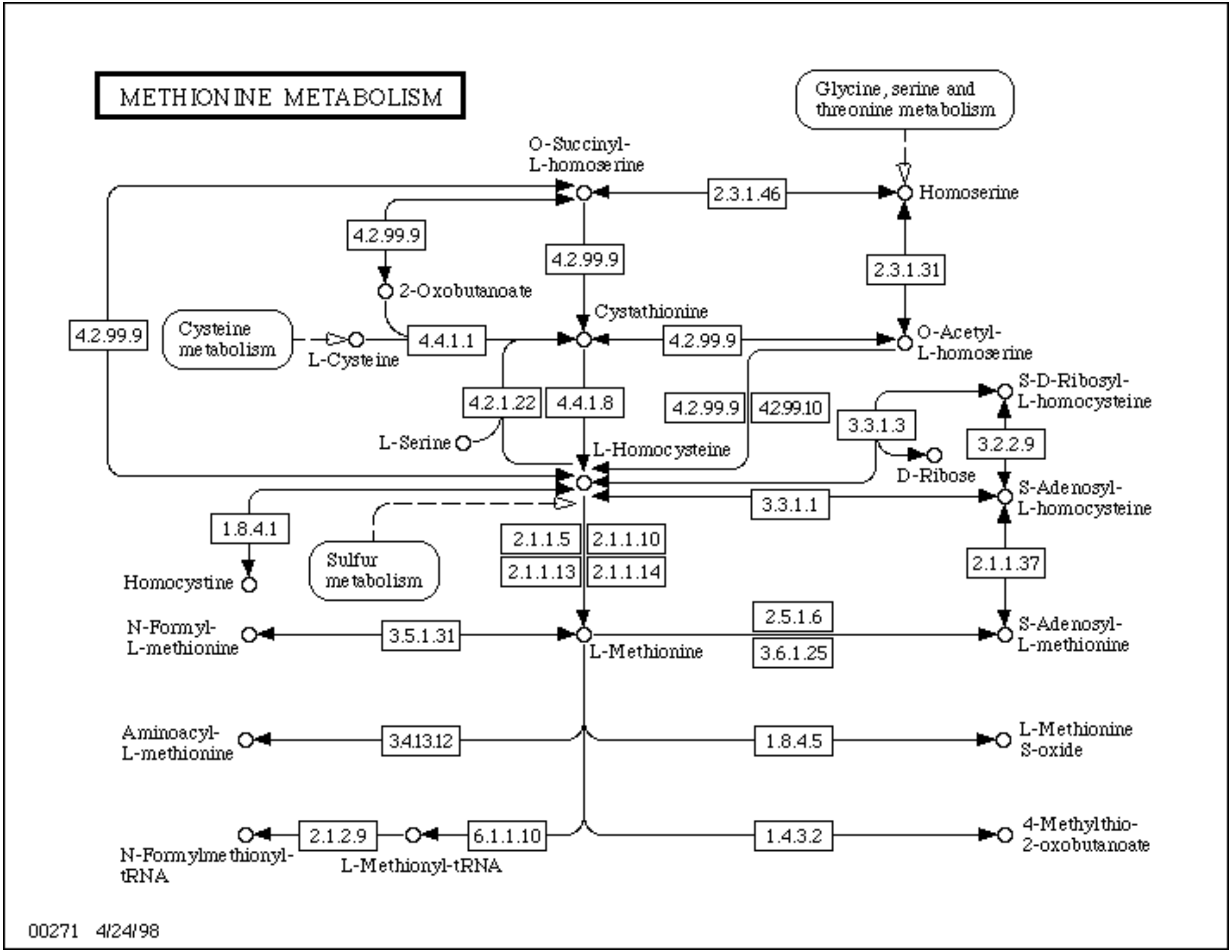
Via methionine (e.g. yeast S.cerevisiae)



Via cysteine (e.g. Bacteria E.coli)

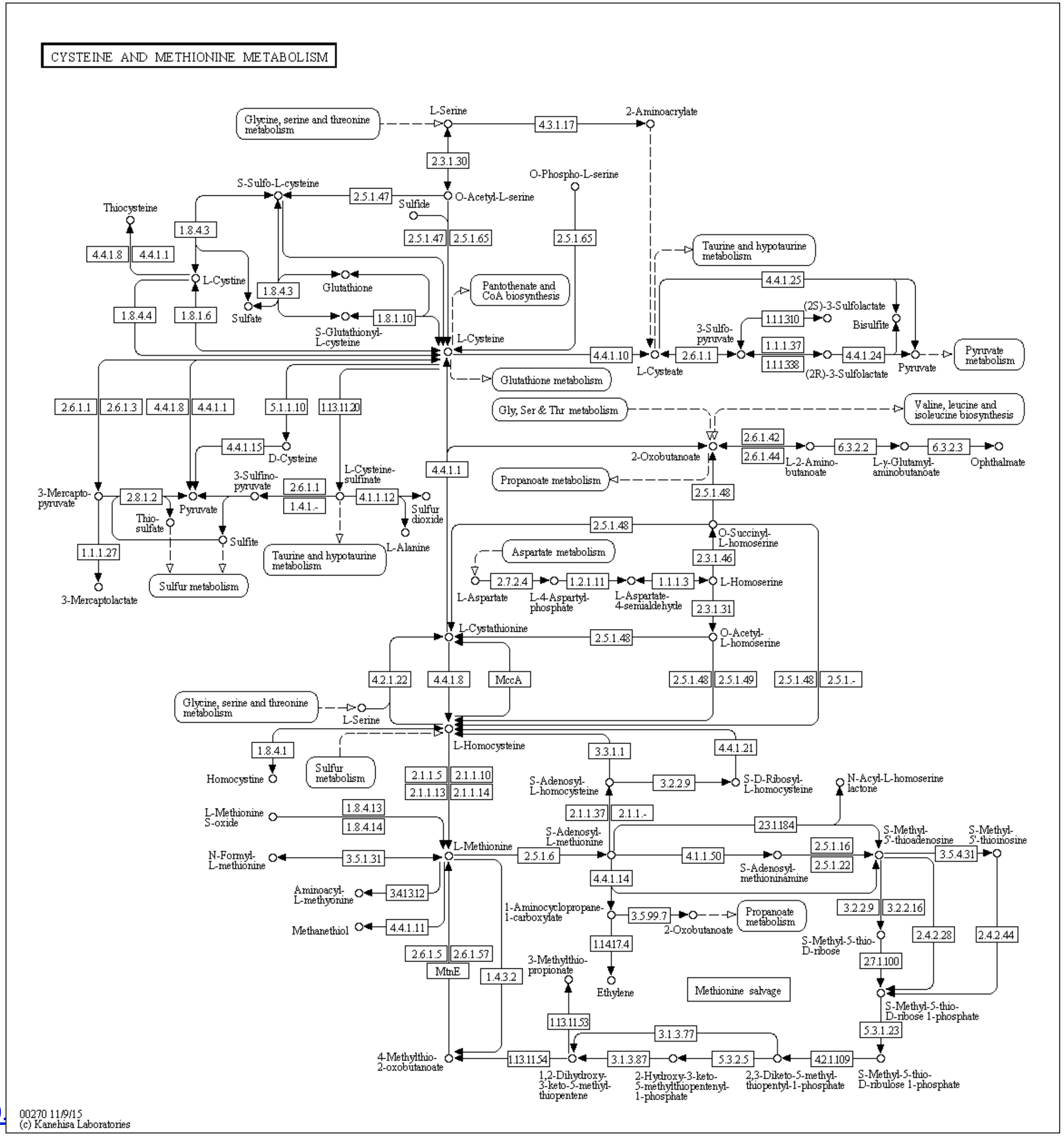


KEGG “reference” pathway - Methionine metabolism (1998)



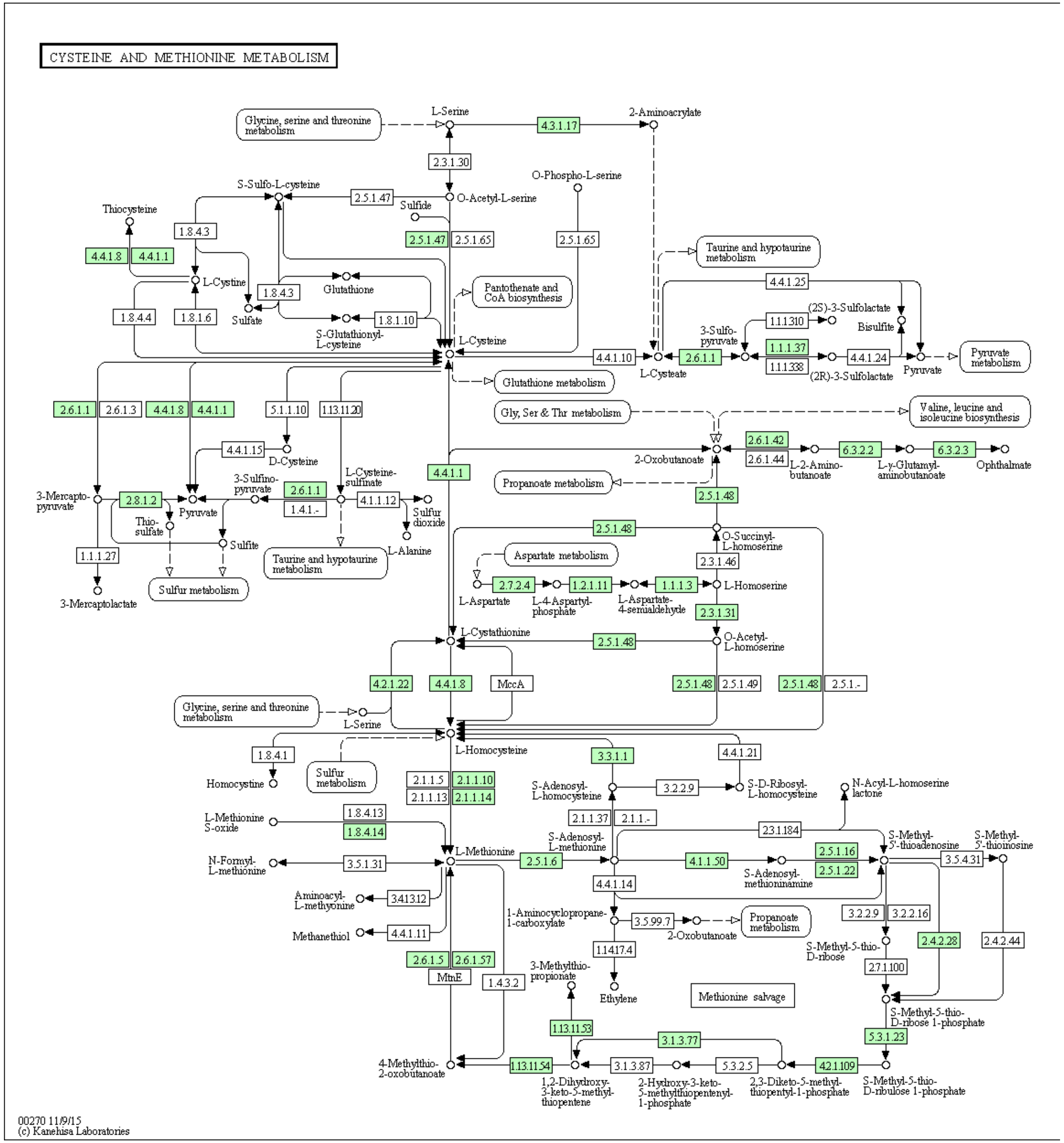
KEGG “reference” map - Cysteine and methionine metabolism (Jan 2016)

- 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 ?



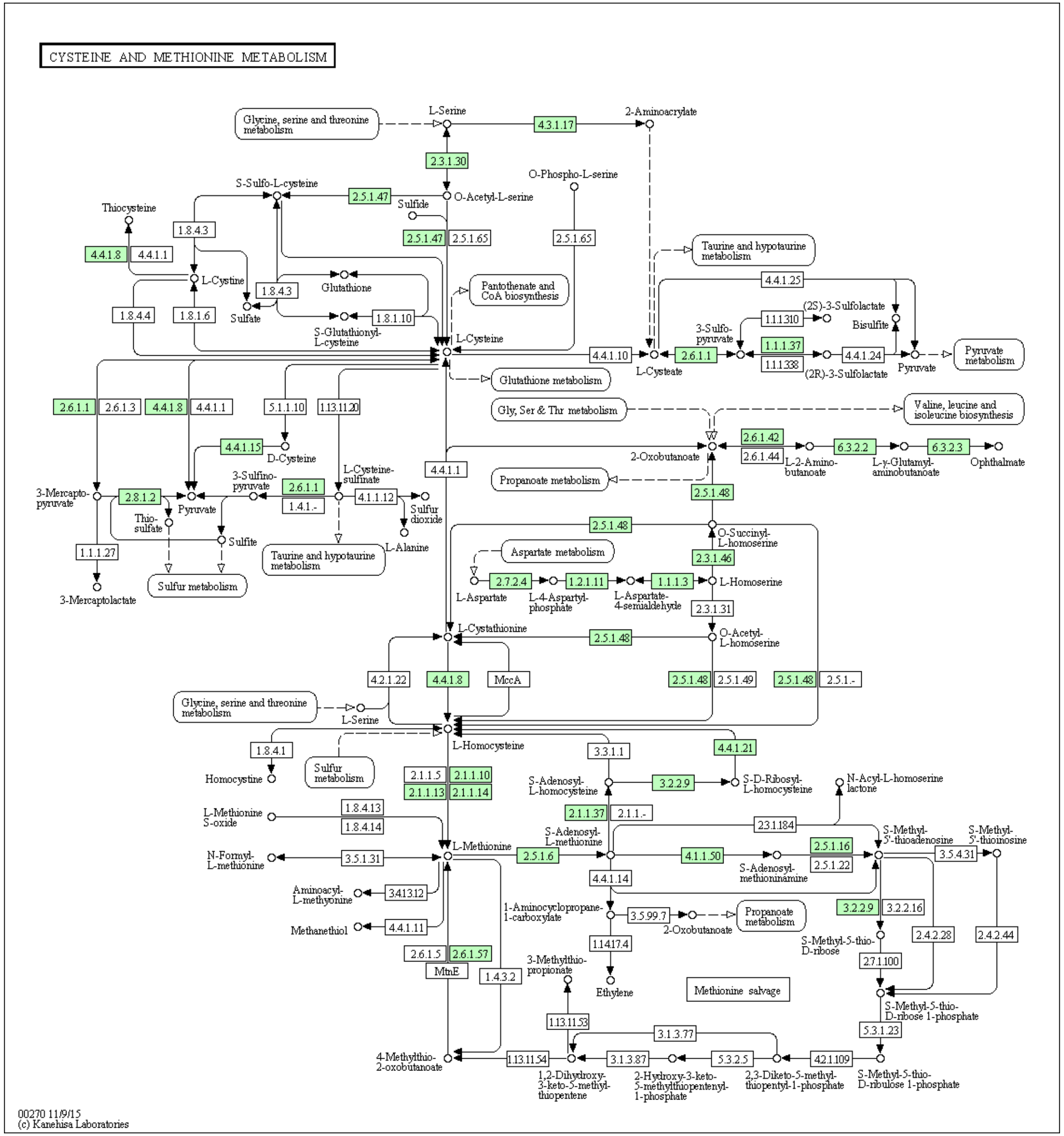
- http://www.genome.jp/kegg/pathway/map/map00270.

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



- http://www.genome.jp/kegg-bin/show_pathway?org_name=sce&mapno=00270

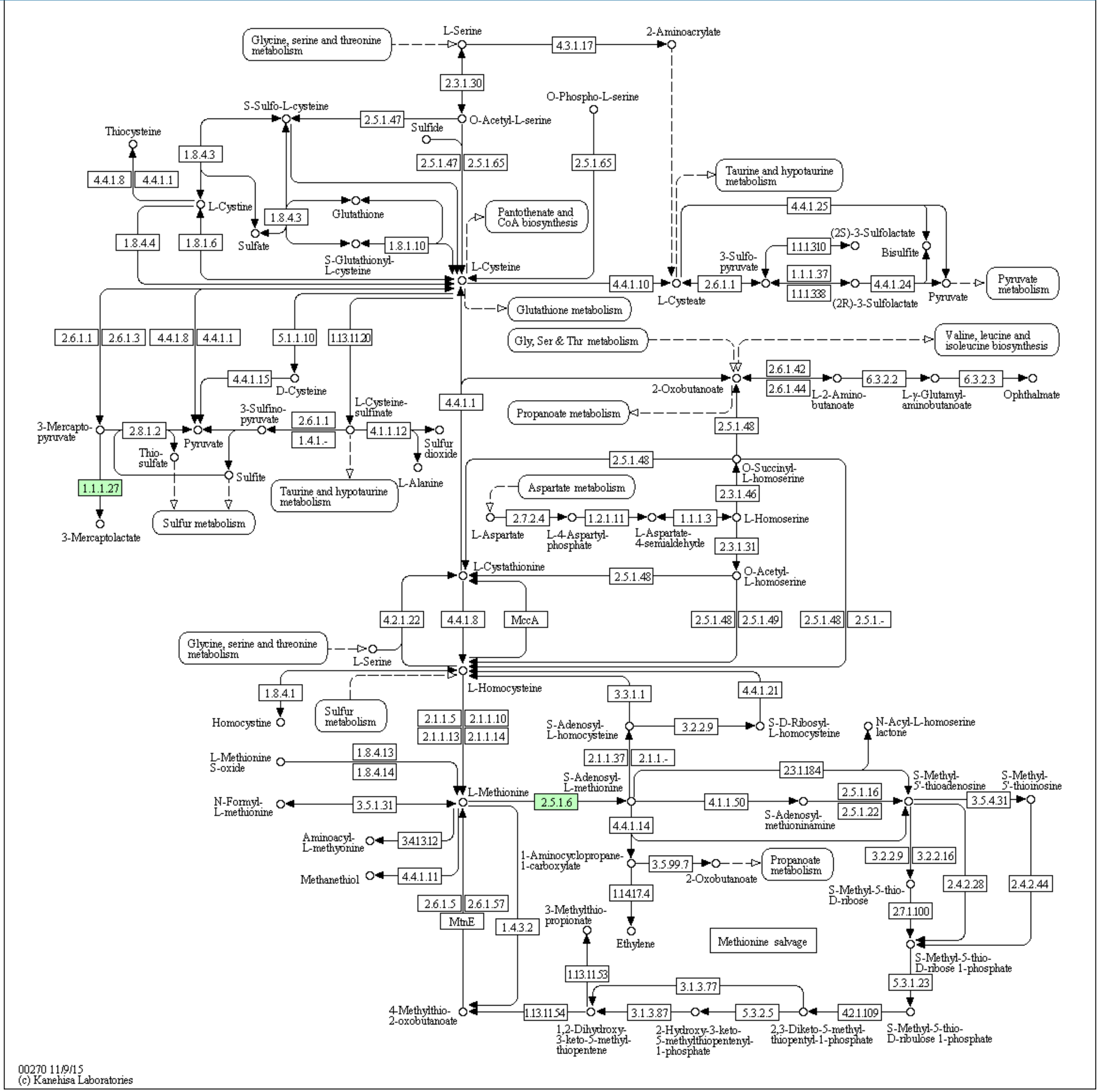
- KEGG cysteine and methionine pathway.
- *Escherichia coli K12.*
- Question
  - How is sulfur incorporated into amino acids in this in this Bacteria?



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.



- http://www.genome.jp/kegg-bin/show_pathway?org_name=mge&mapno=00270

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.

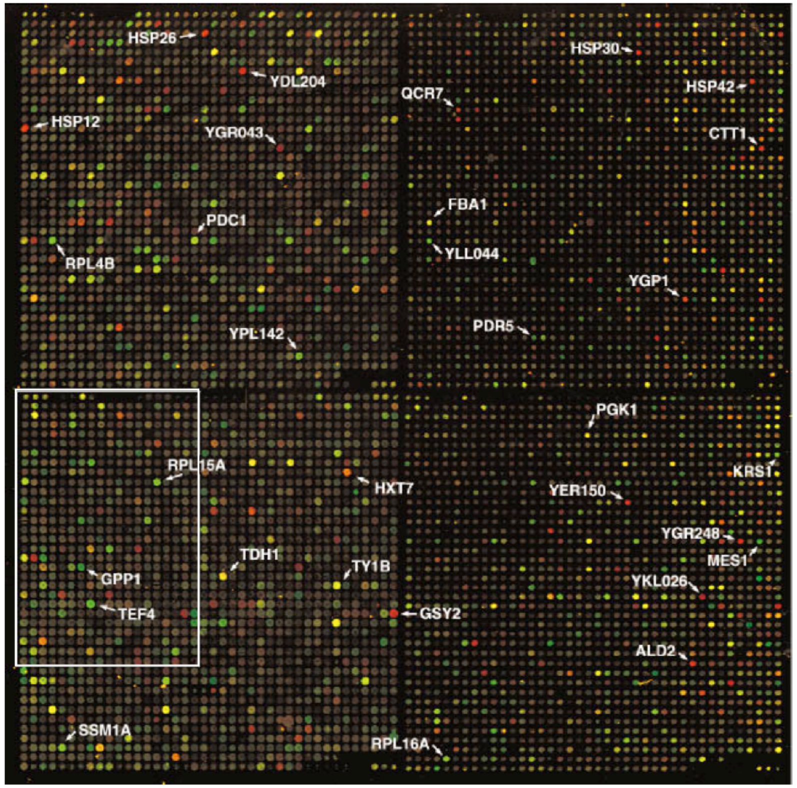
Transcriptome et métabolisme

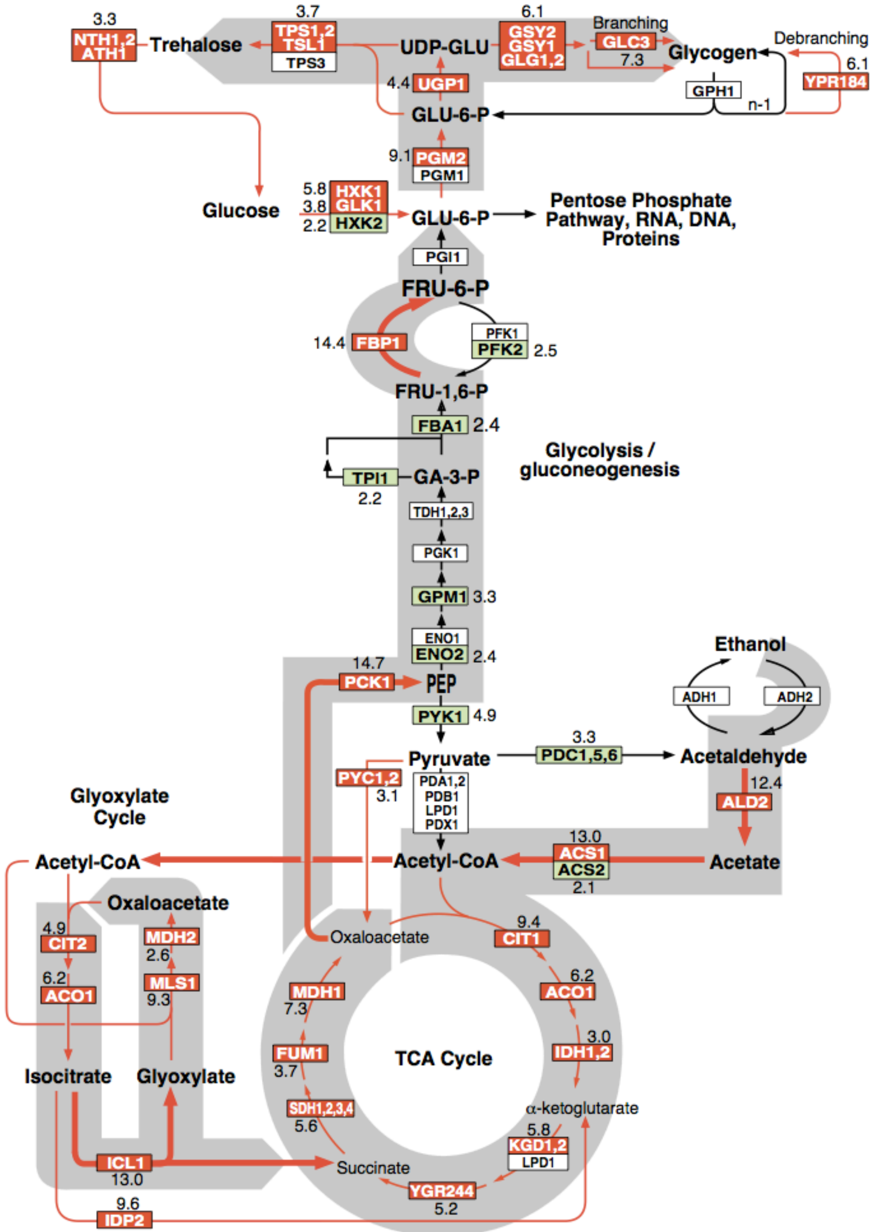
The diauxic shift revisited by DeRisi, Iyer and Brown (1997)

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 genomewide 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 of 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.

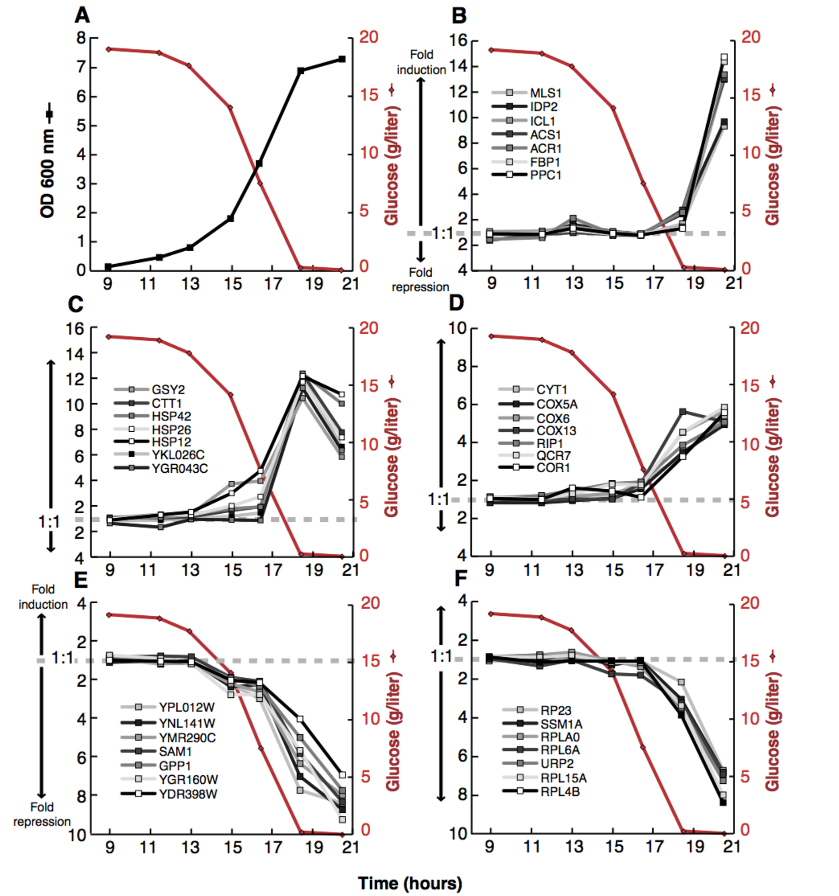


Fig. 5. Distinct temporal patterns of induction or repression help to group genes that share regulatory properties. (**A**) Temporal profile of the cell density, as measured by OD at 600 nm and glucose concentration in the media. (**B**) Seven genes exhibited a strong induction (greater than ninefold) 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.

DeRisi,J., Iyer,V. and Brown,P. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.