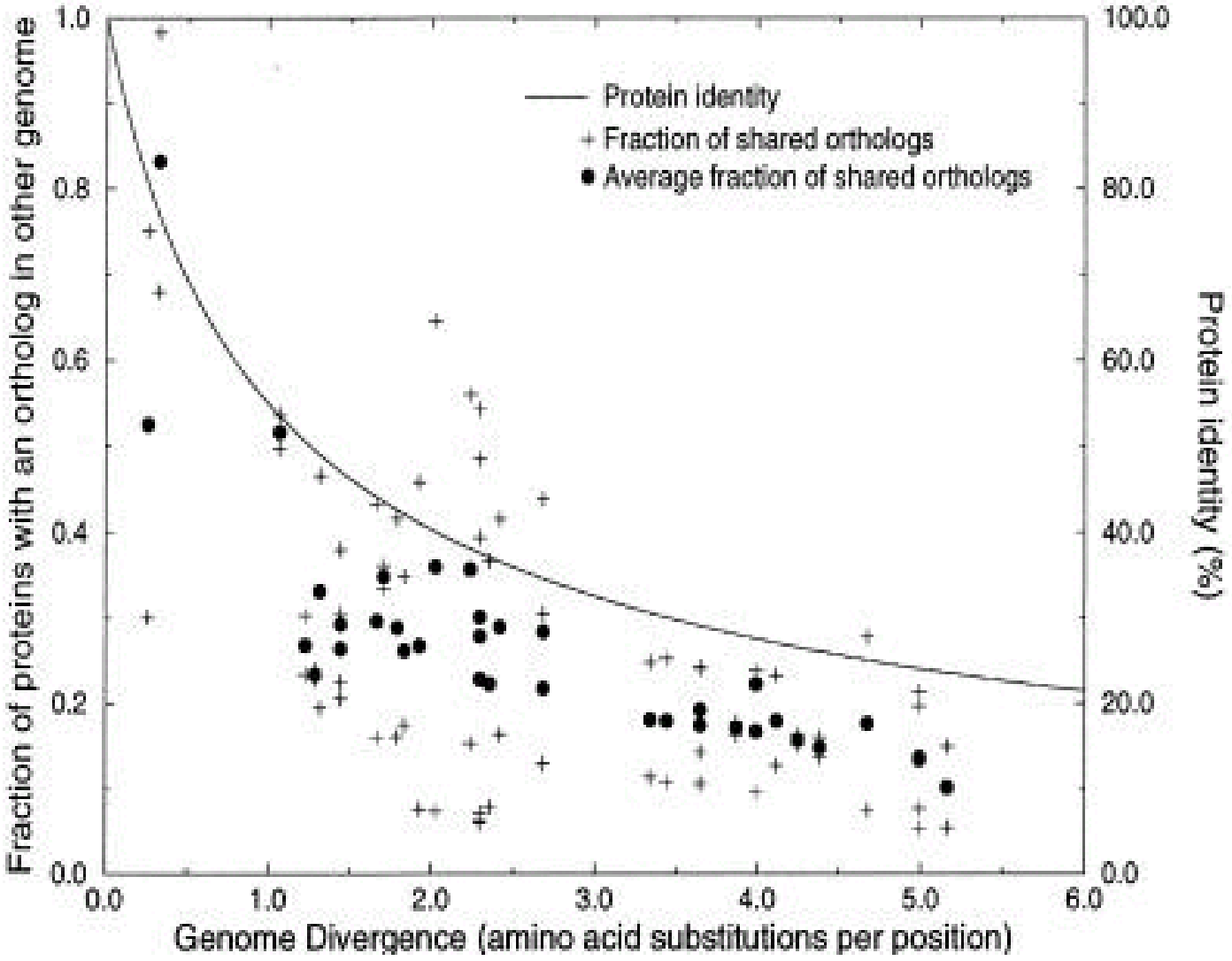
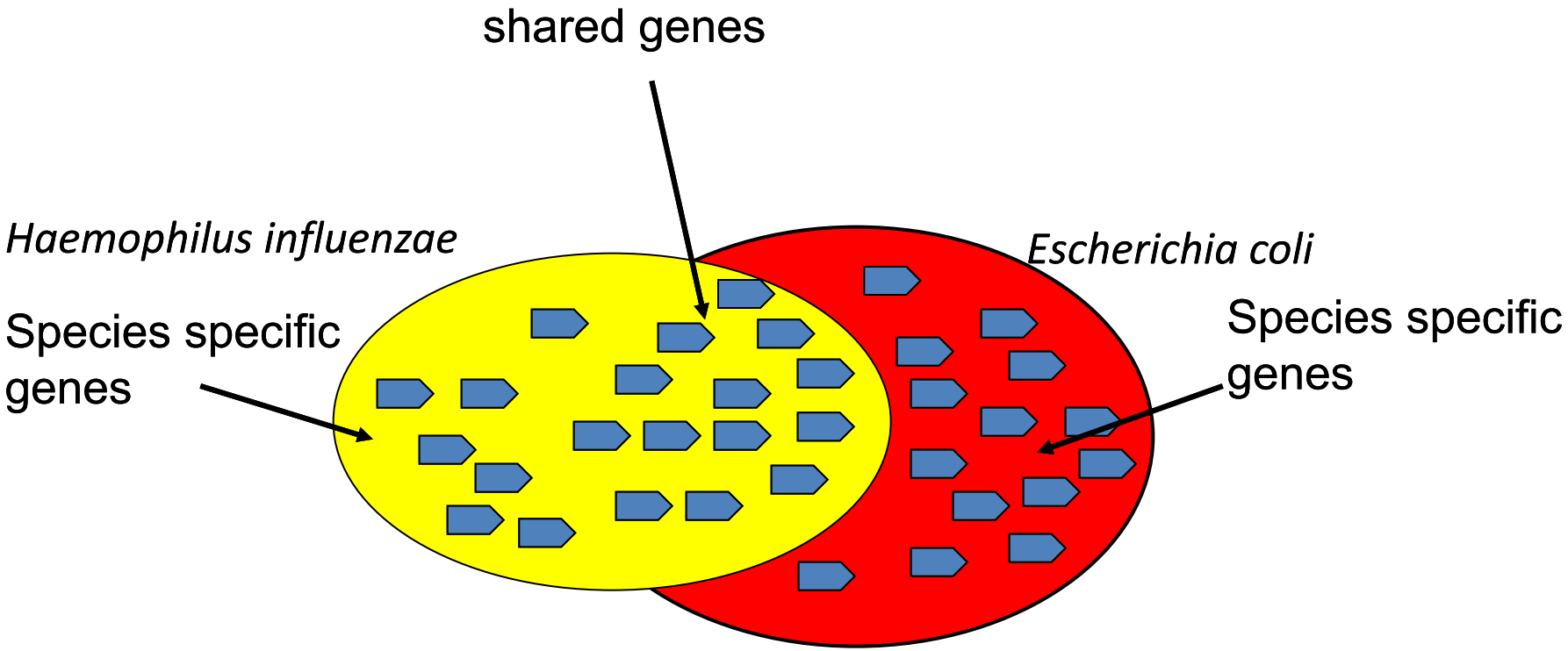
Bioinformatics and Evolutionary Genomics:
Genome Evolution in terms of Gene Content

Gene Content Evolution



What about HGT / genome sizes?
Genome trees based on gene content:



Genome trees based on gene content

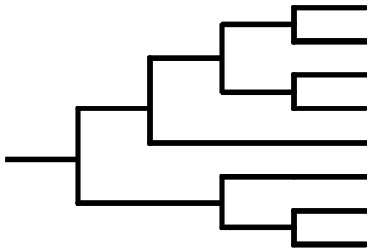
Presence / absence matrix:

	OG1	OG2	OG3	OG4	…
sp1	1	1	0	1	…
sp2	0	1	0	0	…
sp3	0	0	1	1	…
…	…	…	…	…	

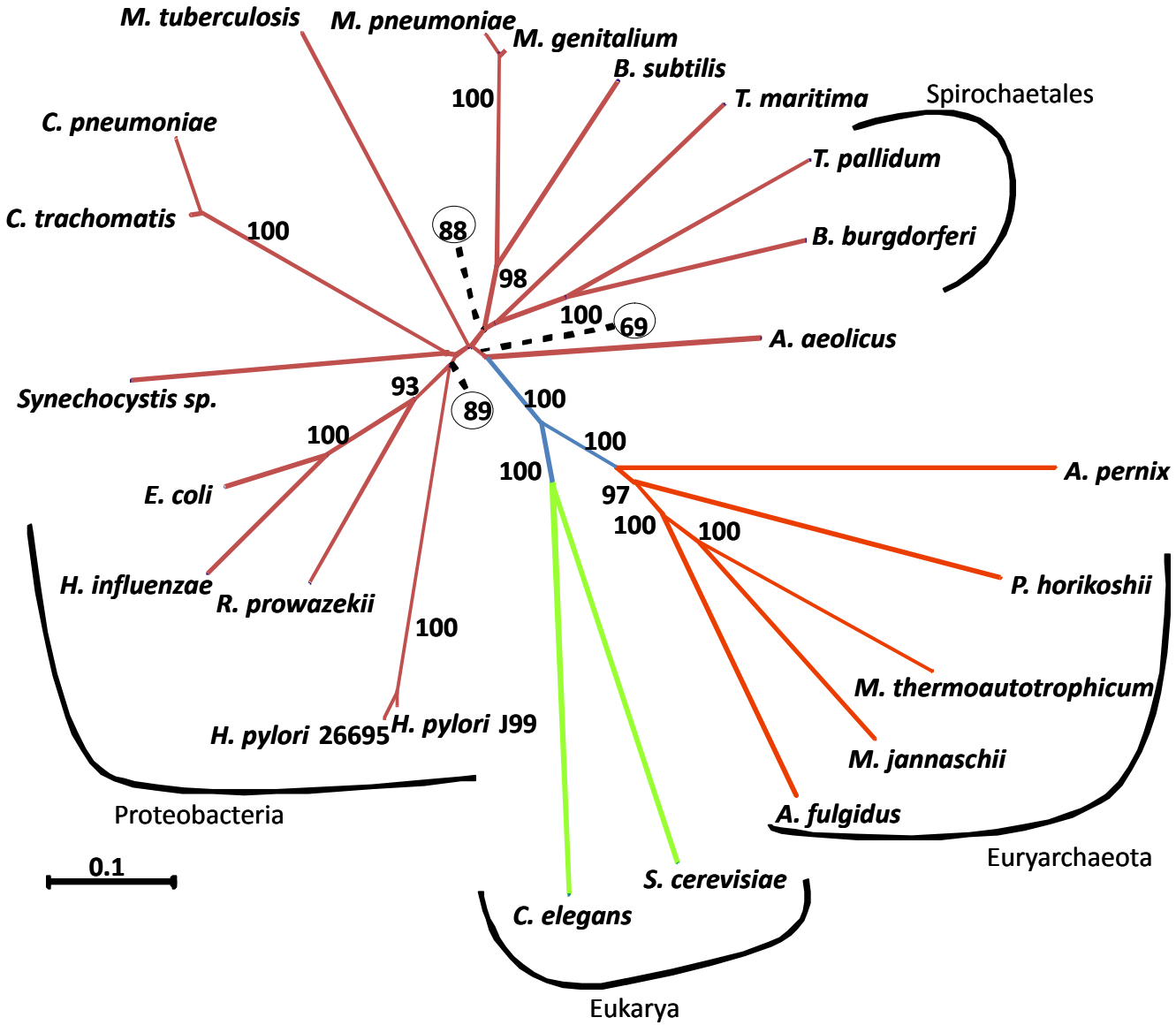
$$dist\,(spA, spB) = 1 - \left(\frac{\#\ shared\ OGs\ (spA, spB)}{\text{Size of the smallest genome}} \right)$$

d\s	sp1	sp2	sp3	sp4	…
sp1	0\1	0.2	0.4	0.2	…
sp2	0.8	0\1	0.9	0.1	…
sp3	0.6	0.1	0\1	0.3	…
sp4	0.8	0.9	0.7	0\1	…
…	…	…	…	…	

Neighbor joining →

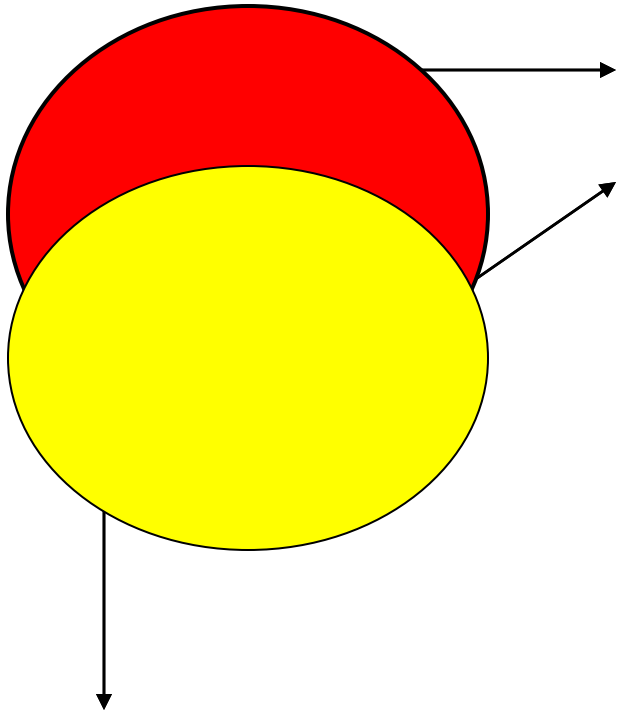


Genome trees based on gene content are remarkably similar to consensus on ToL



Presence / absence of genes

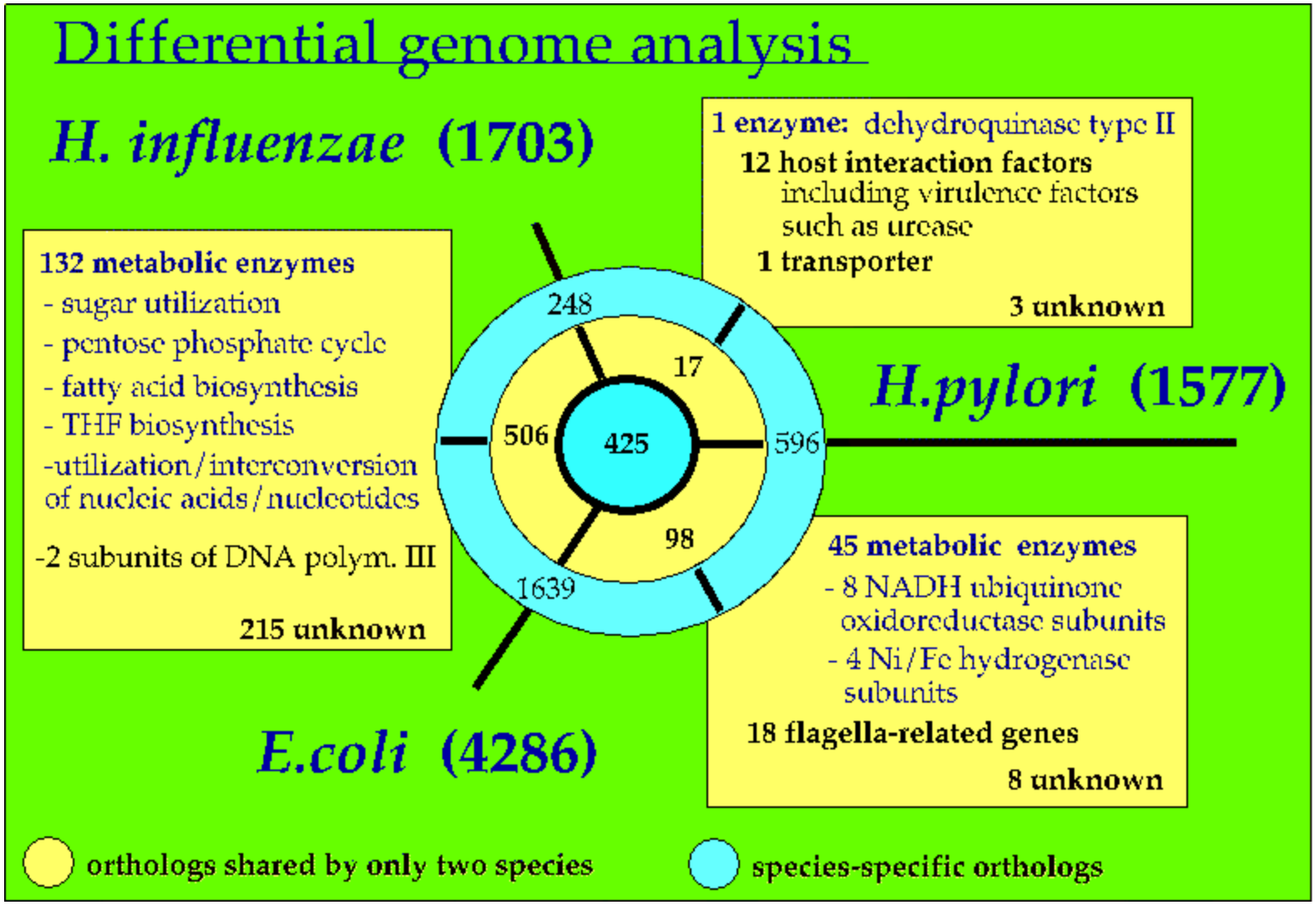
Gene content → co-evolution. (The easy case, few genomes.)



Differences between gene
Content reflect differences in
Phenotypic potentialities

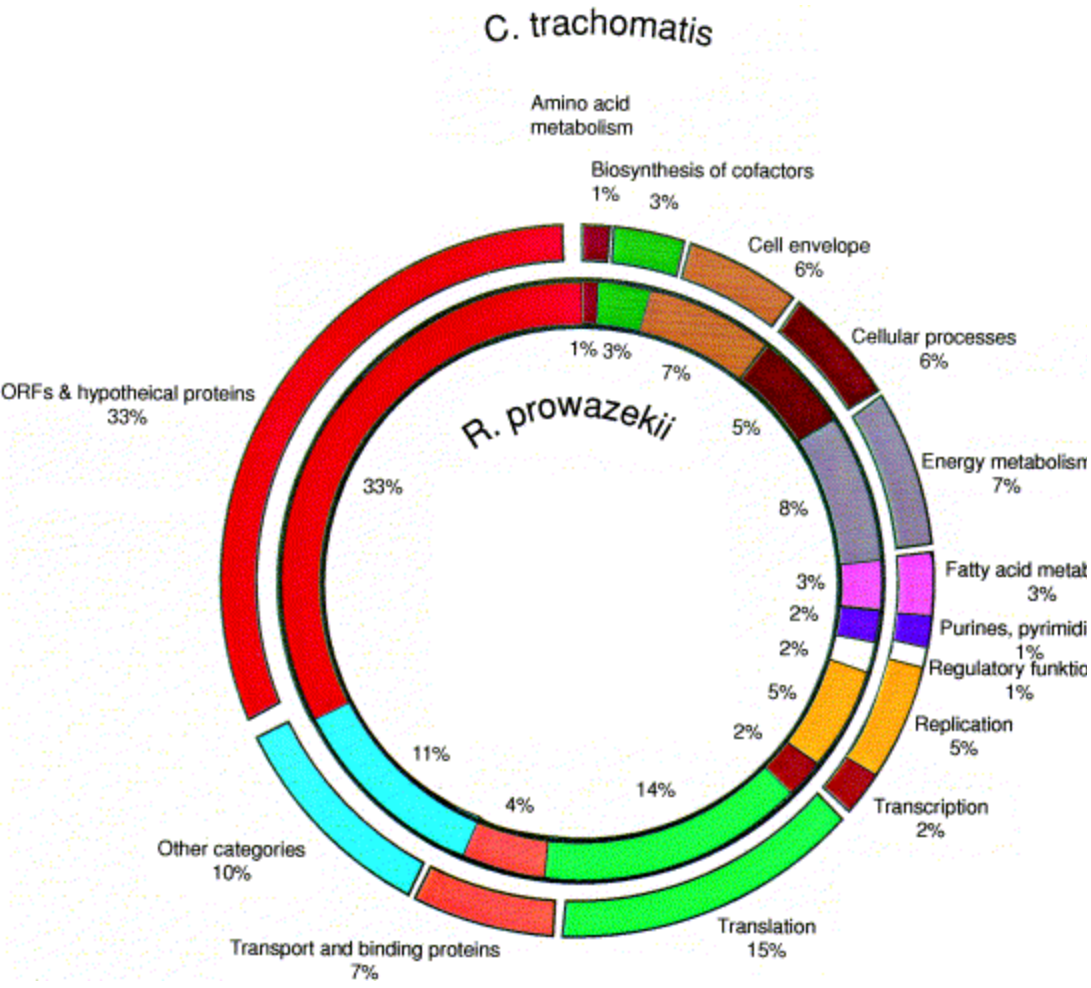
Genomes share genes for phenotypes they have in common

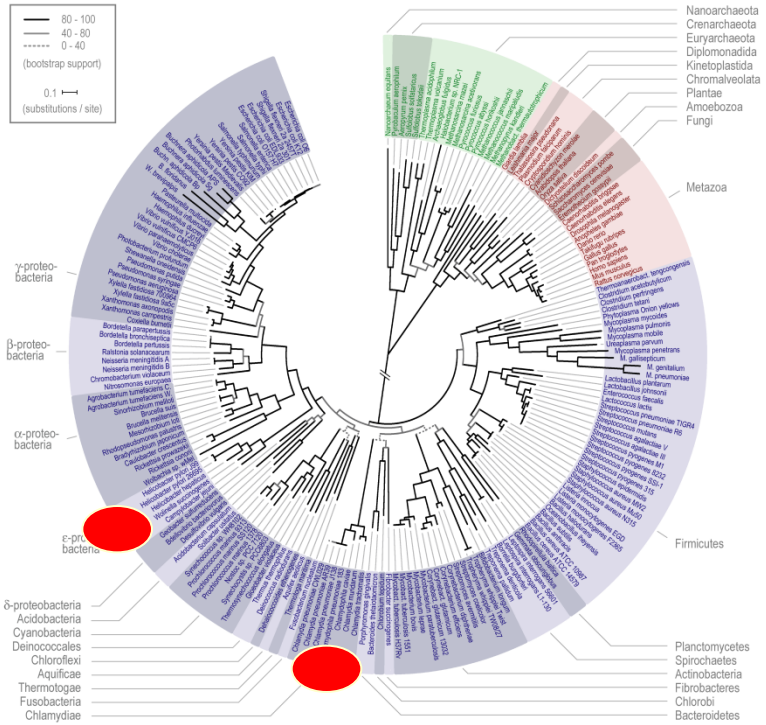
Three-way comparisons



Huynen et al., 1998, FEBS Lett

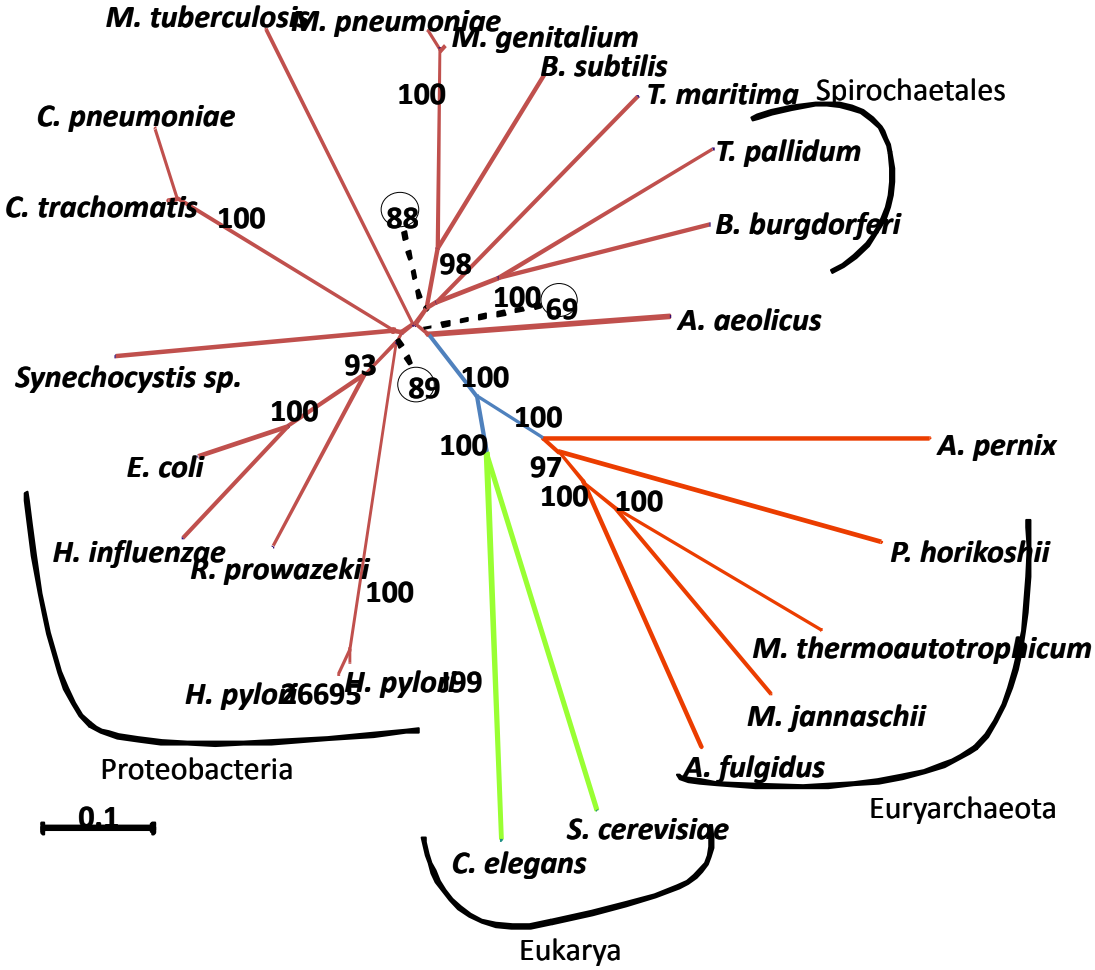
Convergence in functional classes of gene content in small intracellular bacterial parasites



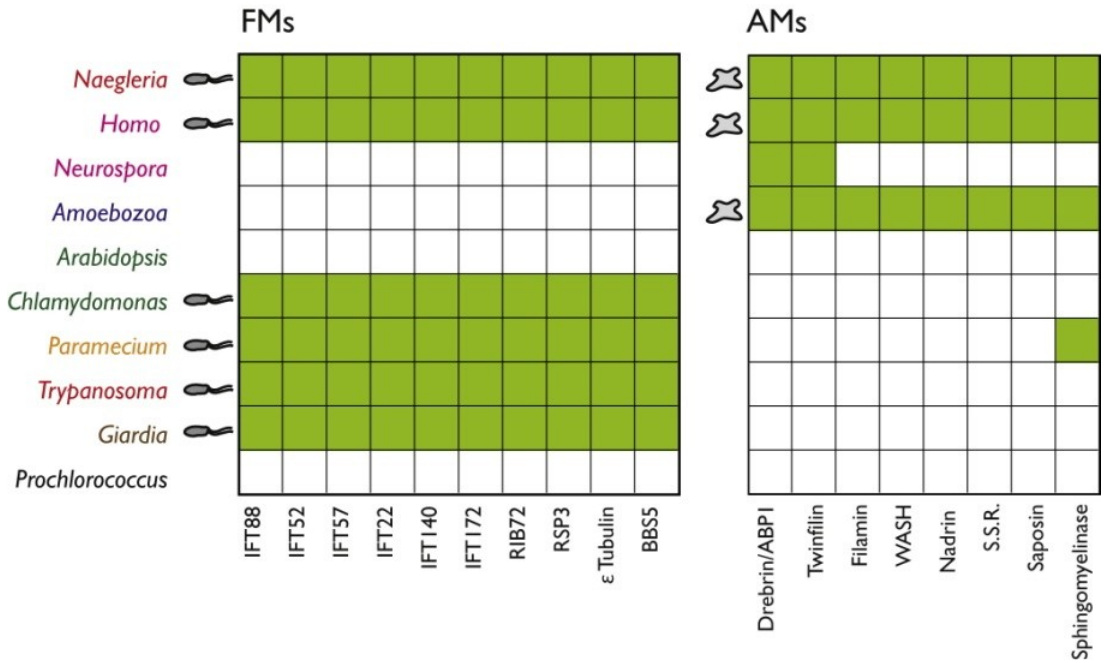


Zomorodipour & Andersson FEBS Letters 1999

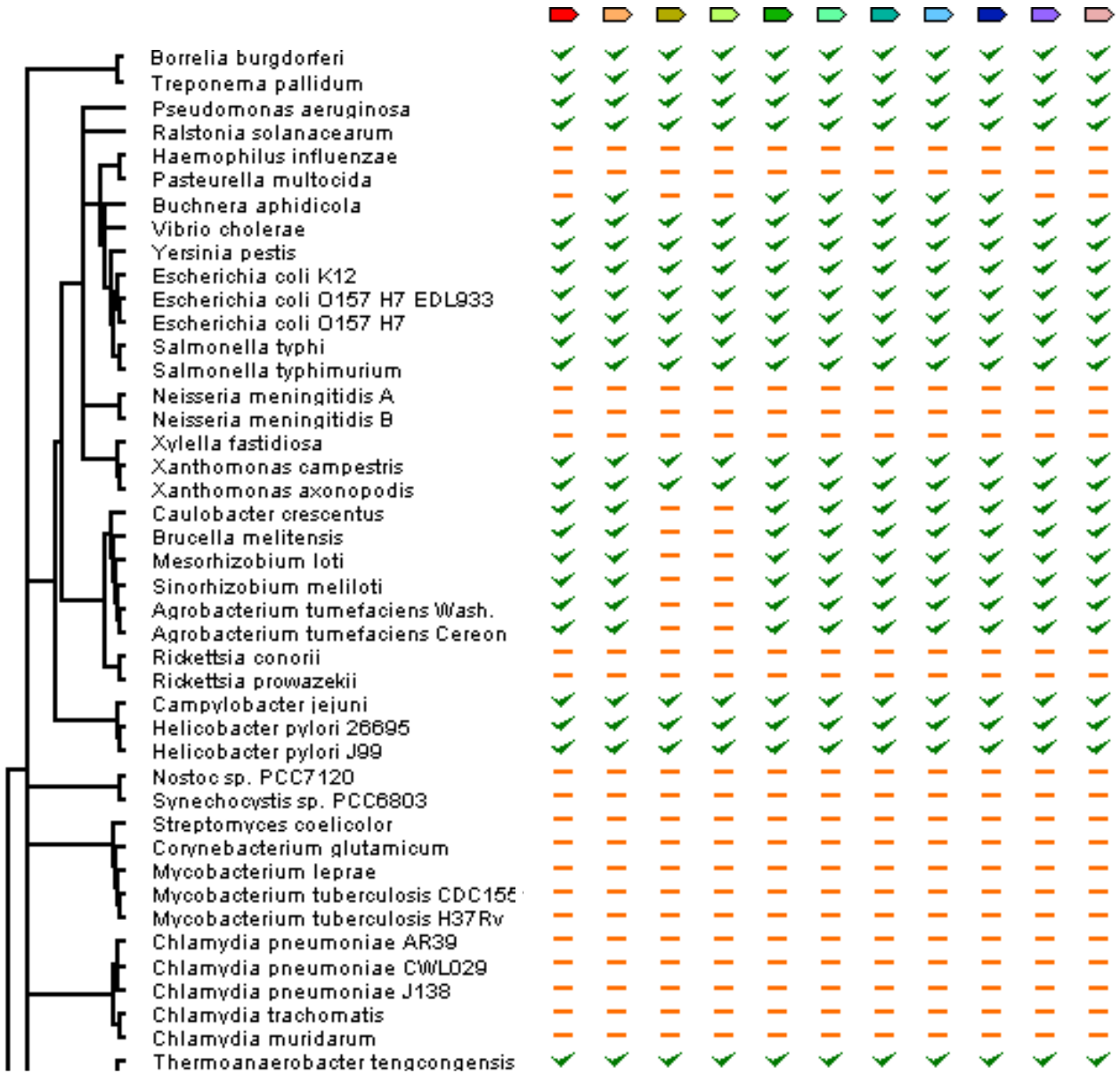
Although we can, qualitatively, interpret the variations in shared gene content in terms of the phenotypes of the species, quantitatively they depend on the relative phylogenetic positions of the species. The closer two species are the larger fraction of their genes they share.



Co-evolving modules



Co-occurrence of genes across genomes as prediction for interaction / association

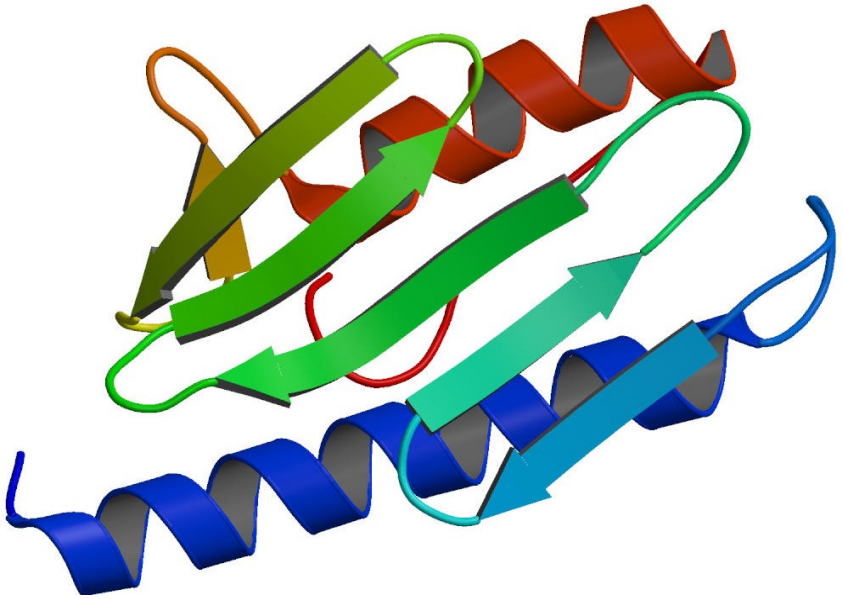


- i.e. two genes have the same presence/ absence pattern over multiple genomes:
- AKA phylogenetic profiles
- NB complete genomes absence -> needed for absence
- Correction for phylogenetic signal needed → events

b

Predicting function of a disease gene protein with unknown function, frataxin, using co-occurrence of genes across genomes / phylogenetic profiles

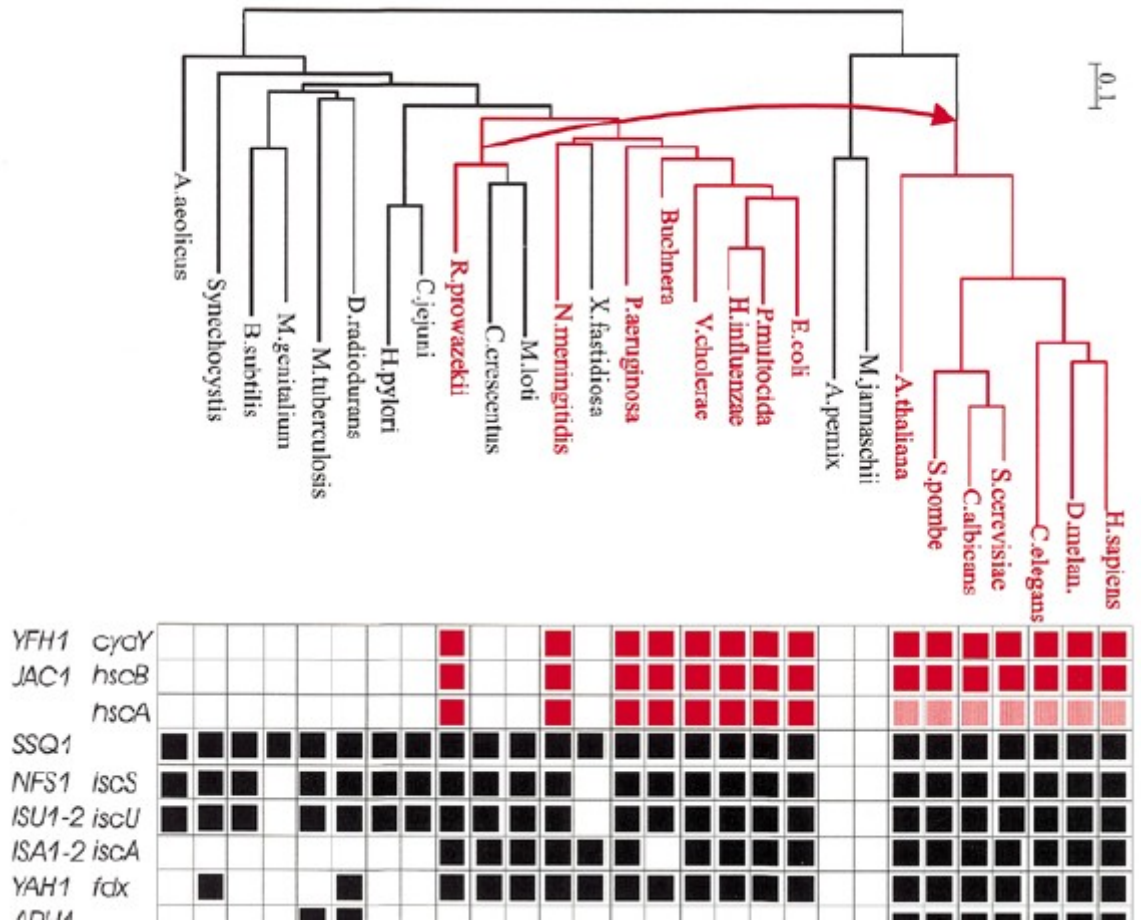
- Friedreich's ataxia
- No (homolog with) known function

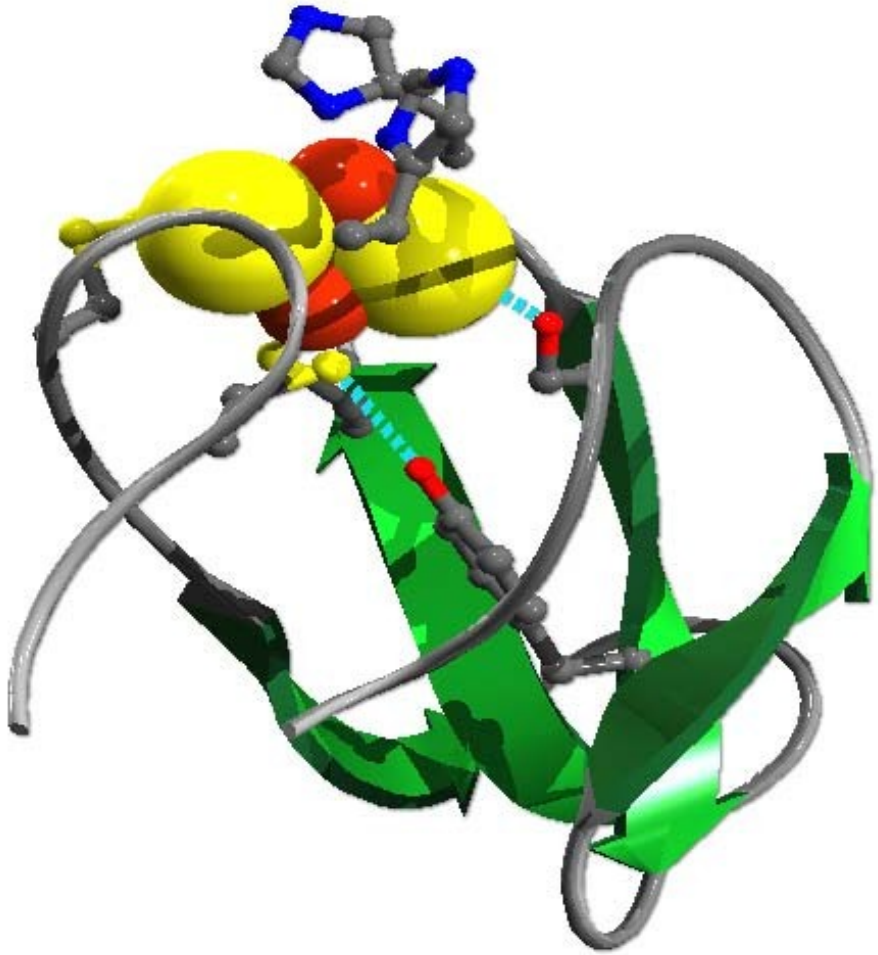


Predicting function of a disease gene protein with unknown function, frataxin, using co-occurrence of genes across genomes

- Friedreich's ataxia
- No (homolog with) known function

Frataxin has co-evolved with hscA and hscB indicating that it plays a role in iron-sulfur cluster assembly





Iron-Sulfur (2Fe-2S) cluster in the Rieske protein

Prediction:

Human Molecular Genetics, 2001, Vol. 10, No. 21 **2463–2468**

The phylogenetic distribution of frataxin indicates a role in iron-sulfur cluster protein assembly

Martijn A. Huynen*, Berend Snel[1], Peer Bork and Toby J. Gibson[1]

Biocomputing, EMBL/Max-Delbrueck-Center fur molecular medicin, Berlin-Buch and [1]Biocomputing, EMBL, Meyerhofstrasse 1, 69117 Heidelberg, Germany

Received July 3, 2001; Revised and Accepted July 30, 2001

~Confirmation:

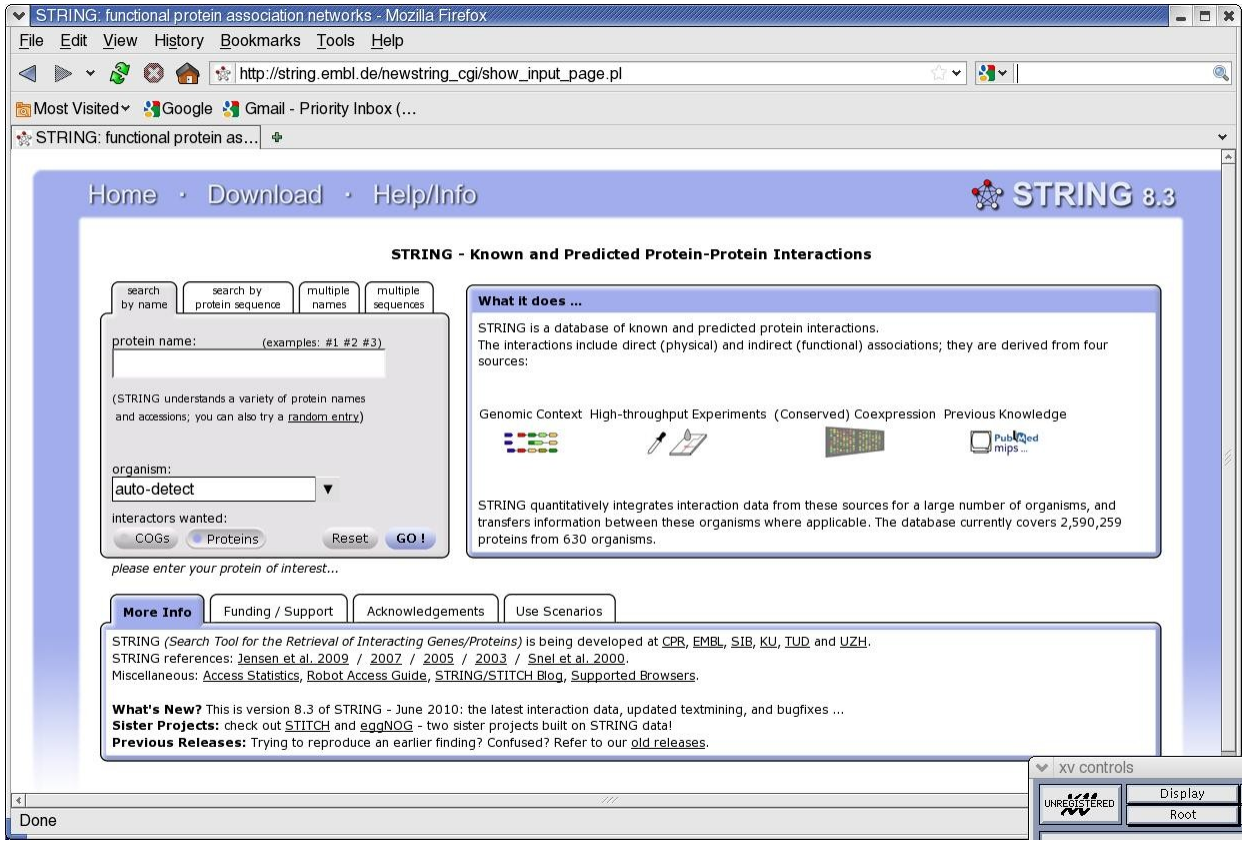
J|A|C|S
ARTICLES
Published on Web 04/26/2003

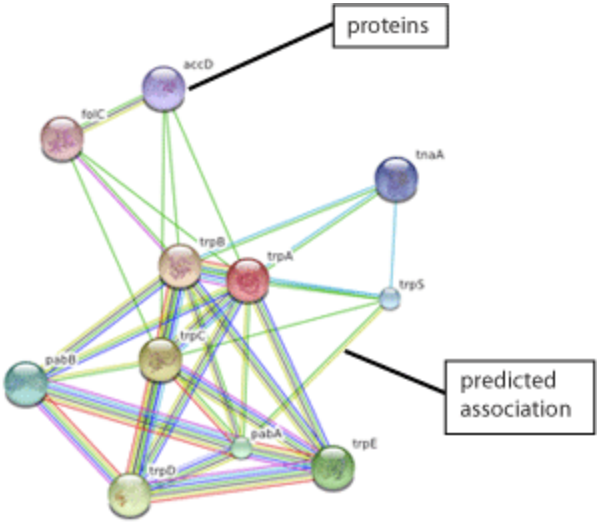
Iron−Sulfur Cluster Biosynthesis. Characterization of Frataxin as an Iron Donor for Assembly of [2Fe-2S] Clusters in ISU-Type Proteins

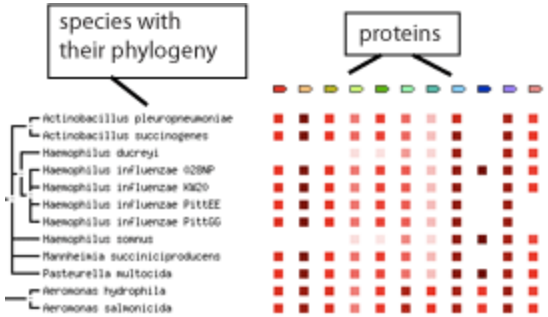
Taejin Yoon and J. A. Cowan*

Contribution from Evans Laboratory of Chemistry, The Ohio State University, 100 West 18th Avenue, Columbus, Ohio 43210

Received August 1, 2002; E-mail: cowan@chemistry.ohio-state.edu

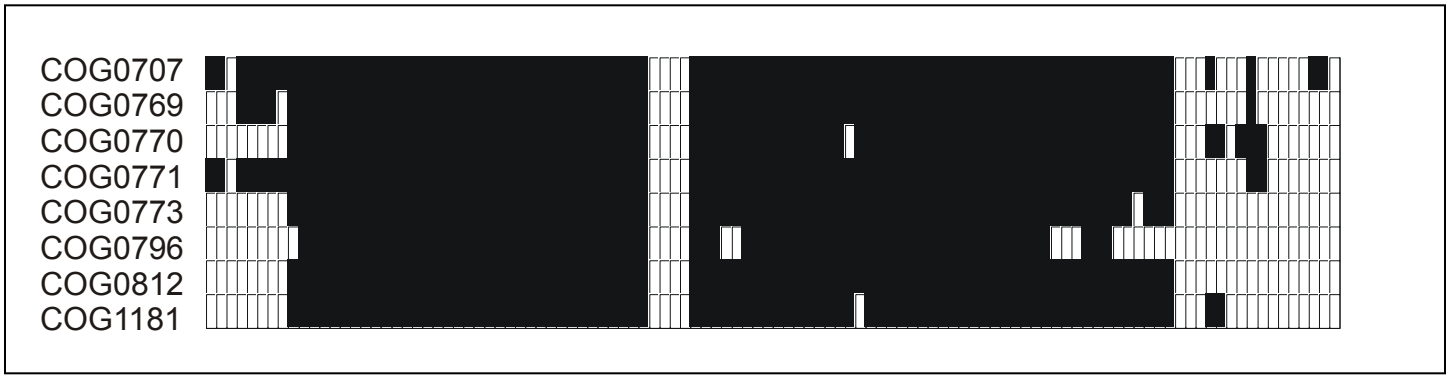

STRING: functional protein association networks - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://string.embl.de/newstring_cgi/show_input_page.pl
Most Visited Google Gmail - Priority Inbox (...
STRING: functional protein as...
Home · Download · Help/Info
STRING 8.3
STRING - Known and Predicted Protein-Protein Interactions
search by name
search by protein sequence
multiple names
multiple sequences
protein name:
(examples: #1 #2 #3)
(STRING understands a variety of protein names and accessions; you can also try a random entry)
organism:
auto-detect
interactors wanted:
COGs
Proteins
Reset
GO !
please enter your protein of interest...
What it does ...
STRING is a database of known and predicted protein interactions.
The interactions include direct (physical) and indirect (functional) associations; they are derived from four sources:
Genomic Context
High-throughput Experiments
(Conserved) Coexpression
Previous Knowledge
STRING quantitatively integrates interaction data from these sources for a large number of organisms, and transfers information between these organisms where applicable. The database currently covers 2,590,259 proteins from 630 organisms.
More Info
Funding / Support
Acknowledgements
Use Scenarios
STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is being developed at CPR, EMBL, SIB, KU, TUD and UZH.
STRING references: Jensen et al. 2009 / 2007 / 2005 / 2003 / Snel et al. 2000.
Miscellaneous: Access Statistics, Robot Access Guide, STRING/STITCH Blog, Supported Browsers.
What's New? This is version 8.3 of STRING - June 2010: the latest interaction data, updated textmining, and bugfixes ...
Sister Projects: check out STITCH and eggNOG - two sister projects built on STRING data!
Previous Releases: Trying to reproduce an earlier finding? Confused? Refer to our old releases.
xv controls
UNREGISTERED
Display
Root
Done


proteins
accD
folC
tnaA
trpB
trpA
trpS
trpC
pabB
pabA
trpE
trpD
predicted association

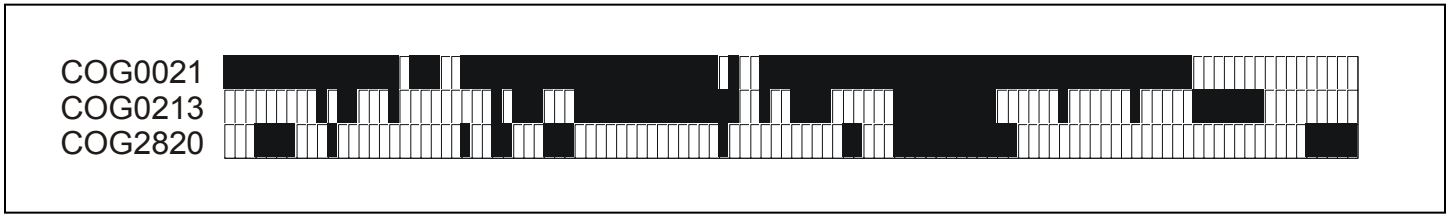

species with their phylogeny
proteins
Actinobacillus pleuropneumoniae
Actinobacillus succinogenes
Haemophilus ducreyi
Haemophilus influenzae 02BNP
Haemophilus influenzae KW20
Haemophilus influenzae PittEE
Haemophilus influenzae PittGG
Haemophilus somnus
Mannheimia succiniciproducens
Pasteurella multocida
Aeromonas hydrophila
Aeromonas salmonicida

- If two genes have a “significantly” similar presence/absence pattern which is different from the phylogenetic signal, than their proteins are likely to interact / be in the same process/pathway
- This pattern can be created by independent loss and/or horizontal gene transfer (of operons)

However



peptidoglycan biosynthesis pathway (highly cohesiveness, far from perfect)



ribose phosphate metabolism (not cohesive at all)

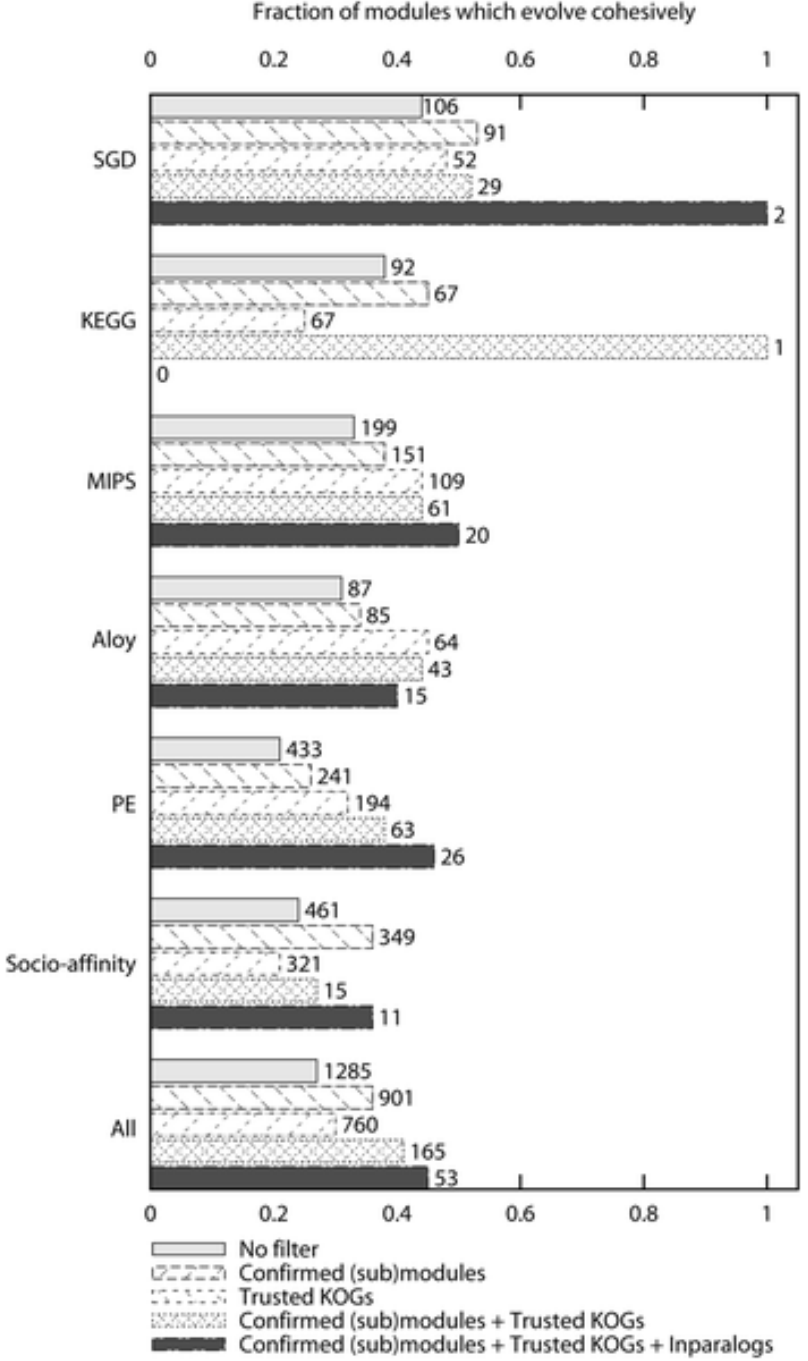
Very few functional modules are perfect; limited cohesiveness; functional units vs evolutionary units

Why?

If the phyletic patterns of two proteins are highly similar they tend to interact, but the reverse is not generally true!

~50% of modules (such as protein complexes) do not have highly similar phyletic patterns.

This seems to not only depend on dataset, noise in orthology detection, or noise in module definition



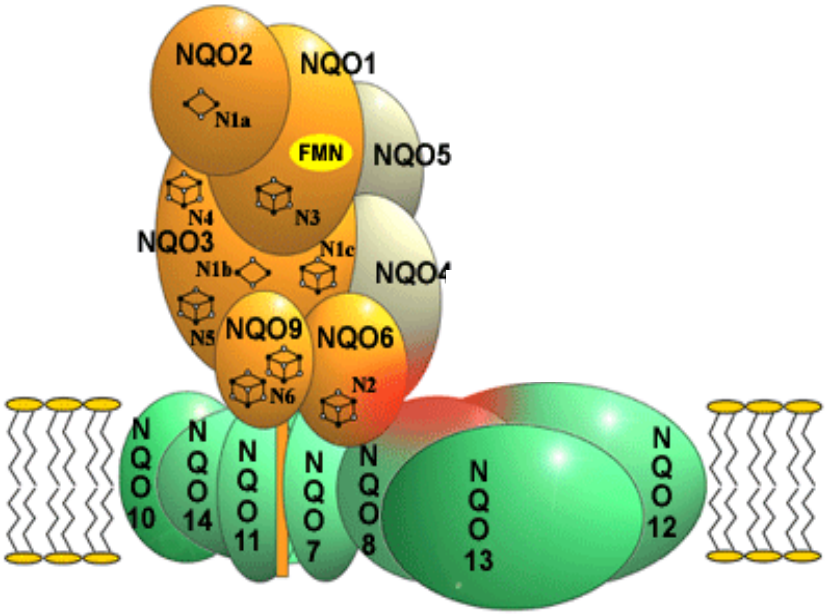
Fokkens and Snel *PLoS Comp Biol* 2009

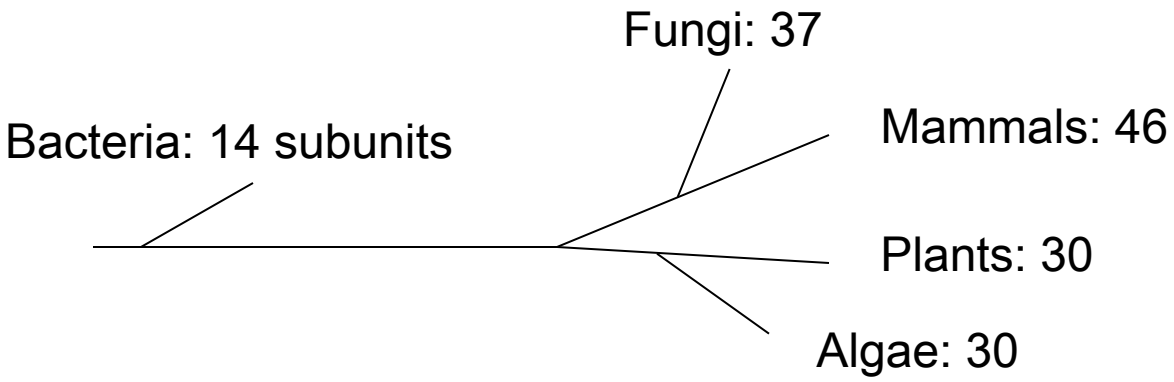
Explaining discordant phyletic patterns of proteins that interact

- Many cases stories and a few large scale studies

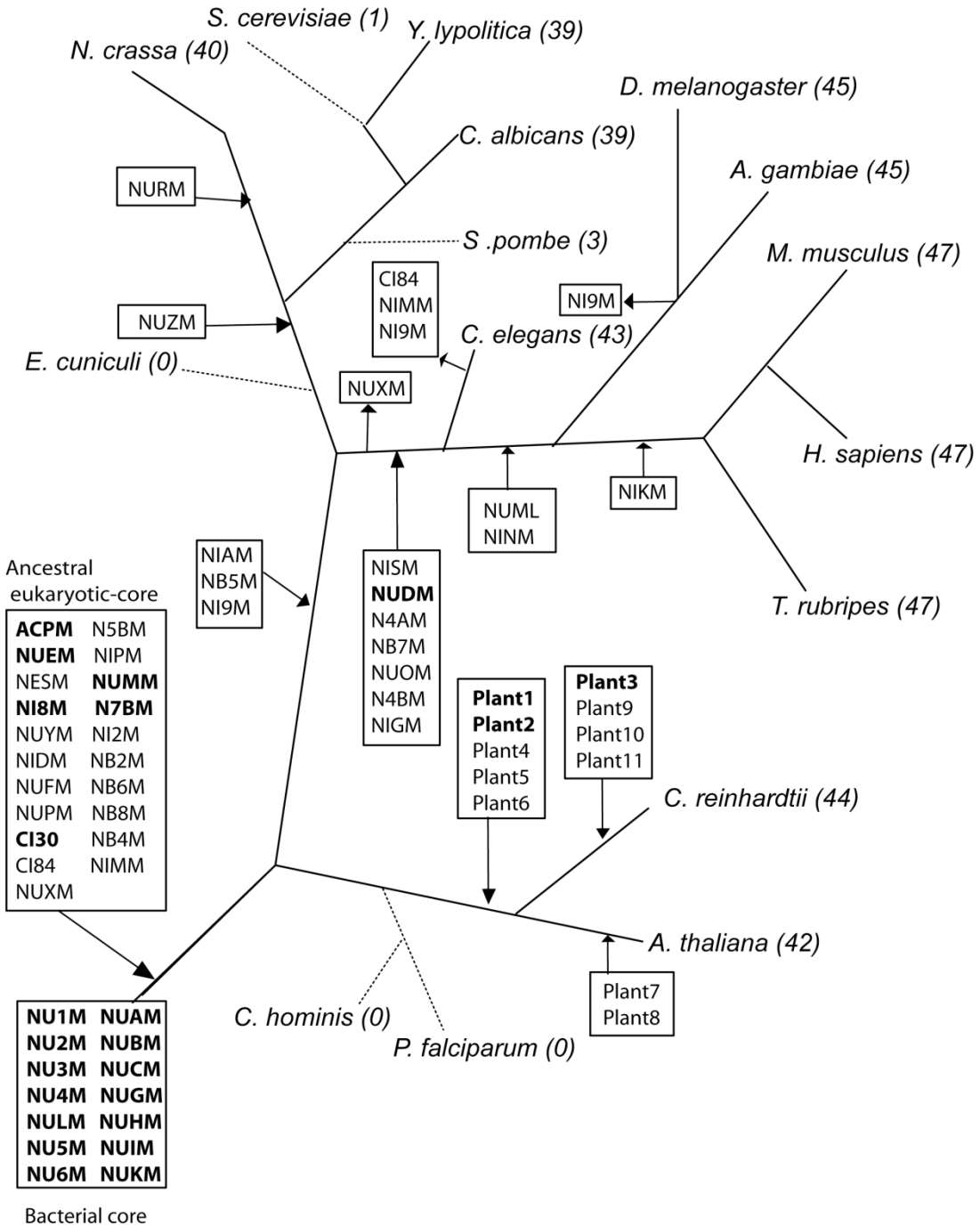
- (we could also just say that evolution is flexible and proteins change function; which I am not going to argue with but (A) conservation of interaction and (B) this is a “just so”, non testable explanation)

Tracing the evolution of NADH:ubiquinone oxidoreductase (Complex I of the oxidative phosphorylation), from 14 subunits (Bacteria) to 46 subunits (Mammals) by comparative genome analysis



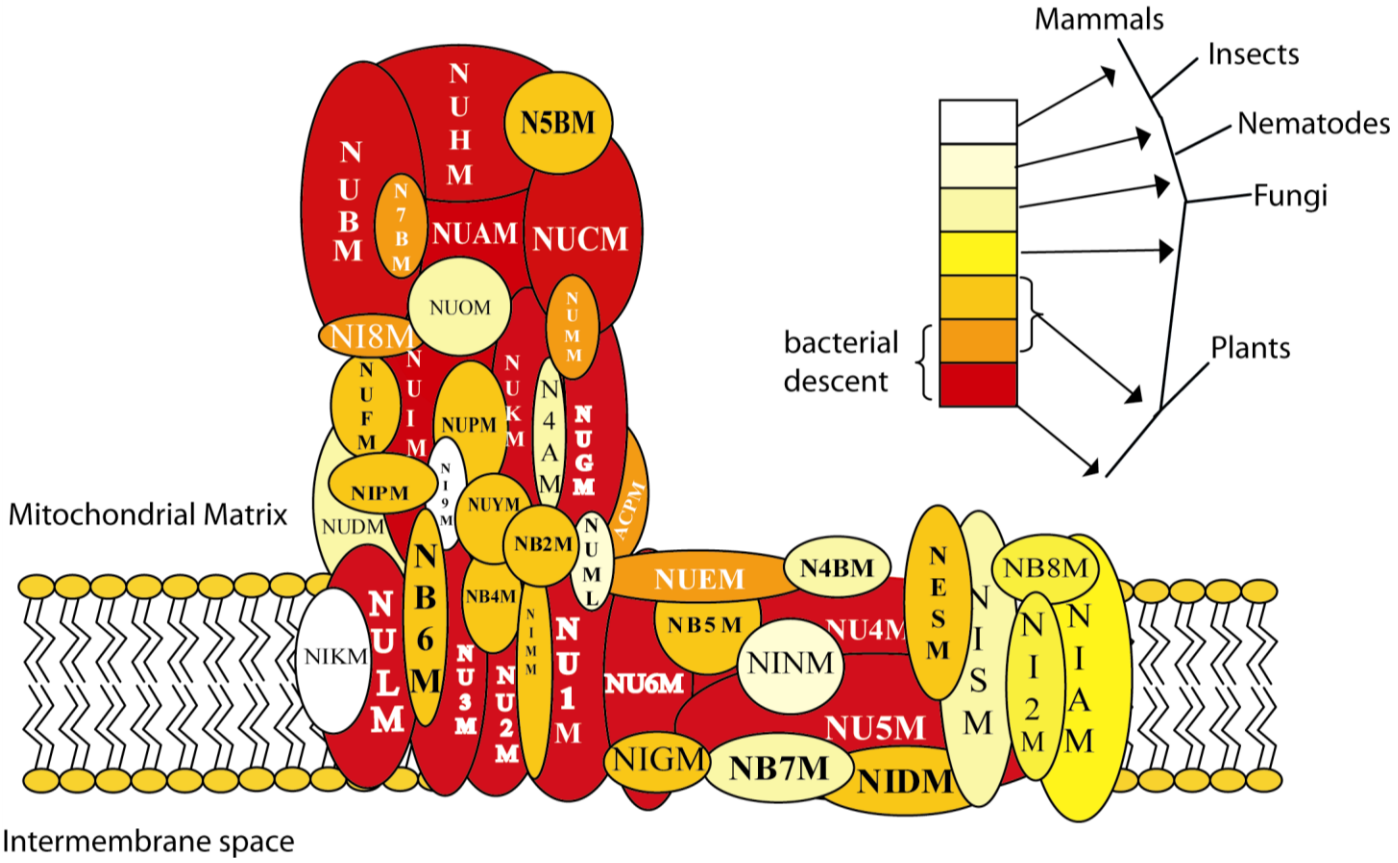


Gabaldon et al, J. Mol. Biol 2005

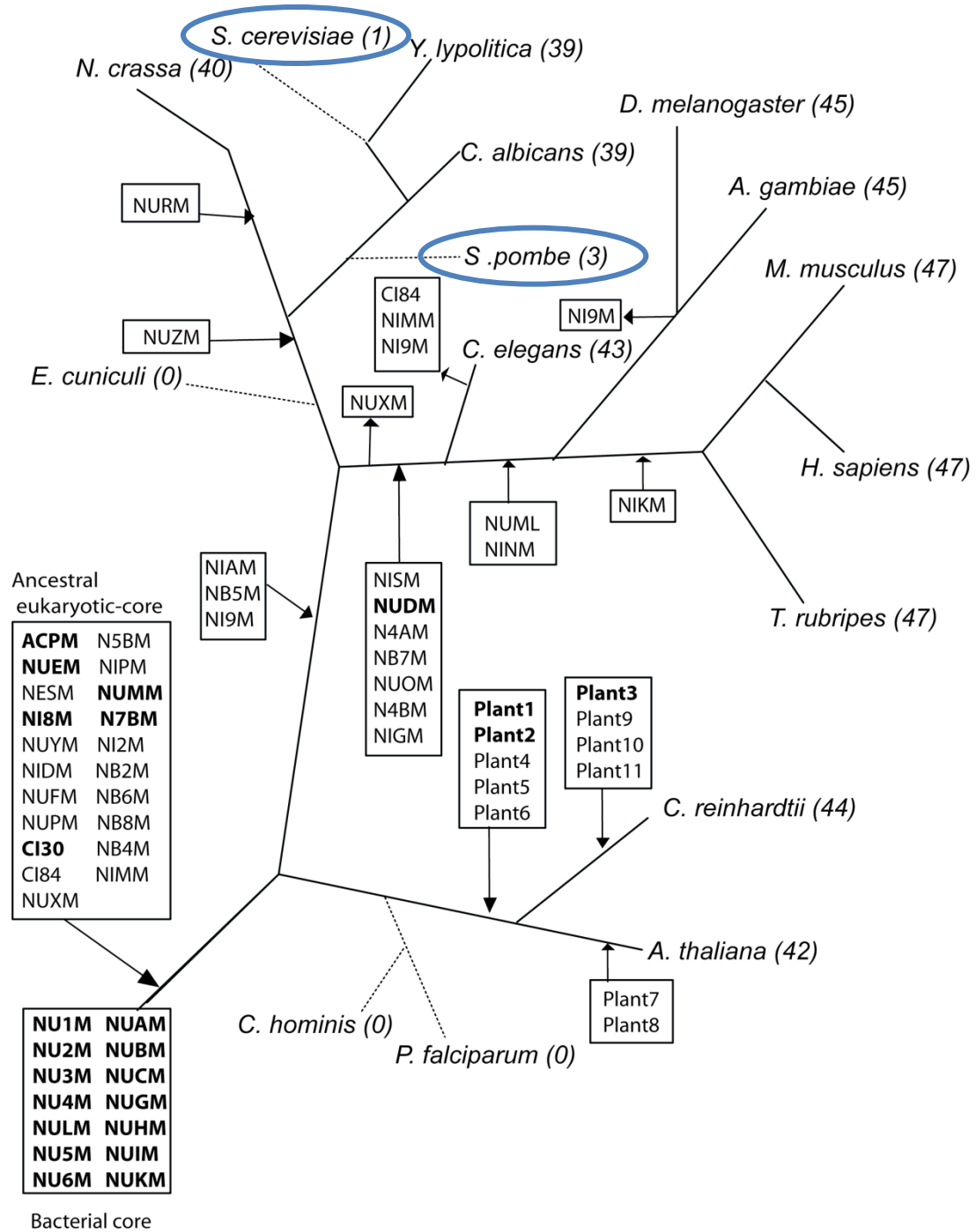


Reconstructing Complex I evolution by mapping the variation onto a phylogenetic tree. After an initial “surge” in complexity (from 14 to 35 subunits in early eukaryotic evolution) new subunits have been gradually added and incidentally lost. most other loss is large scale

In the eukaryotic evolution of Complex I, new subunits have been added “all over” the complex



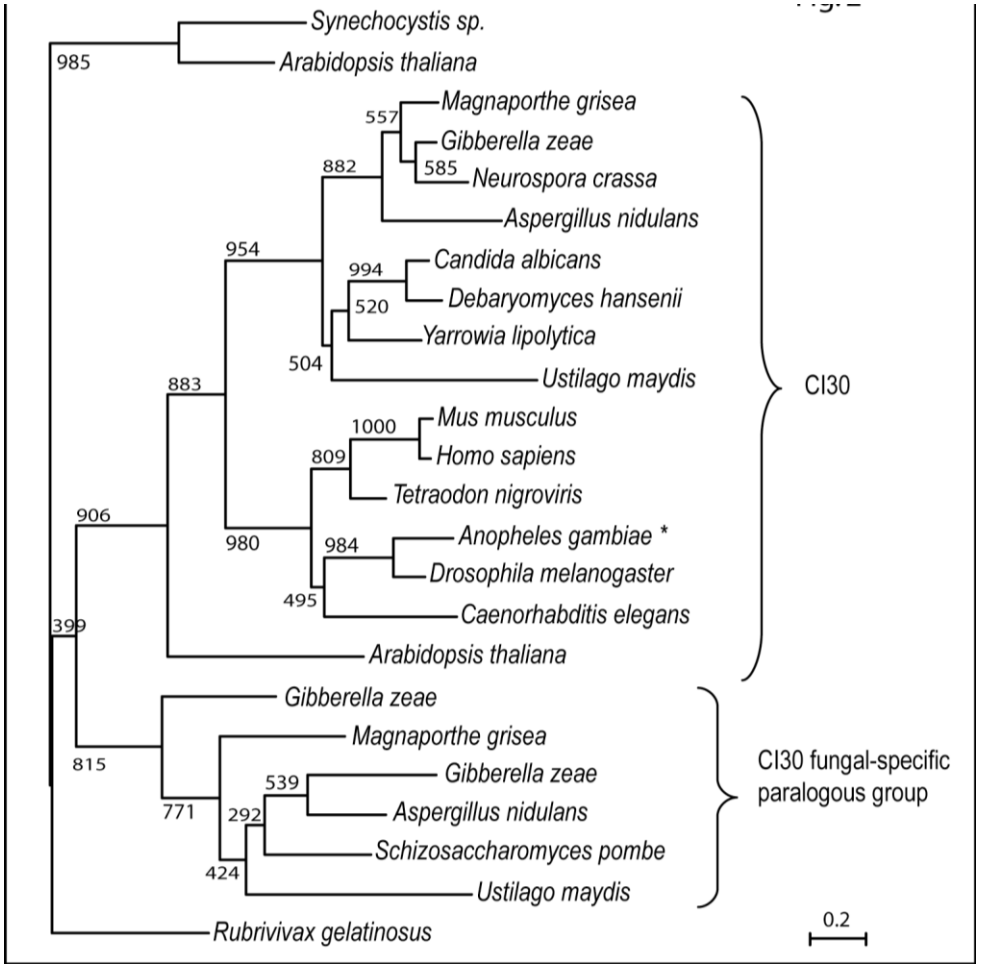
Gabaldon et al, J. Mol. Biol 2005



Reconstructing Complex I evolution by mapping the variation onto a phylogenetic tree. After an initial "surge" in complexity (from 14 to 35 subunits in early eukaryotic evolution) new subunits have been gradually added and incidentally lost., most other loss is large scale

Complex I loss is not always "complete", S.cerevisiae and S.pombe have retained 1 and 3 proteins

Phylogeny of a “remaining” complex I protein in pombe

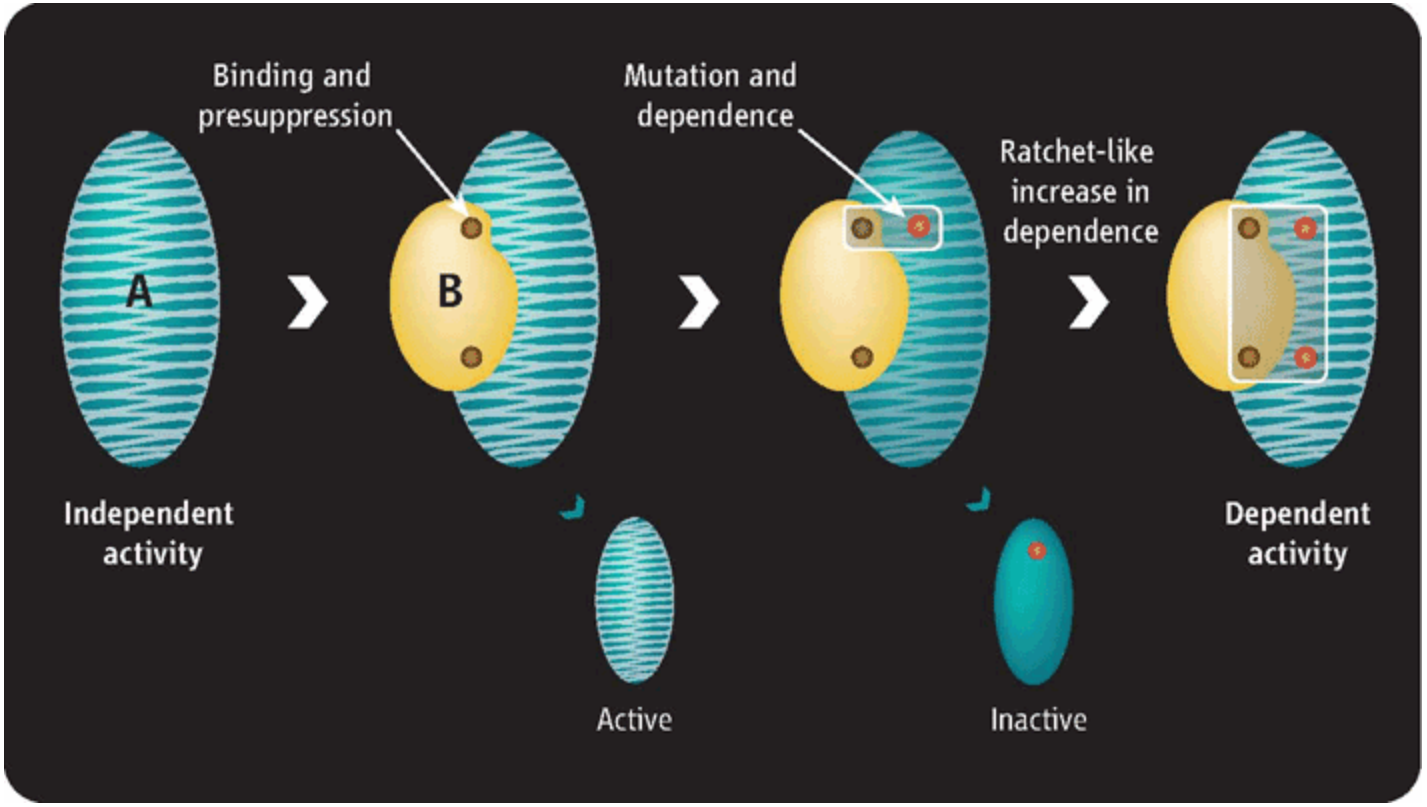


?

“The Complex I assembly protein CI30 has been duplicated in the Fungi.
This can explain the presence of a CIA30-homolog in Complex I-less *S.pombe*”

Gabaldon et al, J. Mol. Biol 2005

Why the accumulation: a neutral explanation



fixation of neutral or slightly deleterious features as a general and unavoidable source of complexity in taxa with small populations

Science. 2010 Nov 12;330(6006):920-1.
Cell biology. Irremediable complexity?
Gray MW, Lukes J, Archibald JM, Keeling PJ, Doolittle WF.

How to falsify?

e.g. Neurospora mito-TyrRS

- *Neurospora* mitochondrial genome encodes several introns which require a tyrosyl tRNA synthetase (TyrRS) to splice.
- “to compensate for structural defects acquired by the intron sequences “
- BUT Introns with defects arising -> negative selection
- ? Reverse: first binding (fortuitously or for reason unrelated to splicing)—> accumulation of mutations in the intron that inactivate splicing, if TyrRS not bound.
- Because the compensatory / suppressive activity exists before mutation “presuppression,”
- the protein dependence by the intron could be selectively neutral (or slightly disadvantageous

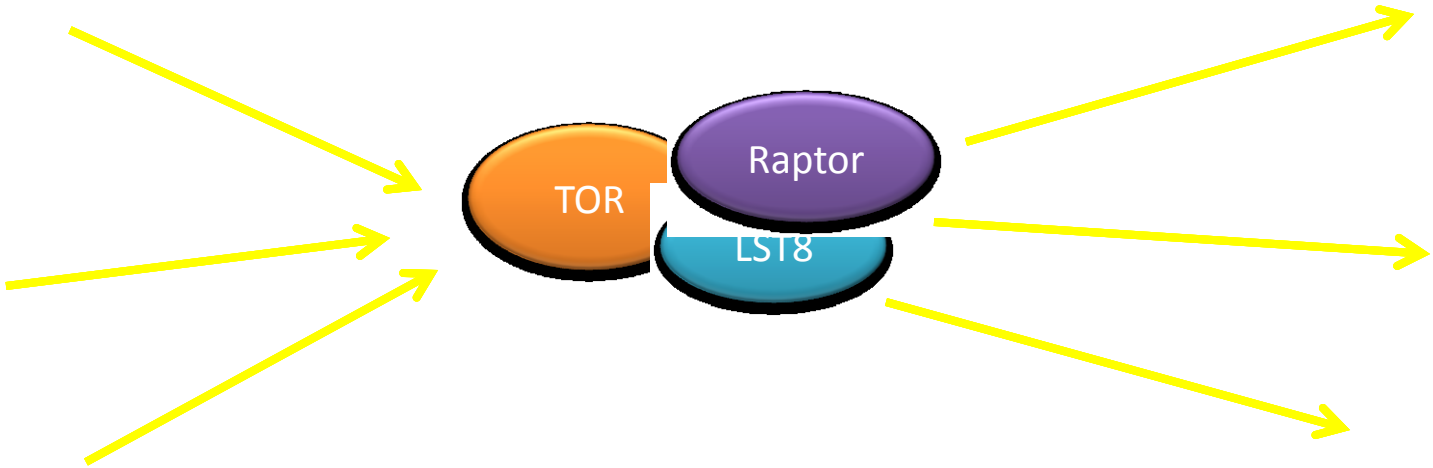
“Constructive neutral evolution”

- Suggested that many taxon specific subunits (taxon specifc proteins that are a subunit in a complex) are regulatory subunits

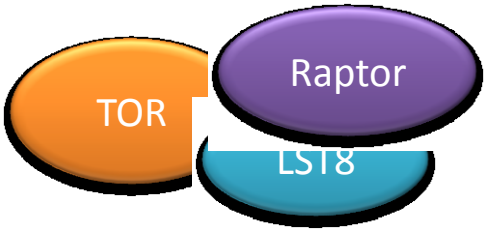
- Hypothesis: neutrally added but necessary subunits could have been appropriated as regulatory subunits?

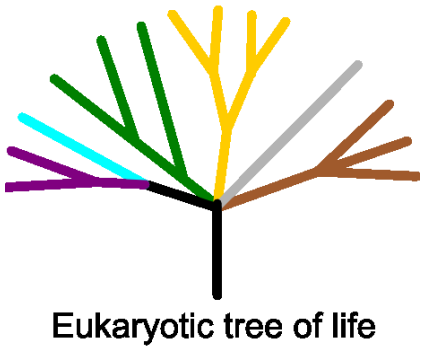
TOR1 complex

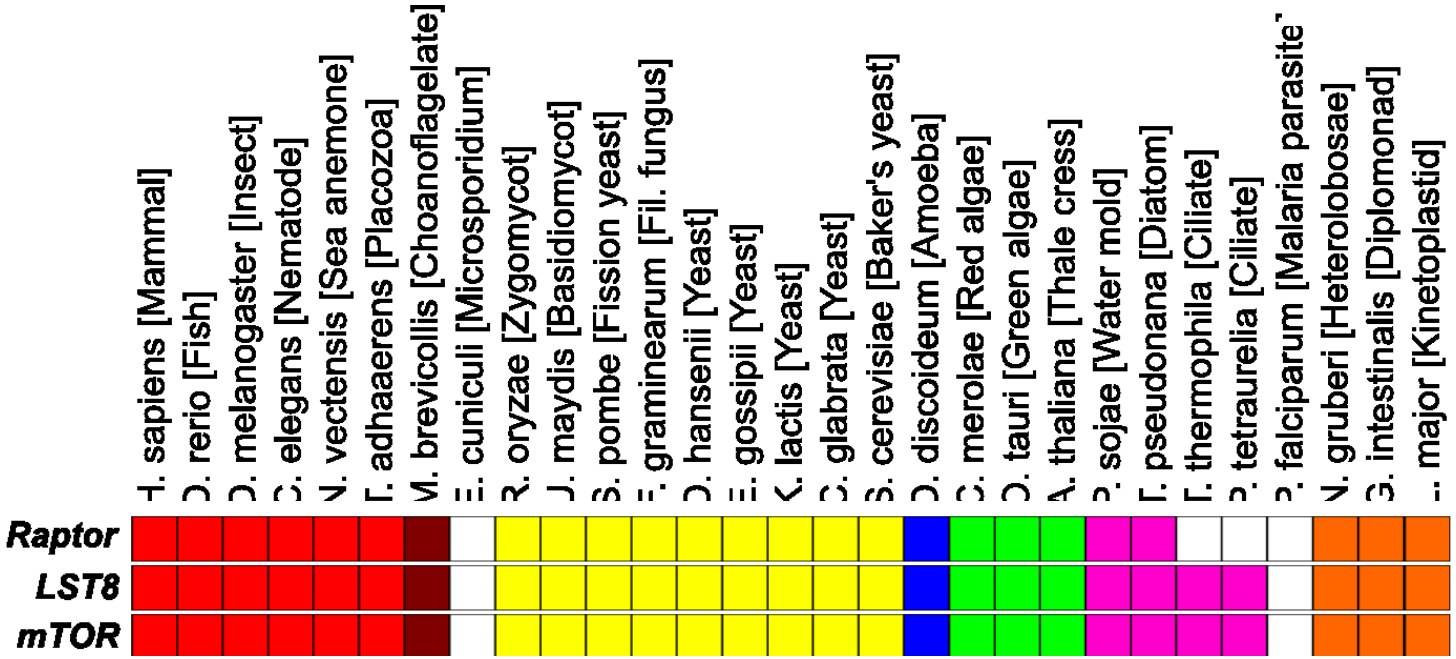
- Kinase
- Regulates growth
- Mutations of TOR1 components involved in Cancer



Evolution of TOR

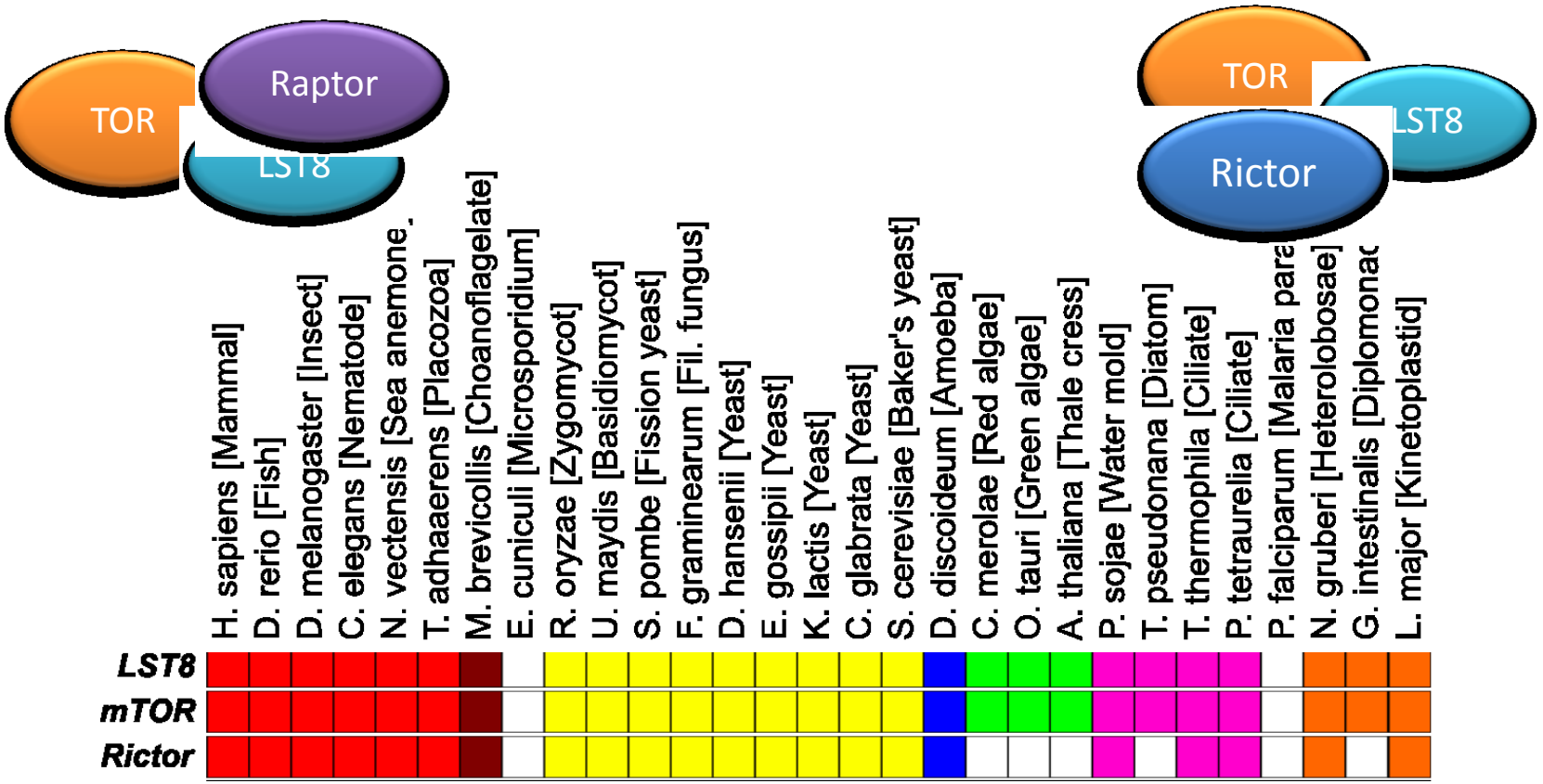


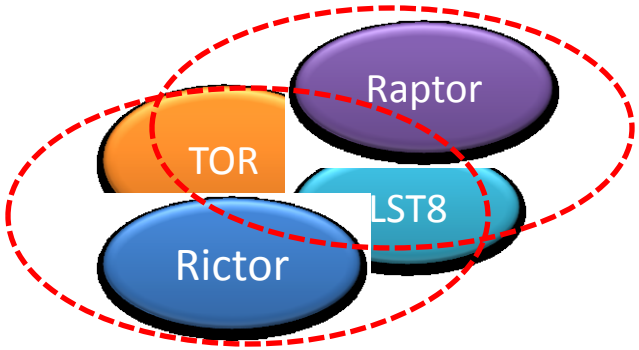


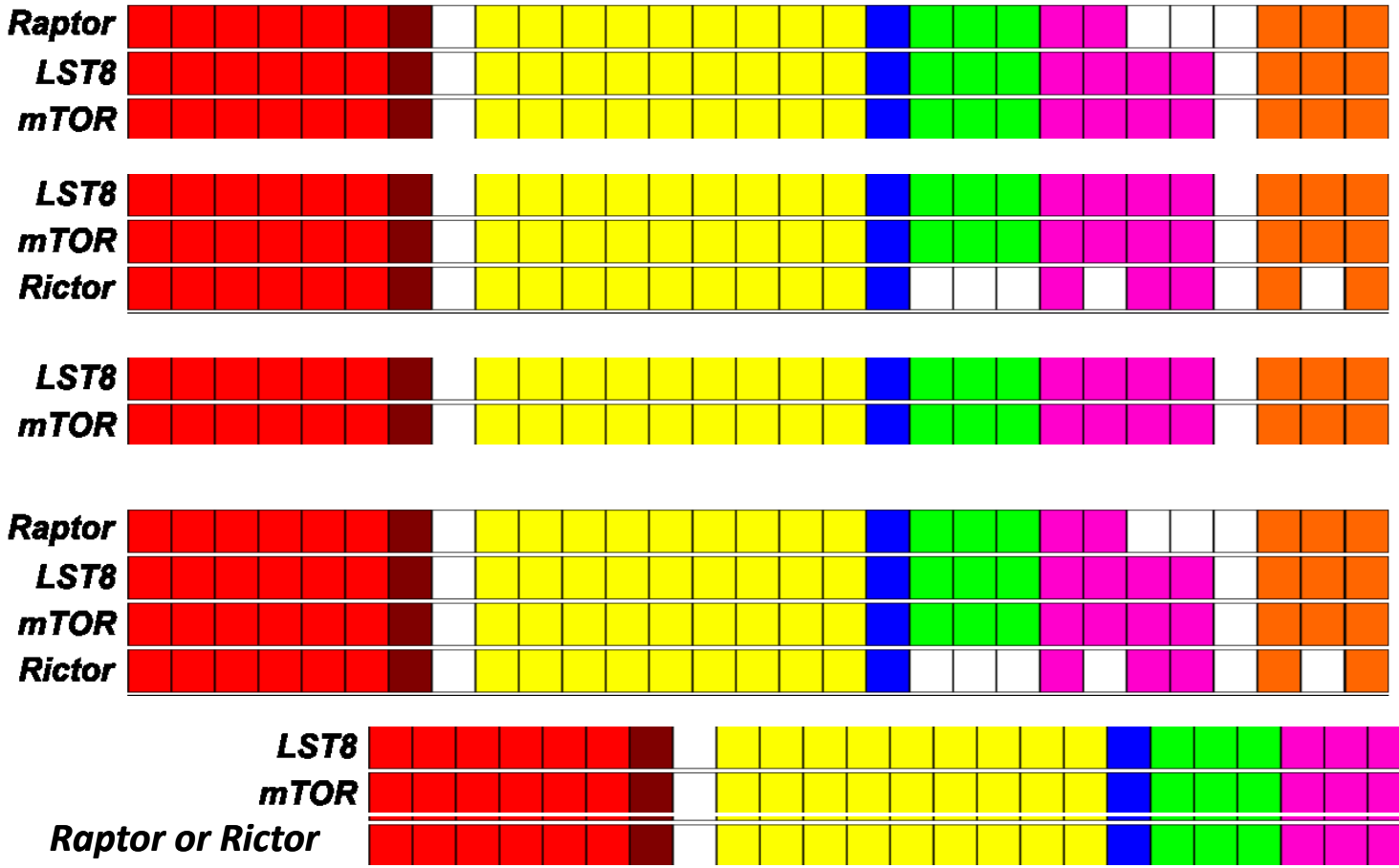


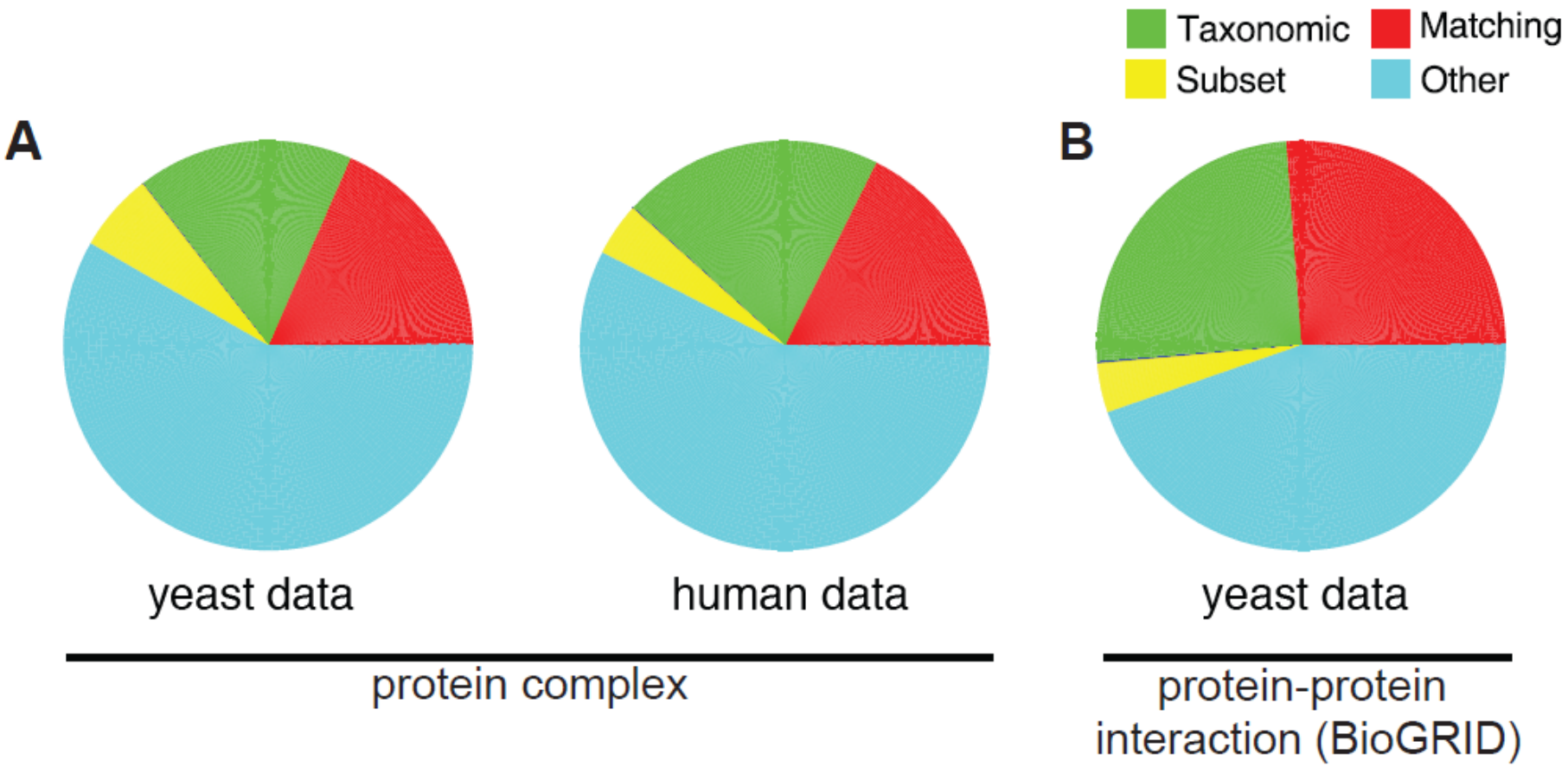
Does evolution of TOR make more sense if we consider the whole network of interactions: TOR2 complex

- TOR2 is involved in rearrangement of cytoskeleton



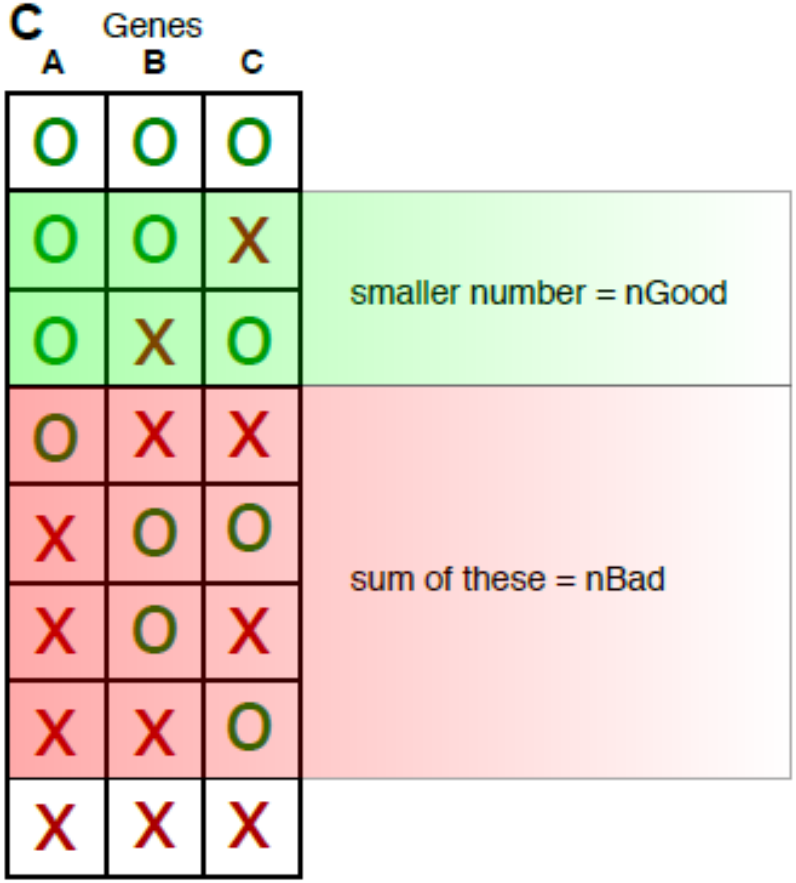

Raptor
TOR
LST8
Rictor


Raptor
LST8
mTOR
LST8
mTOR
Rictor
LST8
mTOR
Raptor
LST8
mTOR
Rictor
LST8
mTOR
Raptor or Rictor



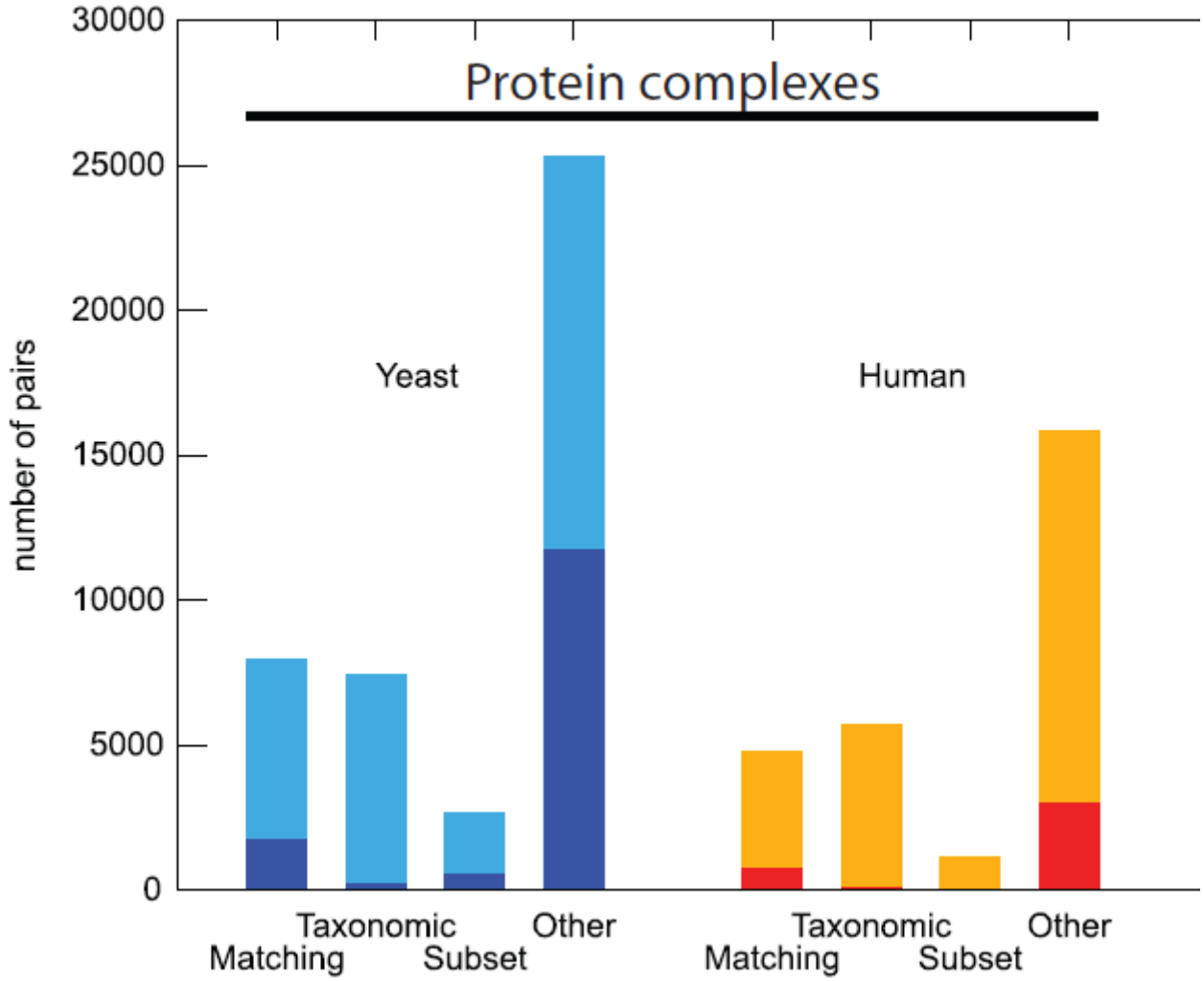
? Taxonomic subunit constructive neutral evolution ?

Considering triangles of interactions instead of pairs: a complementarity score

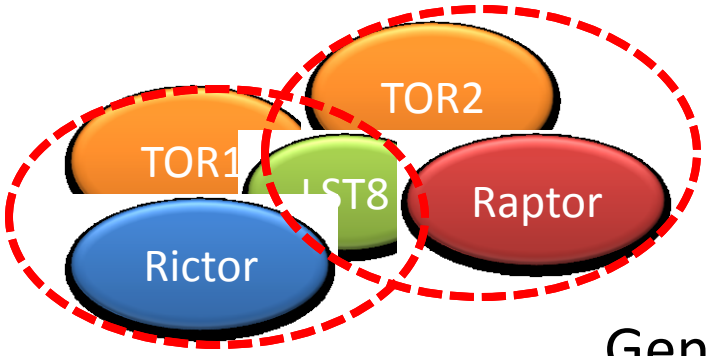


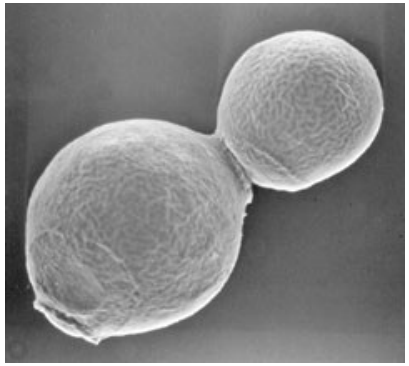
Complementarity = nGood - nBad

A substantial fraction of the “others” has a high complementarity score



Gene duplications are important in evolution and function of TOR complexes





Gene duplication of the tor kinase in *Sacceromyces cerevisiae*, chytrid fungi, oomycetes, poplar.

In yeast molecular biology demonstrated specialization

Why in some species this happens and others not?

Other forms of co-evolution

- Speed/rate of evolution
- Acceleration after loss of binding partner
- Compensatory mutations, co-evolving residues: old problem, never solved, now maybe possibly in reach thanks to evfold (?), not applied to study evolution yet

Non-orthologous gene displacement/analogous proteins

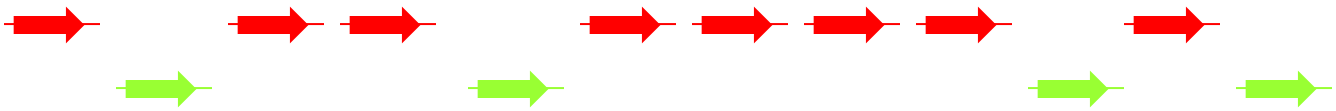
- First systematic analysis on M.genitalium (Koonin et al., Trends Genet. 1997)

TABLE 1. Non-orthologous genes coding for the same function in *Mycoplasma genitalium* and *Haemophilus influenzae*

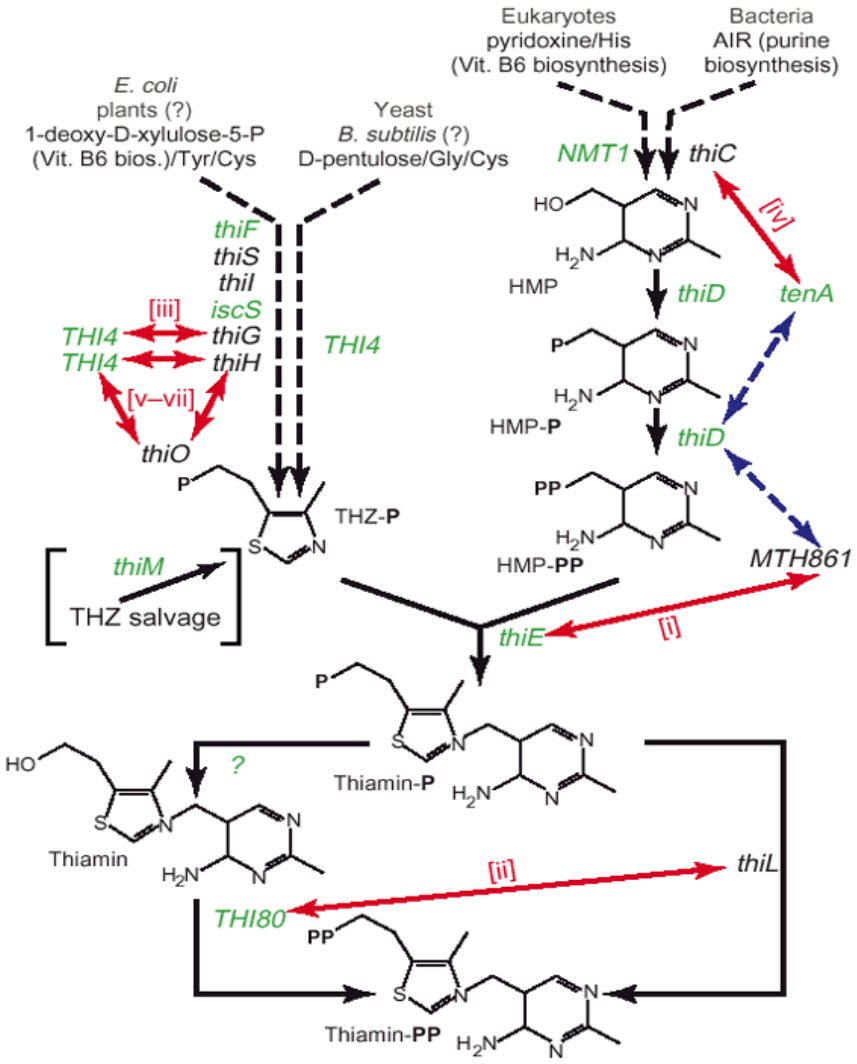
	M. genitalium		*H. influenzae*		
Enzyme	**Gene[a]**	**Orthologs[b]**	**Gene[a]**	**Orthologs[b]**	**Comment[c]**
No sequence similarity between *M. genitalium* and *H. influenzae* proteins					
Phosphoglycerate mutase	MG430 (*ybO*)	PMGI_BACSU PMGI_ECOLI PMGI_MAIZE	HI0757 (*gpmA*)	PMG1_ECOLI PMGM_HUMAN not in G(+)	*Escherichia coli* encodes both types of enzymes
L-lactate dehydrogenase	MG460	LDH_BACSU LDHM_HUMAN	HI1739B (*lctD* or *lldD*)	LLDD_ECOLI G(+)	The HI enzyme is distantly related to eukaryotic cytochrome B2
Lipoate-protein ligase	MG270	LPLA_ECOLI SCYJL046W_1	HI0027 (*lipB*)	LIPB_ECOLI S51458 (yeast)	*E. coli* and yeast encode both types of enzymes
Nucleoside diphosphate kinase	MG264?[d] MG268?[d]	None	HI0876 (*ndk*)	NDK_ECOLI NDKB_HUMAN	The two predicted kinases in MG are candidates for this indispensable activity
DNA polymerase, repair	MG261 (*dnaE*)	DP3A_HAEIN DP3A_ECOLI	HI0856 (*polA*)	DPO1_ECOLI DPO1_MYCTU	MG encodes two homologs of DNA polymerase III. MG261 is the likely repair polymerase as it belongs to a putative repair operon[6]
RNase H	MG262?[d]	DPO1_BACCA DPO1_HAEIN	HI0138 (*rnhA*); HI1059 (*rnhB*)	RNH_ECOLI RNH1_YEAST RNH2_ECOLI MC326_1 (*M. capricol.*) SC23CDS_13 (yeast)	MG262 is homologous to the 5′–3′ exonuclease domain of DNA polymerase I. It is predicted to replace the two unrelated RNases H of HI in primer removal during DNA replication
Glycyl-tRNA synthetase	MG251	SYG_HUMAN	HI0927 (*glyQ*) HI0924 (*glyS*)	SYGA_ECOLI SYGB_ECOLI CTU20547_1 (Chlamydia) G(–)	The MG enzyme contains one subunit, the HI counterpart two
Paralogs in *M. genitalium* and *H. influenzae*					
Prolyl-tRNA synthetase	MG283	YHI0_YEAST	HI0729 (*proS*)	SYP_ECOLI YER7_YEAST	Yeast encodes both types of enzymes
Cytidine deaminase	MG052	CDD_BACSU CDD_HUMAN	HI1350 (*cdd*)	CDD_ECOLI	The MG cytidine deaminase is more closely related to eukaryotic enzymes than to those from G(+) bacteria

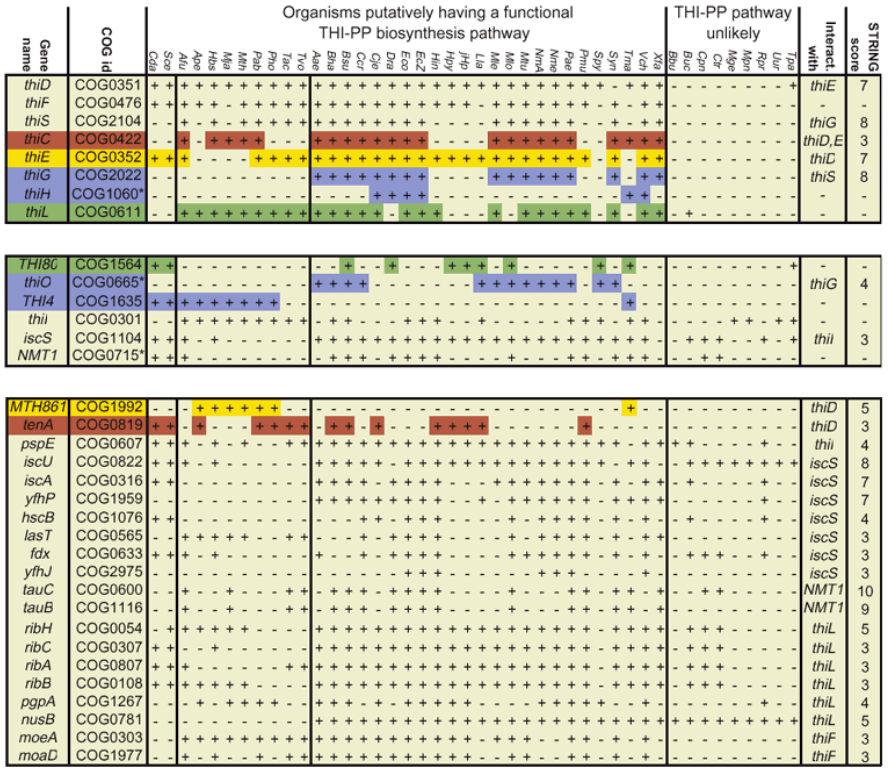
The opposite of co-occurrence: anti-correlation / complementary patterns: predicting analogous enzymes

Genes with complementary phylogenetic profiles could have a similar biochemical function.

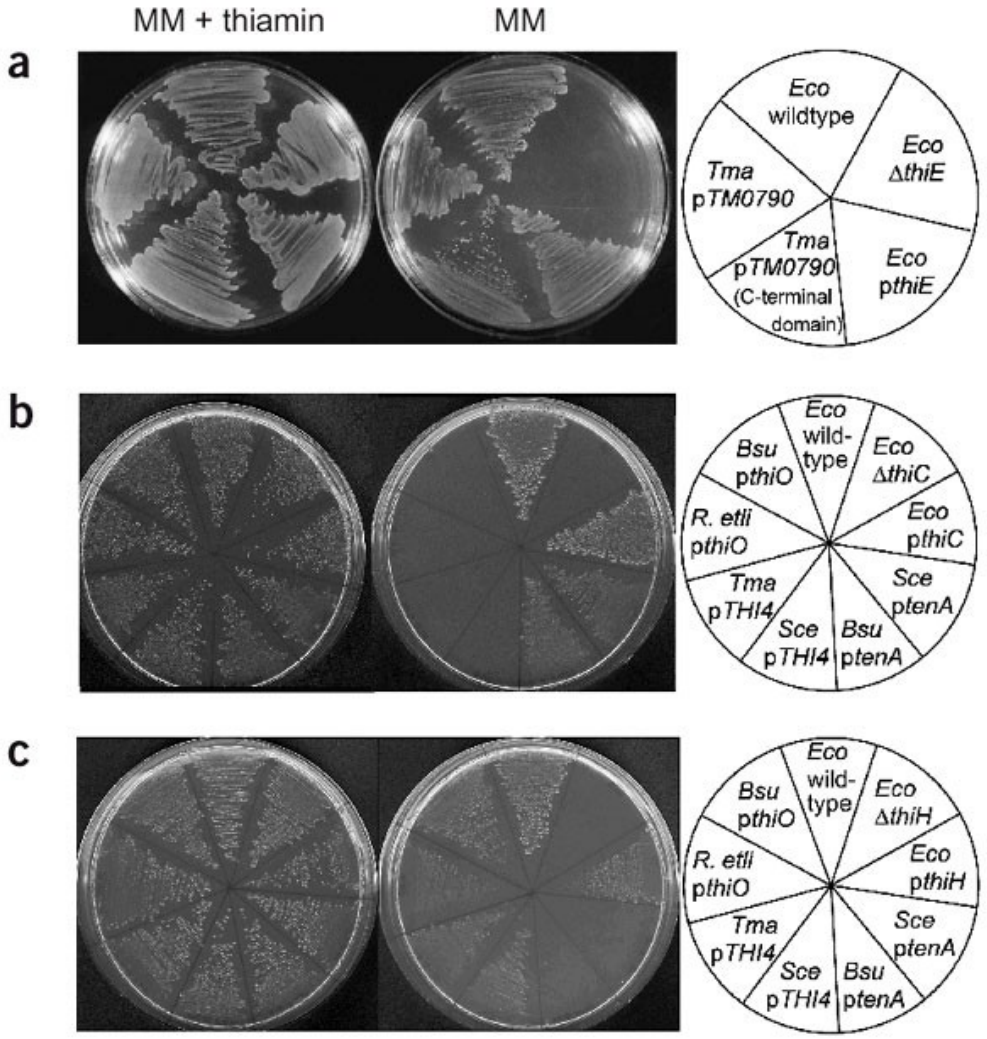


Complementary patterns in thiamin biosynthesis predict analogous

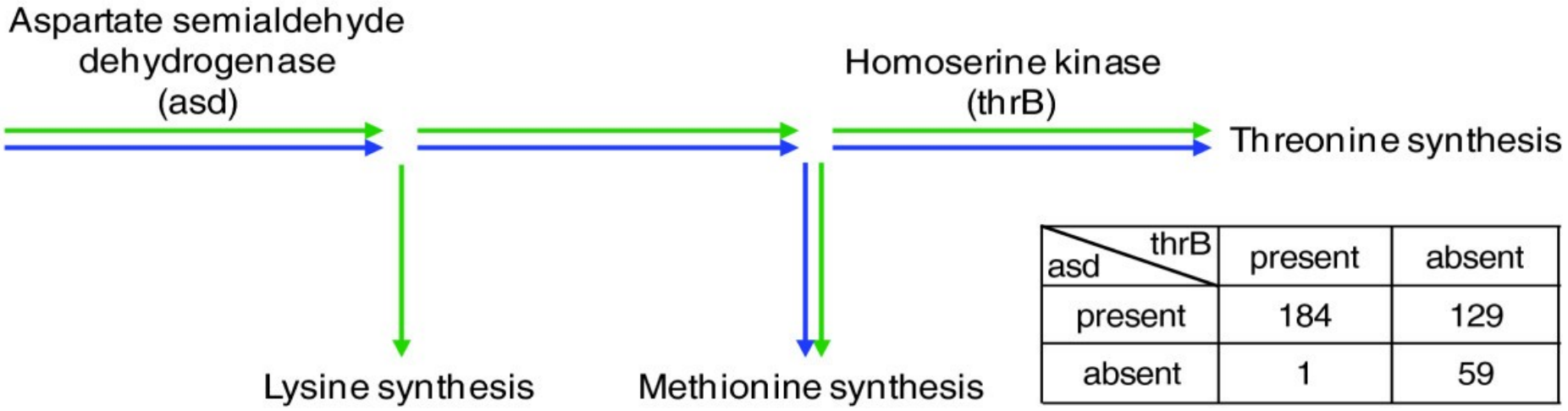




Prediction of analogous enzymes is confirmed

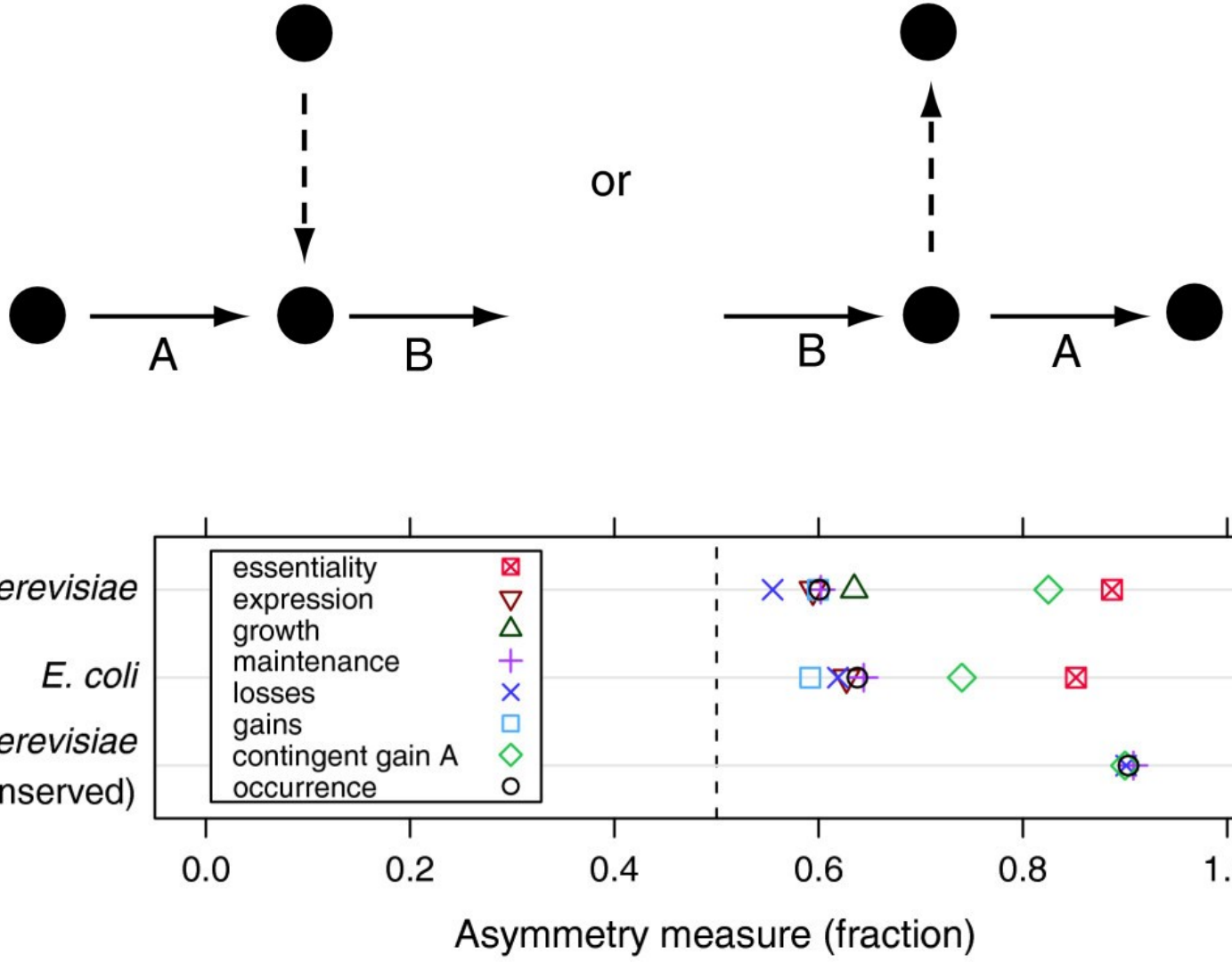


Asymmetric functional/metabolic relations explain non-similar presence absence patterns

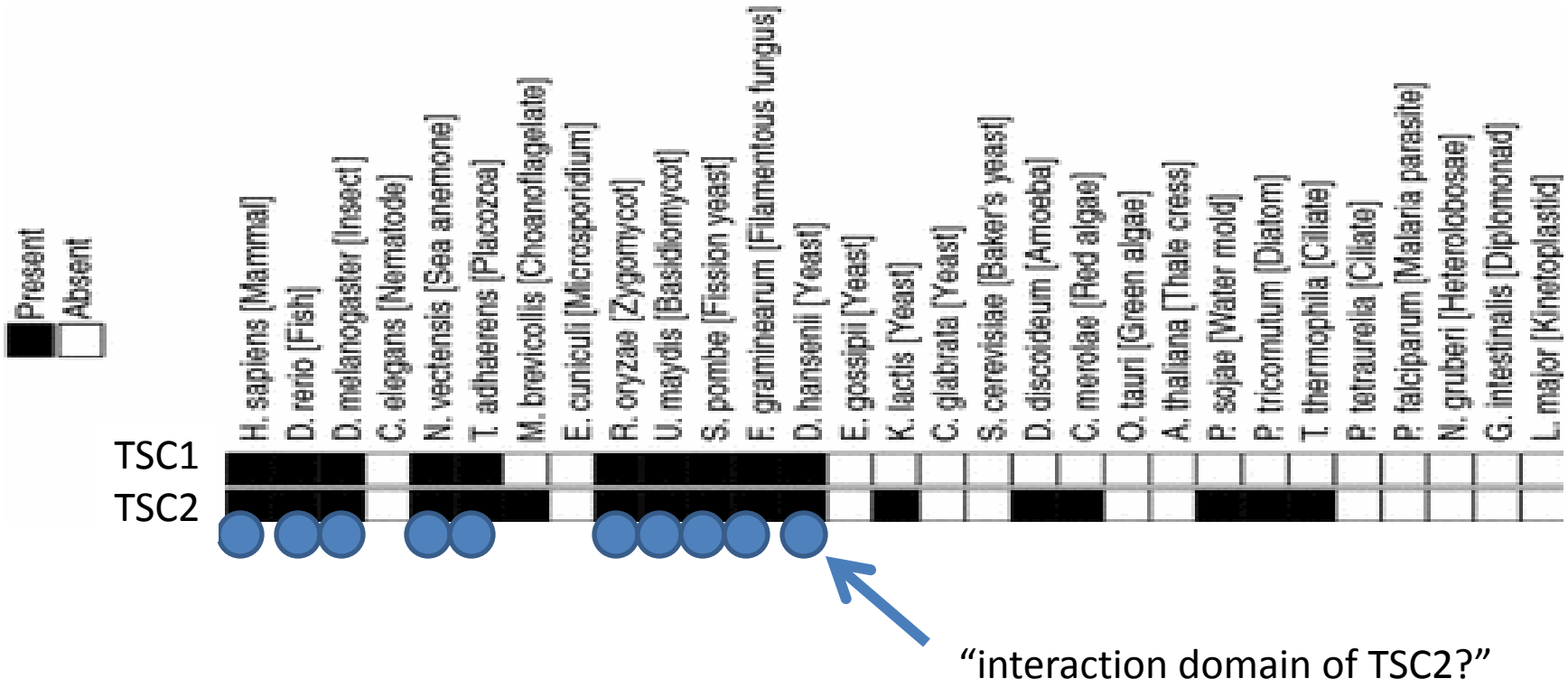


asd \ thrB	present	absent
present	184	129
absent	1	59

Asymmetric relationships between proteins shape genome evolution.
Notebaart RA, Kensche PR, **Huynen MA**, Dutilh BE.
Genome Biol. 2009 Feb 12;10(2):R19.


or
A
B
B
A
S. cerevisiae
E. coli
E. coli and S. cerevisiae
(conserved)
essentiality
expression
growth
maintenance
losses
gains
contingent gain A
occurrence
0.0
0.2
0.4
0.6
0.8
1.0
Asymmetry measure (fraction)

Domains vs proteins?



But why the innovation? Regulatory subunit / neutral accumulation of taxon specific subunits / constructive neutral evolution?

- Explaining the evolution of genes stimulates a better / more focused discussion on what we mean by gene function(al) relationship

- The more/better HTP functional data, the better for studying genome evolution

- Many different plausible, interlocking, reasons for disrupted co-occurrence across genomes of interacting proteins; (role of duplication least systematically researched?)