MCB 372 #13:
Selection, Data Partitioning
Gene Transfer

J. Peter Gogarten

University of Connecticut
Dept. of Molecular and Cell Biology

Edvard Munch, The Dance of Life (1900)

Collaborators:
Olga Zhaxybayeva (Dalhousie)
Jinling Huang (ECU)
Tim Harlow (UConn)
Pascal Lapierre (UConn)
Greg Fournier (UConn)

Funded through the NASA Exobiology and AISR Programs, and NSF Microbial Genetics

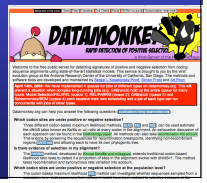
---

Hy-Phy -
Hypothesis Testing using Phylogenies.

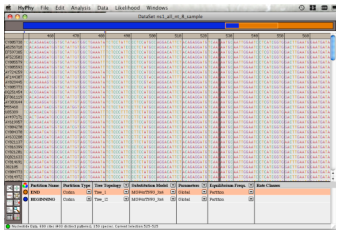
Using Batchfiles or GUI

Information at http://www.hyphy.org/

Selected analyses also can be performed online at http://www.datamonkey.org/



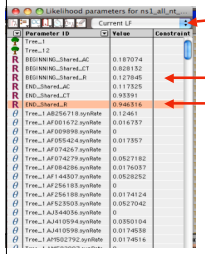
Example testing for dN/dS in two partitions of the data -- John's dataset



Set up two partitions, define model for each, optimize likelihood

---

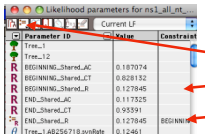
Example testing for dN/dS in two partitions of the data -- John's dataset



Safe Likelihood Function then
select as alternative

The dN/dS ratios for the two partitions are different.

Example testing for dN/dS in two partitions of the data -- John's dataset

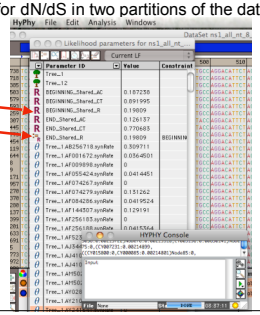


Set up null hypothesis, i.e.:

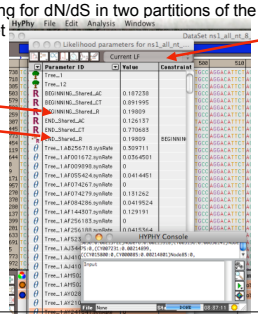
The two dN/dS are equal

(to do, select both rows and then click the define as equal button on top)

Example testing for dN/dS in two partitions of the data -- John's dataset

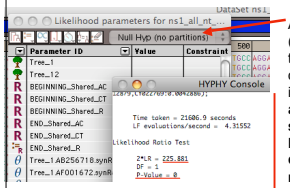


Example testing for dN/dS in two partitions of the data -- John's dataset



Name and save as Null-hyp.

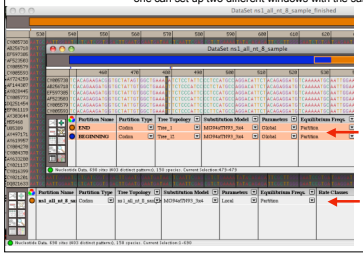
Example testing for dN/dS in two partitions of the data -- John's dataset



After selecting LRT (= Likelihood Ratio test), the console displays the result, i.e., **the beginning and end of the sequence alignment have significantly different dN/dS ratios.**

Example testing for dN/dS in two partitions of the data -- John's dataset

Alternatively, especially if the the two models are not nested, one can set up two different windows with the same dataset:

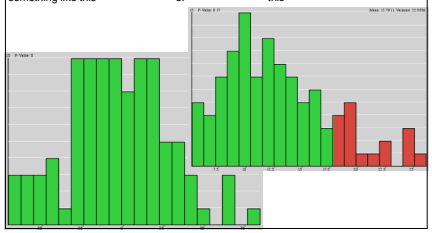


Model 1

Model 2

Example testing for dN/dS in two partitions of the data -- John's dataset

Simulation under model 1, evalutation under model 2, calculate LR
Compare real LR to distribution from simulated LR values. The result might look something like this or this



HGT detection

- **Phylogenetic Incongruence**
  (conflict between gene and species tree)
- **Phyletic Patterns**
  (disjunct/spotty distribution)
- **Surrogate Methods**
  (compositional analyses, violation of clock assumption)

Surrogate Methods - compositional analyses,

Transferred genes often have a different composition compared to the host genome. Especially dinucleotide frequencies provide a useful measure.

Reason A) The transferred gene retains for some time the composition of the donor. (Complete amelioration takes about 100 million years)
http://www.ncbi.nlm.nih.gov/pubmed/9089078?

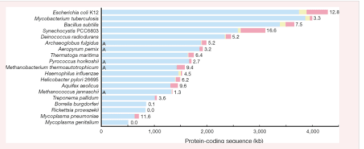
Reason B) The composition reflects the composition of the mobilome, which has a much higher AT content (mutational bias) compared to the genome. (Transferred genes never are AT rich)
http://www.ncbi.nlm.nih.gov/pubmed/15173110?

Surrogate Methods - compositional analyses,

FIGURE 2. Distribution of horizontally acquired (foreign) DNA in sequenced bacterial genomes.

From the following article:
Lateral gene transfer and the nature of bacterial innovation
Howard Ochman, Jeffrey G. Lawrence and Eduardo A. Groisman
Nature **405**, 299-304(18 May 2000)
doi:10.1038/35012500

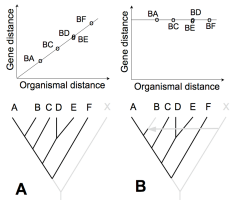
BACK TO ARTICLE



Lengths of bars denote the amount of protein-coding DNA. For each bar, the native DNA is blue; foreign DNA identifiable as mobile elements, including transposons and bacteriophages, is yellow, and other foreign DNA is red. The percentage of foreign DNA is noted to the right of each bar. 'A' denotes an Archaeal genome.

Surrogate Methods

- clocks



Use of an approximate molecular clock to detect horizontally transferred genes. For each gene, the distance between the gene and its orthologs from closely related genomes is calculated and plotted against the evolutionary distance separating the organisms. The latter can be approximated by ribosomal RNAs or by a genome average. If the gene was inherited vertically, and if the substitution rate remained approximately constant, then the points will fall on a straight line through the origin, with a slope depending on the substitution rate of the individual gene (A). If the gene was acquired from outside the organisms considered in the analysis (organism X), then all gene distances will be approximately the same and independent of the distance between the organisms (B). If the transfer occurred to a deeper branch in the tree, part of the points will fall on the diagonal, and part on a parallel line to the abscissa. Modified from Novichkov, P. S., M. V. Omelchenko, M. S. Gelfand, A. A. Mironov, Y. I. Wolf and E. V. Koonin (2004). Genome-wide molecular clock and horizontal gene transfer in bacterial evolution. J. Bacteriol. 186(19): 6575-6585.

Phyletic Methods

- taxonomic distribution of blast hits
- taxonomic position of best blast hit

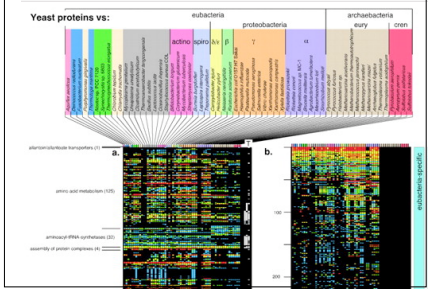
Any non-taxonomic distribution of gene presence and absence can be explained either by

Gene transfer,

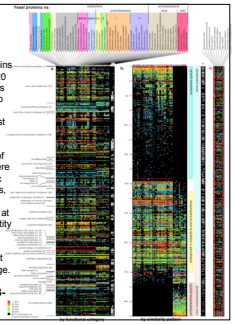
or by

Gene loss. Under the assumption of gene loss any gene present in at least one archaeon and one bacterium would have to be assumed present in the ancestral "Garden of Eden" genome. (Doolittle, W. F., Y. Boucher, C. L. Nesbø, C. J. Douady, J. O. Andersson and A. J. Roger (2003). How big is the iceberg of which organellar genes in nuclear genomes are but the tip? Philosophical Transactions of the Royal Society B: Biological Sciences 358(1429): 39-58.)

Phyletic Patterns (asisde)



Phyletic Patterns (asisde)



Amino acid sequence identity in Smith-Waterman alignments for the 850 yeast proteins that produce a match with an E-value of 10–20 or better in FASTA comparisons to all proteins from the prokaryotic genomes listed at the top of the figure. Color-coding of the percentage identity values is shown at lower left. (a) Yeast proteins grouped by functional category. (b) Yeast proteins sorted by the quotient [15·(sum of eubacterial identities)]/[45·(sum of archaebacterial identities)]; zero quotients were replaced by one. The 383 eubacterial-specific proteins, 111 archaebacterial-specific proteins, and 263 proteins widespread among both groups are indicated by colored bars. Lane T at right is as in (a). (c) Pairwise amino acid identity between yeast homologues and eukaryotic homologues in Blast searches (Altschul et al. 1997), showing that the yeast proteins are not lateral acquisitions specific to the yeast lineage.

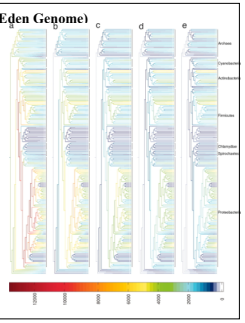
Esser et al. Mol. Biol. Evol. 21(9):1643-1660. 2004

Phyletic Patterns (Garden of Eden Genome)

The colors of nodes and branches correspond to the inferred ancestral genome size, as indicated in the scale. a–e correspond to the SO, LGT≤1, LGT≤3, LGT≤7, and LGT≤15 models, respectively.

From: Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution Tal Dagan and William Martin PNAS | January 16, 2007 | vol. 104 | no. 3 | 870-875

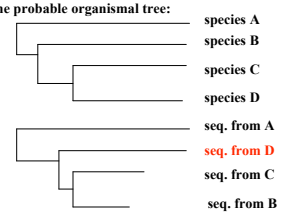
"The results indicate that among 57,670 gene families distributed across 190 sequenced genomes, at least two-thirds and probably all, have been affected by LGT at some time in their evolutionary past."



Trees – what might they mean?

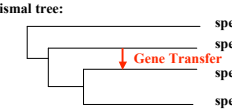
Calculating a tree is comparatively easy, figuring out what it might mean is much more difficult.

If this is the probable organismal tree:

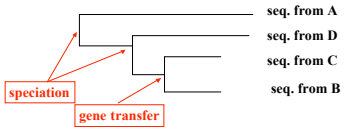


Gene transfer

Organismal tree:

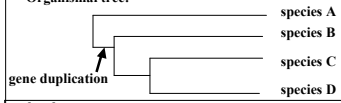


molecular tree:

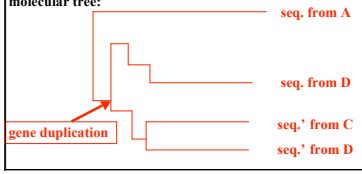


Gene duplication

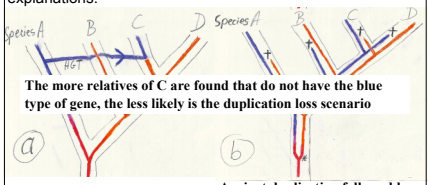
Organismal tree:



molecular tree:



Gene duplication and gene transfer are equivalent explanations.



The more relatives of C are found that do not have the blue type of gene, the less likely is the duplication loss scenario

Horizontal or lateral Gene

Ancient duplication followed by gene loss

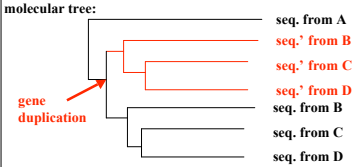
Note that scenario B involves many more individual events than A

1 HGT with orthologous replacement

1 gene duplication followed by 4 independent gene loss events

Function, ortho- and paralogy

molecular tree:



The presence of the duplication is a taxonomic character (shared derived character in species B C D).
The phylogeny suggests that seq' and seq have similar function, and that this function was important in the evolution of the clade BCD.
seq' in B and seq'in C and D are orthologs and probably have the same function, whereas seq and seq' in BCD probably have different function (the difference might be in subfunctionalization of functions that seq had in A. – e.g. organ specific expression)

BIPARTITION OF A PHYLOGENETIC TREE

Bipartition (or split) – a division of a phylogenetic tree into two parts that are connected by a single branch.
It divides a dataset into two groups, but it does not consider the relationships within each of the two groups.

Yellow vs Rest
* * *. . . * *
compatible to illustrated bipartition

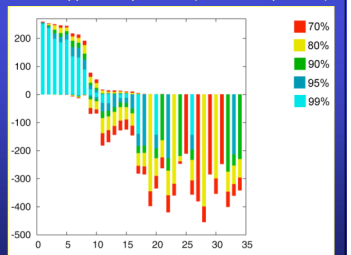
Orange vs Rest
. . * . . . . *
incompatible to illustrated bipartition

"Lento"-plot of 34 supported bipartitions (out of 4082 possible)

13 gamma-proteobacterial genomes (258 putative orthologs):

- E.coli
- Buchnera
- Haemophilus
- Pasteurella
- Salmonella
- Yersinia pestis (2 strains)
- Vibrio
- Xanthomonas (2 sp.)
- Pseudomonas
- Wigglesworthia

There are 13,749,310,575 possible unrooted tree topologies for 13 genomes

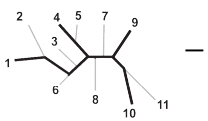


Note: The sixth bipartition (counting from the left) corresponds to the grouping of *Wigglesworthia* with *Buchnera*, a grouping that might reflect an artifact rather than shared ancestry.

PROBLEMS WITH BIPARTITIONS

- No easy way to incorporate gene families that are not represented in all genomes.
- The more sequences are added, the shorter the internal branches become, and the lower is the bootstrap support for the individual bipartitions.
- A single misplaced sequence can destroy all bipartitions.

Bootstrap support values for embedded quartets



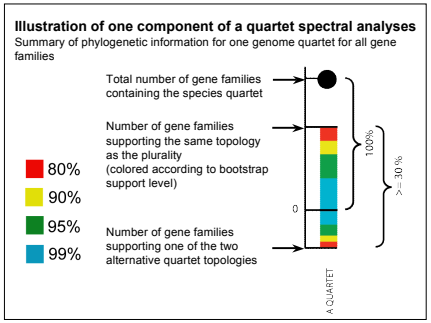
+ — : tree calculated from one pseudo-sample generated by bootstraping from an alignment of one gene family present in 11 genomes

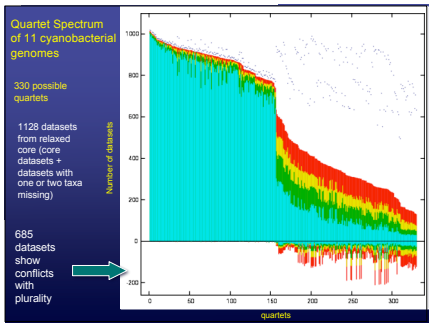
— : embedded quartet for genomes 1, 4, 9, and 10 .
This bootstrap sample supports the topology ((1,4),9,10).

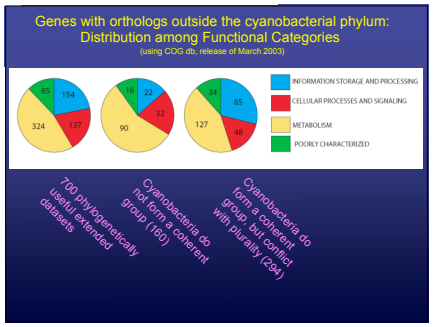
Quartet spectral analyses of genomes iterates over three loops:

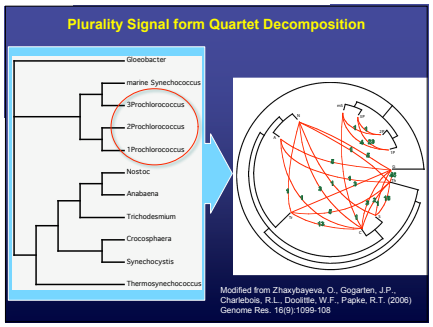
- **Repeat for all bootstrap samples.**
- **Repeat for all possible embedded quartets.**
- **Repeat for all gene families.**

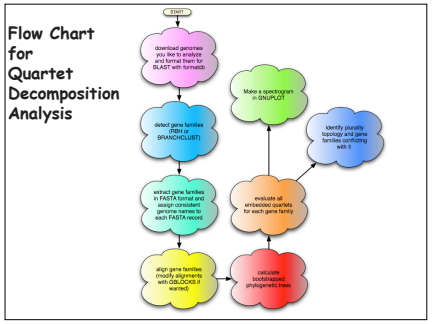
Zharkikhyna et al. 2006 Genome Research. 16:1095-108











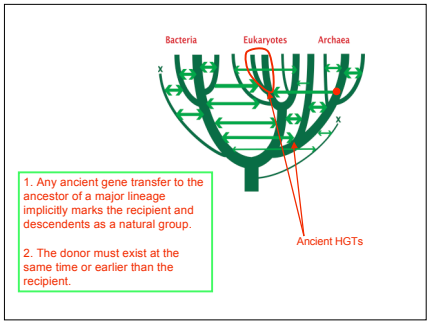
Ancient Gene Transfer and Phylogenetic Reconstruction: Friends or Foes?

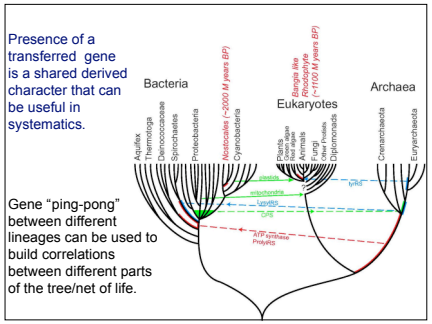
Popular view

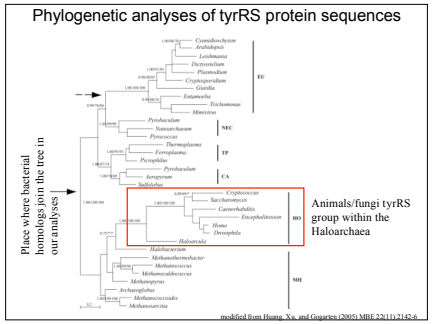
- Gene transfer is a disruptive force in phylogenetic reconstruction.

New view

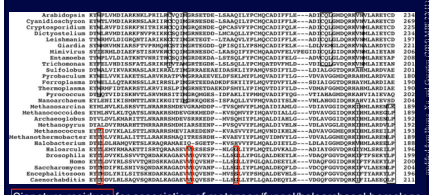
- Events of ancient gene transfer are valuable tools for reconstructing organismal phylogeny.







Multiple protein sequence alignment for tyrRS.



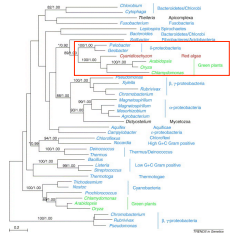
Signature residues for association of metazoan/fungal/haloarchaeal homologs

Transferred tyrRS supports monophyletic opisthokonts

The same conclusion is reached, if haloarchaeal type tyrRS in opisthokonts is explained by ancient paralogy and differential gene loss.

Monophyly of primary photosynthetic eukaryotes is supported by more than 50 ancient gene transfers from different bacterial phyla to the ancestor of the red algae and green plant lineage.

- E.g., ancient gene transfer of frp-gene (florfenicol resistance protein)



Gene from Green plants and Red algae groups with delta proteobacteria

modified from Huang and Gogarten (2006), Trends in Genetics 22, 361- 366

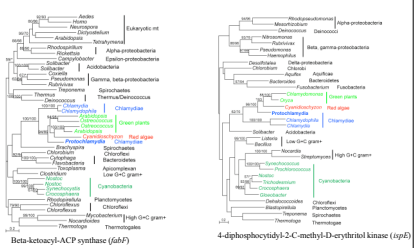
Red Algal and Green Plant Genes of Chlamydial Origin

Gene Name or Gene Product	Presence	Putative Function
ADT/ATP translocase	R and G	ATP/ADP transport
Phosphate transporter	G	Phosphate transport
Sodium hydrogen antiporter	R and G	Ion transport
Cu-ATPase	R and G	Ion transport
4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase (gcpE)	R and G	Isoprenoid biosynthesis
4-diphosphocytidyl-2-C-methyl-D-erythritol kinase (*ispE*)	R and G	Isoprenoid biosynthesis
2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase (*ispD*)	R and G	Isoprenoid biosynthesis
Enoyl-ACP reductase (*fabI*)	R and G	Fatty acid biosynthesis
Beta-ketoacyl-ACP synthase (*fabF*)	R and G	Fatty acid biosynthesis
Glycerol-3-phosphate acyltransferase	R and G	Phospholipid
Polynucleotide phosphorylase	R and G	RNA degradation
Phosphoglycerate mutase	G	Glycolysis
Oligoendopeptidase F	R	Amino acid biosynthesis
Aspartate transaminase	R and G	Amino acid metabolism
Malate dehydrogenase	G	Energy conversion
Tyrosyl-tRNA synthetase	R and G	Translation
23S rRNA (Uracil-5-)-methyltransferase	R and G	RNA modification
Isoamylase	R and G	Starch biosynthesis
Hypothetical protein	R	Unknown
Sugar phosphate isomerase	G	Sugar interconversion
CMP-KDO synthetase	G	Cell envelope formation

Consistent phylogenetic signal links Chlamydiae, red algae and green plants.

modified from Huang and Gogarten 2007. Genome Biol 8: R99

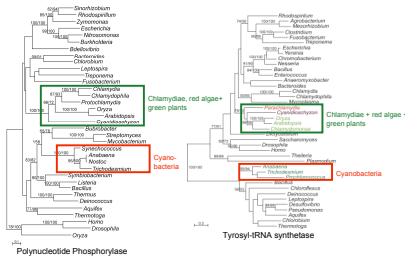
Chlamydial-type genes in red algae and plants are often specifically associated with *Protochlamydia (Parachlamydia)*



Beta-ketoacyl-ACP synthase (*fabF*)

4-diphosphocytidyl-2-C-methyl-D-erythritol kinase (*ispE*)

The chlamydial genes (plant & red algae) group separate from the cyanobacterial homologs



Polynucleotide Phosphorylase

Tyrosyl-tRNA synthetase

Examining possible Hypotheses

1. Plants acquired chlamydial genes via insect feeding activities (Everett et al. 2005).

No. The ancestor of red algae and green plants is much older than insects.

2. Chlamydiae acquired plant-like genes via Acanthamoeba hosts (Stephens et al. 1999; Wolf et al. 1999; Ortutay et al. 2003).

No. All these genes are of bacterial origin. The direction of gene transfer is from bacteria to eukaryotes.

3. Chlamydial and plant sequence similarities reflect an ancestral relationship between chlamydiae and cyanobacteria (Brinkmann et al. 2002; Horn et al. 2004).

No. Genes of chloroplast ancestry should still be more similar to cyanobacterial than to chlamydial sequences. In many instances the cyanobacterial homologs form a clearly distinct, and separate clade.

What do we learn from the data??

(Chlamydial genes in red algal and plant genomes)

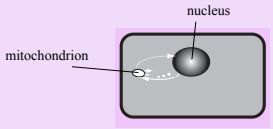
Unless a stable physical association existed, it is highly unlikely for any single donor to transfer such a number of genes to a single recipient.

Our Hypothesis: An ancient, unappreciated symbiotic association existed between chlamydiae and the ancestor of red algae and green plants.

Genes from this chlamydial symbiont might have been crucial to establish communication between host and the cyanobacterial cytoplasms.

Hypothesis: Chlamydiae and the primary plastids

A) The Host



White: α-proteobacterial (mitochondrial) symbiont

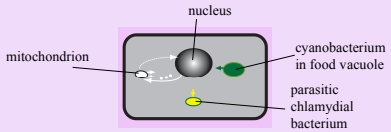
Gene transfer to the nucleus

Transport of nuclear encoded proteins to symbiont

Direction of symbiotic benefit

Hypothesis: Chlamydiae and the primary plastids

B) The Host invaded by a parasite



Yellow: parasitic chlamydial bacterium

Green: cyanobacterium (as food)

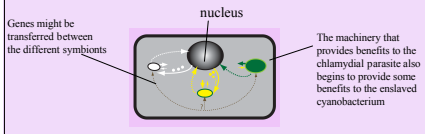
Gene transfer to the nucleus

Transport of nuclear encoded proteins to symbiont

Direction of symbiotic benefit

Hypothesis: Chlamydiae and the primary plastids

C) The chlamydial symbiont becomes mutalistic



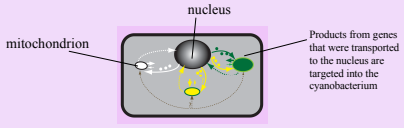
Gene transfer from the symbionts to the nucleus

Transport of nuclear encoded proteins to symbiont

Direction of symbiotic benefit

Hypothesis: Chlamydiae and the primary plastids

D) The cyanobacterium evolves into an organelle



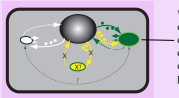
Chlamydial and cyanobacterial type gene products are targeted to the plastid

Gene transfer from the symbionts to the nucleus

Direction of symbiotic benefit

Hypothesis: Chlamydiae and the primary plastids

D) Loss of chlamydial symbiont

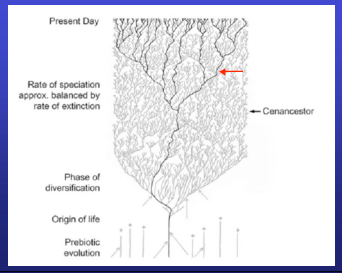


With the help of genes contributed by the chlamydial parasite, the cyanobacterium has evolved into a photosynthetic organelle

The chlamydial symbiont is lost - possibly without trace, except for the genes that facilitated the integration of the metabolism of the host with that of a photoautotroph.

This hypothesis also explains why the evolution of an organelle from a primary endosymbiont is rare. A photoautotroph with a single compartment has few transporters available that would allow integration with the host metabolism. The simultaneous presence of an intracellular parasite allows for the integration of the two cytoplasms.

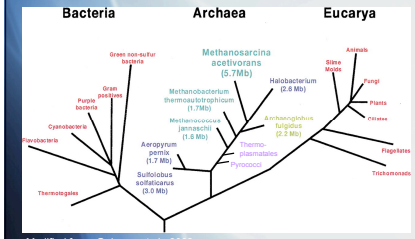
The Coral of Life (Darwin)



Mapping Metabolic Pathways on the Tree of Life

Chris House, Bruce Runnegar, and Sorel Fitz-Gibbon conclude from their analyses of genome based phylogenetic trees (Geobiology **1**, 15-26, 2003):

*"Our results suggest that the last common ancestor of Archaea was **not** a methanogen and **that methanogenesis arose later** during subsequent microbial evolution. This leaves sulphur reduction as the most geochemically plausible metabolism for the base of the archaeal crown group."*



Modified from: Galagan et al., 2002

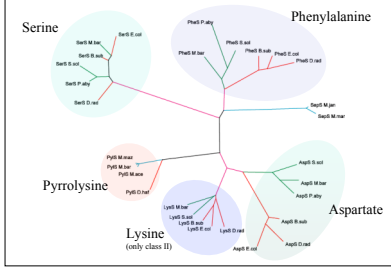
Mapping Metabolic Pathways on the Tree of Life

Chris House, Bruce Runnegar, and Sorel Fitz-Gibbon conclude from their analyses of genome based phylogenetic trees (Geobiology **1**, 15-26, 2003):

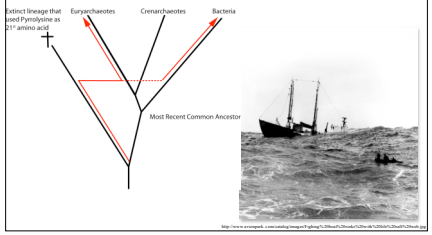
*"Our results suggest that the last common ancestor of Archaea was **not** a methanogen and **that methanogenesis arose later** during subsequent microbial evolution. This leaves sulphur reduction as the most geochemically plausible metabolism for the base of the archaeal crown group."*

This conclusion is at odds with the ancient origin of many of the enzymes specific to methanogens.
Enzymes involved in methylamine reduction use pyrrolysine as a 21 amino acid. The enzyme that charges the pyrrolysine tRNA is as old as the genetic code.

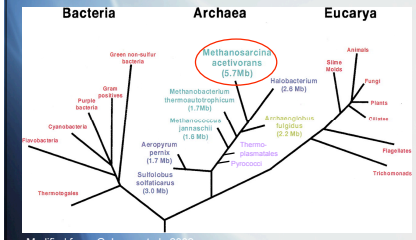
Class II aaRS Phylogeny



Pyrrolysine as 21st amono acid: A genetic "life raft"?

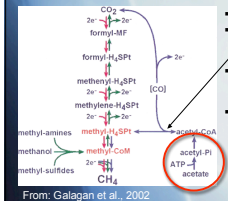


Evolution of Aceticlastic Methanogenesis in *Methanosarcinales* via Horizontal Gene Transfer from *Clostridia*



Modified from: Galagan et al., 2002

Methanogenesis



From: Galagan et al., 2002

- Unique to subset of Archaea
- Energy production via CODH complex (reversible Acetyl-CoA synthesis)
- 1 billion metric tons of biogenic methane produced annually
- Over 66% of biogenic methane originates from acetate, specific pathway

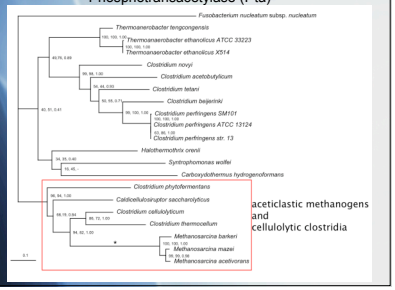
Aceticlastic methanogenesis in *Methanosarcina* genera

- Unique to subset of Archaea
- Energy production via CODH complex for multiple carbon substrates
- 1 billion metric tons of biogenic methane produced annually
- Over 66% of biogenic methane originates from acetate, specific pathway

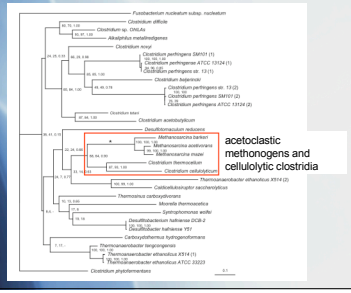
Aceticlastic methanogenesis in *Methanosarcina* genera

acetate kinase (AckA) and phosphotransacetylase (Pta)

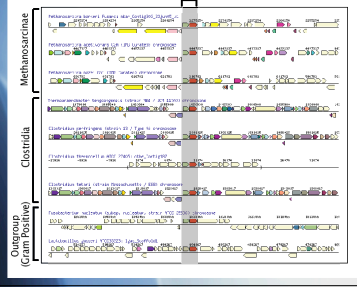
Phosphotransacetylase (Pta)



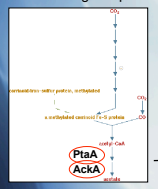
acetate kinase (AckA)



AckA/PtaA

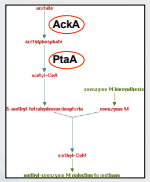


Clostridia acetigenic pathway



HGT

Methanosarcina aceticlastic pathway



Figures drawn with *Metacyc* (www.metacyc.org)

- Methanogenesis from acetate evolved in the *Methanosarcinales* via HGT of AckA and Pta genes from a species closely related to cellulolytic *Clostridia*
- The transfer was a relatively recent singular event involving derived taxa.
- The ancestral state of the recipient likely was a chemoautotrophic methanogen that used the CODH complex for Acetyl-CoA synthesis from C-1 substrates. (Grahame et al. 2005 Arch Microbiol 184 32-43, 2005)
- Gene transfer of AckA/PtaA would provide metabolic reversal of CODH reaction in presence of acetate, with no additional genes required.