Comparative genomics, data, concepts and perspectives

Jacques van Helden

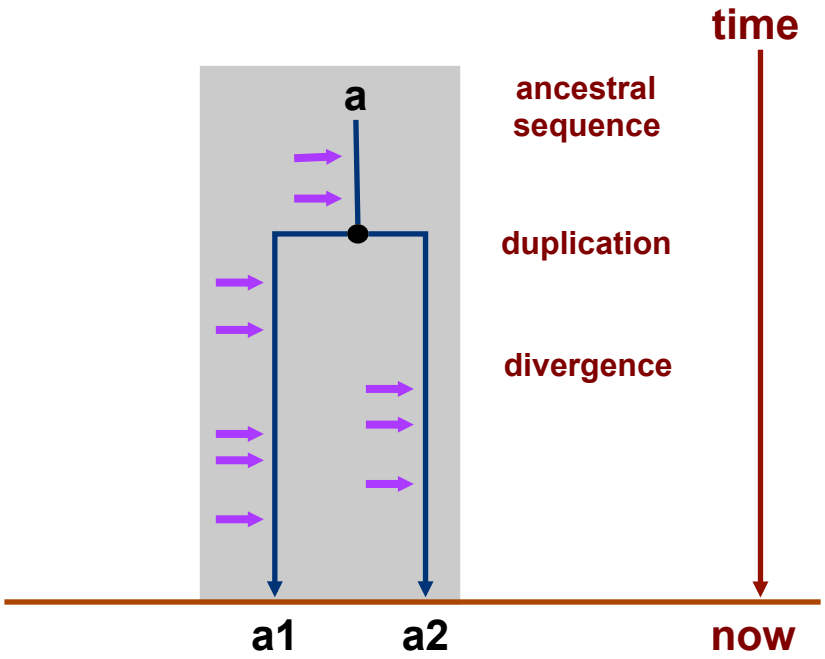
Jacques.van-Helden@univ-amu.fr
Aix-Marseille Université, France
Technological Advances for Genomics and Clinics
(TAGC, INSERM Unit U1090)
http://jacques.van-helden.perso.luminy.univmed.fr/

Ana, homo, ortho, para and other logies

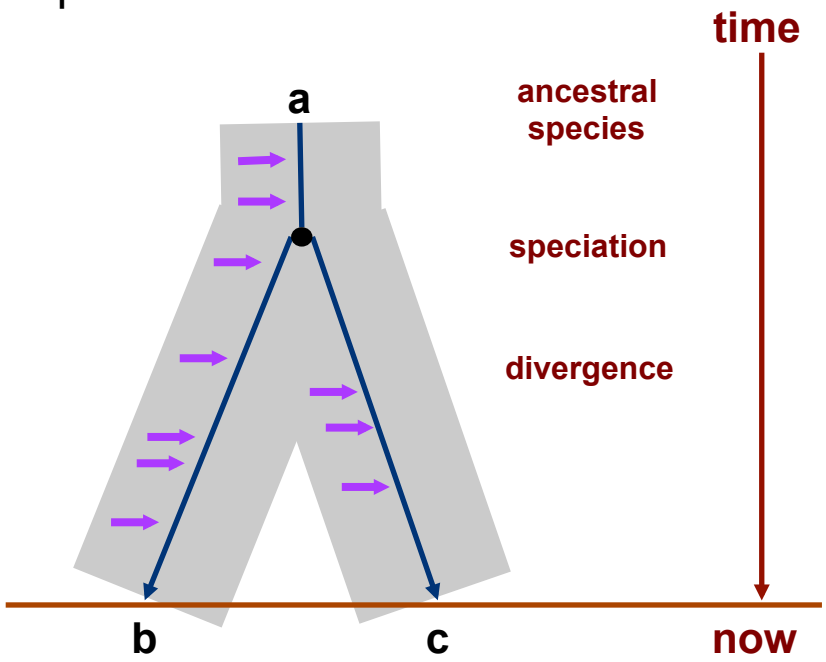
Evolutionary scenarios

- The shaded tree represents the history of the species, the thin black tree the history of the sequences.
- We dispose of similar sequences, and we assume that they diverge from some common ancestor (either by duplication, or by speciation).
- ***Mutational events*** occur during their evolution: substitutions, deletions, insertions.

Gene duplication



Speciation



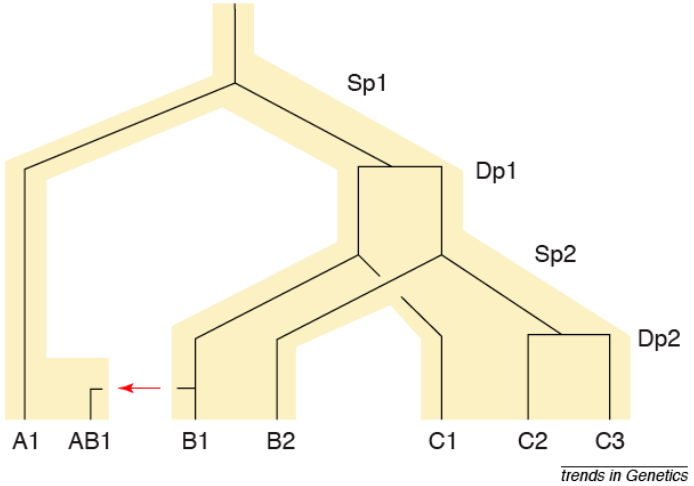
Similarity and homology

- The similarity between two sequences can be interpreted in two alternative ways :
  - ***Homology***: the two sequences diverged from a ***common ancestor***.
  - ***Convergent evolution***: the similar residues appeared ***independently*** in the two sequences, possibly under some selective pressure.
- Inference
  - In order to claim that two sequences are homologous, we should be able to trace their history back to their common ancestor.
  - Since we cannot access the sequence of all the ancestors of two sequences, this is not feasible.
  - The claim that two sequences are homolog thus results from an inference, based on some evolutionary scenario (rate of mutation, level of similarity, …).
  - The inference of homology is always attached to some ***risk of false positive***. Evolutionary models allow to estimate this risk, as we shall see.

- Homology is a Boolean relationship (***true or false*****):** two sequences are homolog, or they are not.
  - It is thus **incorrect** to speak about “percent of homology”.
  - The correct formulation is that we can infer (with a measurable risk of error) that two sequences are homolog, because they share some percentage of identity or similarity.

Concept definitions from Fitch (2000)

- Discussion about definitions of the paper
  - Fitch, W. M. (2000). Homology a personal view on some of the problems. Trends Genet 16, 227-31.
- **Homology**
  - Owen (1843). « the same organ under every variety of form and function ».
  - Fitch (2000). Homology is the relationship of any two characters that have descended, usually with divergence, from a common ancestral character.
    - Note: “character” can be a phenotypic trait, or a site at a given position of a protein, or a whole gene, ...
  - Molecular application: two genes are homologous if diverge from a common ancestral gene.
- **Analogy:** relationship of two characters that have developed convergently from unrelated ancestor.
- **Cenancestor**: the most recent common ancestor of the taxa under consideration
- **Orthology:** relationship of any two homologous characters whose common ancestor lies in the cenancestor of the taxa from which the two sequences were obtained.
- **Paralogy:** Relationship of two characters arising from a duplication of the gene for that character.
- **Xenology:** relationship of any two characters whose history, since their common ancestor, involves interspecies (horizontal) transfer of the genetic material for at least one of those characters.

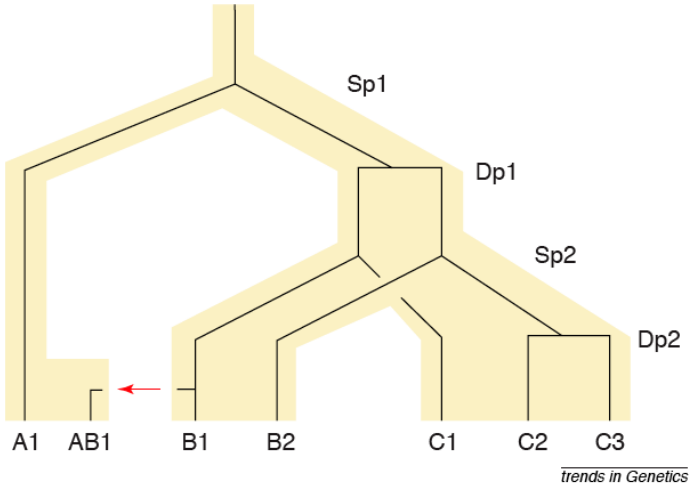


The idealized evolution of a gene (lines) is shown from a common ancestor in an ancestral population (the gray background), descending to three populations labelled A, B and C. There are two speciation events (Sp1 and Sp2), each occurring at the junctions shown as an upside down Y. There are also two 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.

Analogy
Homology
- Paralogy
  - Xenology or not
    (xeonologs from paralogs)
- Orthology
  - Xenology or not

Exercise

- On the basis of Fitch's definitions (previous slide), qualify the relationships between each pair of genes in the illustrative schema.
  - P paralog
  - O ortholog
  - X xenolog
  - A analog



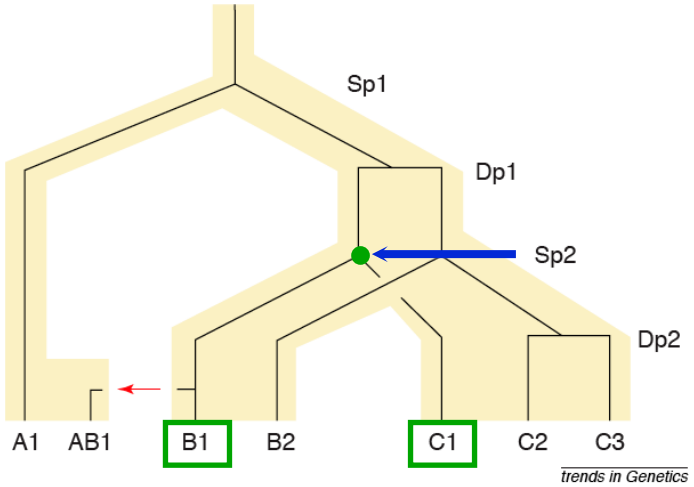
	A1	AB1	B1	B2	C1	C2	C3
A1							
AB1							
B1							
B2							
C1							
C2							
C3							

- **Orthologs** can fomally be defined as a pair of genes whose last common ancestor occurred immediately before a speciation event (ex: a_1 and a_2).
- **Paralogs** can fomally be defined as a pair of genes whose last common ancestor occurred immediately before a gene duplication event (ex: b_2 and $b_{2'}$).

Source: Zvelebil & Baum, 2000

Exercise

- **Example:** B1 versus C1
  - The two sequences (B1 and C1) were obtained from taxa B and C, respectively.
  - The cenancestor (**blue arrow**) is the taxon that preceded the second speciation event (Sp2).
  - The common ancestor gene (**green dot**) coincides with the cenancestor
- -> B1 and C1 are orthologs

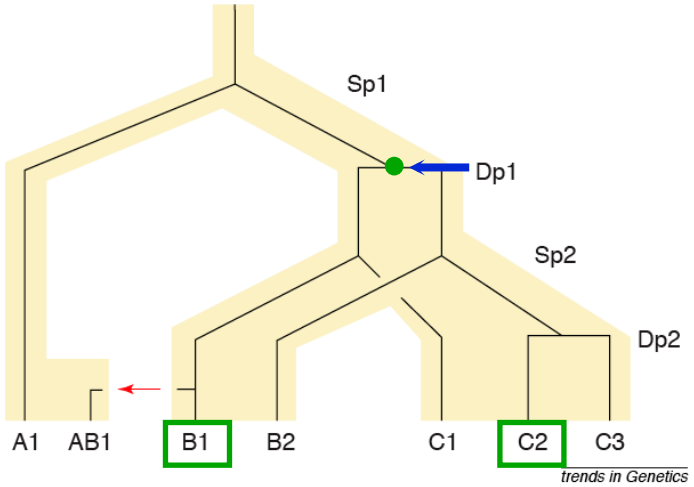


	A1	AB1	B1	B2	C1	C2	C3
A1							
AB1							
B1							
B2							
C1			O				
C2							
C3							

- **Orthologs** can fomally be defined as a pair of genes whose **last common ancestor** occurred immediately before a **speciation event**.
- **Paralogs** can fomally be defined as a pair of genes whose last common ancestor occurred immediately before a gene duplication event.
- **Source: Zvelebil & Baum, 2000**

Exercise

- **Example:** B1 versus C2
  - The two sequences (B1 and C2) were obtained from taxa B and C, respectively.
  - The common ancestor gene (**green dot**) is the gene that just preceded the duplication Dp1.
  - This common ancestor is much anterior to the cenancestor (**blue arrow**).
- -> B1 and C2 are paralogs



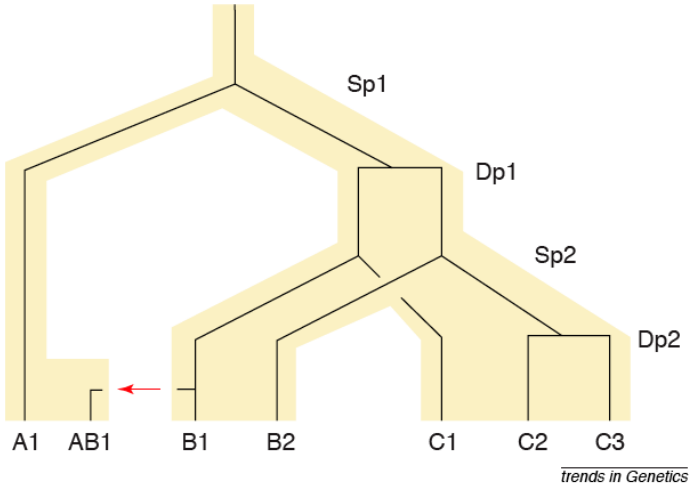
	A1	AB1	B1	B2	C1	C2	C3
A1							
AB1							
B1							
B2							
C1			O				
C2			P				
C3							

- **Orthologs** can fomally be defined as a pair of genes whose last common ancestor occurred immediately before a speciation event.
- **Paralogs** can fomally be defined as a pair of genes whose **last common ancestor** occurred immediately before a **gene duplication event**.
- **Source: Zvelebil & Baum, 2000**

Solution to the exercise

- On the basis of Fitch's definitions (previous slide), qualify the relationships between each pair of genes in the illustrative schema.
  - P paralog
  - O ortholog
  - X xenolog
  - A analog

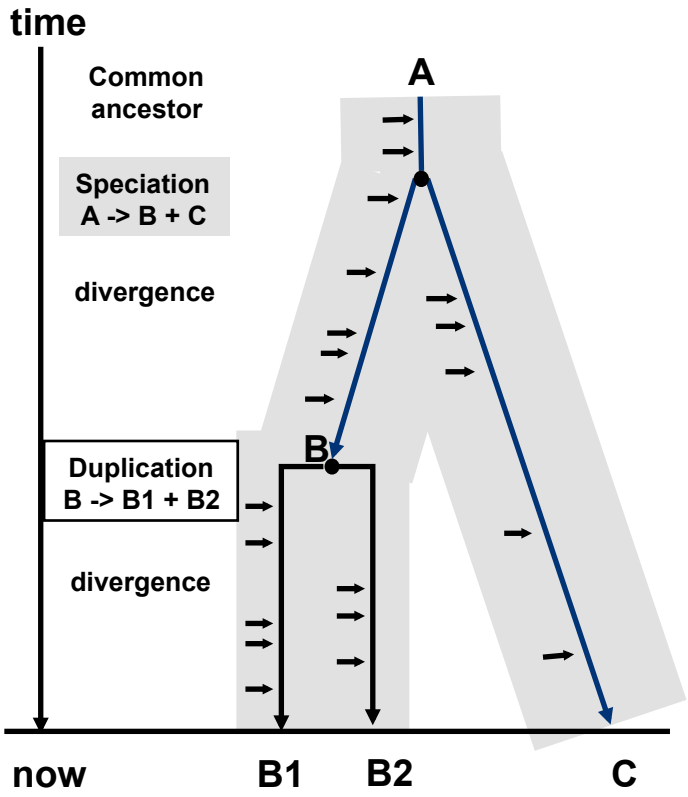
	A1	AB1	B1	B2	C1	C2	C3
A1	I						
AB1	X	I					
B1	O	X	I				
B2	O	X	P	I			
C1	O	X	O	P	I		
C2	O	X	P	O	P	I	
C3	O	X	P	O	P	P	I



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.

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



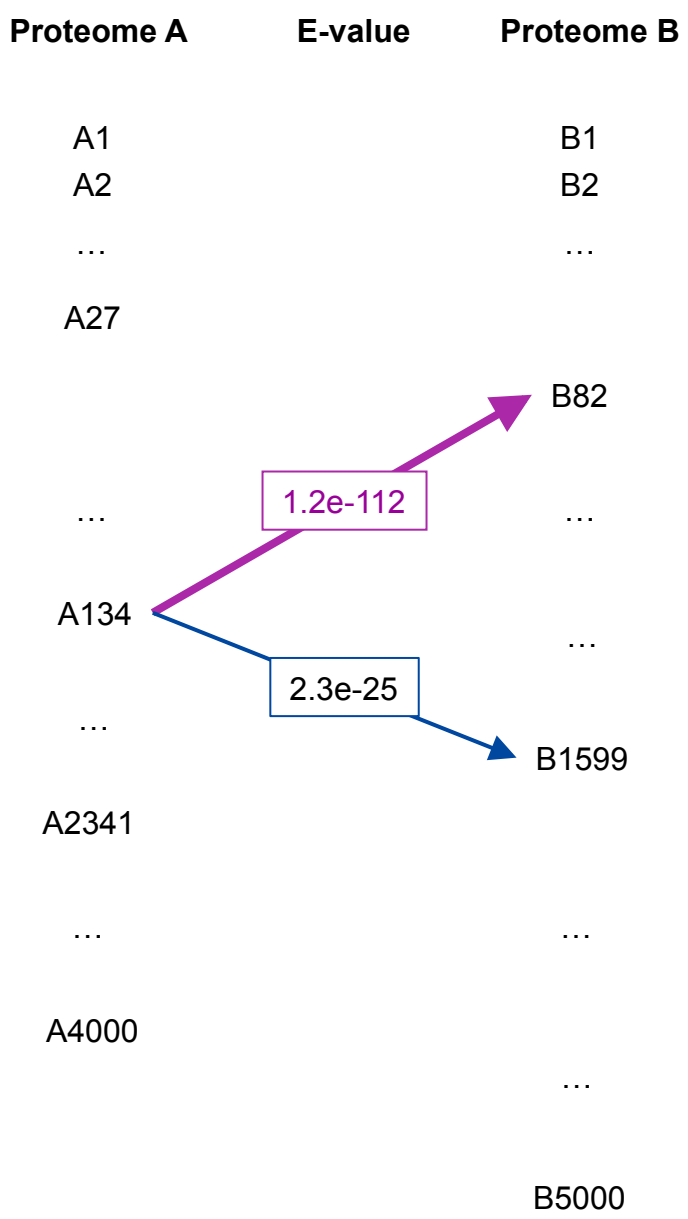
Orthologs are sequences whose **last common ancestor** occurred immediately before a **speciation** event.
Paralogs are sequences whose **last common ancestor** occurred immediately before a **duplication** event.
(Fitch, 1970; Zvelebil & Baum, 2000)

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.

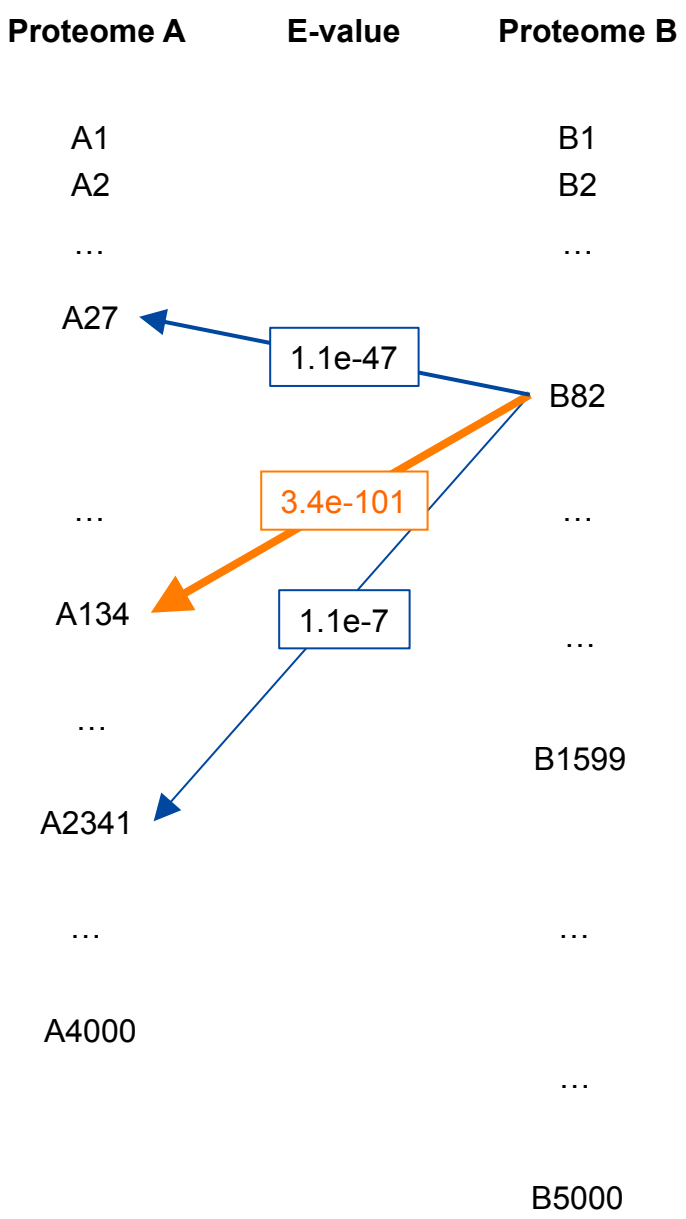
Inferring orthology / paralogy by reciprocal best hits

- 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***.



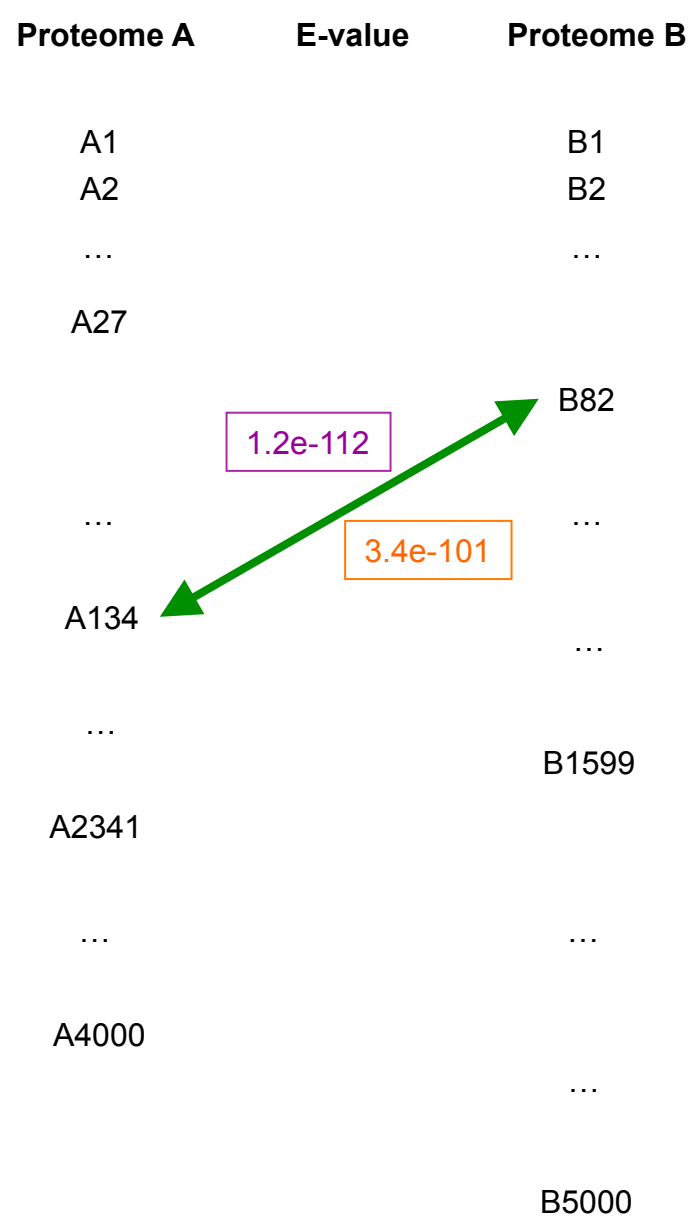
Inferring orthology / paralogy by reciprocal best hits

- 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***.



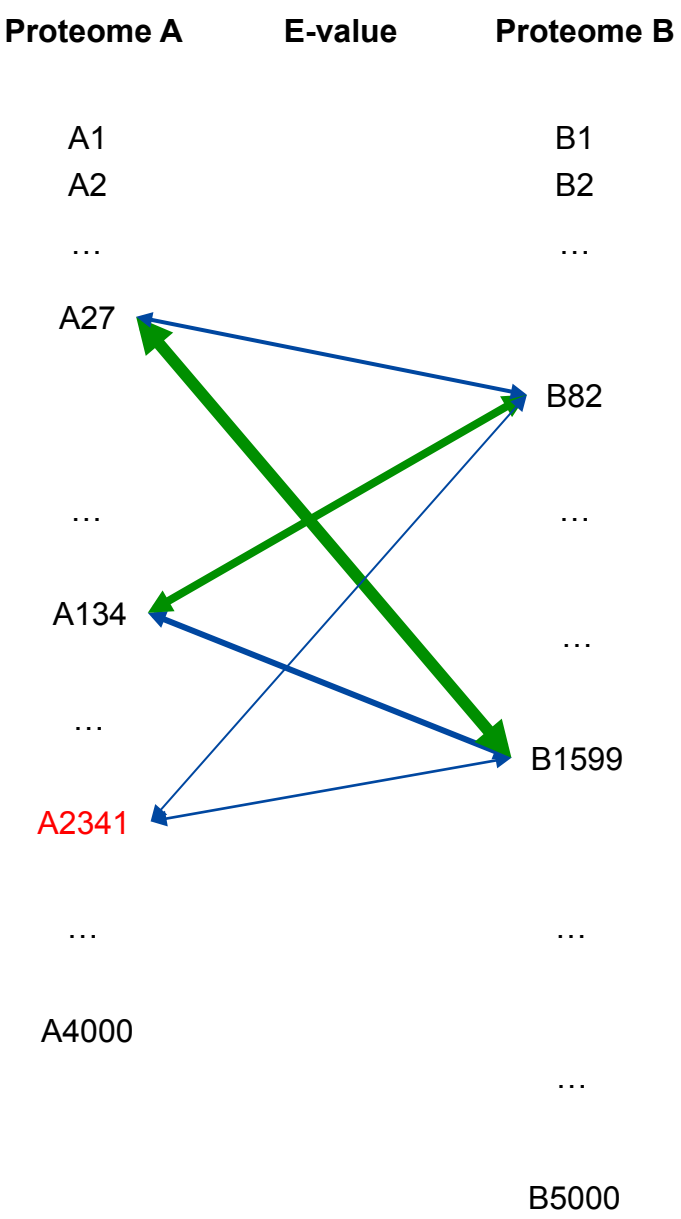
Inferring orthology / paralogy by reciprocal best hits

- 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.



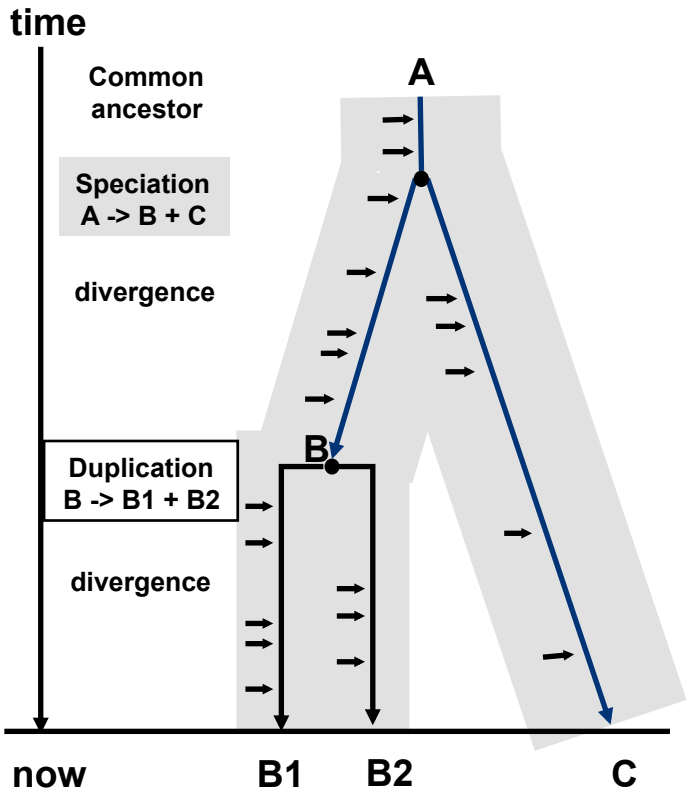
Inferring orthology / paralogy by reciprocal best hits

- 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

Criteria for genome-wise detection of orthologs

- Criterion for detecting paralogy
  - Two genes from a given species (e.g. C) are more similar to each other than to their best hit in genome B.
- Pairs of orthologous genes
  - BeT (Best-scoring BLAST hit)
    - Insufficient to infer orthology
  - Bidirectional best hit (BBH)
    - Better approximation
    - Discuss the problem of gene loss
- Clusters of orthologous genes (COGs)
  - Triangular definition of COGs (Tatusov, 1997)
  - KOG: euKaryotic Orthologous Groups
    - Question: is there any interest of defining a new term for eukaryotes ?
- To discuss
  - theoretical weakness of the COG concept, since orthology is NOT a transitive relationship.
  - Pragmatic value of the concept

Fig. 1. Examples of COGs. Solid lines show symmetrical BeTs. Broken lines show asymmetrical BeTs, with color corresponding to the species for which the BeT is observed. Genes from the same species are adjacent; otherwise the gene names are positioned arbitrarily. A unique COG ID is indicated in the upper left corner. (**A**) Congruent BeTs form a triangle, the minimal COG. Origin of the proteins: KatG, *E. coli*; sll1987, *Synechocystis* sp.; and YKR066c, *S. cerevisiae*. Note that all the BeTs are symmetrical. (**B**) A simple COG with two yeast paralogs. Origin of the proteins: IleS, *E. coli*; HIN0378, *H. influenzae*; MG345, *M. genitalium*; MP322, *M. pneumoniae*; MJ0947, *M. jannaschii*; and YBL076c and YPL040c, *S. cerevisiae*. Note the adjacent triangles with a common side, for example, IleS-MG345-MJ0947 and sll1362-MG345-MJ1362. YPL040c is the yeast mitochondrial isoleucyl-tRNA synthetase; the bacterial orthologs and that from *M. jannaschii* are the BeTs for this yeast protein, but the reverse is true only of the bacterial proteins (symmetrical BeTs). Conversely, for YBL076c, which is the yeast cytoplasmic isoleucyl-tRNA synthetase, the *M. jannaschii* ortholog is a symmetrical BeT, whereas the bacterial BeTs are asymmetrical. (**C**) A complex COG with multiple paralogs. Origin of the proteins: RpoH, RpoS, RpoD, and FliA, *E. coli*; HIN1403 and HIN1655, *H. influenzae*; MG249, *M. genitalium*; MP485, *M. pneumoniae*; sll0184, sll0306, slr0653, sll1689, sll2012, and slr1564, *Synechocystis* sp. RpoD, HIN1655, slr0653, and MG249 are major sigma factors (σ70), whose function is universal in bacteria; note the fully symmetrical relationships between these proteins. The other proteins are specialized sigma factors whose radiation from the ancestral family apparently was accompanied by modification of the function and involved accelerated evolution; note the asymmetrical BeTs.

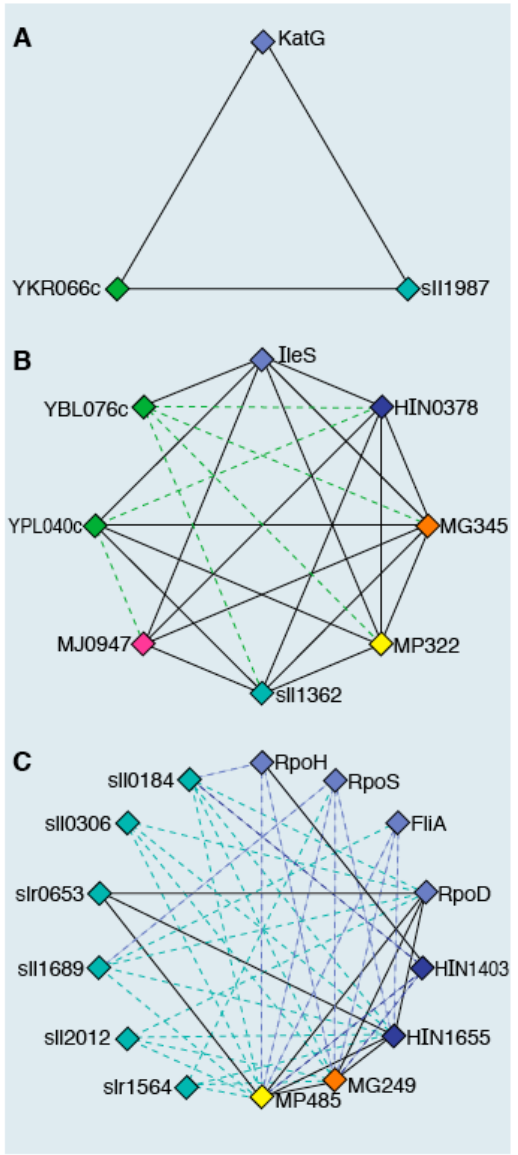


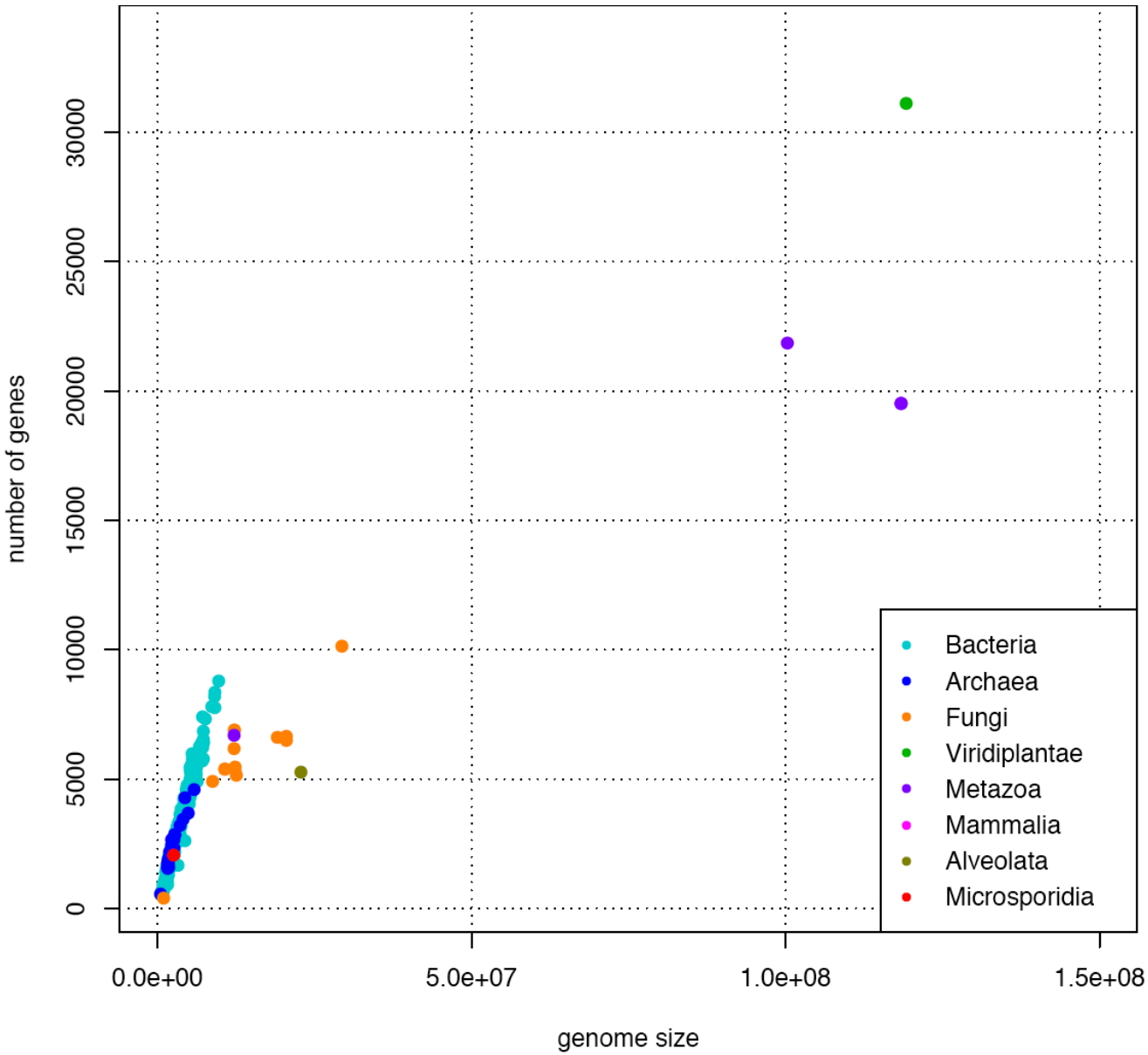
Figure from Tatusov, 1997

Comparative genomics
Genome and proteome sizes

Some milestones

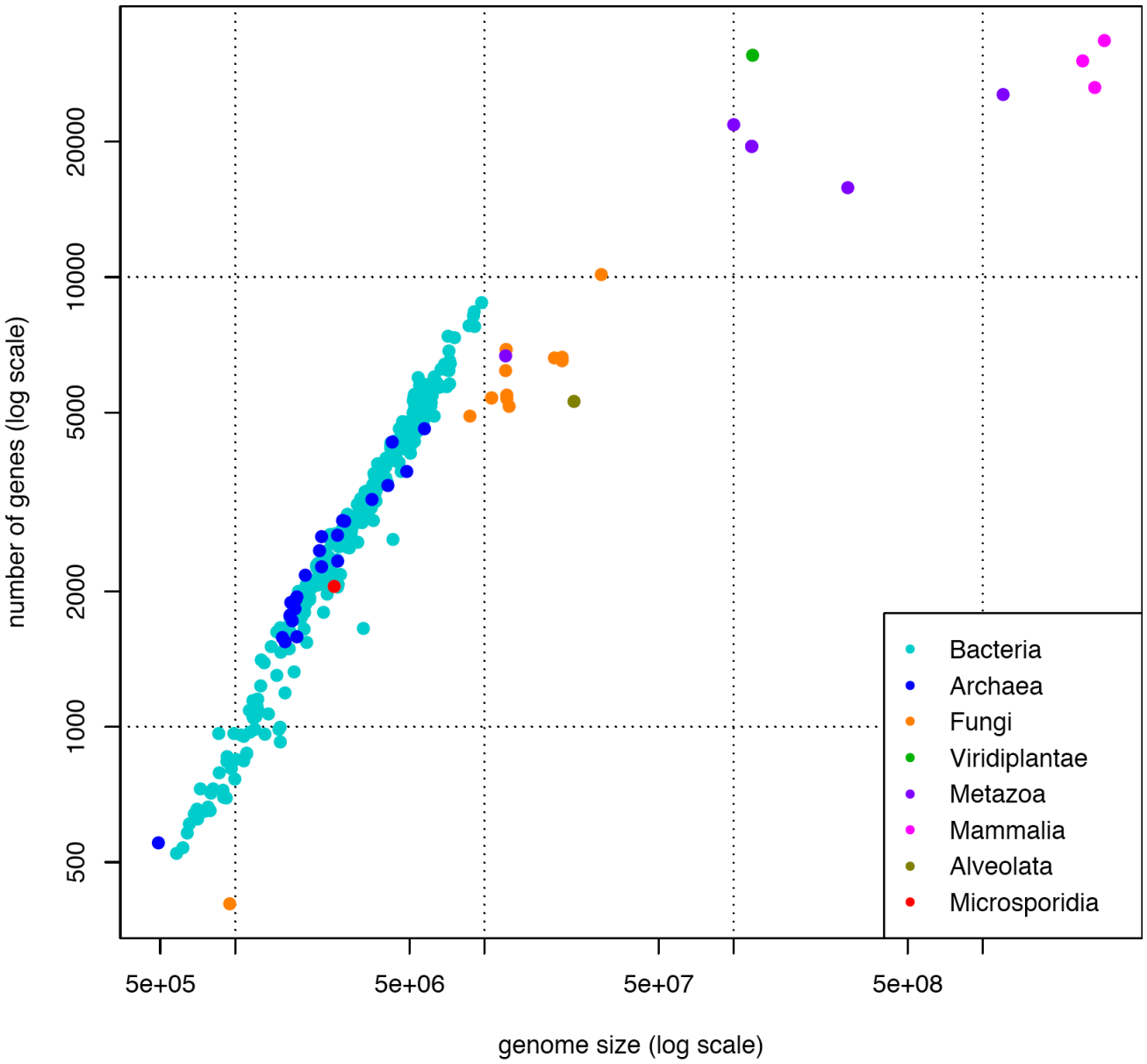
Speices name	Common name	Publication year	Genome size Mb	Gene number	Mean intergenic distance Kb	Fraction of coding sequences %	Non-coding fraction %	Repetitive elements %	Transcribed fraction %	Remarks
Bactérie										
Mycoplasma genitalium	*Mycoplasma*	1995	0.6	481	1.2	90	10			Small genome (intracellular)
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							

Gene numbers as a function of genome sizes



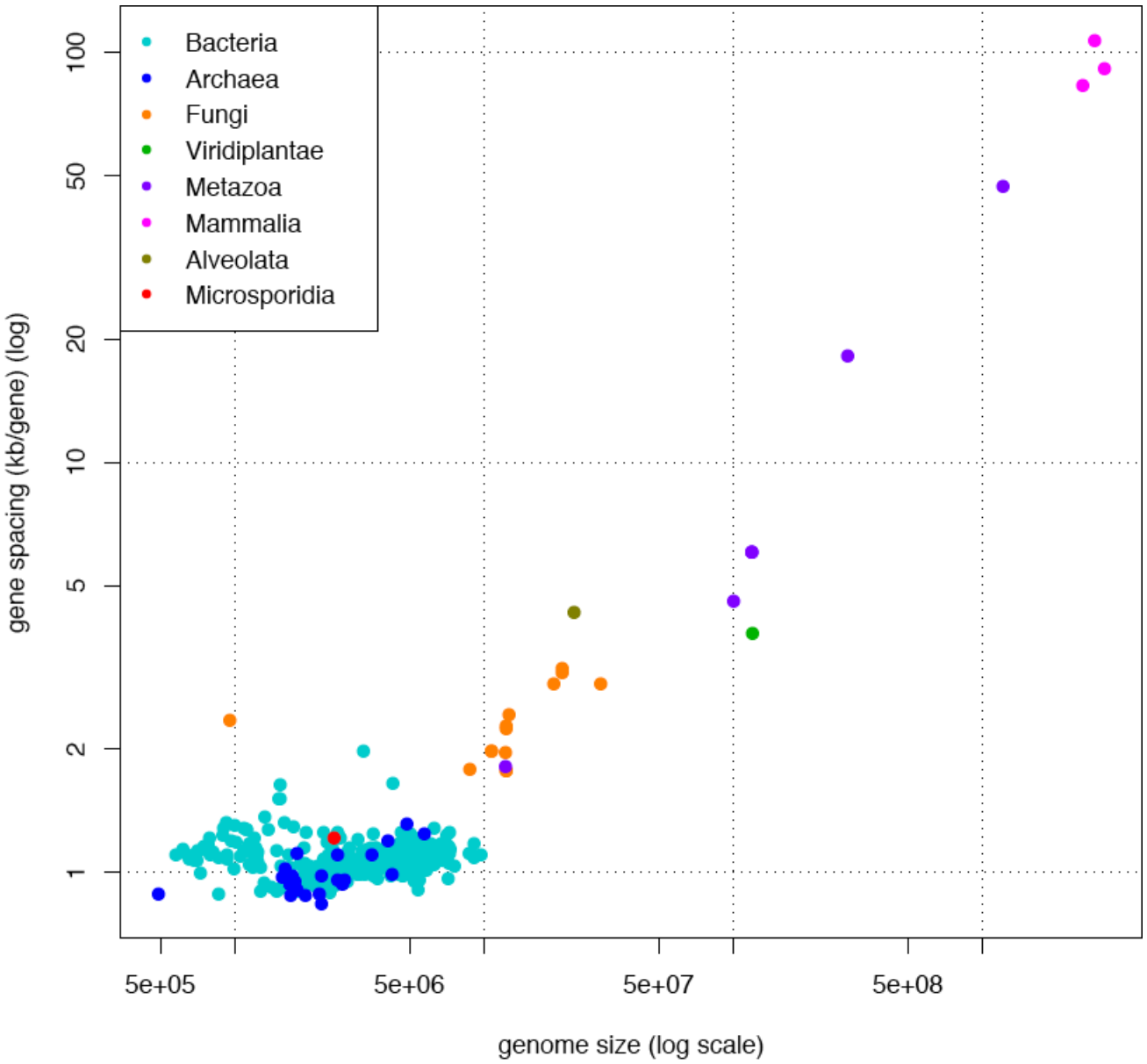
- In prokaryotes, the number of genes increases linearly with genome size
- In eukaryotes, this is not the case: the genome size increases faster than the number of genes

Gene numbers as a function of genome sizes (log-log plot)



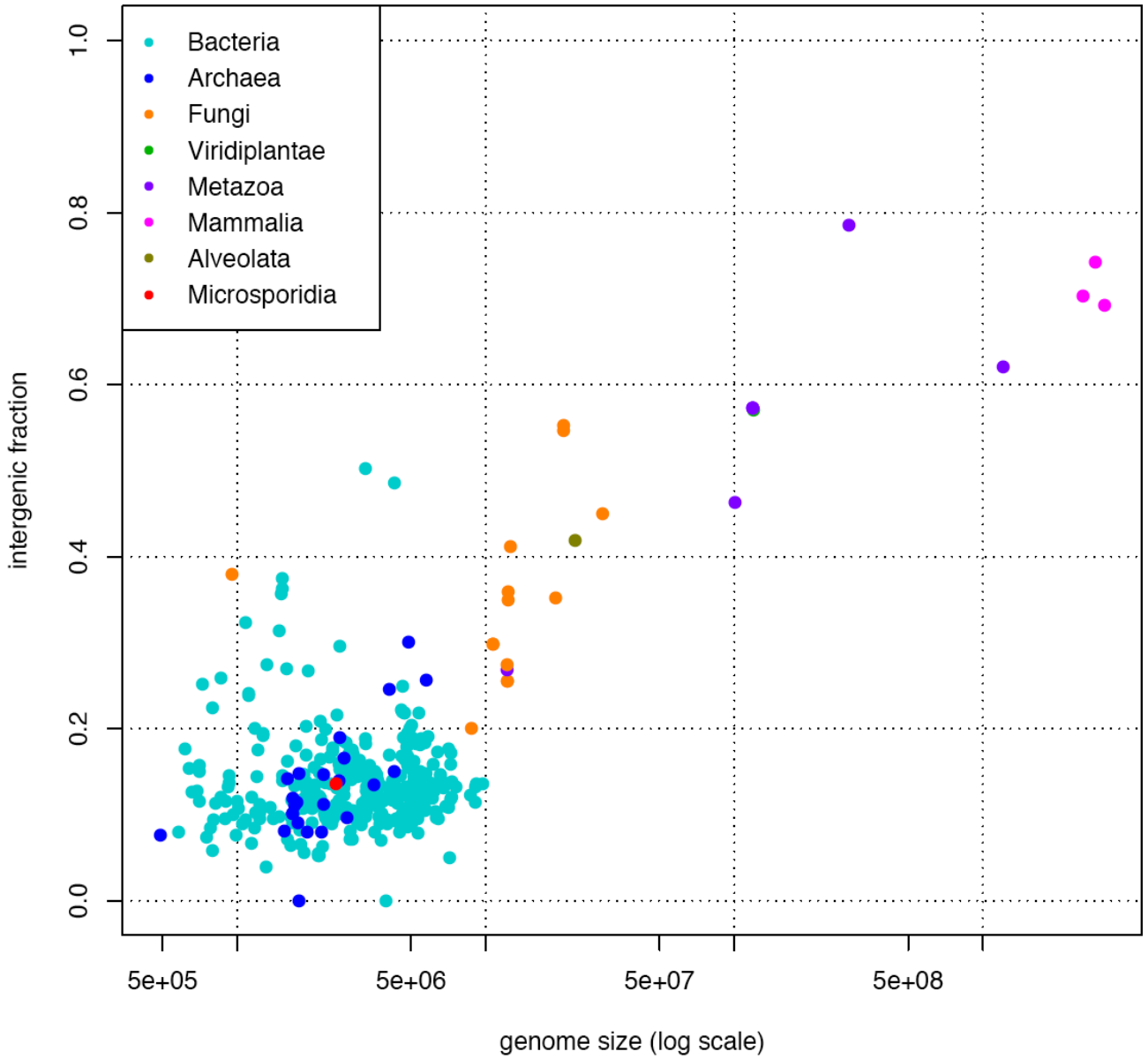
- Beware: the axes are logarithmic.
- This plot represents the same data as the previous one, but in logarithmic scale, in order to see Mammals as well.

Gene spacing



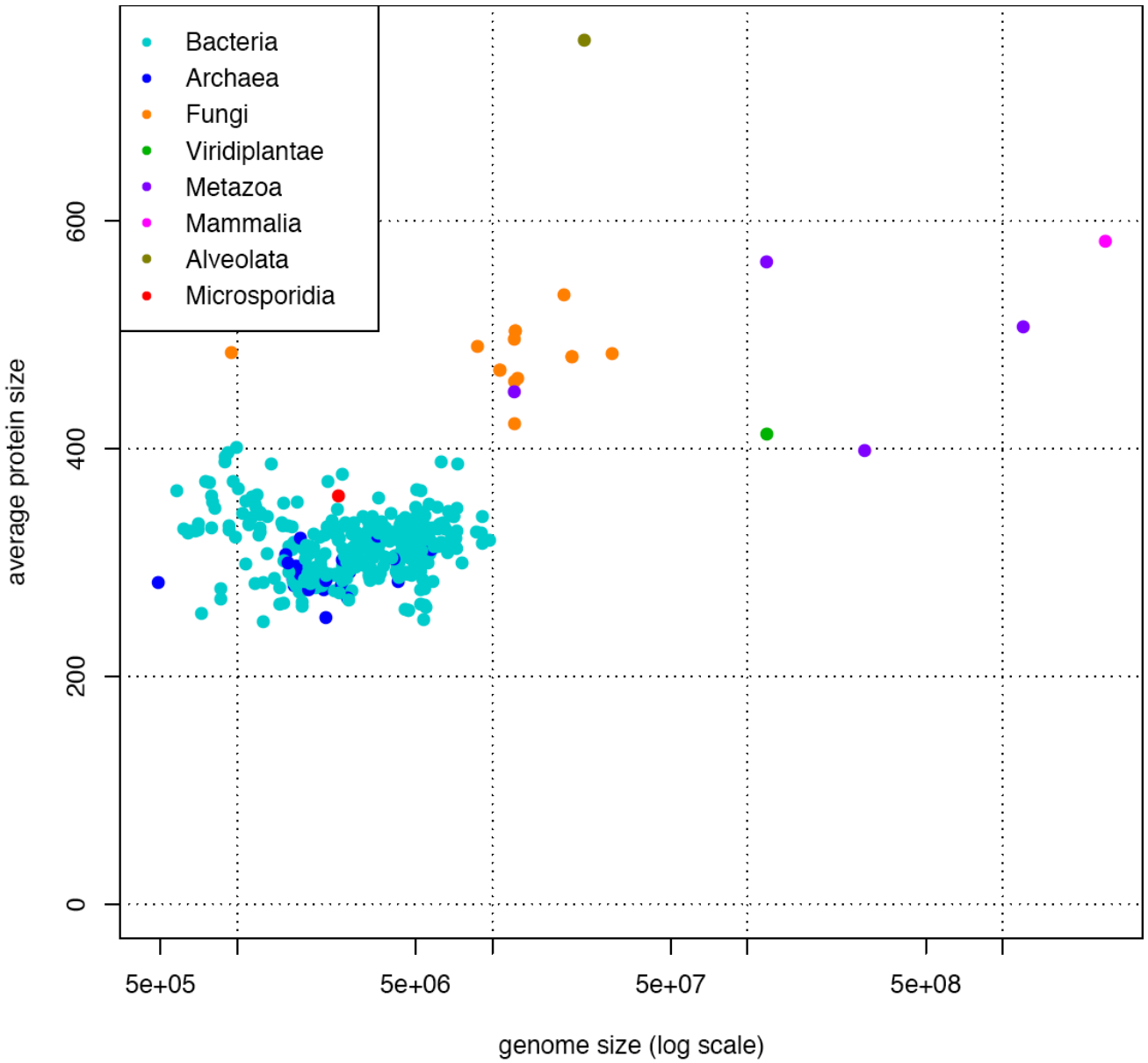
- Gene spacing increases considerably with the complexity off the organisms.
- Note: the X axis si logarithmic, not the Y axis -> the increase seems grossly exponential.

Proportion of intergenic regions



- Beware: the X axis is logarithmic.
- The proportion of intergenic regions increases with the complexity of an organism.
- In addition (not shown here), introns represent an increasing fraction of the genome.
- For example, the exonic fraction represents <5% of the human genome.

Protein size versus genome size



- Protein sequences are shorter in prokaryotes than in eukaryotes.
- Among eukaryotes, the increase in genome size is not correlated to an increase in protein size
  - higher eukaryotes have a much larger genome than fungi, without increase in protein size

Comparative genomics methods
Phylogenetic profiles

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 called ***phylogenetic profiles***
  - For each protein of the reference organism, detect all orthologs in a set of other genomes (phylogenetic profiles of occurrence).
  - Detect groups of co-occurring proteins: similar profiles of presence / absence across proteomes.
- Today, this method can be applied to several thousands of genomes. Its power increases with the number of genomes.

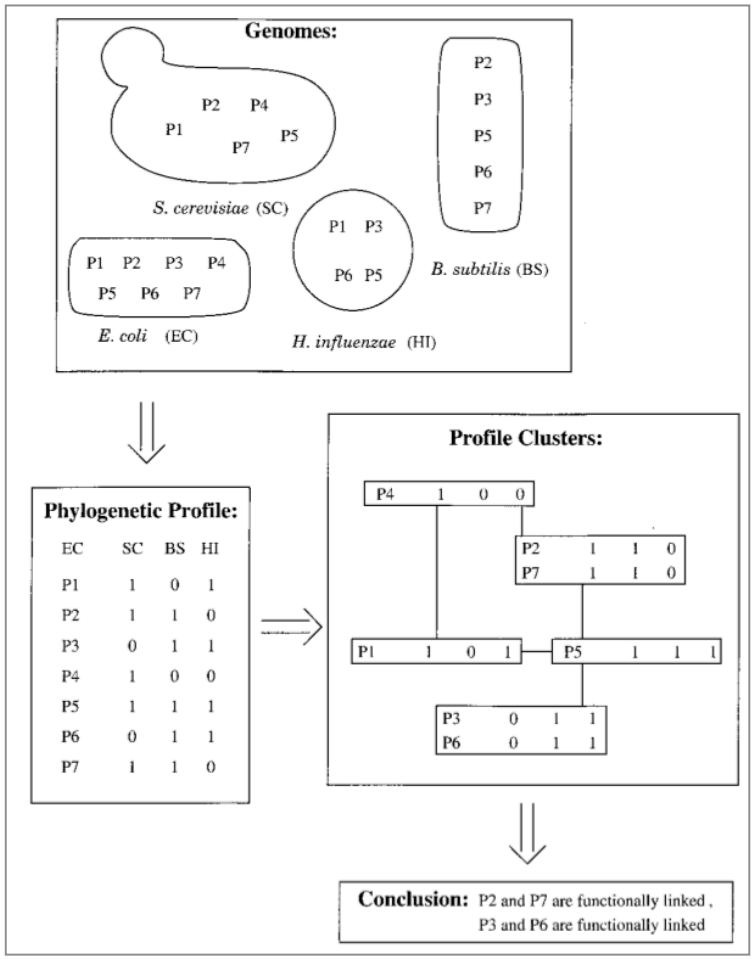


Table 1. Phylogenetic profiles link protein with similar keywords

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

- Approach
  - For each protein of the reference organism (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
- Note
  - This approach is not properly speaking “phylogenetic”, since there is no attempt to retrace the phylogeny (history of descent) of the proteins.

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	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.meliloti
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
16127996	1	0	0	1	1	1	1	1	1	1	1	1	0	1	0	0	1	1	1	0	1	0	0	1	1	1	1	1	1	1	1	1	1	0	1	1
16127997	1	0	0	1	1	1	1	1	1	0	1	1	0	1	0	0	1	1	1	0	1	0	0	0	1	0	0	1	1	1	0	0	0	0	0	0
16127998	0	0	0	0	1	0	0	1	1	0	0	1	0	1	0	0	0	0	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	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	1	0	0	0	1	0	1	0	0	1	1	1	1	0
16128001	0	1	1	0	0	1	1	0	1	1	0	0	0	0	0	0	1	1	1	1	0	0	1	0	0	1	0	1	1	1	0	1	1	0	0	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	0	1	1	0	0	1	0	0	0
16128003	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	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	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	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	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
16128008	1	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
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
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

Phylogenetic profiles reveal groups of functionally related genes

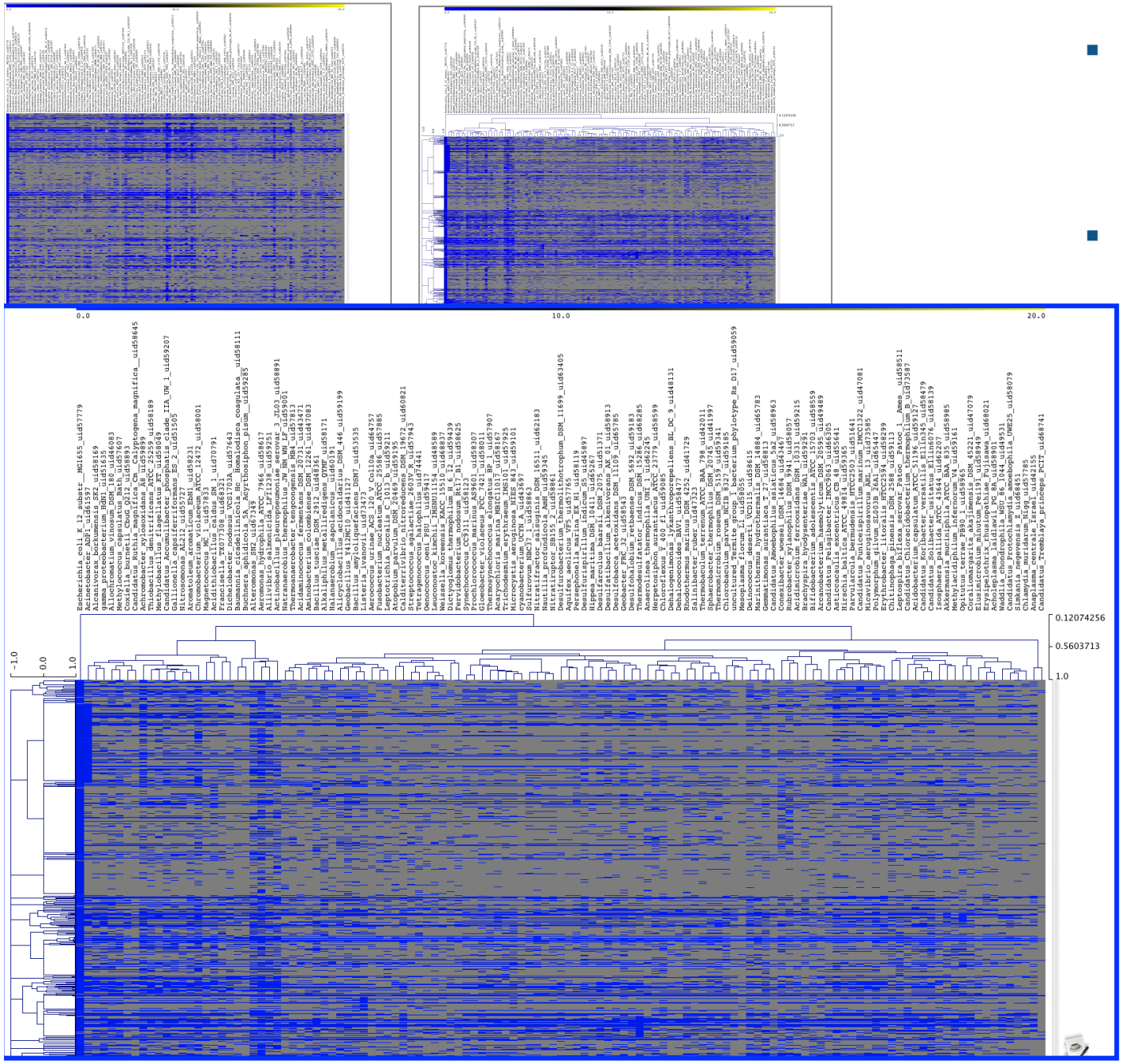
- The phylogenetic profile table indicates the presence/absence of genes (one row per gene) in a set of genomes (one column per cross-species comparison).
- Reference organism: *Escherichia coli K-12 substrain MG1665*
- Query genomes
  - Selected 154 Bacteria among 2065 (1 species for each group at depth 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.
- Analysis done 2013-05-02

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		Stats on orthologs per organism	
min	1	min	73
max	123	max	3968
mean	23.3	mean	758.2
median	10.0	median	677

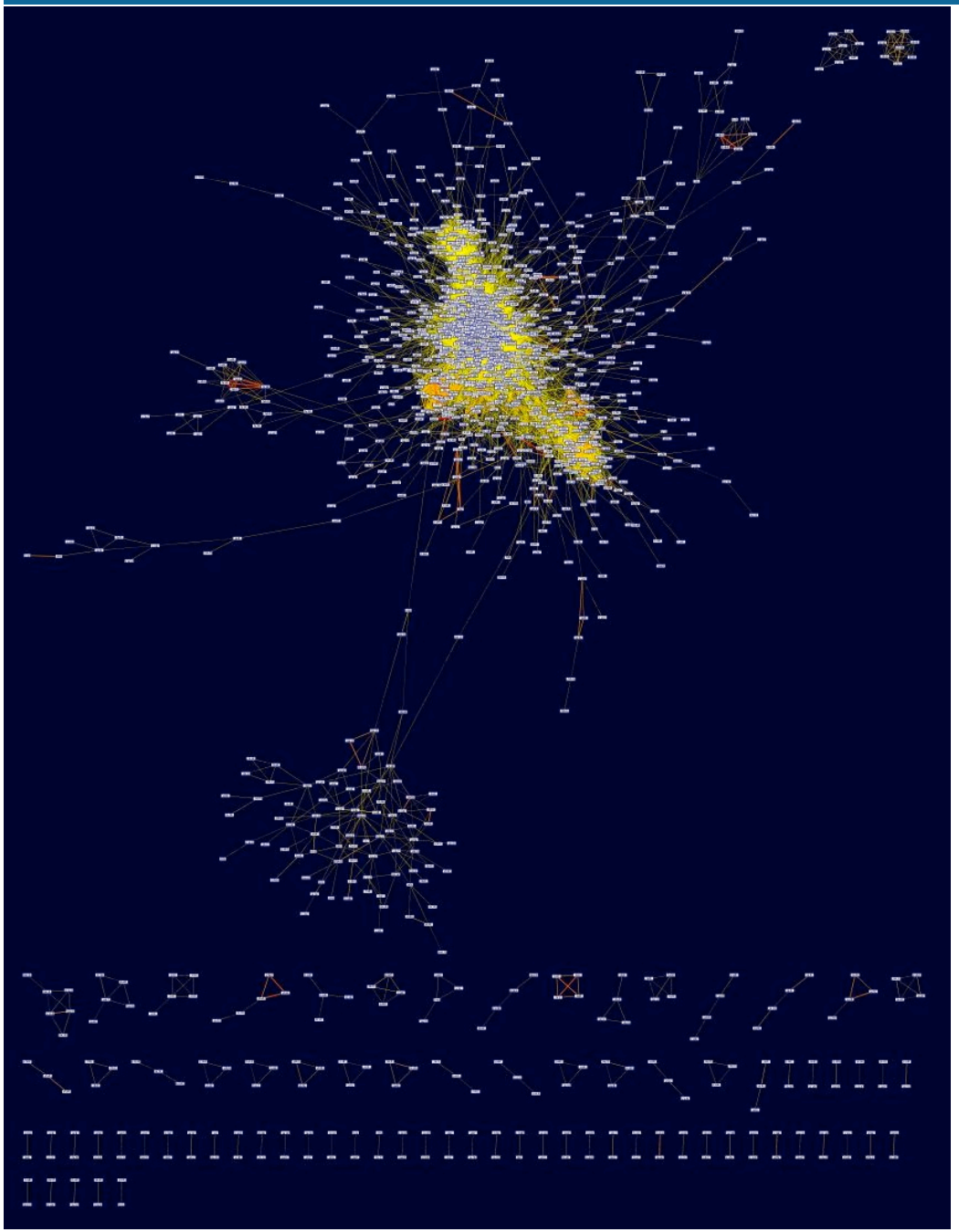
Gene name	Gene ID	Number of organisms per reference gene	Escherichia_coli_K_12_substr__MG1655_uid57779	Acidobacterium_capsulatum_ATCC_51196_uid59127	Candidatus_Chloracidobacterium_thermophilum_B_uid73587	Candidatus_Koribacter_versatilis_Ellin345_uid58479	Candidatus_Solibacter_usitatus_Ellin6076_uid58139	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	0	1	0	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	0	1	0	0	0	0	1	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	0	1	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	1	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	0	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

Heatmap of phylogenetic profiles for Escherichia coli K-12 MG1665



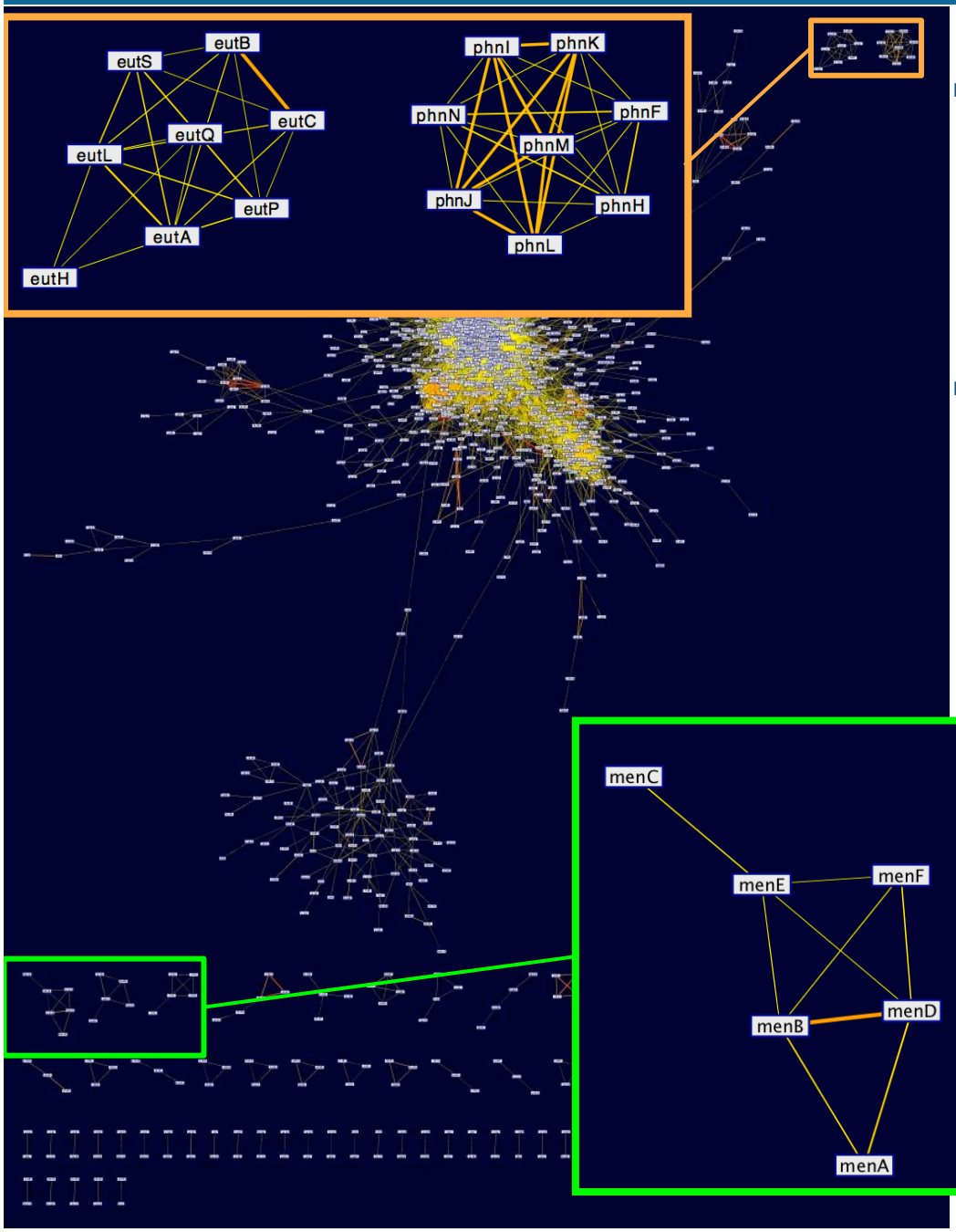
- Phylogenetic profiles can be visualized as a heatmap, with one row per gene, and one column per organism.
- We can apply clustering
  - on rows, to regroup genes having similar profiles;
  - on columns, to regroup organisms having similar genes in the whole genome.
  - The heatmap remains however difficult to interpret.

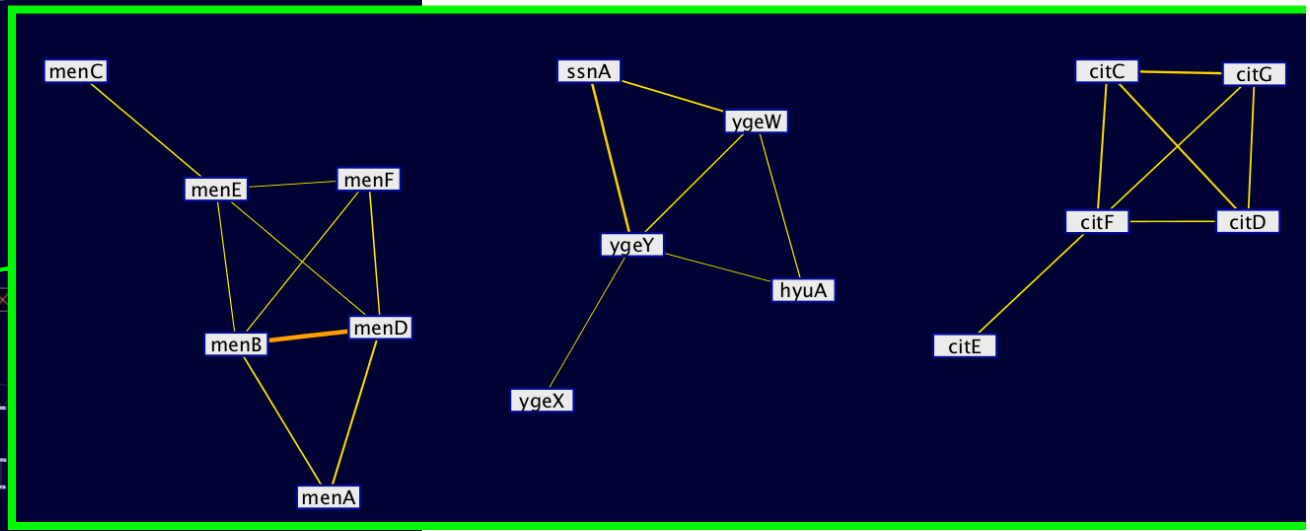
Co-occurrence network extracted from phylogenetic profiles



- Co-occurrence network extracted from phylogenetic profiles
  - Reference organism: *Escherichia coli K-12 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

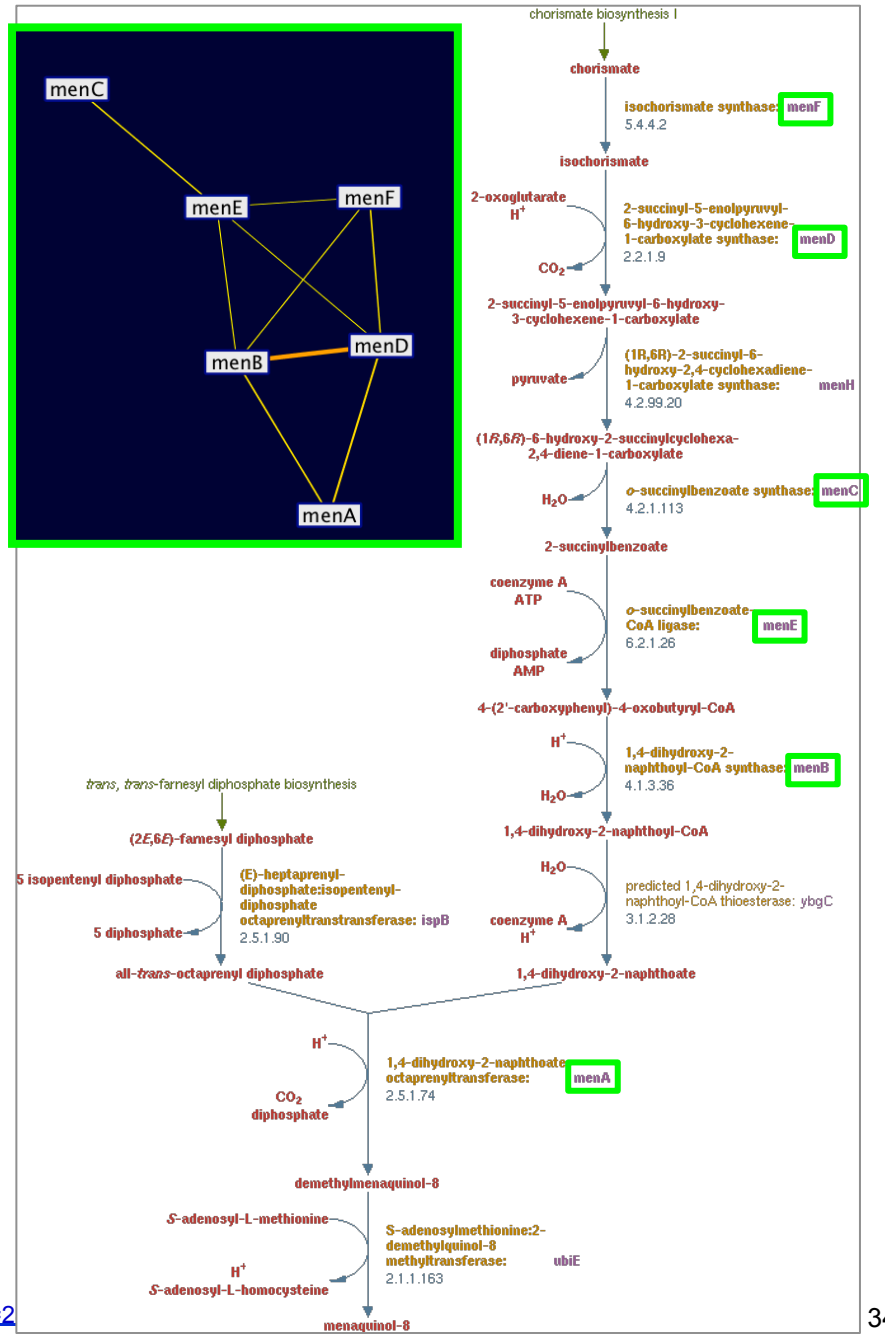




- Groups of inter-connected genes are generally involved in a common function.
  - eut ethanol utilisation
  - men menaquinol-8 biosynthesis
  - cit citrate metabolism
  - phn phosphonate metabolism
- These groups of genes appear clustered on the co-occurrence network, because they are either present together, or absent together in genomes.

Clusters of co-occurring genes reveal pathways

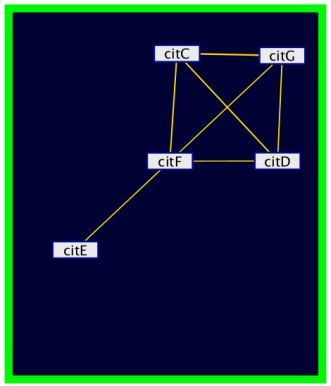
- Phylogenetic profiles reveal a group of co-occurring genes whose name starts by “men”.
- The product of these genes catalyse 6 among 10 reactions of the superpathway “menaquinol-8 biosynthesis I”.
- Phylogenetic profiles revealed the associations between these genes without any indication of their function.



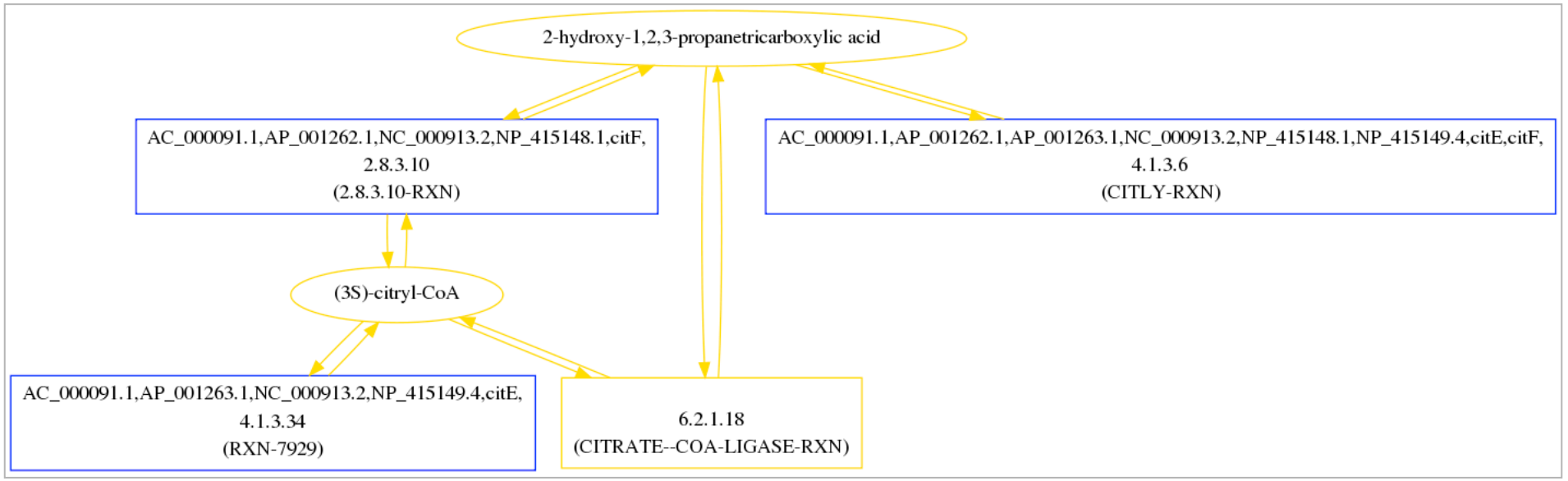
- http://ecocyc.org/ECOLI/NEW-IMAGE?type=PATHWAY&object=PWY-5838&detail-level=2

Discovering pathways from clusters of co-occurrence genes

- Rather than comparing the co-occurrence cluster to annotated pathways, we can run a pathway extraction algorithm to identify metabolic pathways that can be catalysed by clusters of co-occurring enzymes.



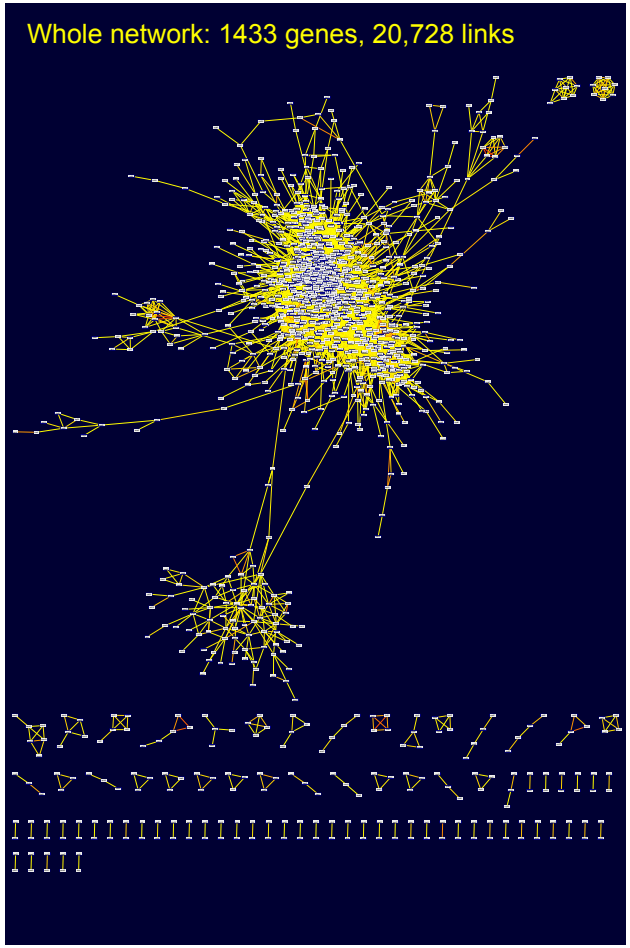
	#Matches	EC	node ids	node type	description
☑	citC	6.2.1.22	CITC-RXN	GENE_NAME	[citrate (*pro*-3*S*)-lyase] ligase
☑	citE	4.1.3.34	CITRYLY-RXN, RXN-7929	GENE_NAME	Citryl-CoA lyase, citryl-CoA lyase
☑	citE	4.1.3.6	CITLY-RXN	GENE_NAME	citrate (*pro*-3*S*)-lyase
☑	citF	2.8.3.10	2.8.3.10-RXN, CITTRANS-RXN	GENE_NAME	citrate CoA-transferase, Citrate CoA-transferase
☑	citF	4.1.3.6	CITLY-RXN	GENE_NAME	citrate (*pro*-3*S*)-lyase
☑	citG	2.7.8.25	2.7.8.25-RXN	GENE_NAME	triphosphoribosyl-dephospho-CoA synthase

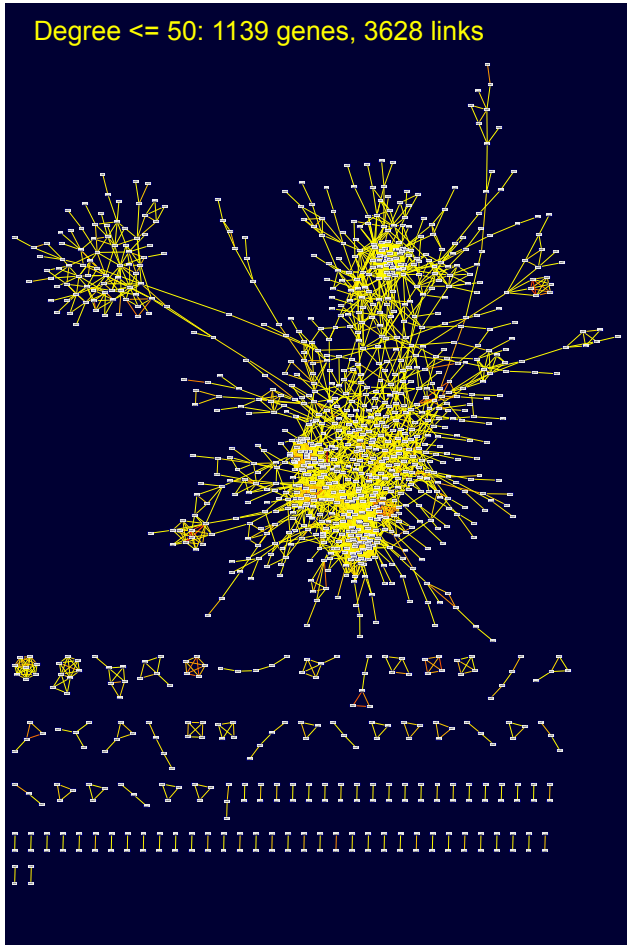


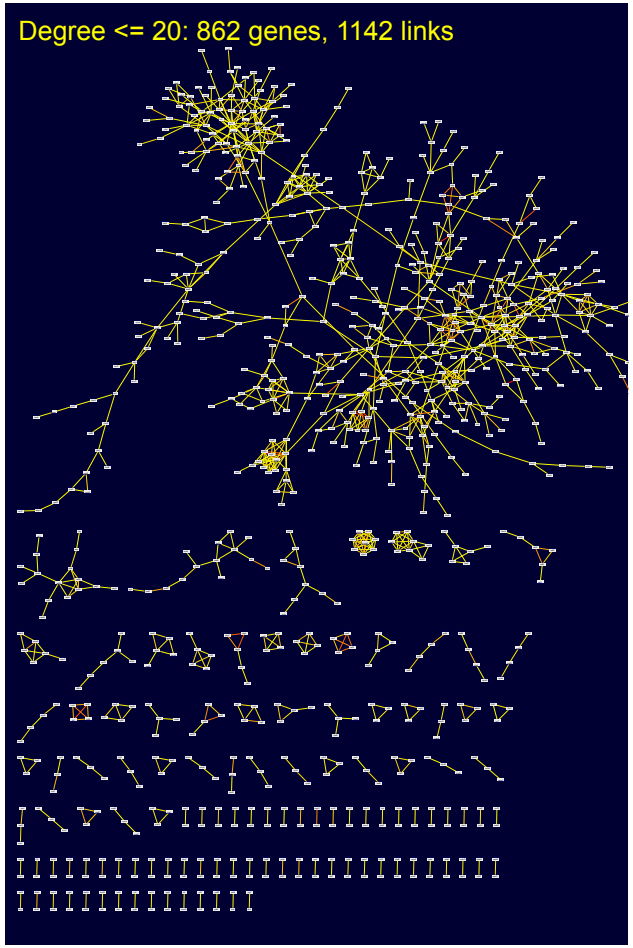
- Pathway extraction tool: http://neat.rsat.eu/
- Method
  - Faust and van Helden. Predicting metabolic pathways by sub-network extraction. Methods Mol Biol (2012) vol. 804 pp. 107-30
  - Faust et al. Prediction of metabolic pathways from genome-scale metabolic networks. BioSystems (2011) vol. 105 (2) pp. 109-21
  - Faust et al. Pathway discovery in metabolic networks by subgraph extraction. Bioinformatics (2010) pp.

Filtering out the hubs

- The network inferred from phylogenetic profiles generally contains a large clump of genes, in which it is difficult to distinguish specific clusters.
- One approach is to filter out the “hubs” of this network: discard genes whose degree (number of links to other genes) exceeds a given threshold.
- The network becomes more “readable”, but we probably loose a part of meaningful information.



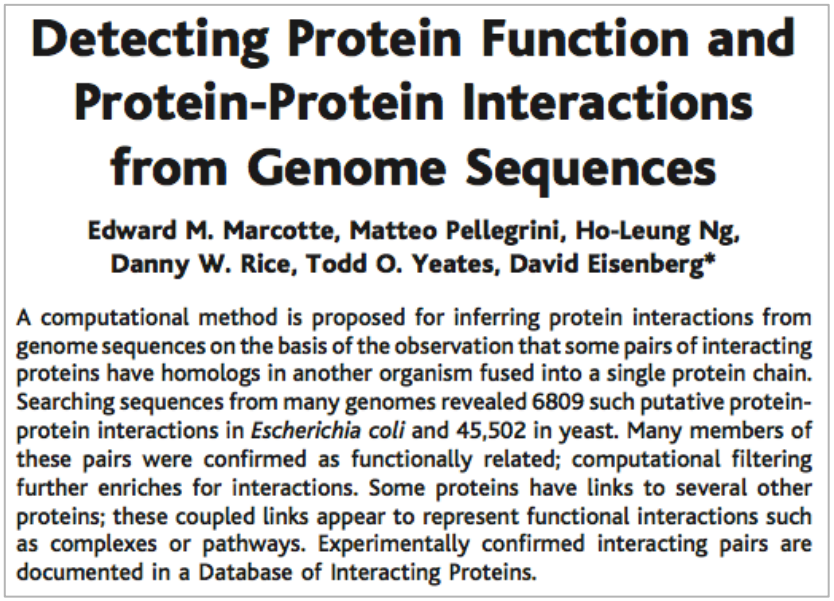




Comparative genomics methods

Gene fusions

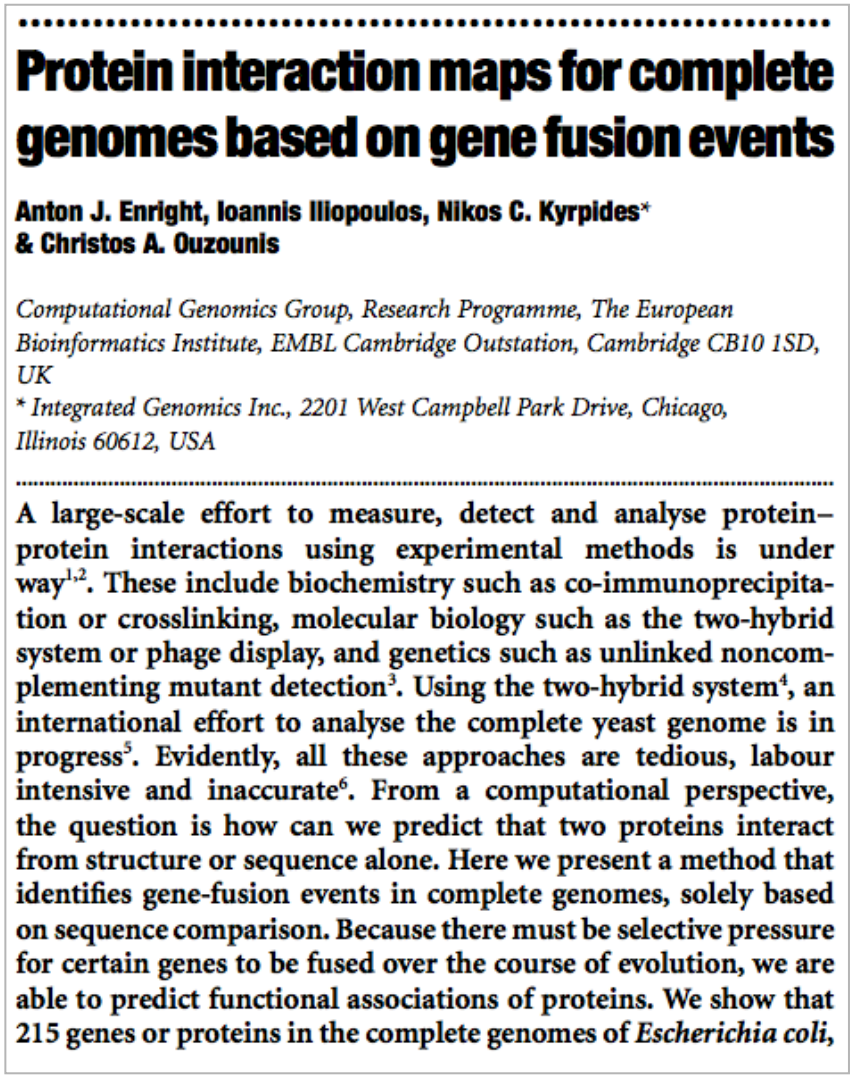
- In 1999, two groups propose a method to predict functional interactions between genes based on cross-genome identification of gene fusions.
  - Marcotte et al., Science, 1999
  - Enright et al., Nature, 1999



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.



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

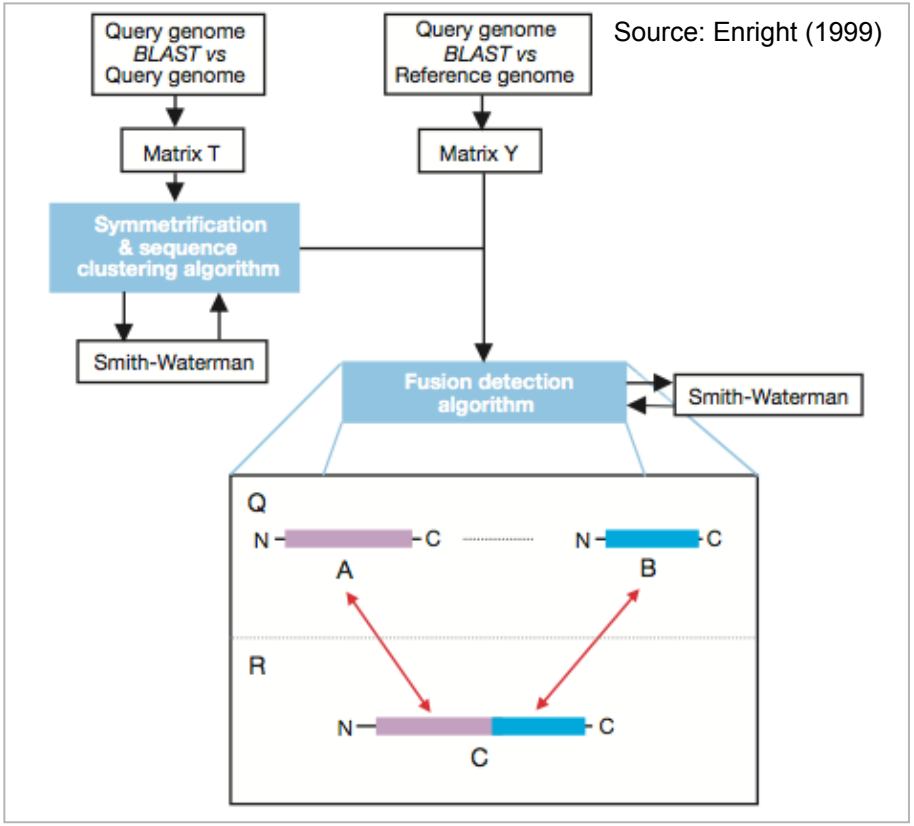
** Integrated Genomics Inc., 2201 West Campbell Park Drive, Chicago, Illinois 60612, USA*

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*,

- 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

Gene fusion – principle of the method

- Principle: identify in a query genome (Q) pairs of genes (A,B) which match non-overlapping segments of a single gene (C) in a reference genome (R).



Examples of gene fusions

- Marcotte et al. illustrate the relevance of gene fusions by discussing a few selected examples.

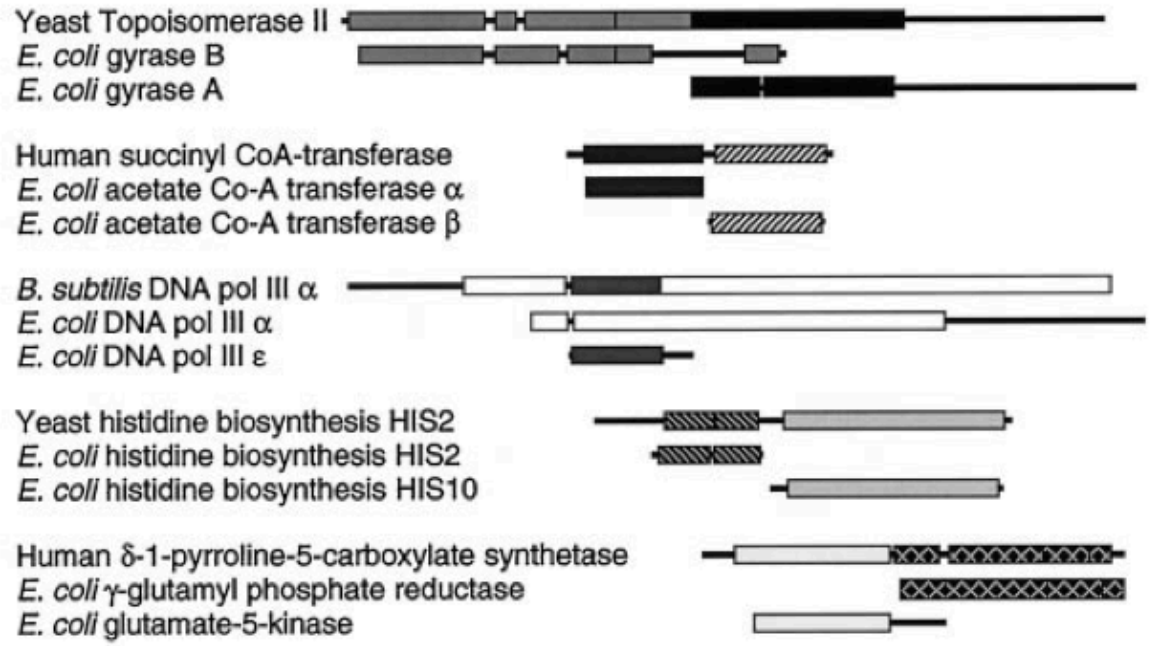


Fig. 1. Five examples of pairs of *E. coli* proteins predicted to interact by the domain fusion analysis. Each protein is shown schematically with boxes representing domains [as defined in the ProDom domain database (*17*)]. For each example, a triplet of proteins is pictured: The second and third proteins are predicted to interact because their homologs are fused in the first protein (called the Rosetta Stone protein in the text). The first three predictions are known to interact from experiments (*18*). The final two examples show pairs of proteins from the same pathway (two nonsequential enzymes from the histidine biosynthesis pathway and the first two steps of the proline biosynthesis pathway) that are not known to interact directly.

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

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***").

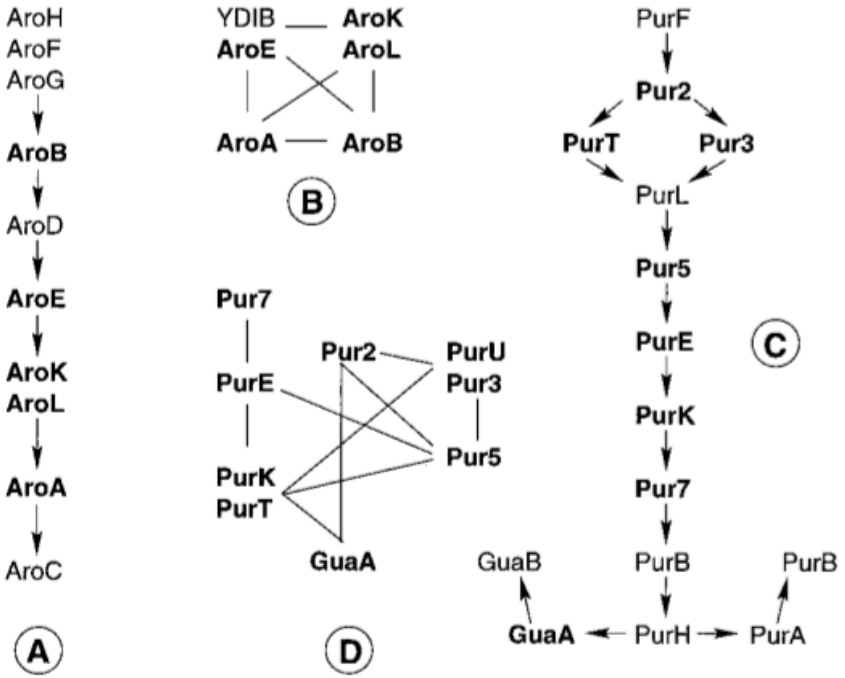


Fig. 2. Reconstruction of two metabolic pathways in *E. coli*, with only interactions predicted by the domain fusion method. Pathways A and C are the known pathways for biosynthesis of shikimate and purine, respectively; they are ordered by the traditional method of successive action of the enzymes on 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

Can gene fusions predict physical interactions between proteins ?

- In their original publication, Marcotte et al. also propose to consider gene fusions as predictions of physical interactions between the proteins coded by these genes.
- They develop a model for the evolution of protein-protein interactions, based on transient interactions between domains in single-protein produced by fused genes.
- This model is rather speculative, and is not supported by the results of their own study.
- It is now agreed that gene fusions can generally reveal functional interactions, but that only a fraction of these would also involve physical interactions between gene products.

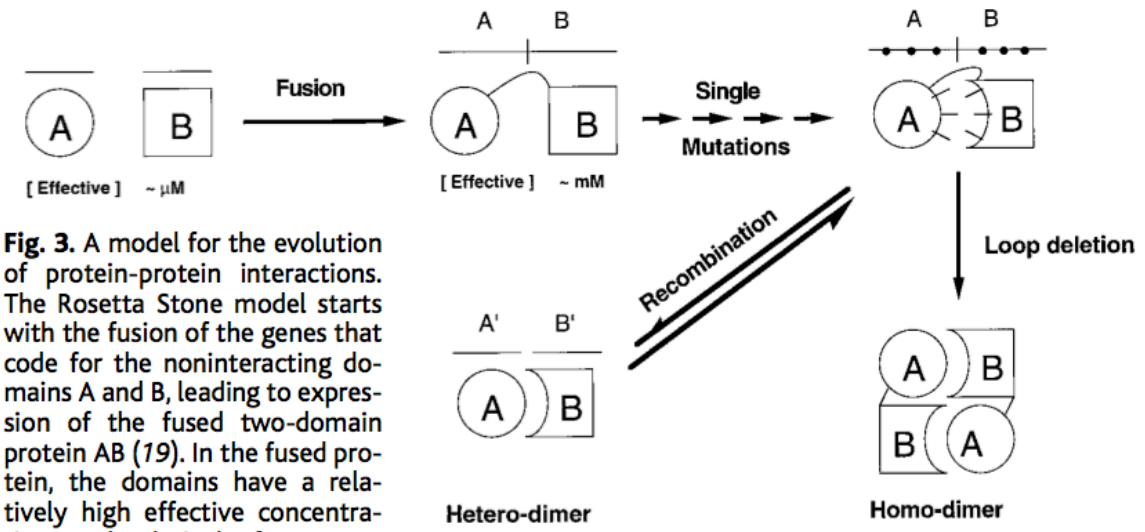
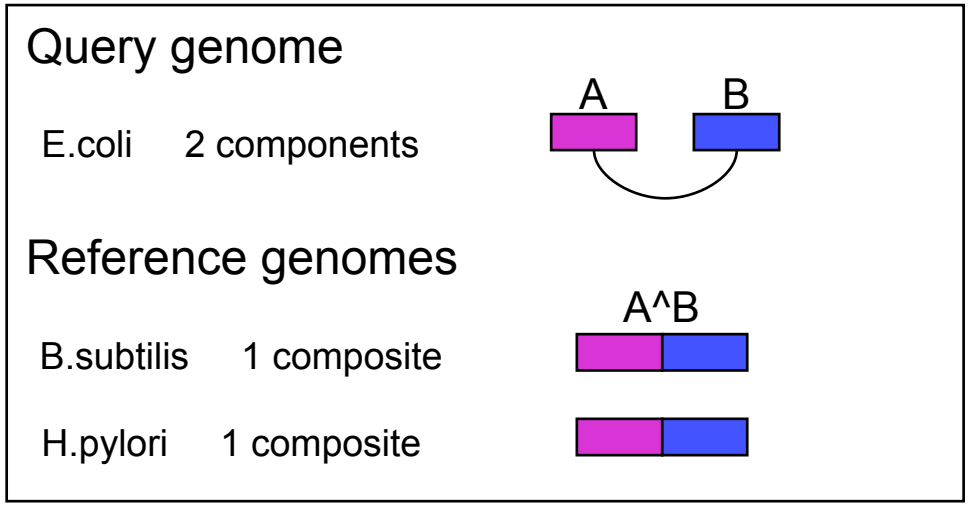


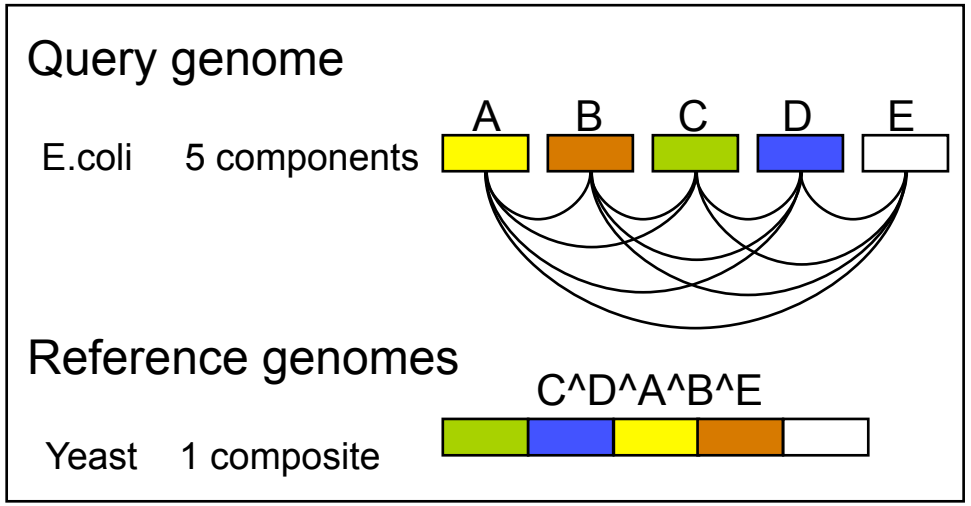
Fig. 3. A model for the evolution of protein-protein interactions. The Rosetta Stone model starts with the fusion of the genes that code for the noninteracting domains A and B, leading to expression of the fused two-domain protein AB (*19*). In the fused protein, the domains have a relatively high effective concentration, and relatively few mutations create a primitive binding site between the domains that is optimized by successive mutations. In the second line, the interacting domains are separated by recombination with another gene to create an interacting pair of proteins A and B. An interacting pair of proteins A and B can be created by fission of a protein, so that the preliminary fusion step is not essential to the Rosetta Stone hypothesis. The lower righthand step shows another possible mutation, a loop deletion that leads to a domain-swapped homodimer. This evolutionary path to homooligomers has been discussed earlier (*20*) and is the analog for homooligomers of the evolutionary path suggested here for heterooligomers.

- 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.

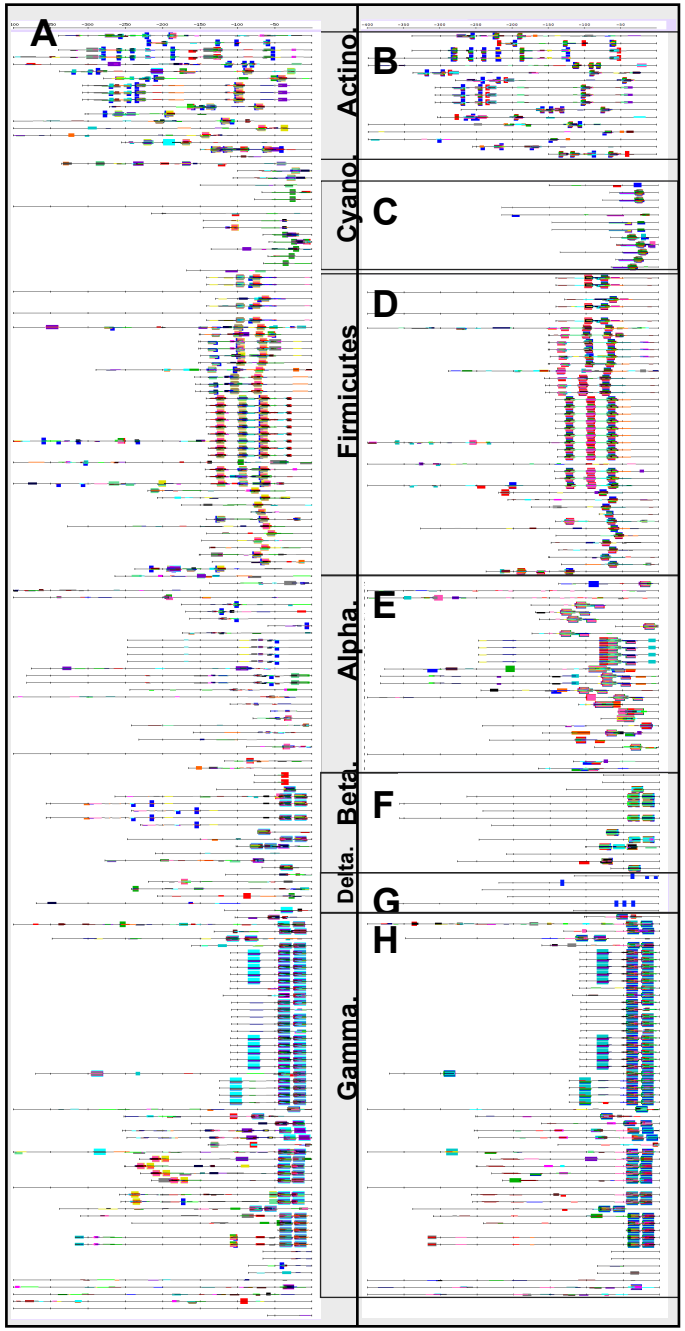
Operons and directons

Predicting operons in Bacterial genomes

- In Bacterial genomes, genes are organized in operons (poly-cistronic transcription units): a single mRNA contains a series of coding sequences.
- Operons thus contain groups of co-expressed genes, which are frequently (but not always) functionally related.

Phylogenetic footprints

Significant dyads in promoters of lexA orthologs in Bacteria



- When all the bacterial promoters are analyzed together, the program dyad-analysis detects most of taxon-specific motifs discussed before, and the feature-map highlights their taxon-specific locations.
- This illustrates the robustness of the method: the motifs can be detected even if present in a subset of the sequences only.
- The significance is however lower when all sequences are analyzed together than with the taxon-per-taxon analysis.

Janky, R. and van Helden, J. Evaluation of phylogenetic footprint discovery for predicting bacterial cis-regulatory elements and revealing their evolution BMC Bioinformatics **9**, 37 (2008).

Suggested readings

- Homology
  - Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997). A genomic perspective on protein families. Science 278, 631-7.
  - Fitch, W. M. (2000). Homology a personal view on some of the problems. Trends Genet 16, 227-31.
  - Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39, 309-38.
  - Zvelebil, M. J. and Baum, J. O. (2008). Understanding Bioinformatics. Garland Science: New York and London.
- Phylogenetic profiles
  - Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997). A genomic perspective on protein families. Science 278, 631-7.
  - Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96, 4285-8.
- Gene fusion analysis
  - Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O. & Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. Science 285, 751-3.