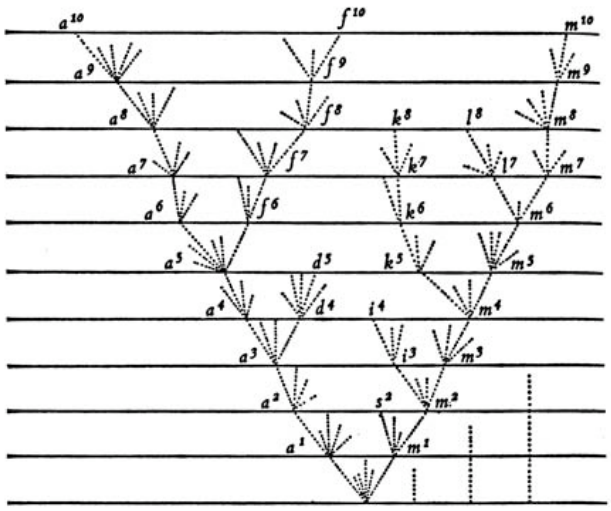
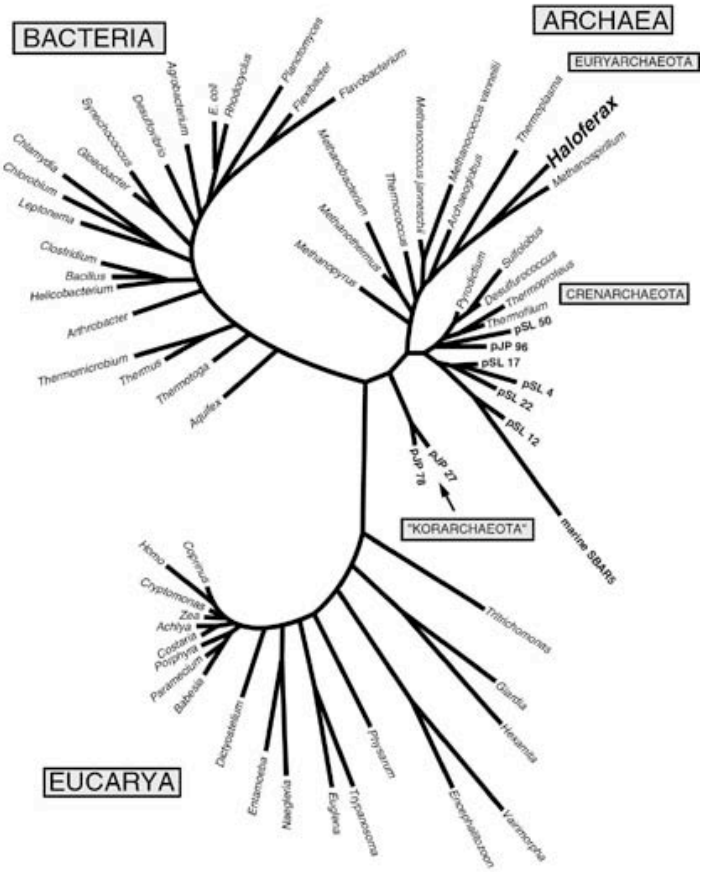
Phylogenomics

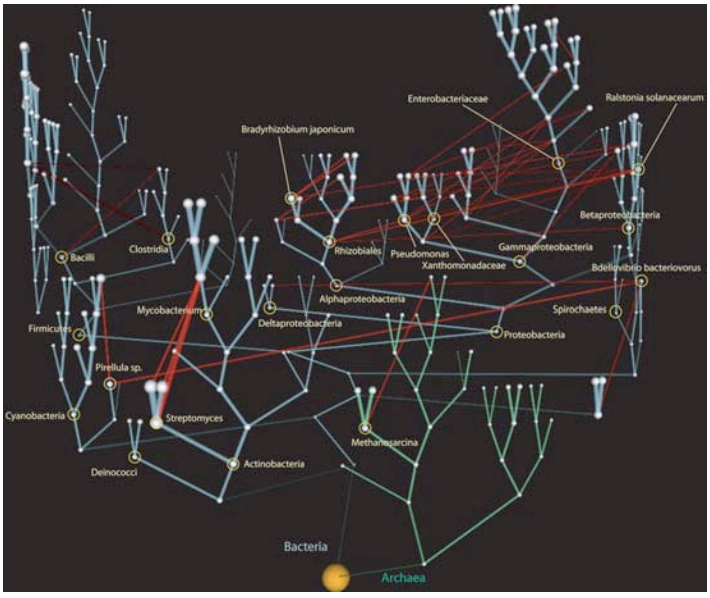
Gene History, Genome History and Organismal Phylogeny (genealogy of cellular life)





Barnes, S.M. *et al.*, 1996, Proc. Natl. Acad. Sci. USA, **93**: 9188-9193.

Woese CR, Fox GE.
PNAS 1977 Nov;74(11):5088.



MB&B 452 Genomics and Bioinformatics
Lecturer: Patrick O'Donoghue Oct. 16, 2006.

MB&B 452

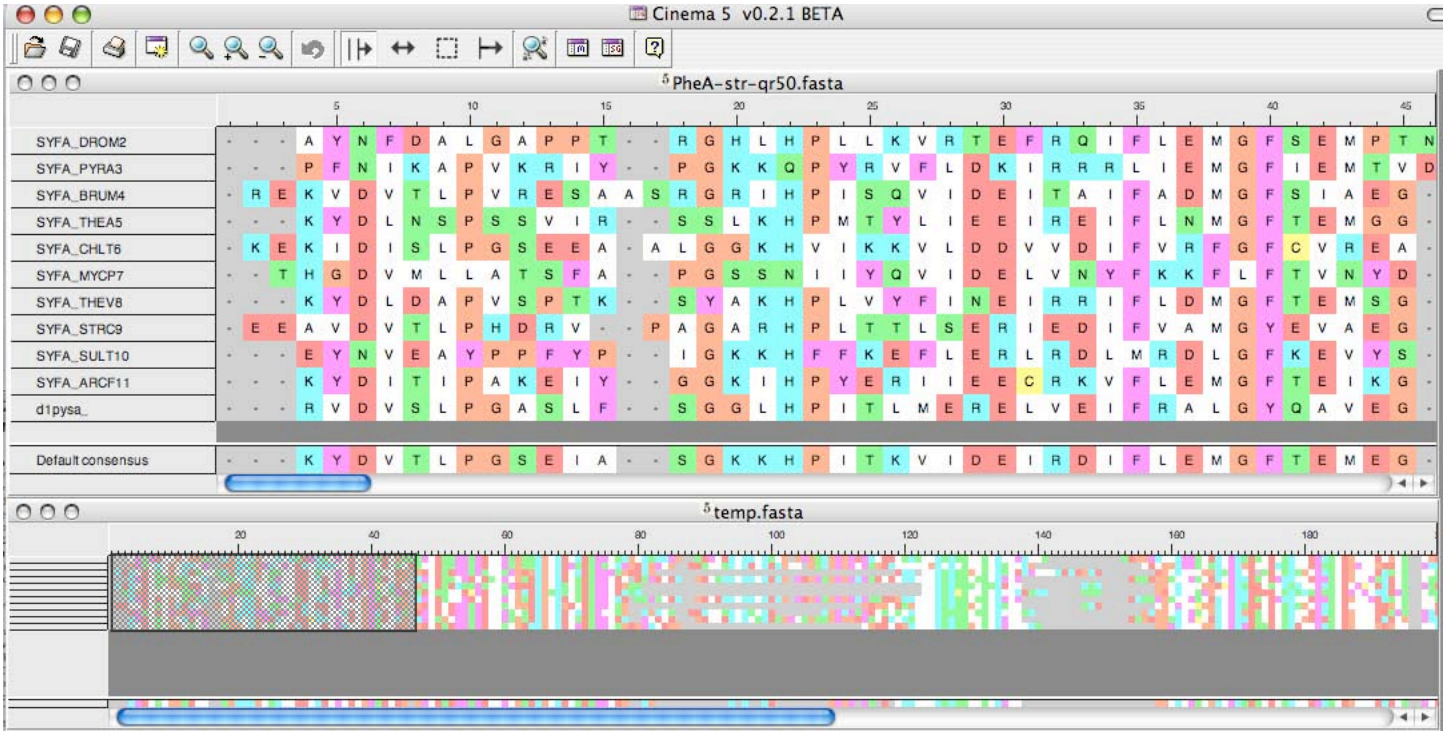
Overview

1. Molecular Phylogenetics
   - Basis of Molecular Phylogenetics
     - Homologous proteins, nucleic acids
     - Sequence or structure based alignment
   - Vertical Gene Transfer - Canonical Phylogenetic Pattern
   - Horizontal Gene Transfer

2. Inferring (Computing) Phylogeny
   - Algorithmic methods
   - Optimization methods

3. Phylogenomics
   - What is it?
   - Three applications: tree of life
     - pathogen evolution
     - gene function prediction
   - Advantages & Disadvantages of Phylogenomic Approach
   - How Genome Evolution, HGT relate to cellular character.

Sequence Alignment
the basis of molecular phylogeny



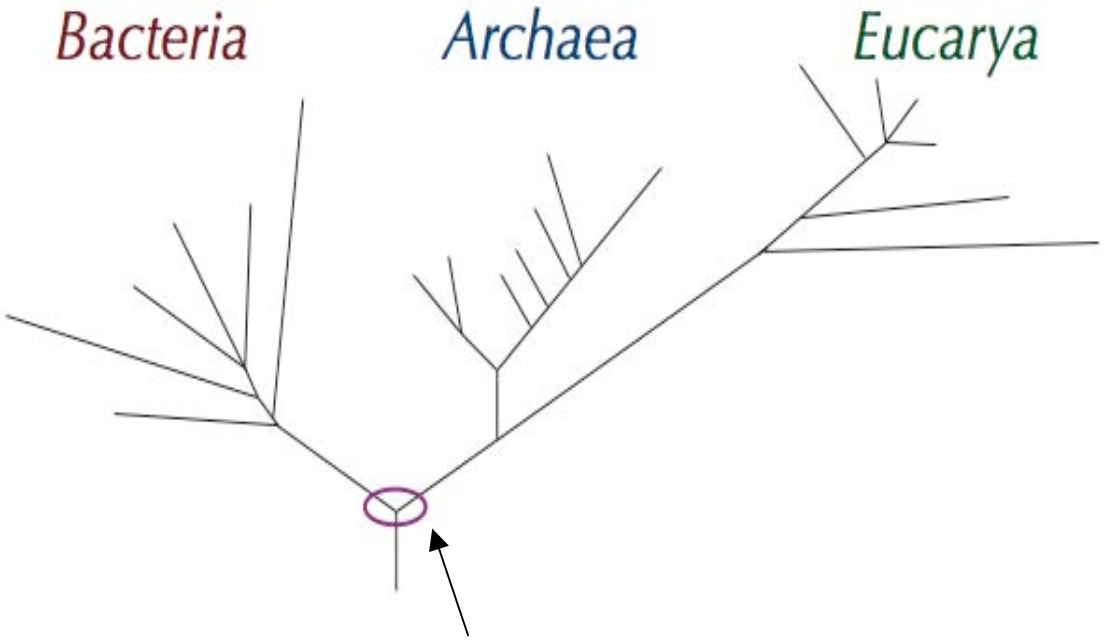
Homologous protein or nucleic acid sequences from different organisms can be aligned.

Sequence similarity is assumed to be proportional to evolutionary distances.

Screen shot from the sequence alignment editor Cinema 5
http://aig.cs.man.ac.uk/research/utopia/cinema/cinema.php

Canonical Phylogenetic Pattern

Universal Phylogenetic Tree (UPT) based on the ribosome (rRNA).



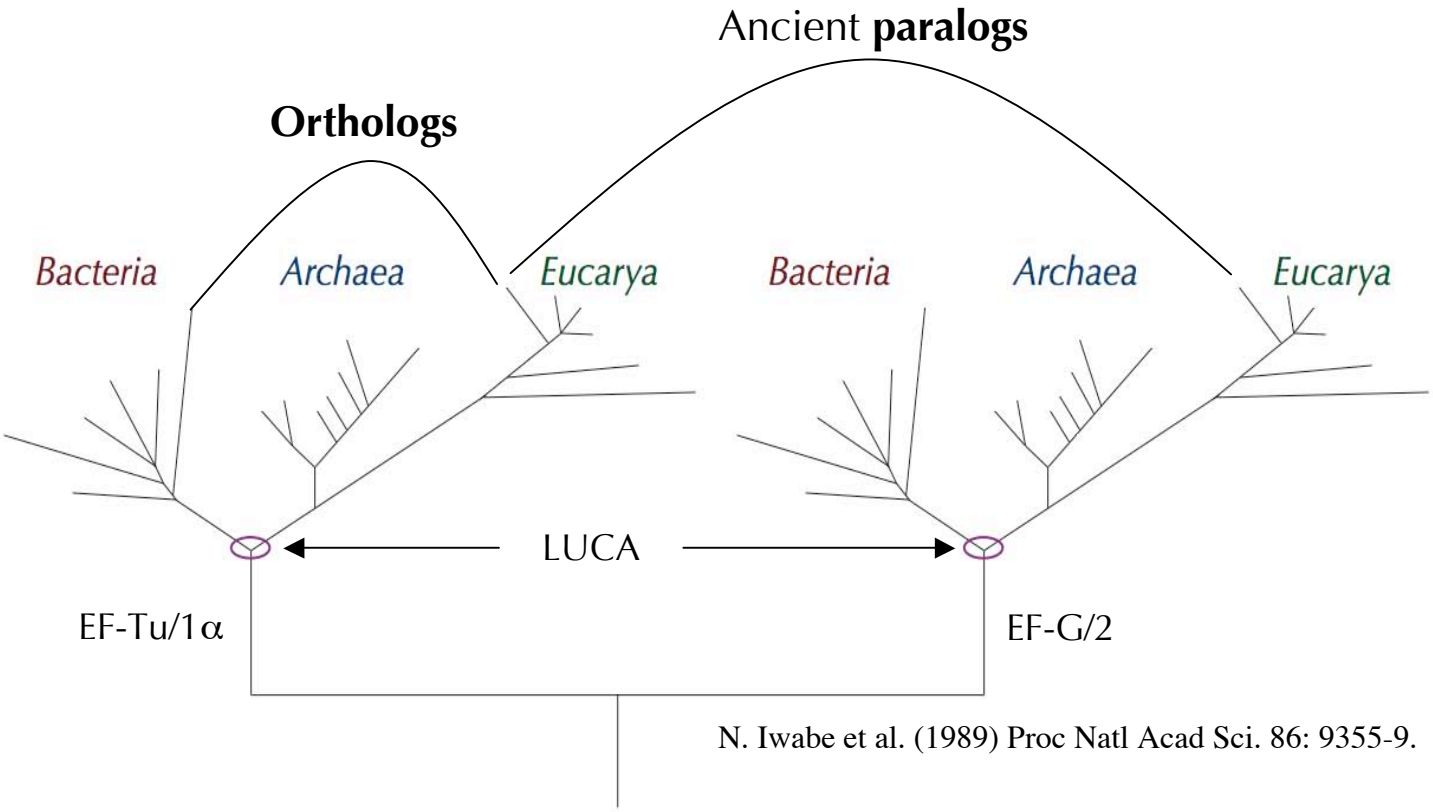
The Last Universal Common Ancestor (LUCA) is represented by the base of the UPT.

C. R. Woese (2000) Proc Natl Acad Sci. 97: 8392-6.

Forterre, Fondation des Treilles 1996.

Gene History

Gene Duplication Prior to LUCA



N. Iwabe et al. (1989) Proc Natl Acad Sci. 86: 9355-9.

Paralogs homologous proteins in the same genome.
Orthologs homologous proteins in different genomes.

Orthologous relationships can reveal organismal phylogeny.

Ancient **paralogous** relationships indicate gene history that extends earlier than LUCA, *i.e.*, prior to the origin of species.

Recent gene duplications can result in paralogs, which are also uninformative regarding organismal phylogeny.

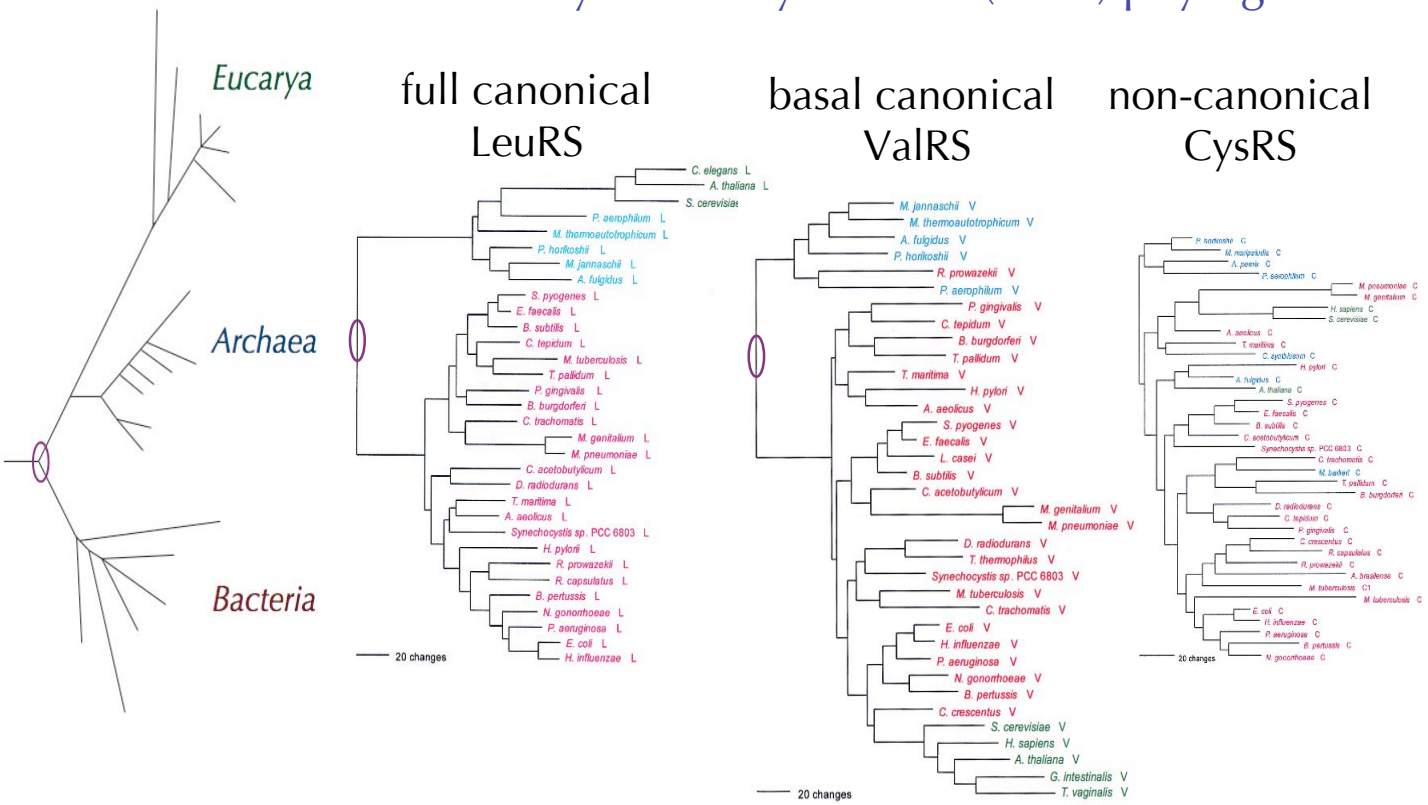
Organismal phylogeny is a subset of gene history, determining which part of the genetic record tells of organismal relationships is a challenge.

Gene History

Recurrence and Erosion of Canonical Phylogenetic Pattern

rRNA tree

Aminoacyl-tRNA Synthetase (aaRS) phylogenies



A number of gene phylogenies, *e.g.*, universal components of translation, transcription, protein secretory pathway (SecY), are congruent with rRNA.

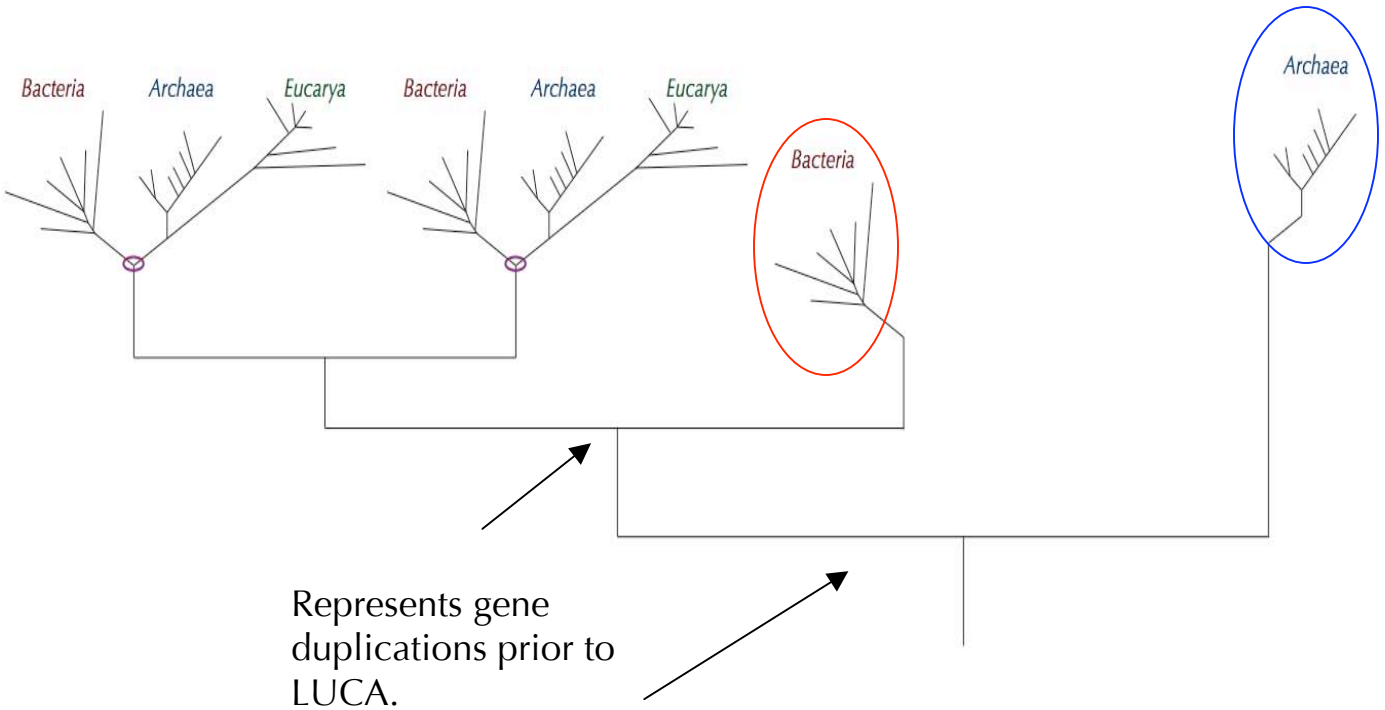
2/3 of aaRSs show at least basal canonical pattern.

Canonical pattern recurs in aaRS phylogenies, HGT patterns are unique.

C. R. Woese, G. J. Olsen, M. Ibba & D. Söll (2000) *MMBR*. 64, 202-236.
See also, Y. Wolf, L. Aravind, N. Grishin, and E. Koonin. (1999) *Genome Res*. 9:689–710.

Gene History

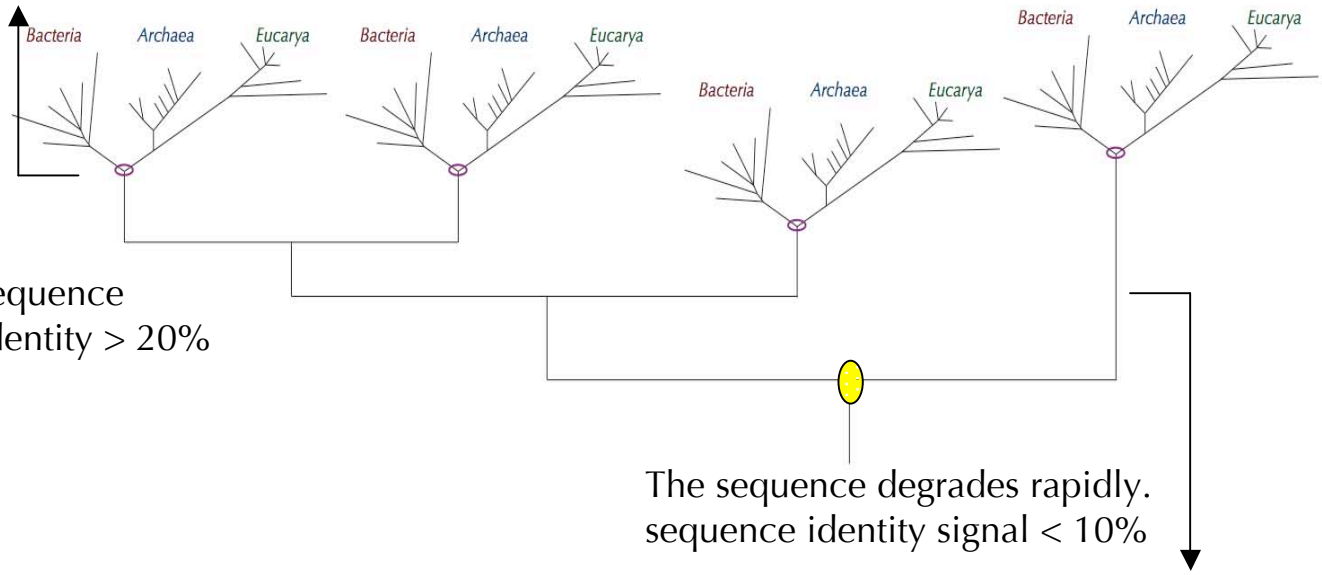
Phylogeny of Protein Families

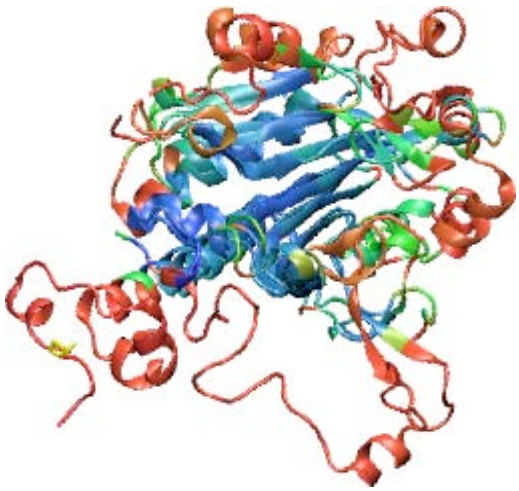


Although the phylogenetic distribution is limited for the circled genes, we can infer that these gene must have been extant prior to & in LUCA.

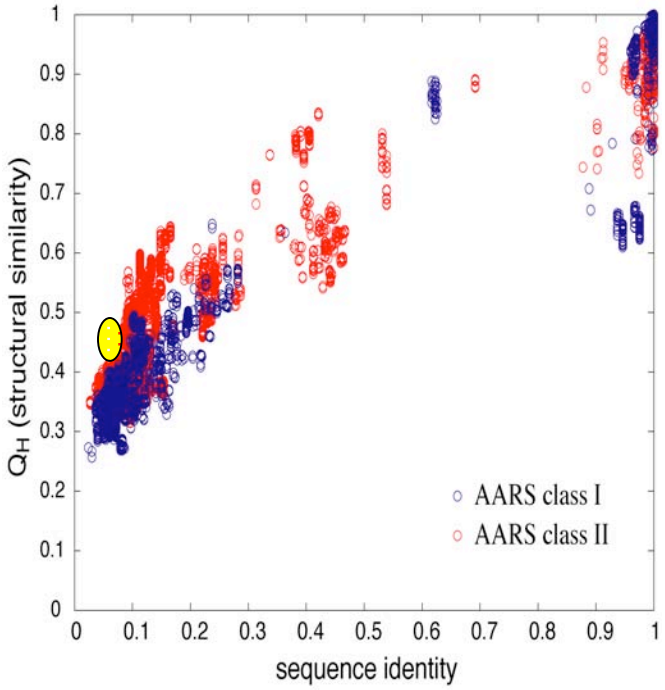
P. O'Donoghue, A. Sethi, C. R. Woese & Z. Luthey-Schulten. (2005) *PNAS* 102:19003-8.

The Relationship Between Sequence & Structure



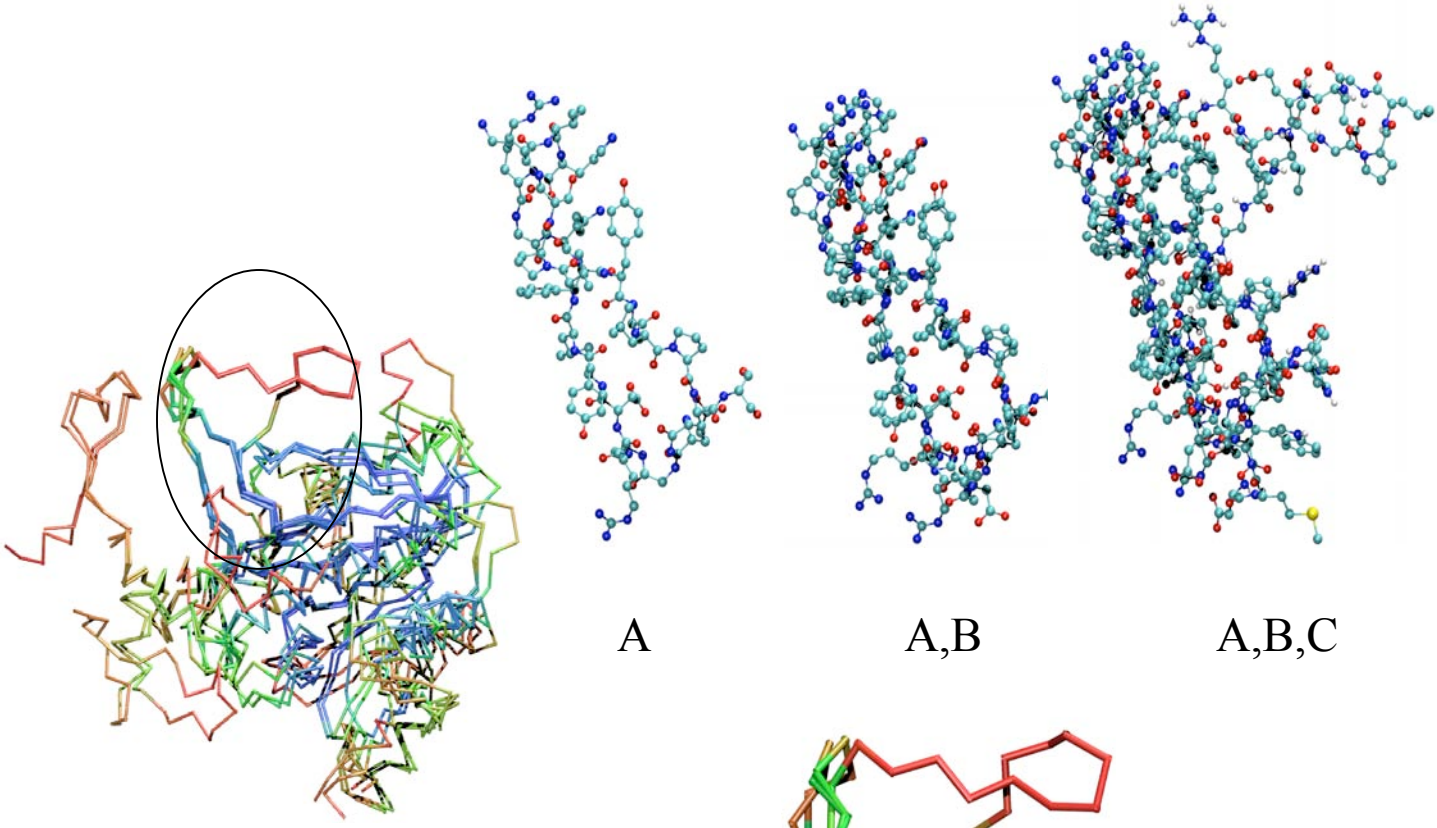


Structural superposition of AlaRS & AspRS.
Sequence id = 0.055, Q_H= 0.48

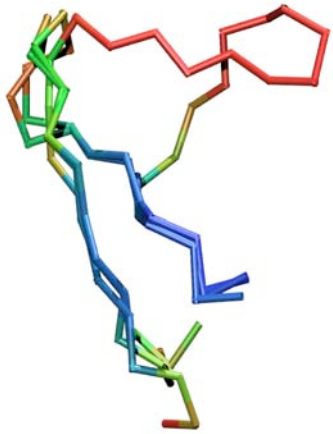


O'Donoghue & Luthey-Schulten (2003) *MMBR* 67: 550–73.

Protein Homology in Structure and Sequence



3 homologous structures;
2 closely related (A,B),
1 more distant (C).

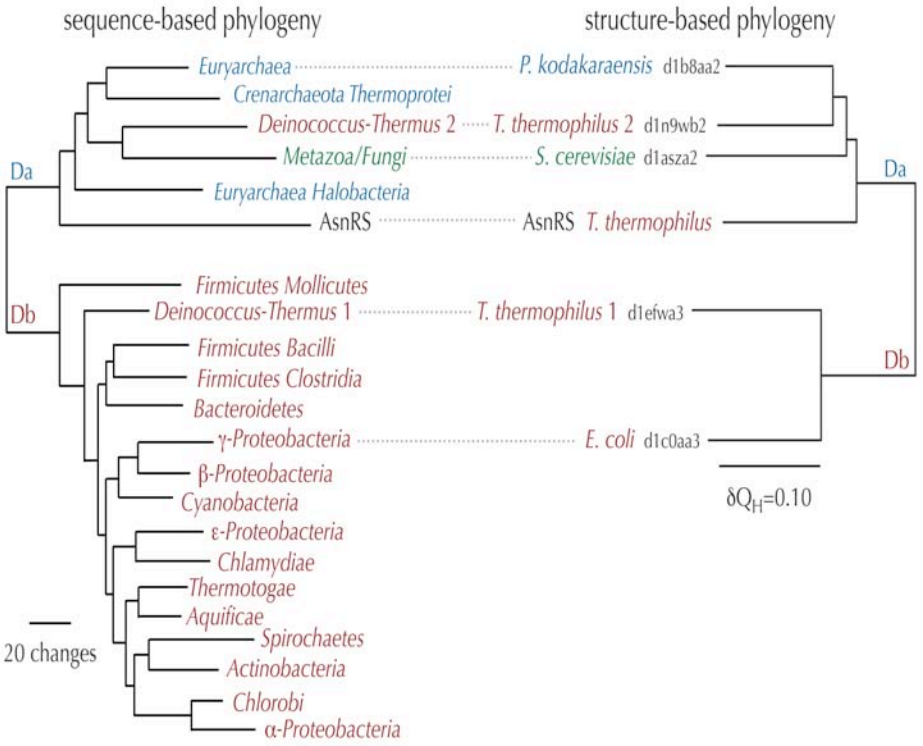


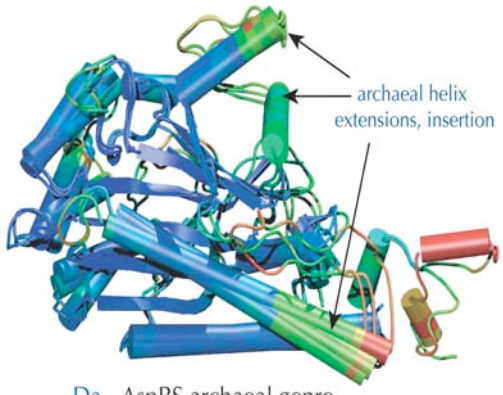
Overlap of protein backbones.

```
A      E---GARDFLV-PYRHE----------PGLFYALPQS
B      -E--GARDYLV-PSRVH----------KGKFYALPQS
C      ---DMWDTFWLT-GE--GFRLEGPLGEEVEGRLLLRTH
```

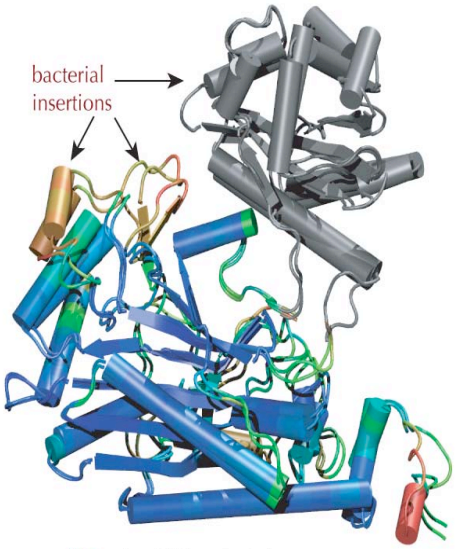
Evolutionary History in Sequence & Structure

aspartyl-tRNA synthetase





Da - AspRS archaeal genre

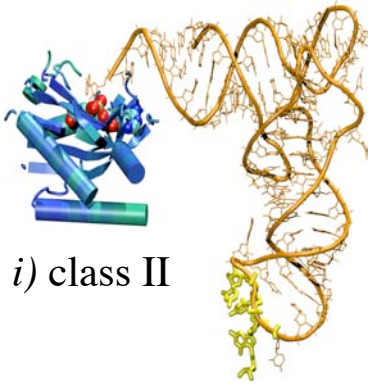


Db - AspRS bacterial genre

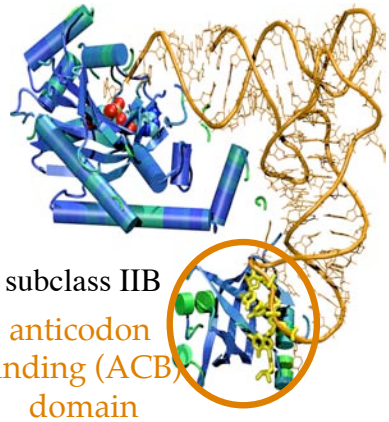
Congruence justifies using structure to trace back evolutionary events beyond the reach of sequence phylogeny.

O'Donoghue & Luthey-Schulten (2003) *MMBR* 67: 550–73.

Evolution of Structure and Function in AspRS

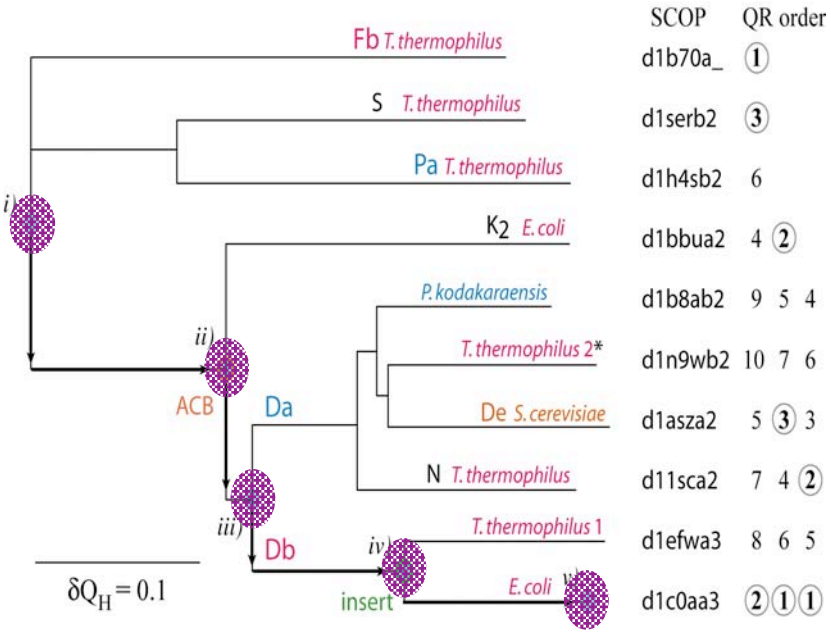


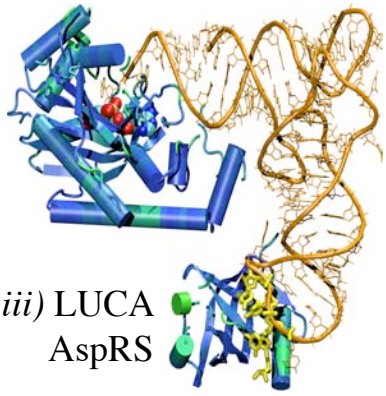
i) class II



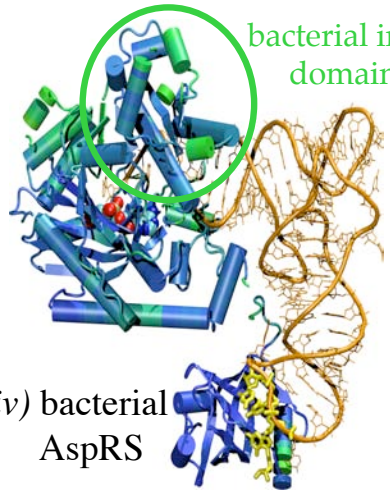
ii) subclass IIB

anticodon binding (ACB) domain



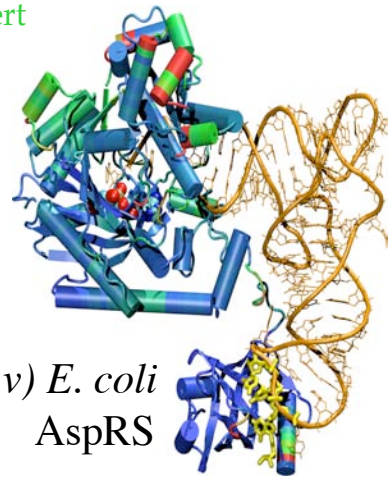


iii) LUCA AspRS



bacterial insert domain

iv) bacterial AspRS



v) E. coli AspRS

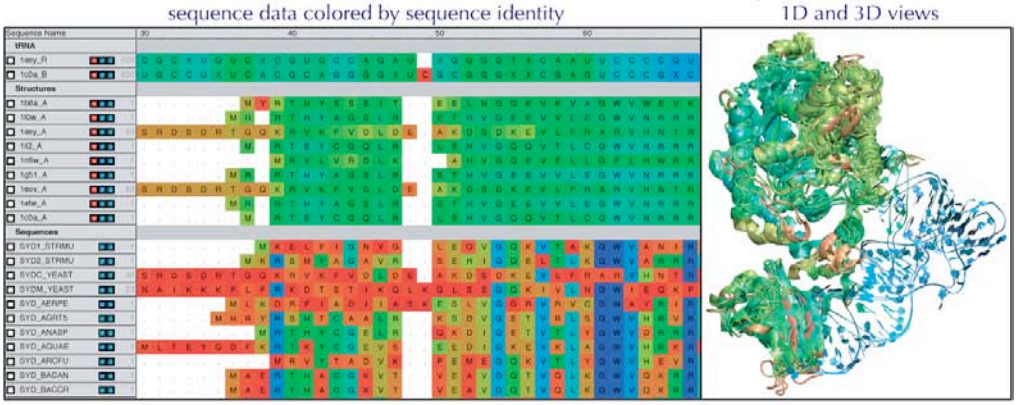
O'Donoghue & Luthey-Schulten (2005) *JMB* 346:875-894.

Multiseq 2.0 in VMD 1.8.5

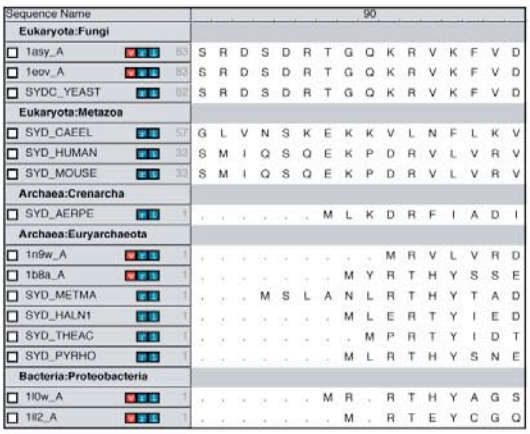
http://www.scs.uiuc.edu/~schulten/multiseq

View structural data colored by structural conservation and sequence data colored by sequence identity

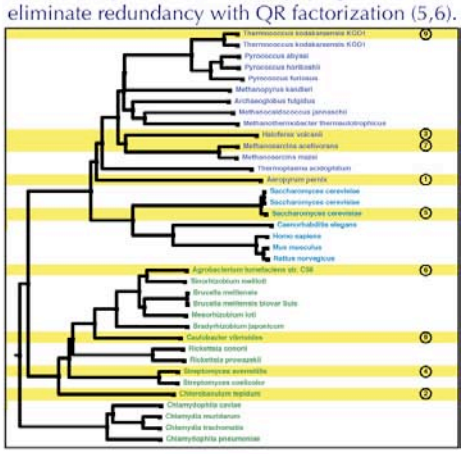
Synchronization between 1D and 3D views



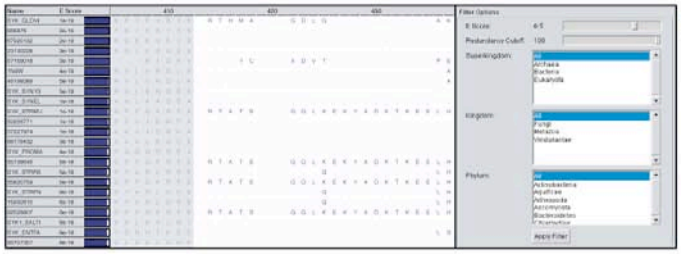
Group data by taxonomic classification



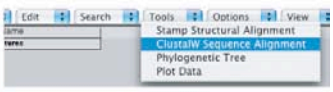
View sequence or structure phylogenies and eliminate redundancy with QR factorization (5,6).



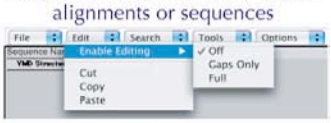
Import data directly from BLAST databases



Align sequences with Clustal



Sequence Editor: manually adjust alignments or sequences

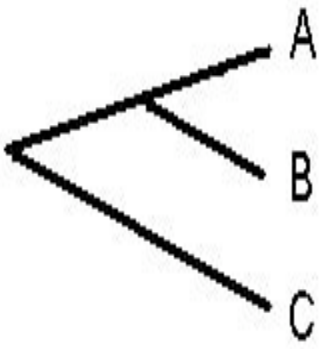


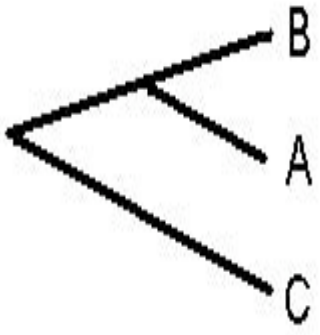
Elijah Roberts, John Eargle, Dan Wright & Zaida Luthey-Schulten. *BMC Bioinformatics* 2006, **7**:382

Tree representations

These 2 trees are equivalent, line lengths represent the evolutionary distance.

The trees indicate that A, B & C have evolved at equal rates since their divergence from a common ancestor, i.e., a constant molecular clock is assumed.

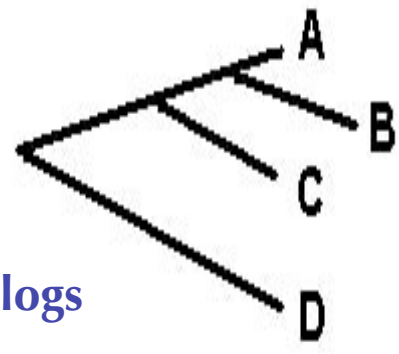




Not all genes (or organisms) evolve at the same rate.

A & B share the most recent common ancestry, but B has evolved with a faster "clock" than A.

This is one reason that organisms can share a recent common ancestor, but have more distantly related genes than expected. (HGT, and loss of orthologs are other reasons.)

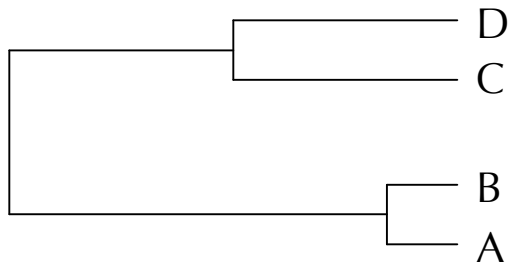


Adapted From Gary Olsen's notes on classification and phylogeny:
http://www.bact.wisc.edu/Bact303/phylogeny

Algorithms & Programs

Algorithmic or Clustering Methods

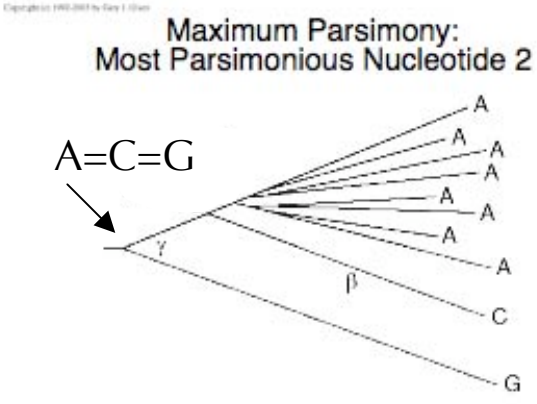
Add sequences to a tree according to similarity relationships.
Produces one tree.



Optimality Methods

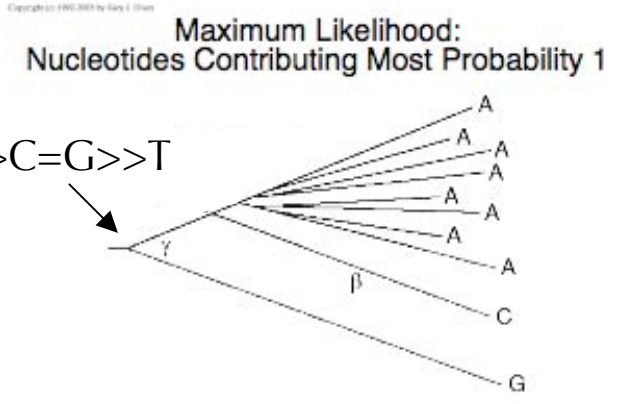
Heuristic search through the space of possible trees.
One tree is optimal according to:

Parsimony
(fewest changes)



A, C and G are equally parsimonious as the second ancestral nucleotide (requires 2 changes).

Maximum Likelihood
(most probable tree)



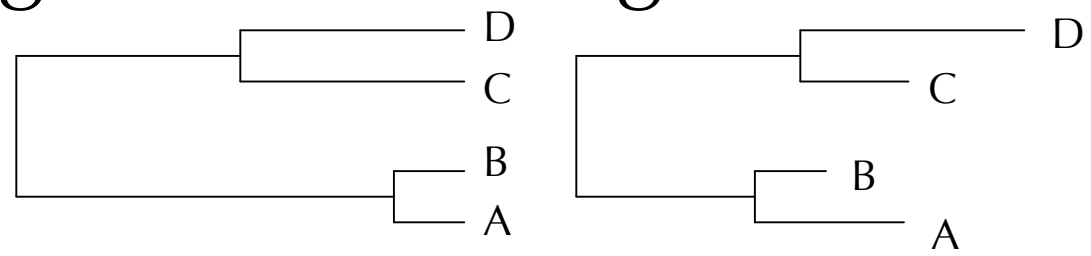
A is much more likely to give rise to the observed descendants of ancestor 1 than are C or G, and T is particularly bad.

Each alignment position is considered independently on a test tree.

Branch topologies and lengths are sampled until the best tree is found.

Adapted from Gary Olsen at http://geta.life.uiuc.edu/~gary/

Algorithms & Programs



	Methods That Impose a Molecular Clock	Methods That Do Not Impose a Molecular Clock
Clustering Methods	UPGMA WPGMA Single-linkage Complete-linkage	Neighbor-joining Phylip's neighbor
Objective Criterion-Based Methods	Least-squares distance (*e.g.*, KITSCH) Maximum likelihood (*e.g.*, dnamlk)	Least-squares distance (*e.g.*, FITCH) Minimum evolution Maximum parsimony Maximum likelihood (*e.g.*, dnaml, fastDNAml, protml) Bayesian (*e.g.*, MrBayes)

Other Considerations

Substitution cost matrix. (for distance & parsimony)

Cost of replacing one nucleic acid (or amino acid) for another.

Evolutionary models (for likelihood).

Invariant positions, evolutionary rate heterogeneity among positions, can estimate rates of change from one base (or amino acid) to another from the alignment data.

Programs

PHYLIP http://evolution.genetics.washington.edu/phylip.html
PAUP http://paup.csit.fsu.edu/
http://paup.csit.fsu.edu/paupfaq/faq.html
PHYML http://atgc.lirmm.fr/phyml/

Adapted from Gary Olsen at http://geta.life.uiuc.edu/~gary/

Phylogenomics

What is Phylogenomics?

Inference of phylogeny, for some group of taxa,based on the comparative analysis of some property of genomes or gene clusters/groups.

The first paper to mentioned "Phylogenomics" is a review.
Eisen JA. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis.
Genome Res. 1998 Mar;8(3):163-7.

Only 135 citations in pubmed for "phylogenomic*".

What is being compared?

Genome Identity
Gene Presence/Absence
"Genome Conservation"
Gene expression

Why phylogenomics?

Hypothesis: Additional information from genome sequences should help resolve evolutionary histories.

i. Tree of life
ii. Tracking pathogenic lineages across a population. Identify infection source & virulence genes.
iii. Gene annotation, protein function prediction.

Tree of Life 1

PNAS

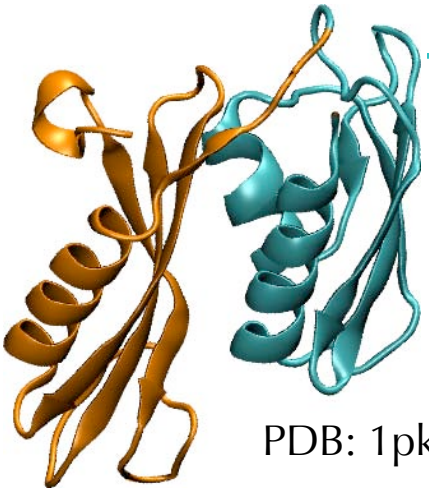
Phylogeny determined by protein domain content

Song Yang*, Russell F. Doolittle*, and Philip E. Bourne†‡

Departments of *Chemistry and Biochemistry and †Pharmacology and San Diego Supercomputer Center, University of California at San Diego, La Jolla, CA 92093

Contributed by Russell F. Doolittle, November 26, 2004

Protein Domains and Superfamilies



PDB: 1pkp

4. Superfamily: Ribosomal protein S5 domain 2-like [54211]

Families:

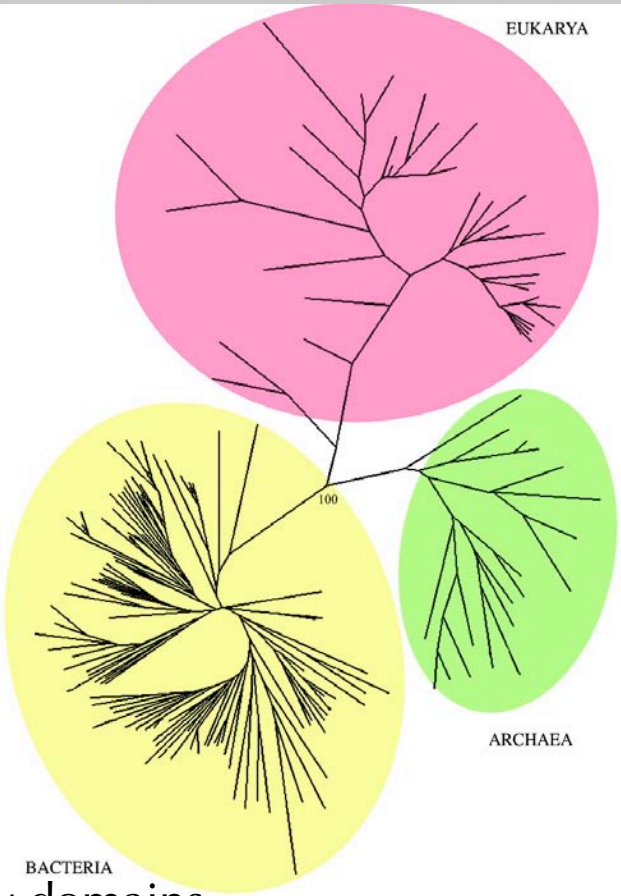
http://scop.mrc-lmb.cam.ac.uk/scop/

1. Translational machinery components [54212] (5)
2. RNase P protein [54220] (3)
3. DNA gyrase/MutL, second domain [54224] (7)
4. Hsp90 middle domain [102755] (1)
   related to the DNA gyrase/MutL family; contains extra C-terminal alpha/beta subdomain
5. Ribonuclease PH domain 1-like [54229] (4)
6. GHMP Kinase, N-terminal domain [54232] (9)
7. Early switch protein XOL-1, N-terminal domain [89824] (1)
   diverged from the GHMP Kinase family; lost the ATP-binding site
8. UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase LpxC [89827] (1)
   duplication; there are two structural repeats of this fold; each repeat is elaborated with additional structures forming the active site
9. Imidazole glycerol phosphate dehydratase [102766] (1)
   duplication; there are two structural repeats of this fold
10. ATP-dependent protease Lon (La), catalytic domain [102769] (1)
    contains extra C-terminal alpha/beta subdomain
11. Hypothetical protein YigZ, N-terminal domain [102772] (1)
    modification of the common fold; contains extra alpha-beta unit after strand 2, the extra strand is inserted between strands 3 and 4

Presence/Absence Matrix

```
   Protein Domain Superfamilies
              1 2 3 4 5 6 7 … M

Genome 1 = [0 0 0 1 1 0 1 … 1]
Genome 2 = [0 0 0 1 1 0 0 … 1]
…
Genome N = [1 1 1 0 1 1 1 … 0]
```



Evolutionary Distance

D = A′ /(A′+AB)

A′ is number of superfamily domains.

AB is the number of shared superfamily domains.

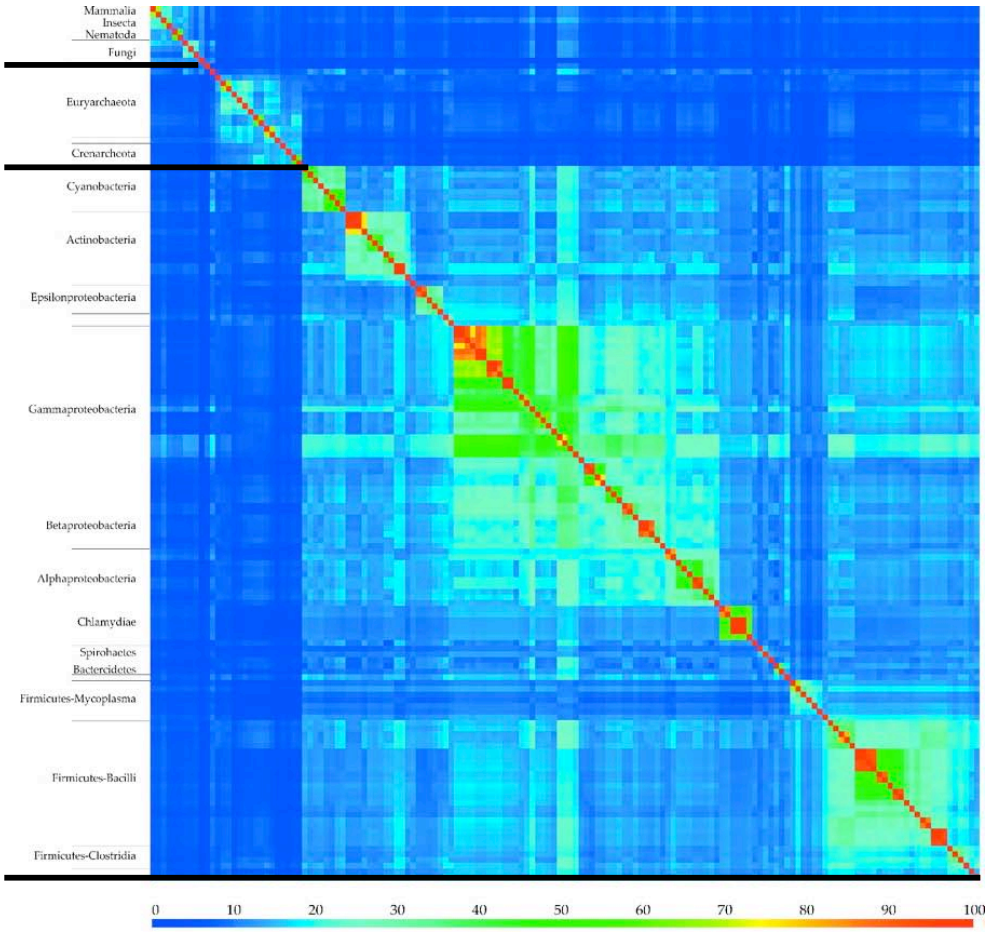
Tree of Life 2

Genome Conservation

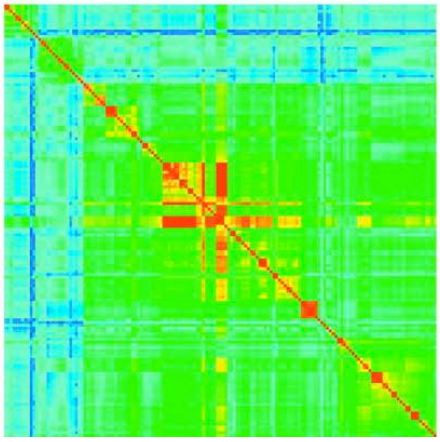
Eucarya

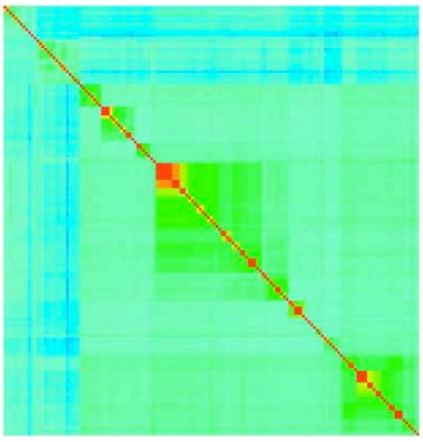
Archaea

Bacteria



Gene presence/ absence





average ortholog similarity

Genome Conservation weights the average sequence similarity with the number of homologs between two genomes.

Genome conservation produces a tree that is mainly congruent with rRNA.

Kunin V, Ahren D, Goldovsky L, Janssen P, Ouzounis CA.Nucleic Acids Res. 2005;33:616-21.

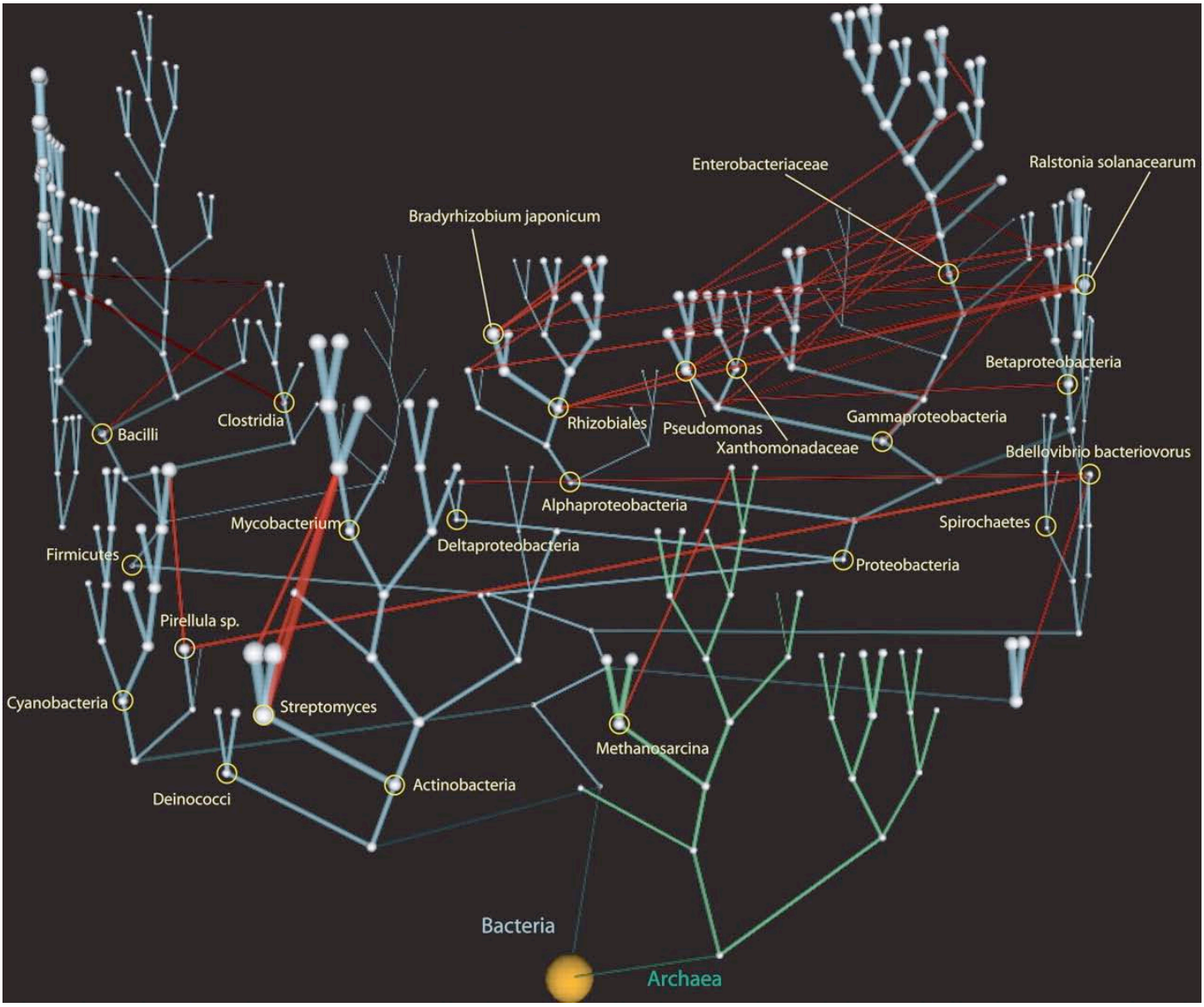
Tree (net) of Life 2

Letter

The net of life: Reconstructing the microbial phylogenetic network

Victor Kunin,[1] Leon Goldovsky, Nikos Darzentas, and Christos A. Ouzounis[2]

Computational Genomics Group, The European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge CB10 1SD, United Kingdom



Blue branches - vertical inheritance

Red branches - horizontal gene transfer

Genome Res. 2005 Jul;15(7):954-9.

Pathogen Evolution

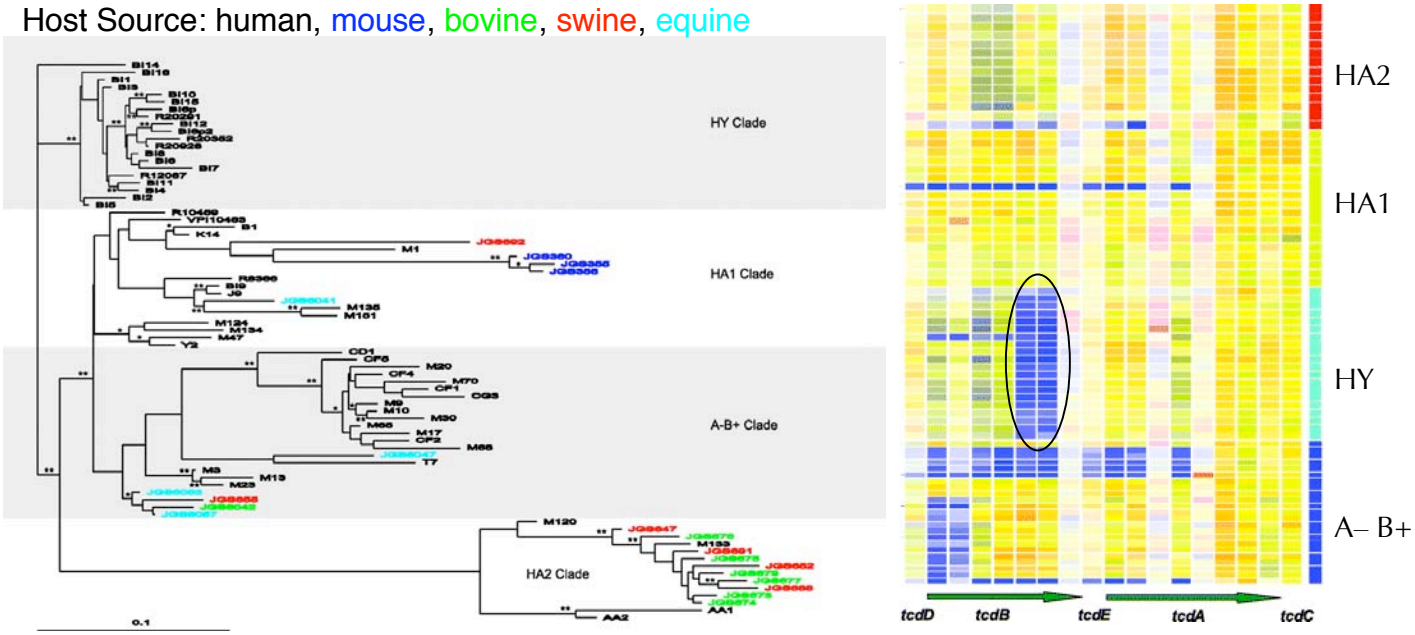
Identify Virulence Factors

tcdA & tcdB encode proteins that synthesize toxin A & B, respectively.

Apparent deletion or highly divergent sequences at the end of *tcdB* is specific to the hypervirulent (HY) strains.

Microarray experiment tests if DNA from other strains hybridizes (yellow) or not (blue) to *C. difficile* 630.

Host Source: human, mouse, bovine, swine, equine



Geographic Spread of Pathogen Lineages

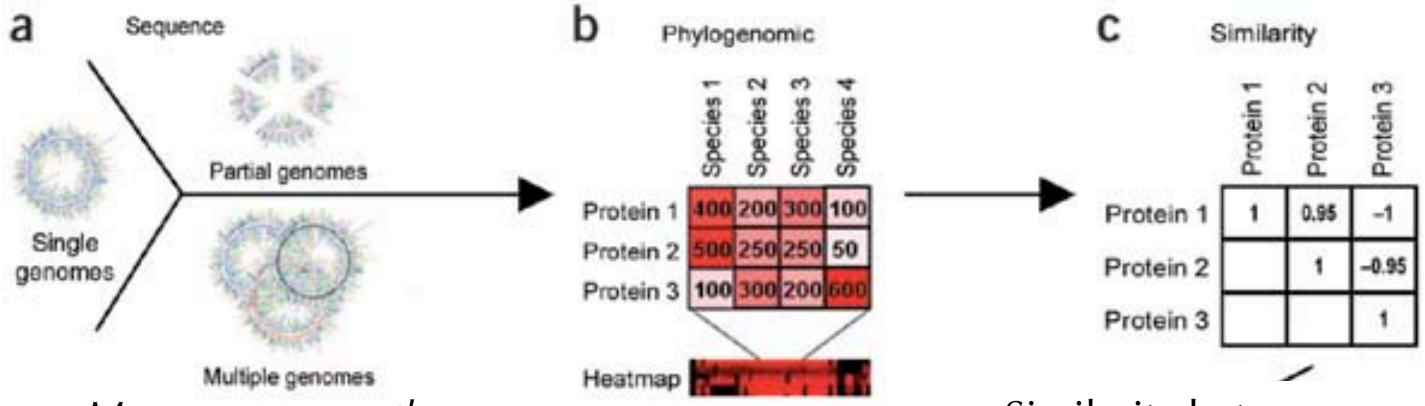
> "The 20 [hypervirulent] strains were from diverse locations in the United States, Canada, and the United Kingdom, confirming their transcontinental spread."

Microarrays can be used to "type" strains.

Stabler et al. J Bacteriol. 2006 Oct;188(20):7297-305.

Function Prediction

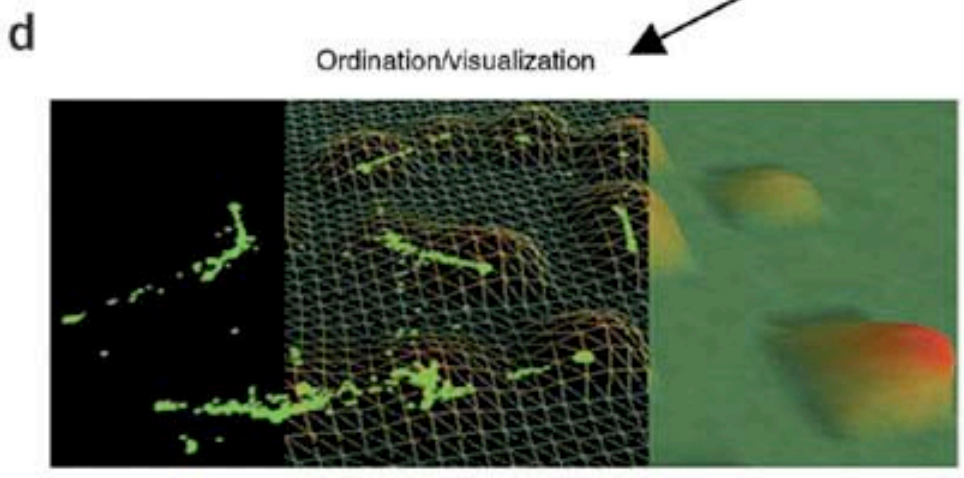
"proteins that function together in a pathway or structural complex are likely to evolve in a correlated fashion"
Eisenberg & co-workers. Proc. Natl. Acad. Sci. USA 96, 4285-288 (1999).



Myxococcus xanthus ORFs blasted against genomes.

Blast bit scores

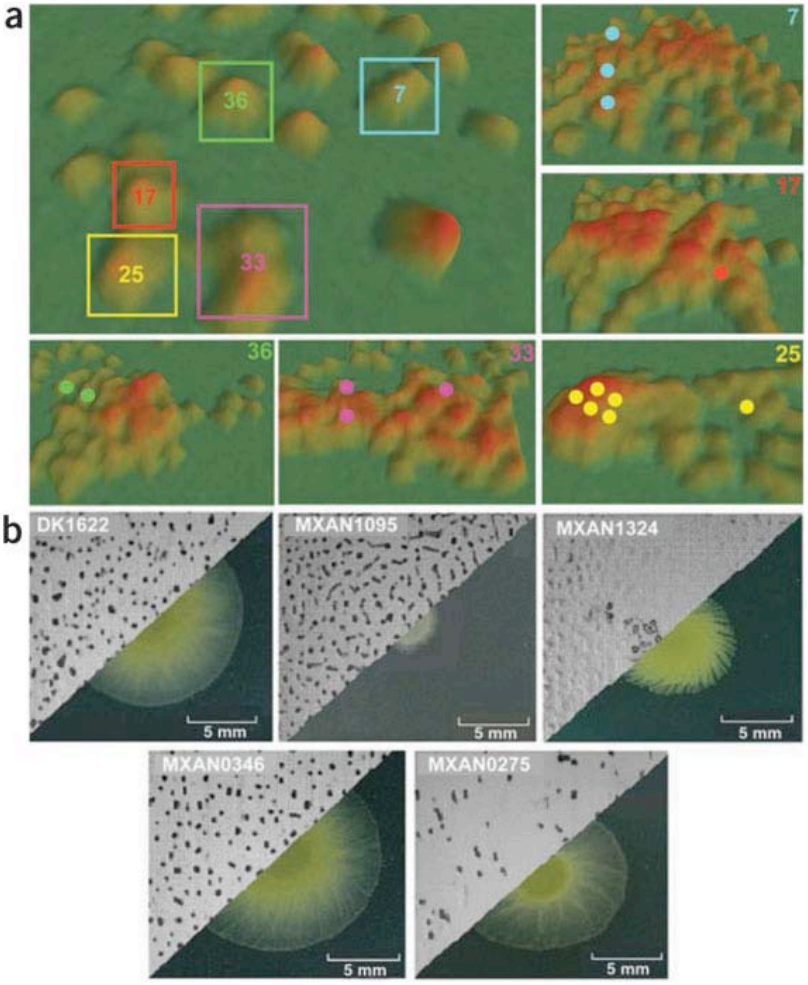
Similarity between Blast hit lists.



"the products of genes with similar evolutionary histories cluster together in mountains, and where local height is proportional to the density of proteins within an area."

Srinivasan BS, Caberoy NB, Suen G, Taylor RG, Shah R, Tengra F, Goldman BS,Garza AG, Welch RD. Functional genome annotation through phylogenomic mapping. Nat Biotechnol. 2005 Jun;23(6):691-8.

Function Prediction



Phylogenetic clustering of all genes in the *Myxococcus xanthus* genome.

"Mountains" include genes with similar evolutionary histories.

"M. xanthus (δ-proteo) can exist as both a single-species biofilm and a free-living cell.

The biofilm is a self-organizing predatory swarm that has many of the characteristics of a multicellular organism."

Putative annotation	Swarm expansion	Aggregation
MXAN1095 TonB system transport protein ExbB/TolQ	-	+
MXAN1324 TPR domain protein	-	-
MXAN0346 TolB protein, putative	+	+
MXAN0275 Twitching mobility protein	-	+

Genes with the same evolutionary history produce a similar phenotype when disrupted by plasmid insertion.

In summary, 12 of 15 ORFs targeted for disruption on the basis of phylogenomic proximity to known motility proteins produced obvious defects in swarm expansion and/or aggregation when inactivated.

Srinivasan BS, Caberoy NB, Suen G, Taylor RG, Shah R, Tengra F, Goldman BS,Garza AG, Welch RD. Functional genome annotation through phylogenomic mapping. Nat Biotechnol. 2005 Jun;23(6):691-8.

Assessing Phylogenomics

Advantages

The complete information in the genome sequence can be used for constructing the tree of life, monitoring pathogen evolution, locating virulence genes and predicting gene function.

When rRNA is nearly identical, whole genome comparisons provide a more sensitive measure of relationships.

Disadvantages

Not all genes have the same history.

Incongruent gene histories result from various sources:

1. loss of close orthologs
2. horizontal gene transfer
3. lineage specific evolutionary rate acceleration.

Data Selection

> "selecting only data that contain minimal nonphylogenetic signals takes full advantage of phylogenomics and markedly reduces incongruence." (Jeffroy et al. 2006 *Trends in Genetics* v22)
> Remove genes with different histories.

Explicit representation of Horizontal,Vertical Gene Transfer.

HGT & Cellular Character

J. Mol. Microbiol. Biotechnol. (2002) 4(4): 453-461. JMMB Research Article

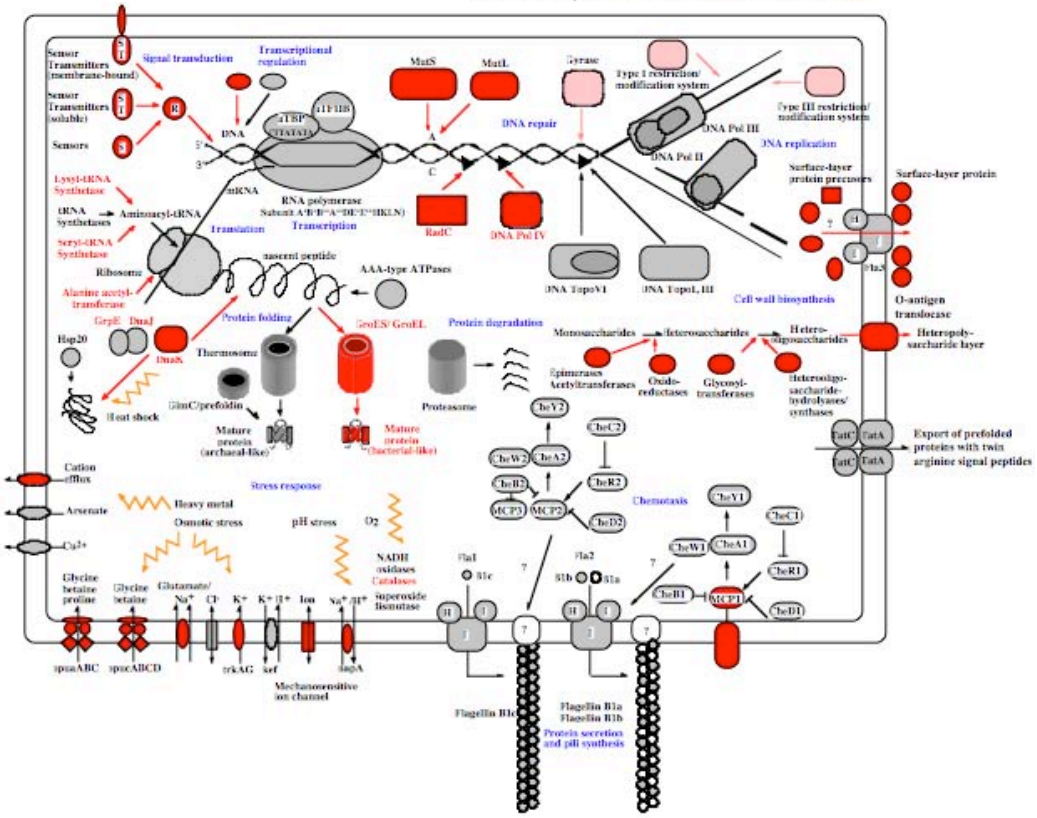
The Genome of *Methanosarcina mazei*: Evidence for Lateral Gene Transfer Between Bacteria and Archaea

Uwe Deppenmeier[1,2], Andre Johann[1], Thomas Hartsch[1,4], Rainer Merkl[3], Ruth A. Schmitz[2], Rosa Martinez-Arias[1], Anke Henne[1], Arnim Wiezer[1], Sebastian Bäumer[1], Carsten Jacobi[1,5], Holger Brüggemann[1], Tanja Lienard[2], Andreas Christmann[3], Mechthild Bömeke[1], Silke Steckel[1], Anamitra Bhattacharyya[4], Athanasios Lykidis[4], Ross Overbeek[4], Hans-Peter Klenk[1,7], Robert P. Gunsalus[5], Hans-Joachim Fritz[1,3], Gerhard Gottschalk[1,2*]

system and the presence of tetrahydrofolate-dependent enzymes. These findings might indicate that lateral gene transfer has played an important evolutionary role in forging the physiology of this metabolically versatile methanogen.

Introduction

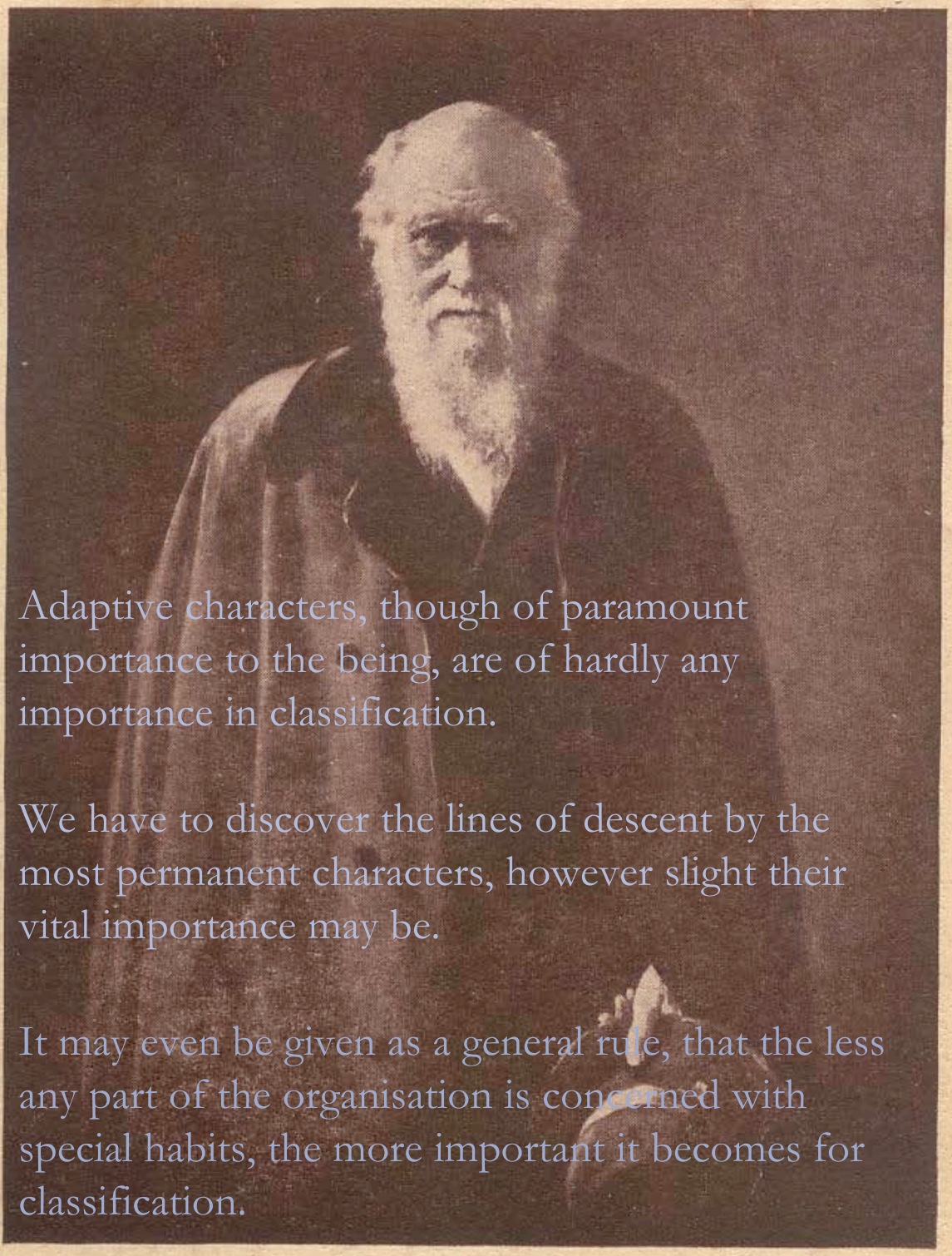
Methanosarcina species are obligate anaerobic Archaea which are indispensable members of anaerobic food chains



1/3 of *Methanosarcina mazei*'s genome is of bacterial origin (red).

(Black) Core energetic (methanogenesis), information processing genes, lipid membranes, and gene order are all characteristic of its archaeal relatives.

A genome may acquire a large fraction of foreign genes, without fundamentally transforming the core cellular subsystems or the cell itself into another type.



Adaptive characters, though of paramount importance to the being, are of hardly any importance in classification.

We have to discover the lines of descent by the most permanent characters, however slight their vital importance may be.

It may even be given as a general rule, that the less any part of the organisation is concerned with special habits, the more important it becomes for classification.

DARWIN