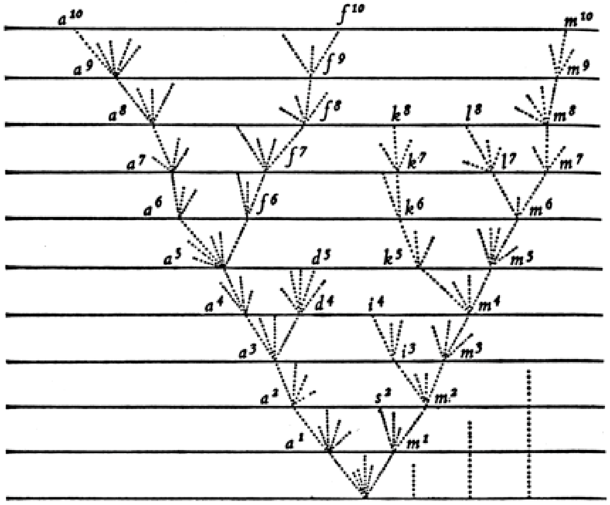
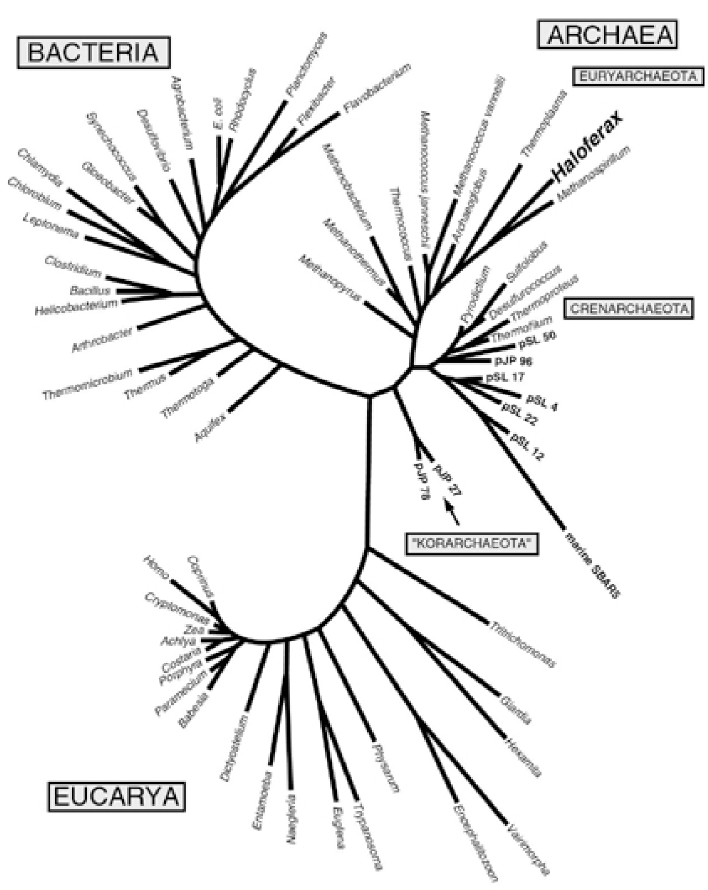
Phylogenomics

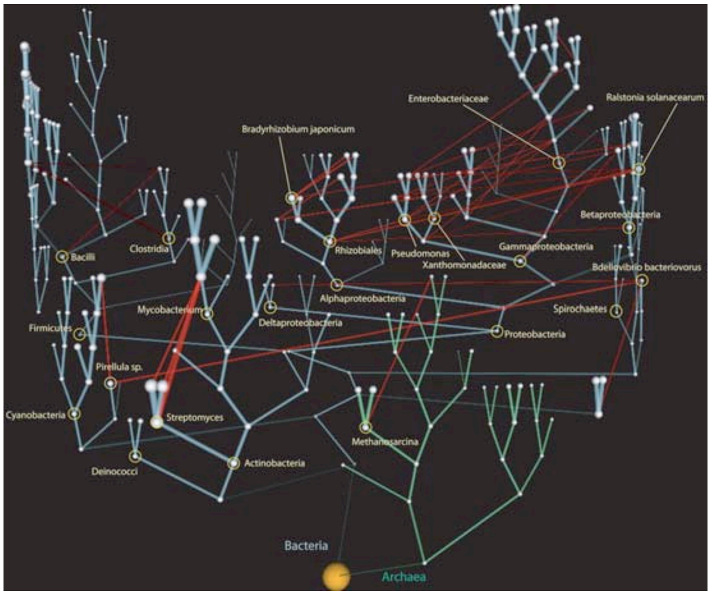
Gene History, Genome History and Organismal Phylogeny (genealogy of cellular life)



"Universal" Unrooted Phylogenetic Tree



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 & Bioinformatics
Patrick O'Donoghue

Feb 25, 2009

MB&B 452

Overview

1. Molecular Phylogenetics

Basis of Molecular Phylogenetics

Evolutionary relationships between organisms can be deduced from the comparison of homologous gene or protein sequences.

More ancient evolutionary events can be mapped by comparing three-dimensional structures of proteins.

Different genes have different histories

Two types of gene flow shape evolution:
Vertical Gene Transfer versus Horizontal Gene Transfer

2. Inferring (Computing) Phylogeny

Algorithmic methods

Optimization methods

3. Phylogenomics

What is it?

Three applications: tree of life
pathogen evolution
Macaca mulatta genome sequence

Assessing phylogenomic approaches

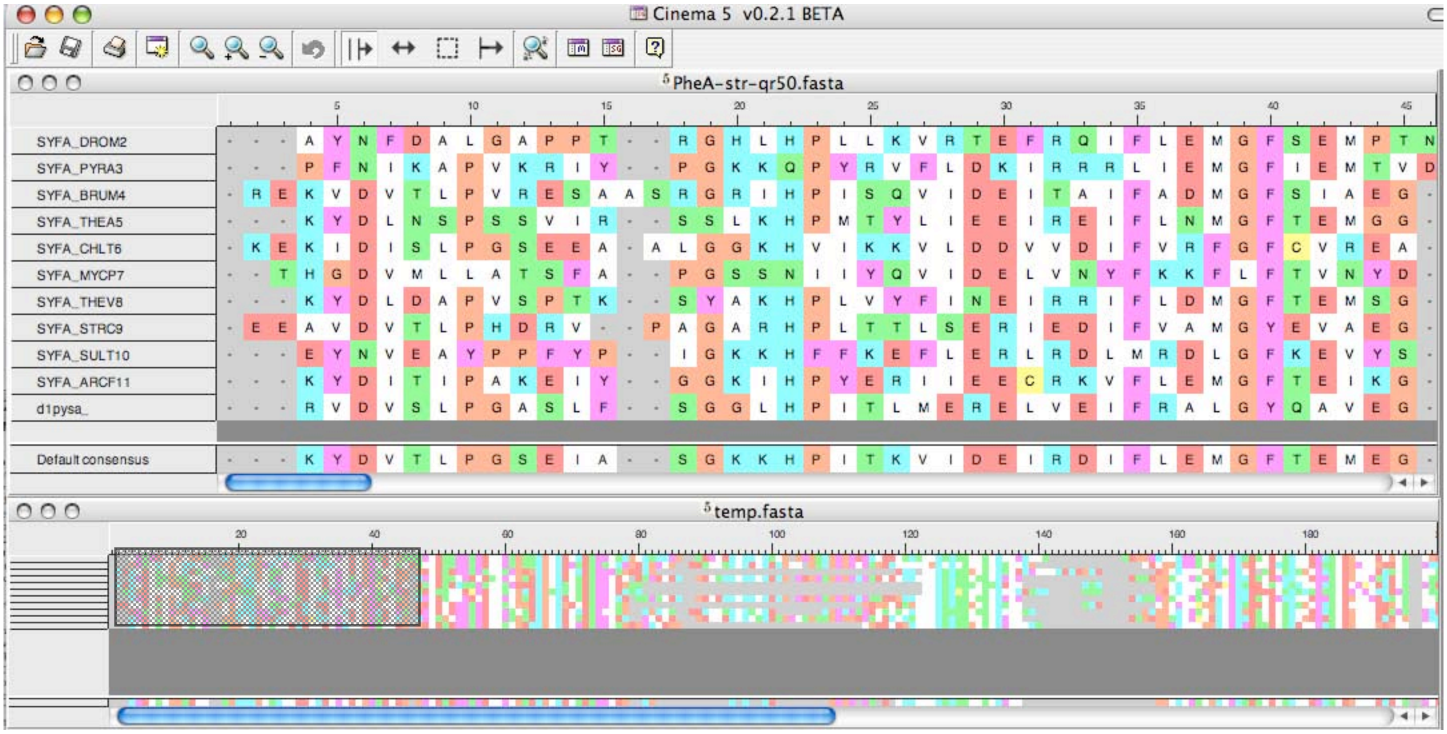
How genome evolution, HGT relate to cellular character.

Sequence Alignment

the basis of molecular phylogeny

Zuckerkandl, E. & Pauling, L. J. “Molecules as Documents of Evolutionary History” Theor. Biol. 8, 357366 (1965).

Comparing evolutionarily related hemoglobin sequences can be used to infer phylogeny.



```
                  5    10   15   20   25   30   35   40   45
SYFA_DROM2        ---AYNFDALGAPPT--RGHLHPLLKVRTEFRQIFLEMGFSEMPTN
SYFA_PYRA3        ---PFNIKAPVKRIY--PGKKQPYRVFLDKIRRRLIEMGFIEMTVD
SYFA_BRUM4        -REKVDVTLPVRESAASRGRIHPISQVIDEITAIFADMGFSIAEG-
SYFA_THEA5        ---KYDLNSPSSVIR--SSLKHPMTYLIEEIREIFLNMGFTEMGG-
SYFA_CHLT6        -KEKIDISLPGSEEA-ALGGKHVIKKVLDDVVDIFVRFGFCVREA-
SYFA_MYCP7        --THGDVMLLATSFA--PGSSNIIYQVIDELVNYFKKFLFTVNYD-
SYFA_THEV8        ---KYDLDAPVSPTK--SYAKHPLVYFINEIRRIFLDMGFTEMSG-
SYFA_STRC9        -EEAVDVTLPHDRV--PAGARHPLTTLSERIEDIFVAMGYEVAEG-
SYFA_SULT10       ---EYNVEAYPPFYP--IGKKHFFKEFLERLRDLMRDLGFKEVYS-
SYFA_ARCF11       ---KYDITIPAKEIY--GGKIHPYERIIEECRKVFLEMGFTEIKG-
d1pysa_           ---RVDVSLPGASLF--SGGLHPITLMERELVEIFRALGYQAVEG-

Default consensus ---KYDVTLPGSEIA--SGKKHPITKVIDEIRDIFLEMGFTEMEG-
```

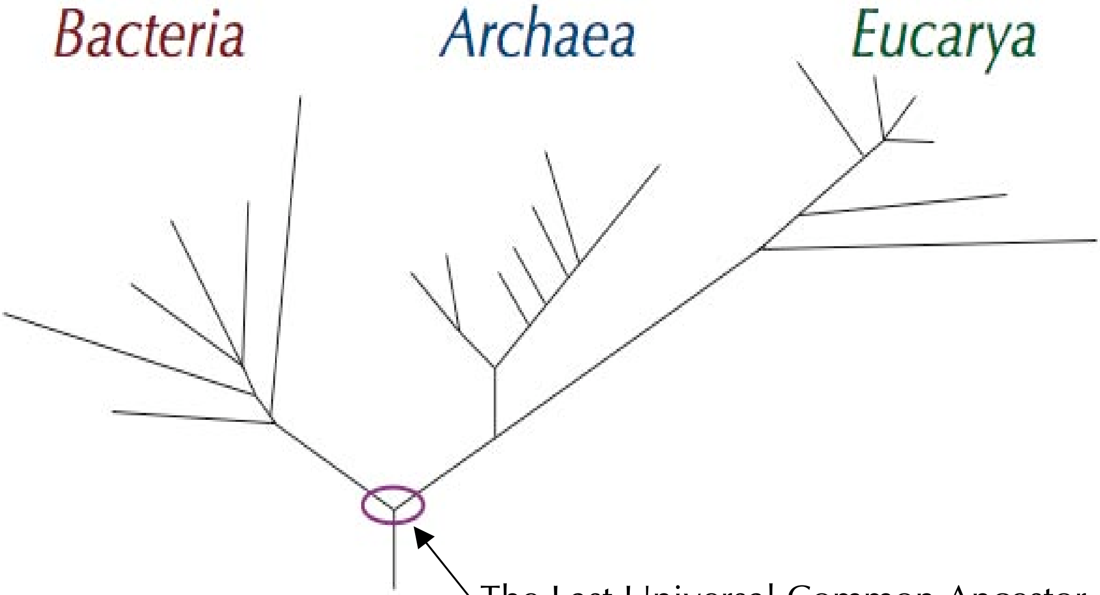
Homologous protein or nucleic acid sequences from different organisms can be aligned.

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.

In 1977, Carl Woese and colleagues used sequence differences in rRNA sequences to show how to draw a phylogenetic tree that included all cellular life.

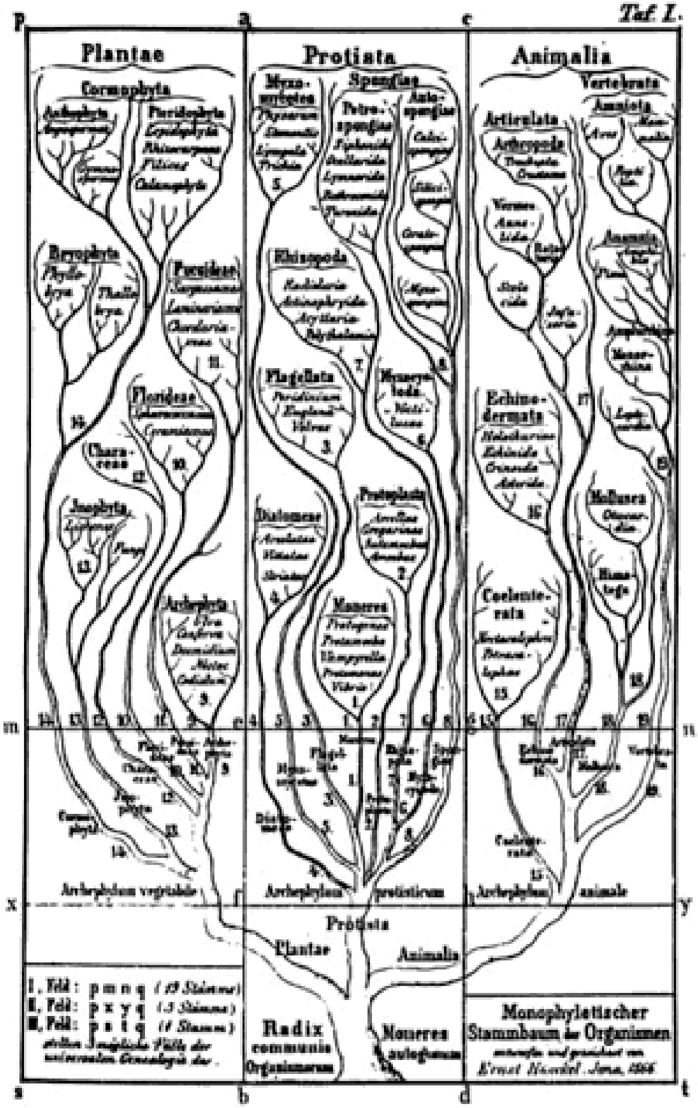
Evolutionary relationships between organisms as different as E. coli and humans can only be established at the molecular level using genes (such as the ribosome) that are found in all cellular organisms.

Morphological similarities, which can be used to determine evolutionary relationships among higher eukaryotic groups, e.g., primates, are not useful in classifying microorganisms.

C. R. Woese, G. E. Fox (1977) PNAS 74: 5088-5090.
C. R. Woese (2000) Proc Natl Acad Sci. 97: 8392-6.

Tree of life

Haeckel's 1866 tree



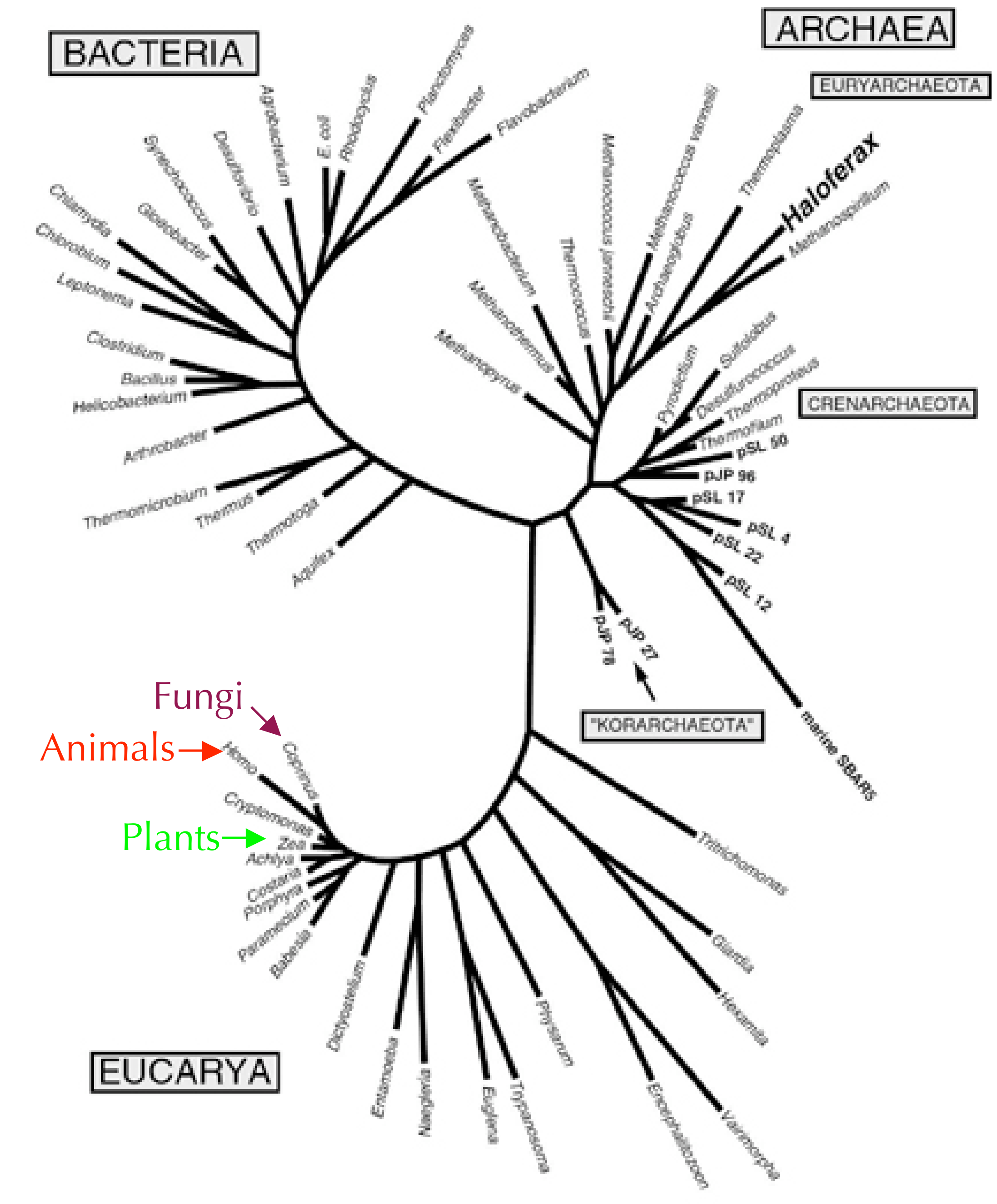
How are modern organisms related to one another?

How did the diversity of life come into being?

Does the five kingdom concept hold up in the genomic era?

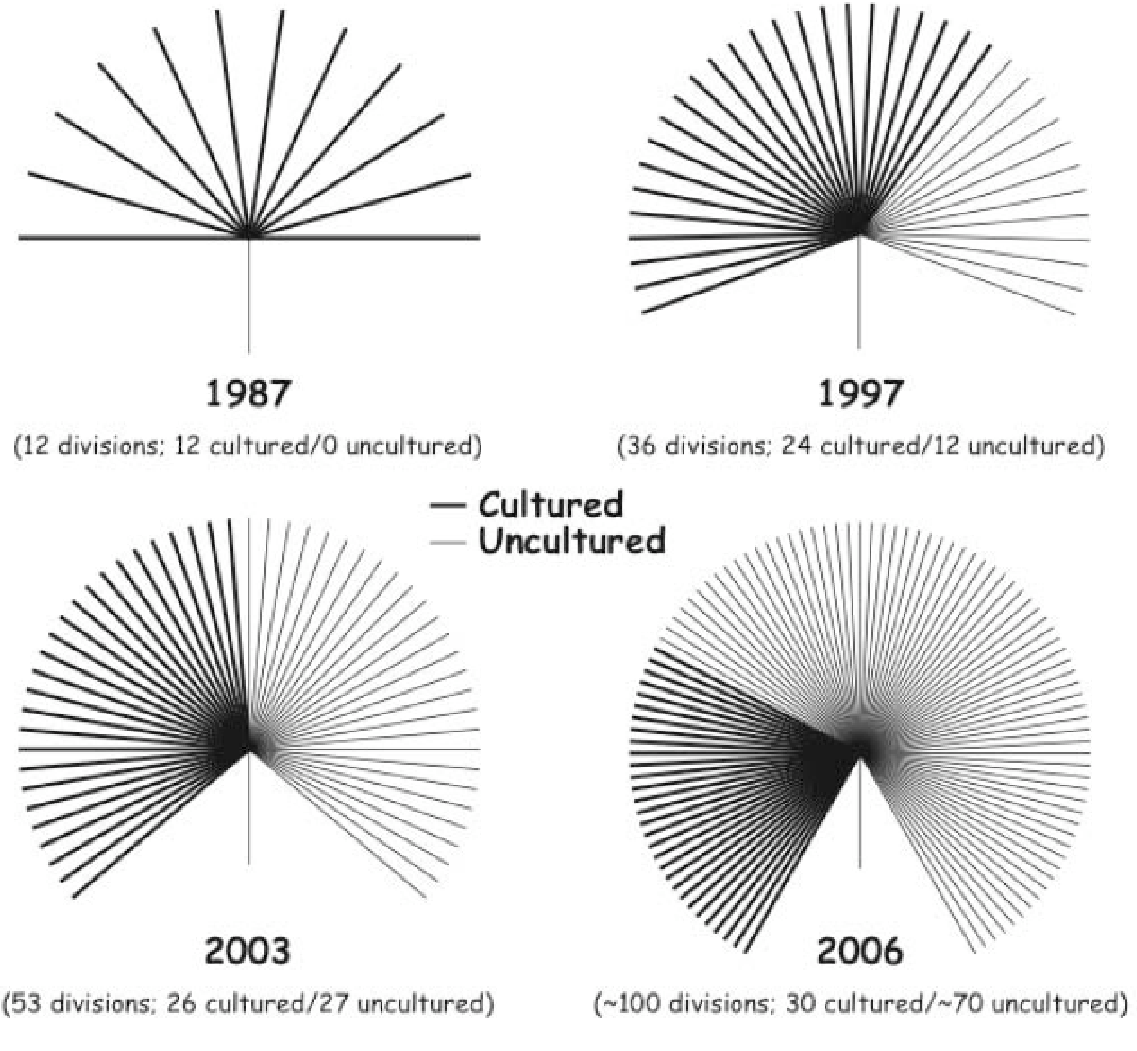
Three domain tree of life

"Universal" Unrooted Phylogenetic Tree



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

A glimpse of microbial diversity



1987
(12 divisions; 12 cultured/0 uncultured)

1997
(36 divisions; 24 cultured/12 uncultured)

2003
(53 divisions; 26 cultured/27 uncultured)

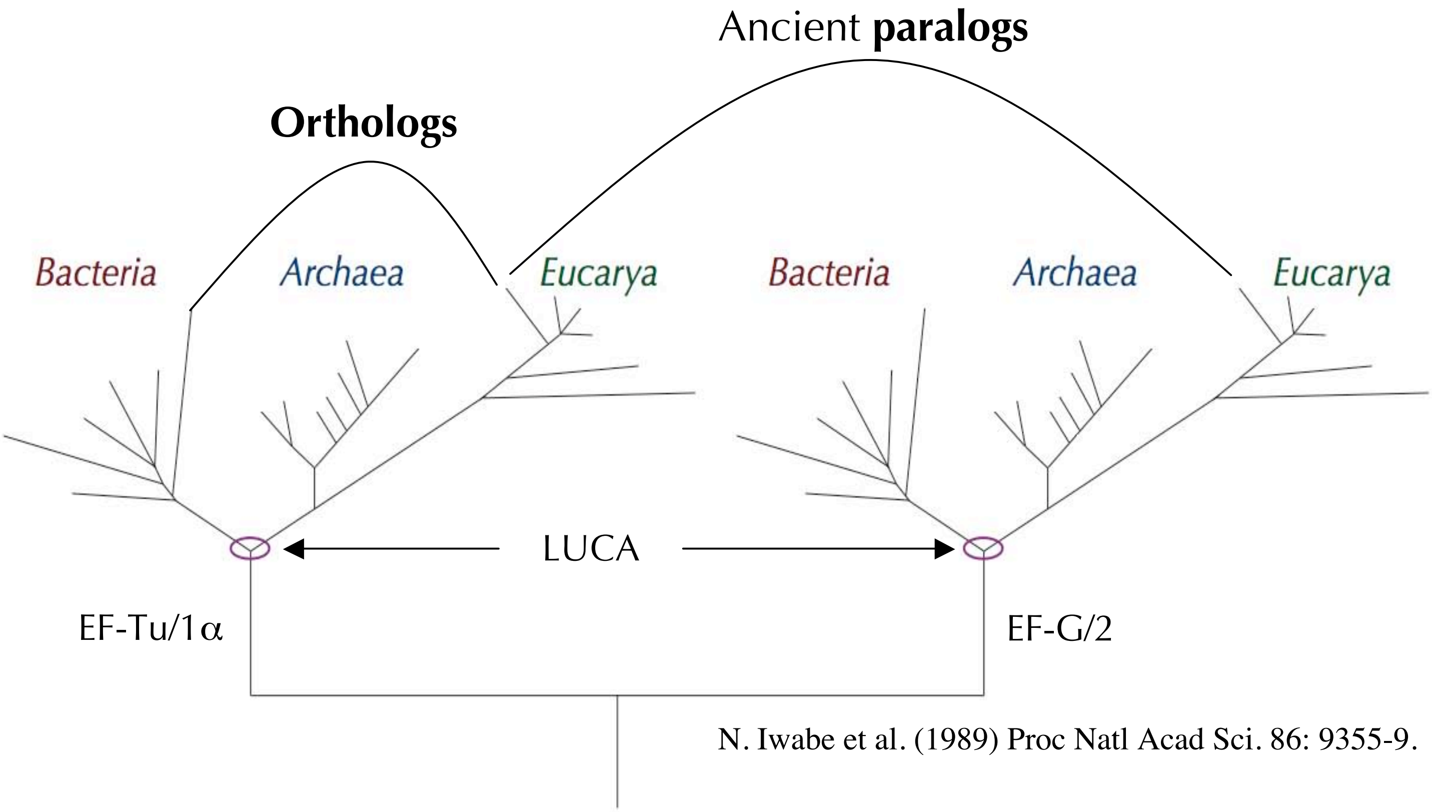
2006
(~100 divisions; 30 cultured/~70 uncultured)

environmental sequenceing reveals an explosion of primary bacterial divisions

From Norman Pace’s Microbial Diversity Course
http://mcdb.colorado.edu/courses/4350/

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.

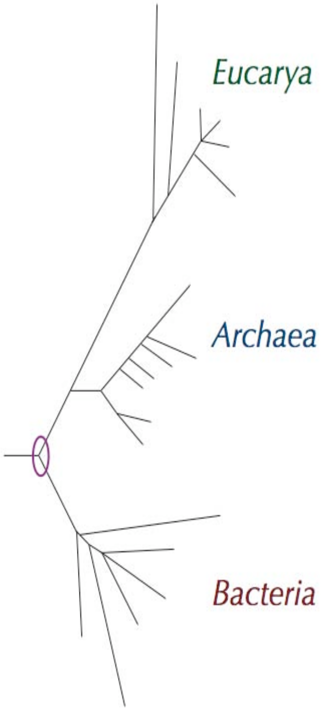
Recent gene duplications can result in paralogs, which are also uninformative regarding organismal phylogeny.

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

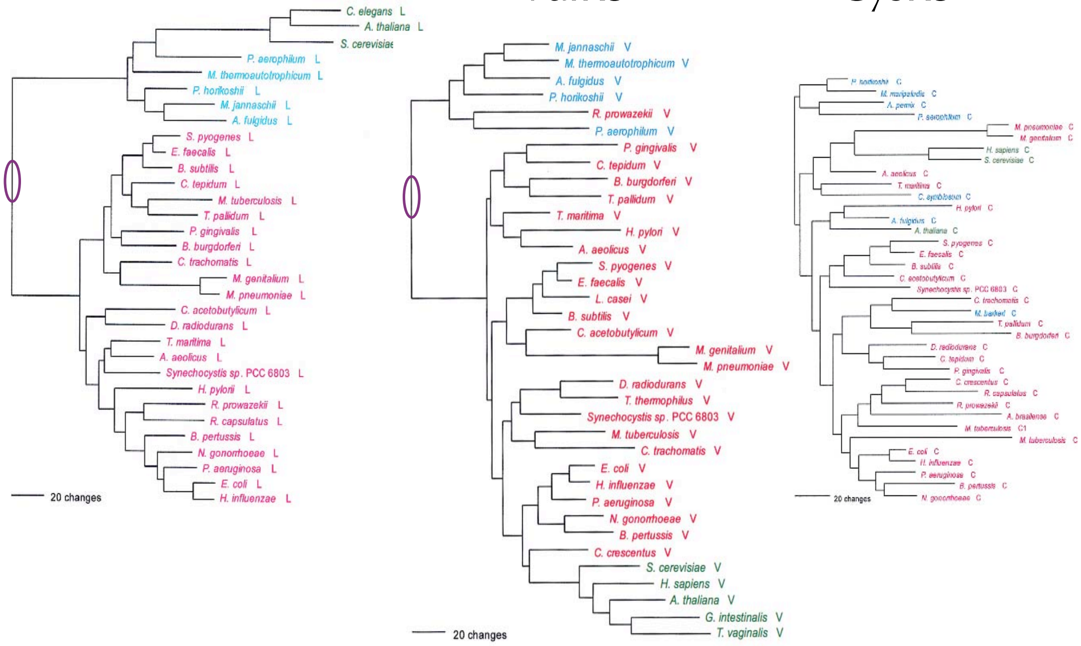


Aminoacyl-tRNA Synthetase (aaRS) phylogenies

full canonical LeuRS

basal canonical ValRS

non-canonical CysRS



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

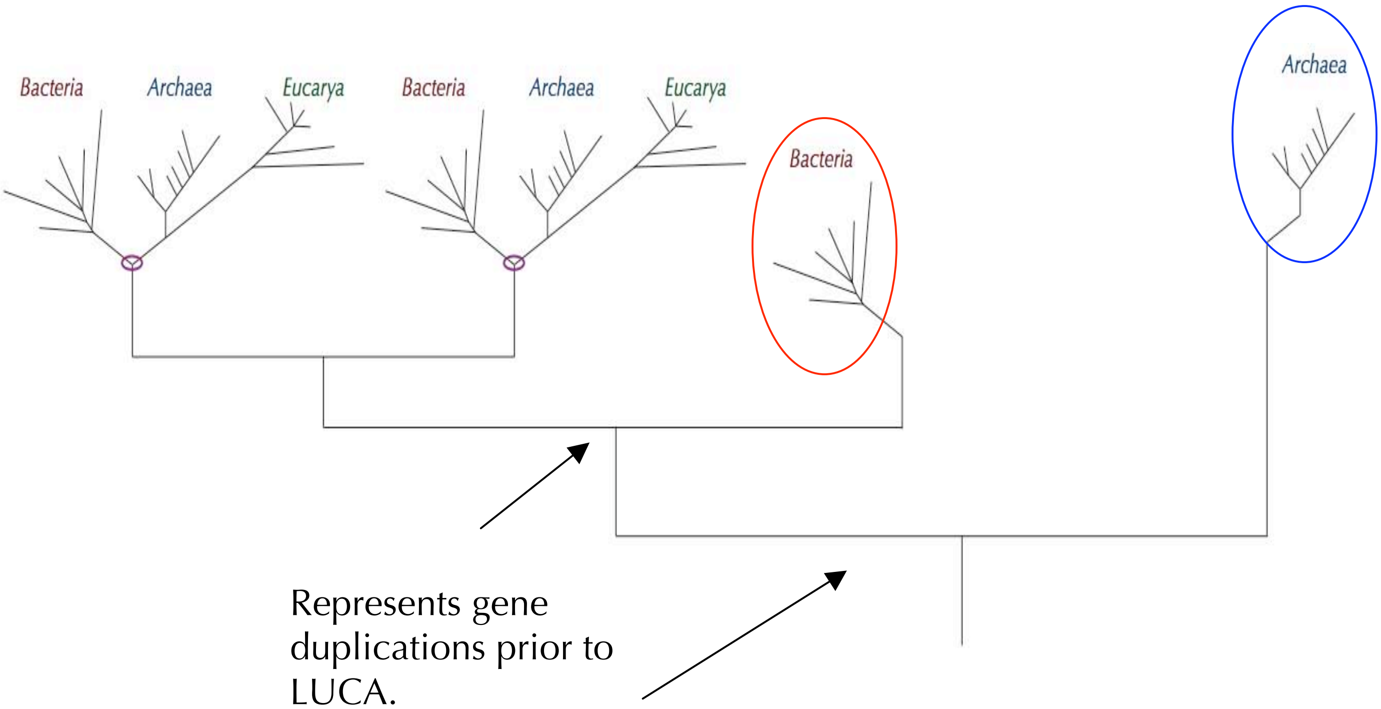
2/3 of aaRSs show at least basal canonical pattern.

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

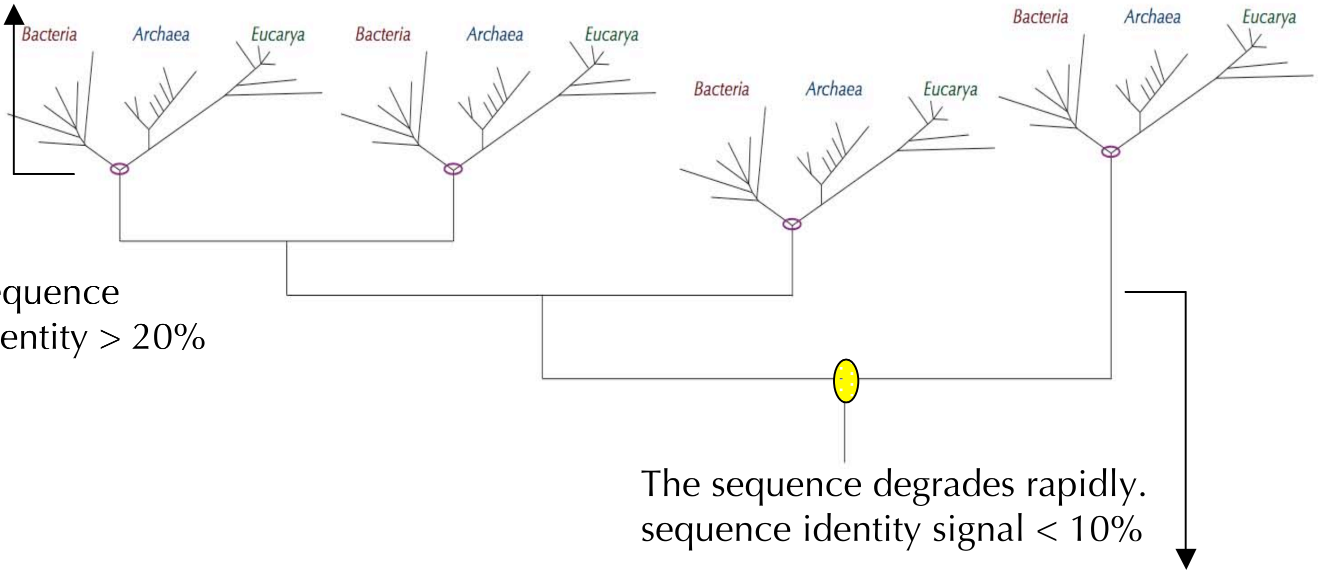


Represents gene duplications prior to LUCA.

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

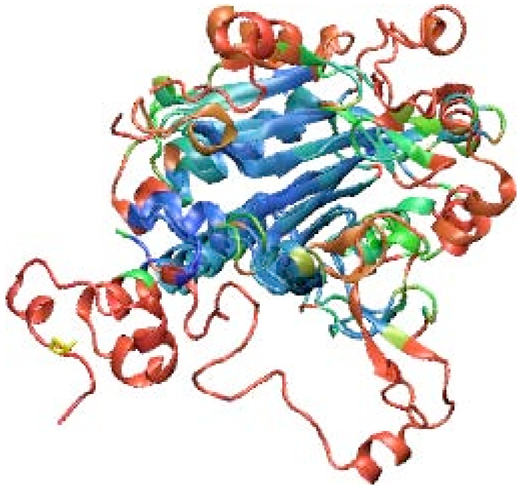
P. O'Donoghue, A. Sethi, C. R. Woese & Z. Luthey-Schulten. (2005) *PNAS* 102:19003-8.

The Relationship Between Sequence & Structure



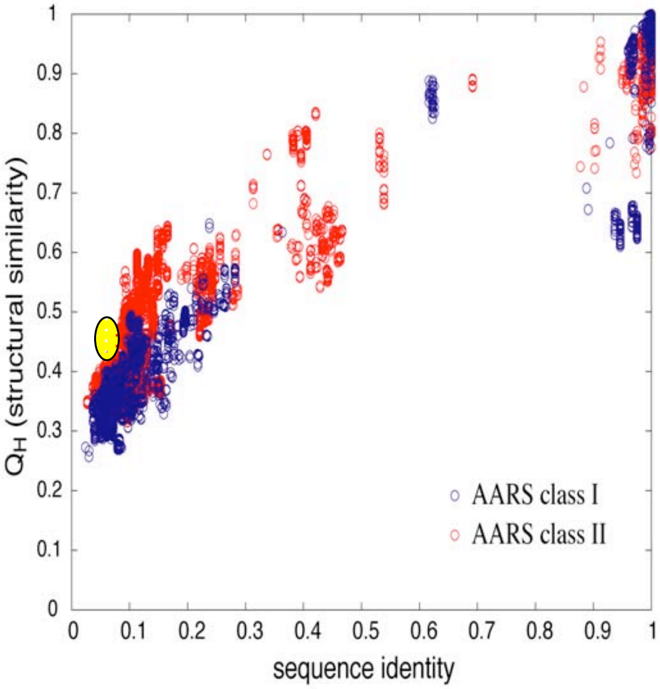
sequence
identity > 20%

The sequence degrades rapidly.
sequence identity signal < 10%



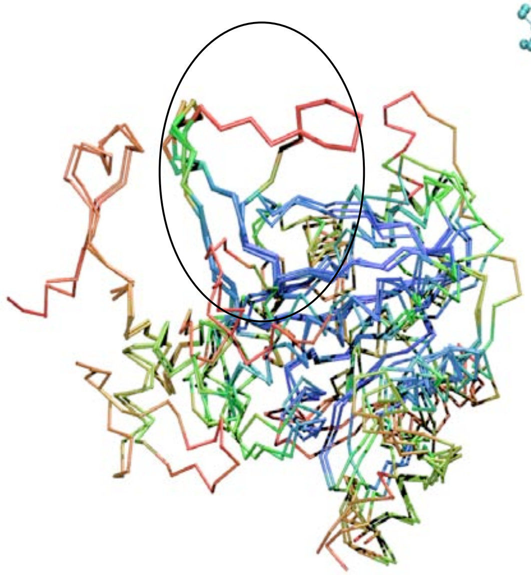
Structural superposition of AlaRS & AspRS.

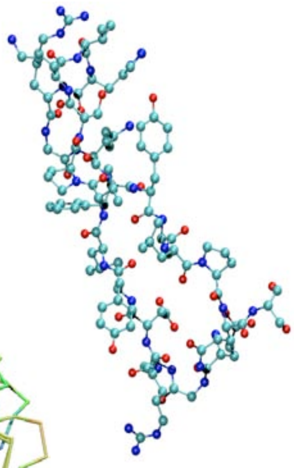
● Sequence id = 0.055, Q_H= 0.48



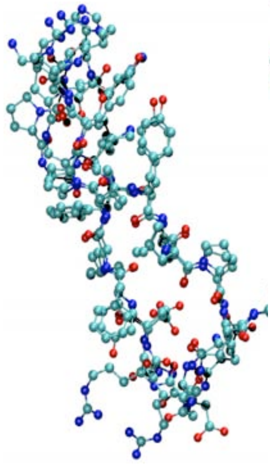
O'Donoghue & Luthey-Schulten (2003) *MMBR* 67: 550–73.

Protein Homology in Structure and Sequence

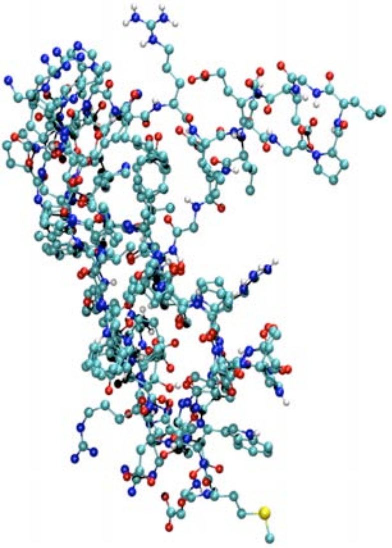




A

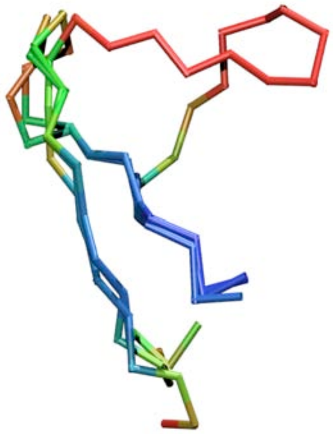


A,B



A,B,C

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

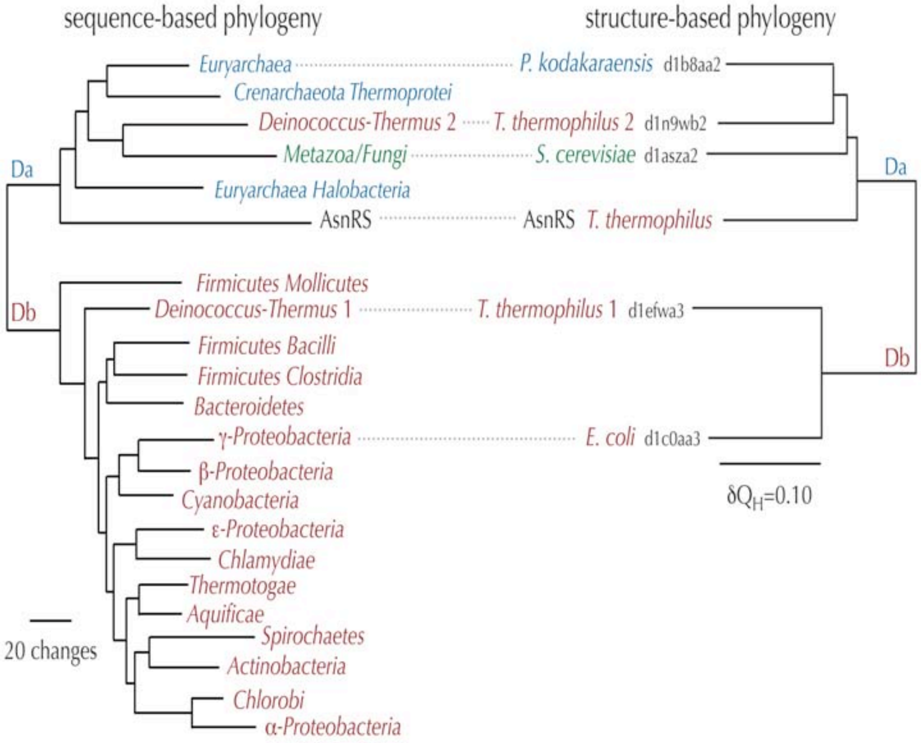


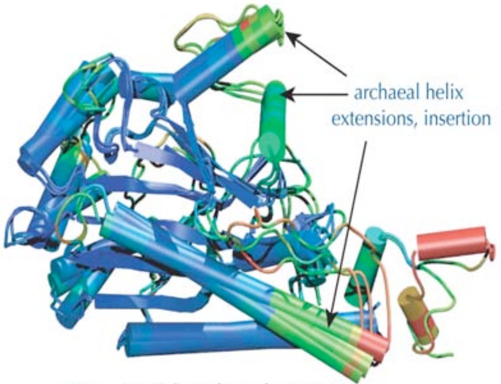
Overlap of protein backbones.

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

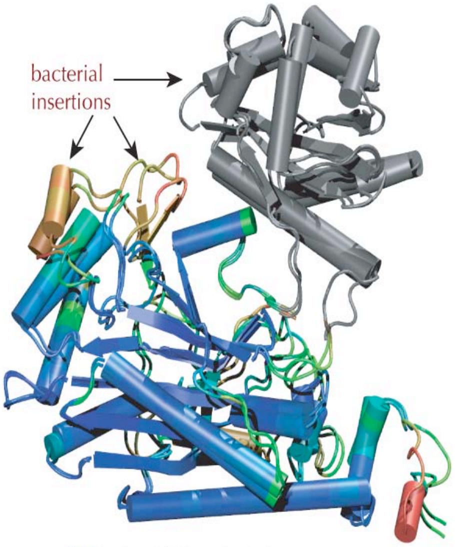
Evolutionary History in Sequence & Structure

aspartyl-tRNA synthetase





Da - AspRS archaeal genre



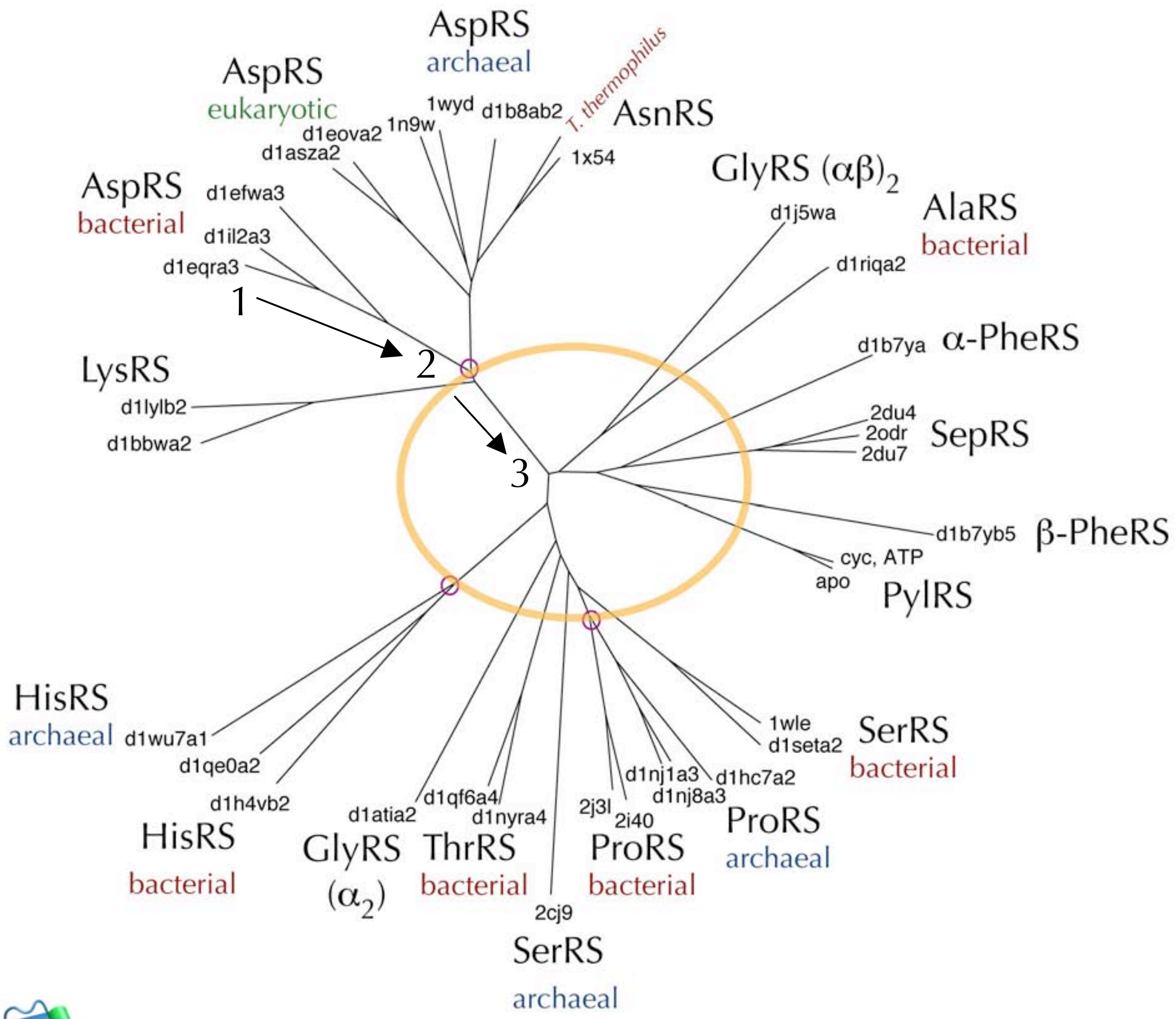
Db - AspRS bacterial genre

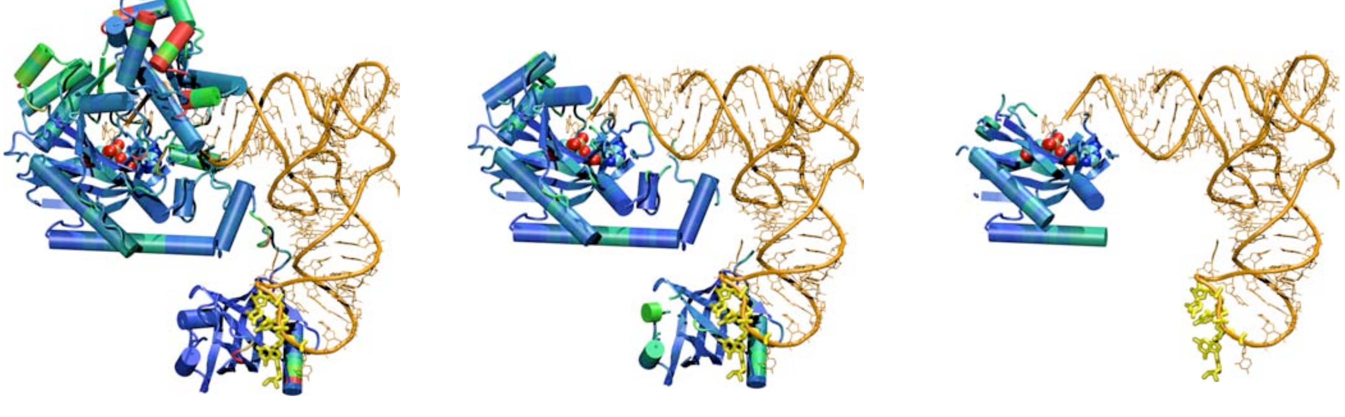
Congruence justifies using structure to trace back evolutionary events beyond the reach of sequence phylogeny.

O'Donoghue & Luthey-Schulten (2003) *MMBR* 67: 550–73.
Elijah Roberts, John Eargle, Dan Wright & Zaida Luthey-Schulten. *BMC Bioinformatics* 2006, **7**:382
Structural phylogeny software: http://www.scs.uiuc.edu/~schulten/multiseq

Evolution of Structure in the class II aaRSs

structural phylogeny reveals distant evolutionary events



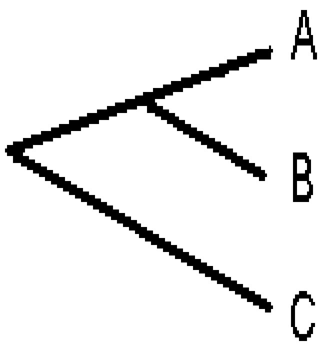


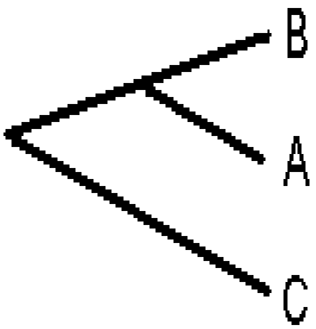
1. Modern AspRS ⟶ 2. AspRS in LUCA ⟶ 3. Ancestral class II aaRS

Tree representations

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

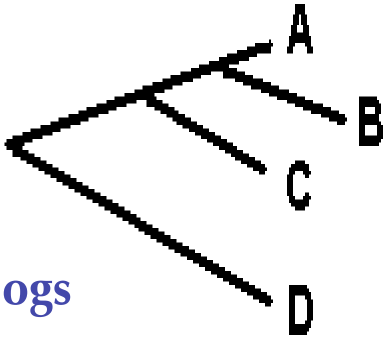
The trees indicate that A, B & C have evolved at equal rates since their divergence from a common ancestor, i.e., a constant molecular clock is assumed.





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.



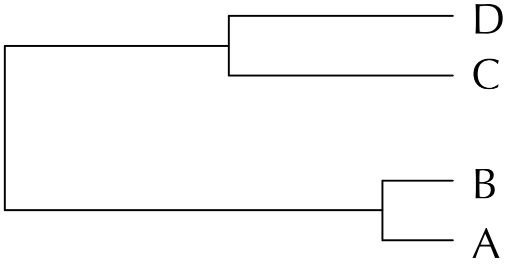
This is one reason that organisms can share a recent common ancestor, but have more distantly related genes than expected. (HGT, and loss of orthologs are other reasons.)

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

Algorithms & Programs

Algorithmic or Clustering Methods

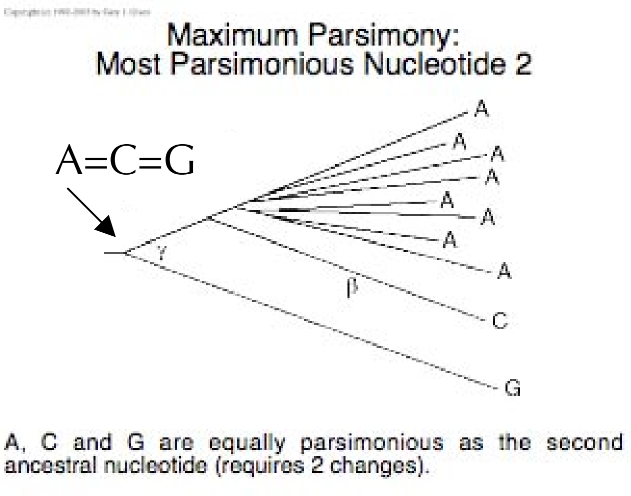
Add sequences to a tree according to similarity relationships.
Produces one tree.



Optimality Methods

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

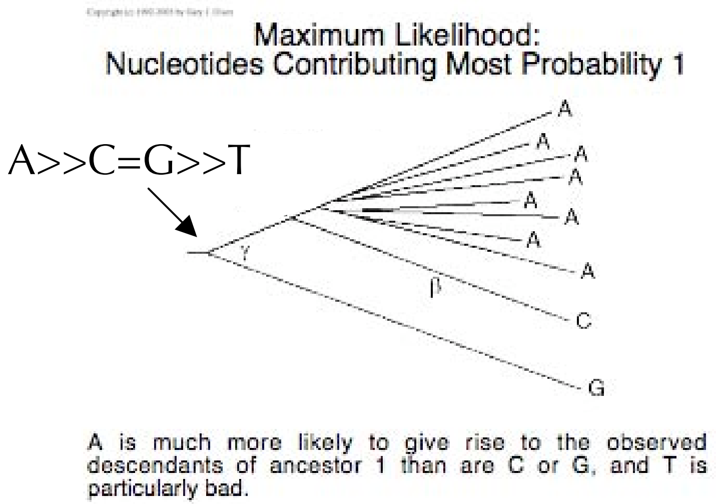
Parsimony
(fewest changes)



Maximum Parsimony:
Most Parsimonious Nucleotide 2

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

Maximum Likelihood
(most probable tree)



Maximum Likelihood:
Nucleotides Contributing Most Probability 1

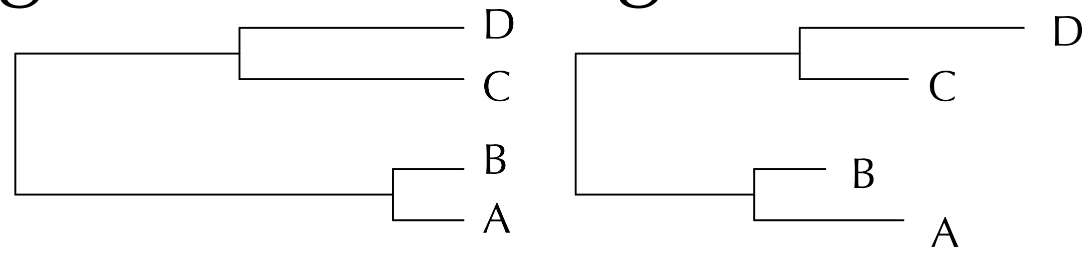
A is much more likely to give rise to the observed descendants of ancestor 1 than are C or G, and T is particularly bad.

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.

> "functional predictions can be greatly improved by focusing on *how* the genes became similar in sequence (i.e., evolution) rather than on the sequence similarity itself. "

Only 321 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. Molecular basis of human evolution.

Tree of Life 1

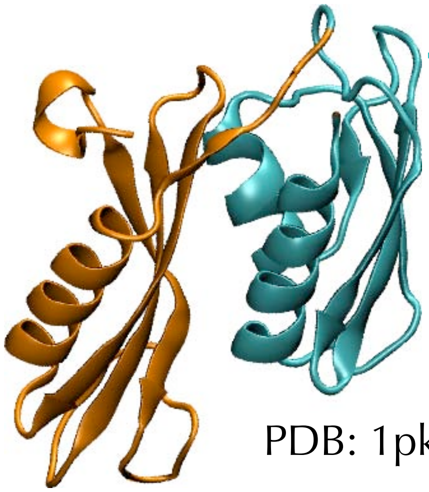
Phylogeny determined by protein domain content

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

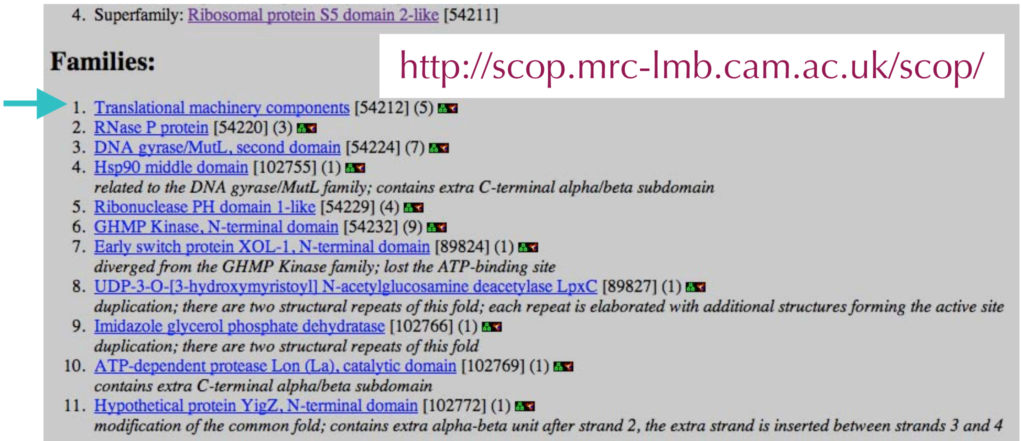
Departments of *Chemistry and Biochemistry and †Pharmacology and San Diego Supercomputer Center, University of California at San Diego, La Jolla, CA 92093

Contributed by Russell F. Doolittle, November 26, 2004

Protein Domains and Superfamilies



PDB: 1pkp



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

Families:

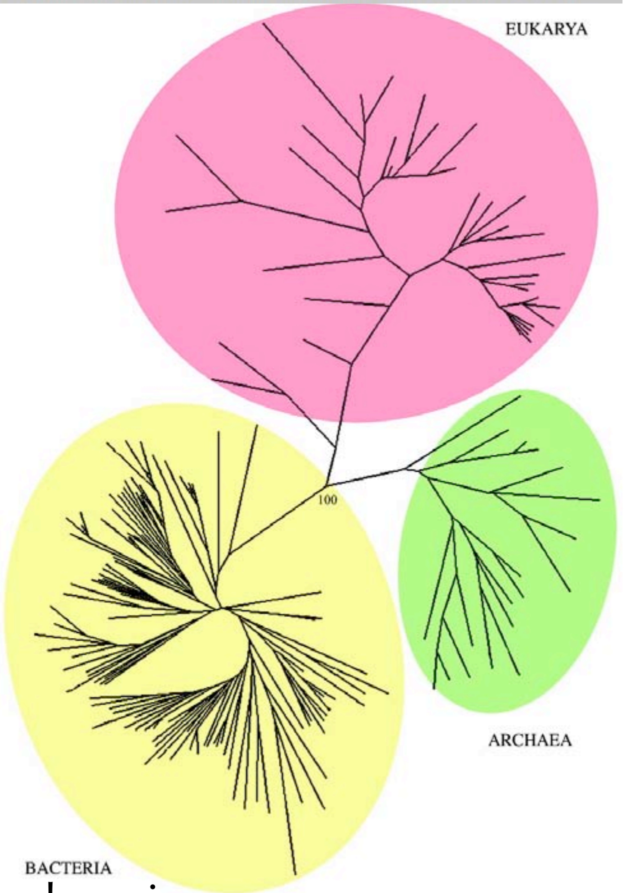
http://scop.mrc-lmb.cam.ac.uk/scop/

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

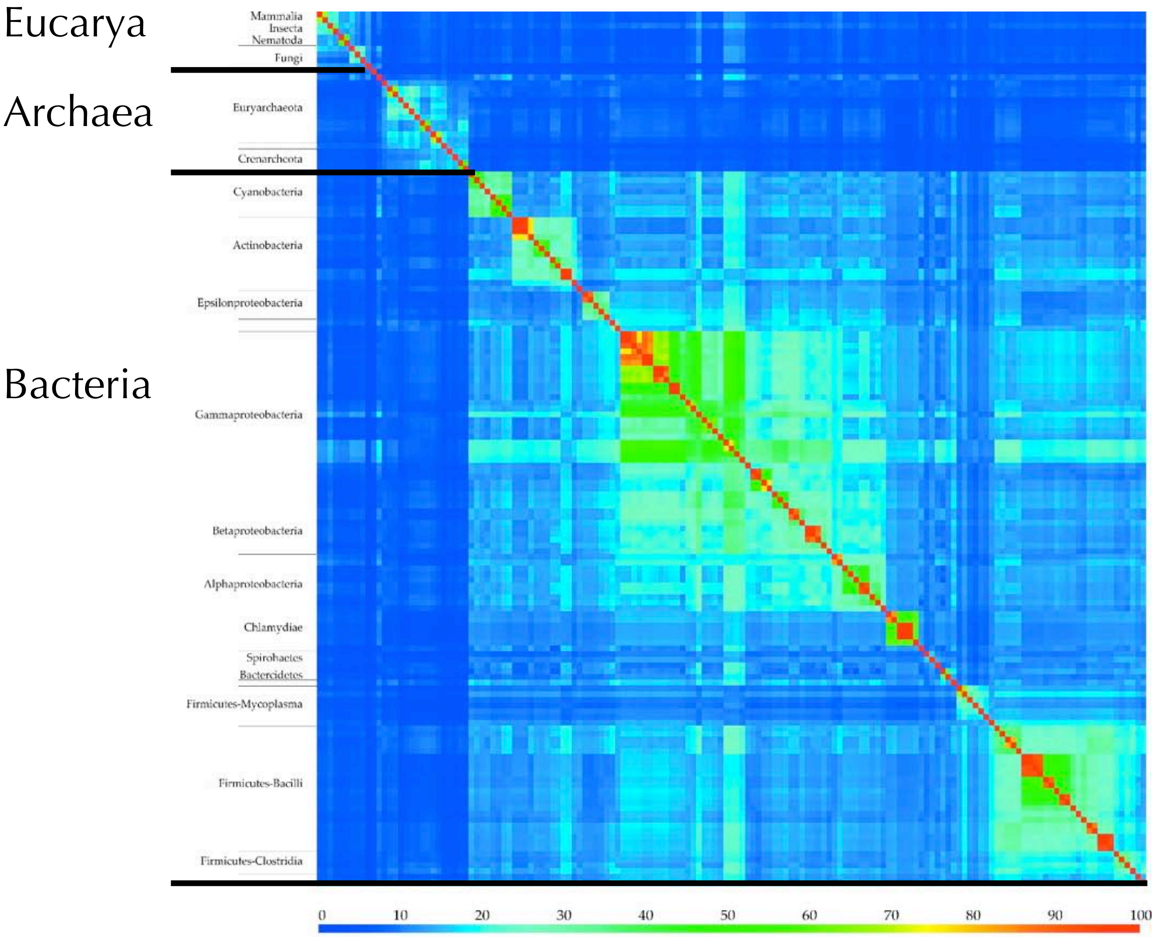
D = A′ /(A′+AB)

A′ is number of superfamily domains.

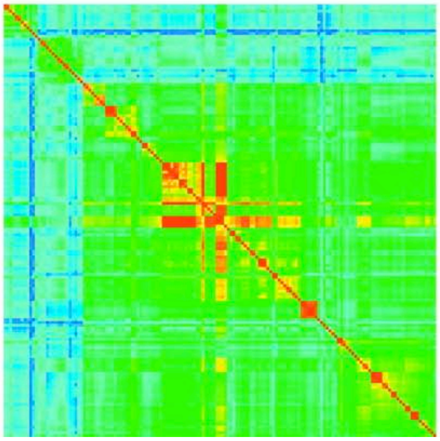
AB is the number of shared superfamily domains.

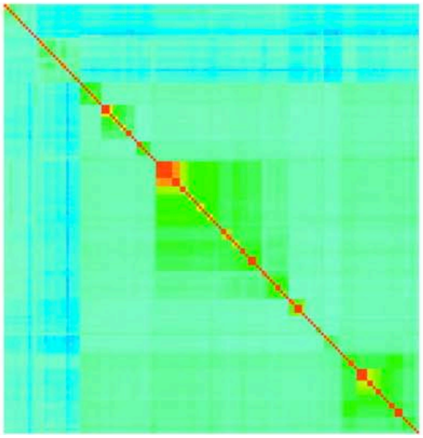
Tree of Life 2

Genome Conservation



Gene presence/ absence





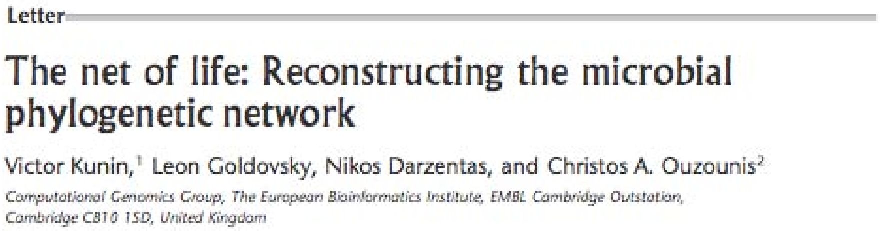
average ortholog similarity

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

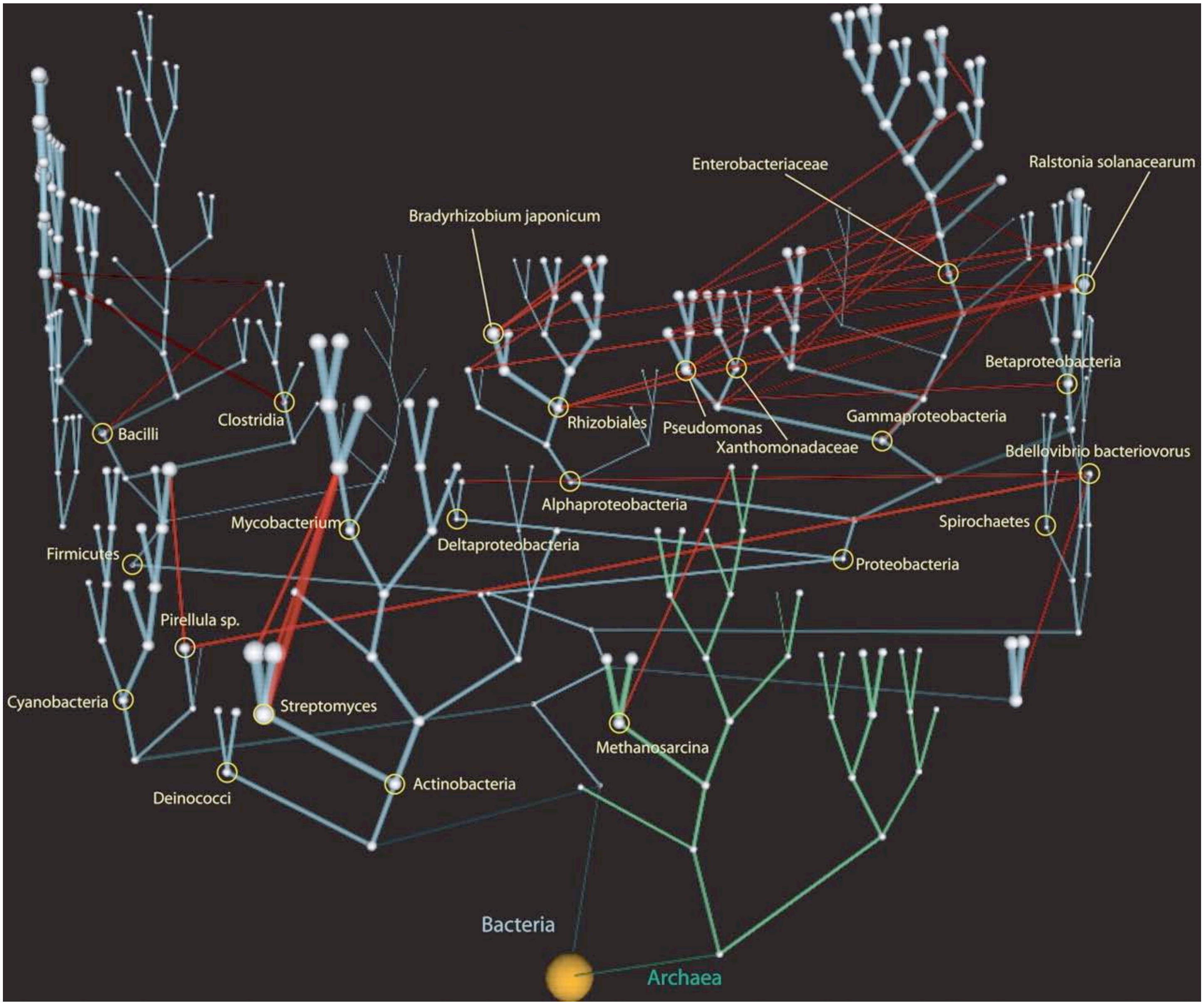


Letter

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

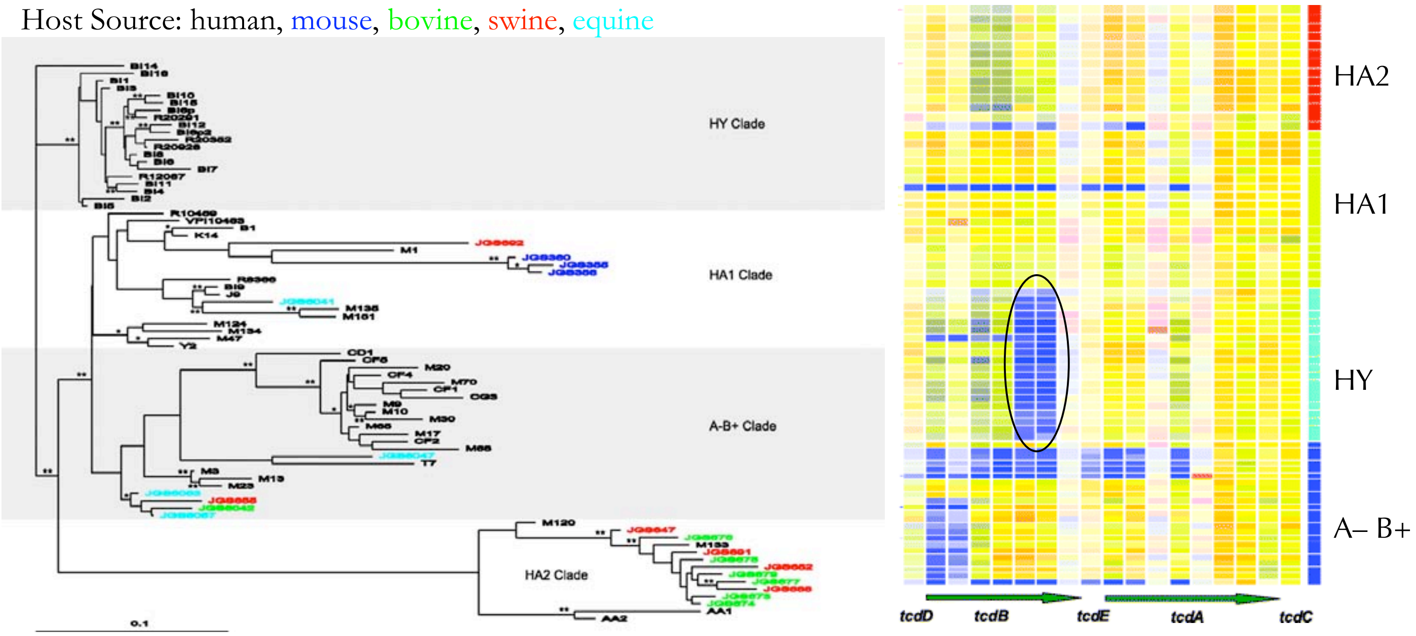
Identify Virulence Factors

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



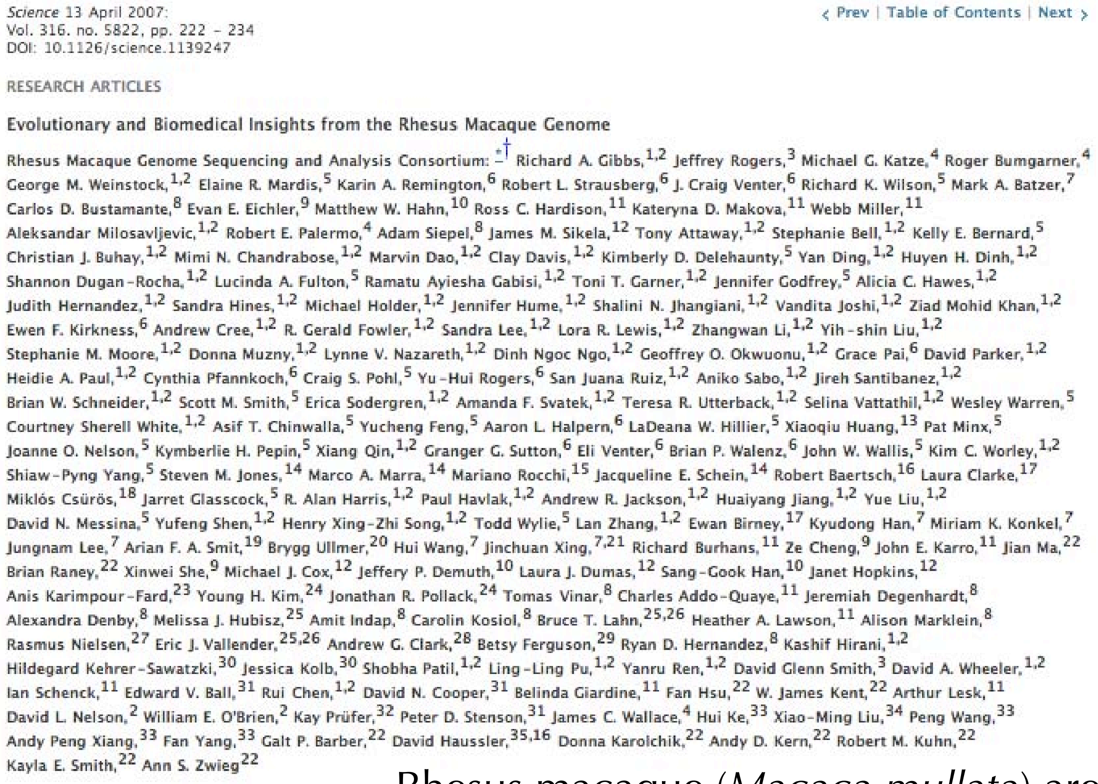
Geographic Spread of Pathogen Lineages

> "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.

Macaca Genome & Human Evolution



Science 13 April 2007:
Vol. 316. no. 5822, pp. 222 - 234
DOI: 10.1126/science.1139247

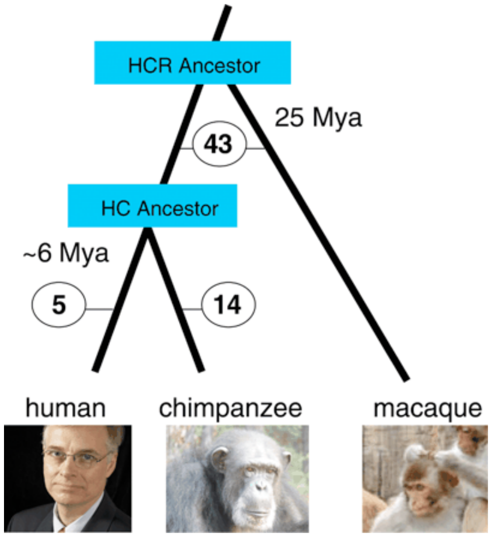
‹ Prev | Table of Contents | Next ›

RESEARCH ARTICLES

Evolutionary and Biomedical Insights from the Rhesus Macaque Genome

Rhesus Macaque Genome Sequencing and Analysis Consortium: [*,†] Richard A. Gibbs,[1,2] Jeffrey Rogers,[3] Michael G. Katze,[4] Roger Bumgarner,[4] George M. Weinstock,[1,2] Elaine R. Mardis,[5] Karin A. Remington,[6] Robert L. Strausberg,[6] J. Craig Venter,[6] Richard K. Wilson,[5] Mark A. Batzer,[7] Carlos D. Bustamante,[8] Evan E. Eichler,[9] Matthew W. Hahn,[10] Ross C. Hardison,[11] Kateryna D. Makova,[11] Webb Miller,[11] Aleksandar Milosavljevic,[1,2] Robert E. Palermo,[4] Adam Siepel,[8] James M. Sikela,[12] Tony Attaway,[1,2] Stephanie Bell,[1,2] Kelly E. Bernard,[5] Christian J. Buhay,[1,2] Mimi N. Chandrabose,[1,2] Marvin Dao,[1,2] Clay Davis,[1,2] Kimberly D. Delehaunty,[5] Yan Ding,[1,2] Huyen H. Dinh,[1,2] Shannon Dugan-Rocha,[1,2] Lucinda A. Fulton,[5] Ramatu Ayiesha Gabisi,[1,2] Toni T. Garner,[1,2] Jennifer Godfrey,[5] Alicia C. Hawes,[1,2] Judith Hernandez,[1,2] Sandra Hines,[1,2] Michael Holder,[1,2] Jennifer Hume,[1,2] Shalini N. Jhangiani,[1,2] Vandita Joshi,[1,2] Ziad Mohid Khan,[1,2] Ewen F. Kirkness,[6] Andrew Cree,[1,2] R. Gerald Fowler,[1,2] Sandra Lee,[1,2] Lora R. Lewis,[1,2] Zhangwan Li,[1,2] Yih-shin Liu,[1,2] Stephanie M. Moore,[1,2] Donna Muzny,[1,2] Lynne V. Nazareth,[1,2] Dinh Ngoc Ngo,[1,2] Geoffrey O. Okwuonu,[1,2] Grace Pai,[6] David Parker,[1,2] Heidie A. Paul,[1,2] Cynthia Pfannkoch,[6] Craig S. Pohl,[5] Yu-Hui Rogers,[6] San Juana Ruiz,[1,2] Aniko Sabo,[1,2] Jireh Santibanez,[1,2] Brian W. Schneider,[1,2] Scott M. Smith,[5] Erica Sodergren,[1,2] Amanda F. Svatek,[1,2] Teresa R. Utterback,[1,2] Selina Vattathil,[1,2] Wesley Warren,[5] Courtney Sherell White,[1,2] Asif T. Chinwalla,[5] Yucheng Feng,[5] Aaron L. Halpern,[6] LaDeana W. Hillier,[5] Xiaoqiu Huang,[13] Pat Minx,[5] Joanne O. Nelson,[5] Kymberlie H. Pepin,[5] Xiang Qin,[1,2] Granger G. Sutton,[6] Eli Venter,[6] Brian P. Walenz,[6] John W. Wallis,[5] Kim C. Worley,[1,2] Shiaw-Pyng Yang,[5] Steven M. Jones,[14] Marco A. Marra,[14] Mariano Rocchi,[15] Jacqueline E. Schein,[14] Robert Baertsch,[16] Laura Clarke,[17] Miklós Csürös,[18] Jarret Glasscock,[5] R. Alan Harris,[1,2] Paul Havlak,[1,2] Andrew R. Jackson,[1,2] Huaiyang Jiang,[1,2] Yue Liu,[1,2] David N. Messina,[5] Yufeng Shen,[1,2] Henry Xing-Zhi Song,[1,2] Todd Wylie,[5] Lan Zhang,[1,2] Ewan Birney,[17] Kyudong Han,[7] Miriam K. Konkel,[7] Jungnam Lee,[7] Arian F. A. Smit,[19] Brygg Ullmer,[20] Hui Wang,[7] Jinchuan Xing,[7,21] Richard Burhans,[11] Ze Cheng,[9] John E. Karro,[11] Jian Ma,[22] Brian Raney,[22] Xinwei She,[9] Michael J. Cox,[12] Jeffery P. Demuth,[10] Laura J. Dumas,[12] Sang-Gook Han,[10] Janet Hopkins,[12] Anis Karimpour-Fard,[23] Young H. Kim,[24] Jonathan R. Pollack,[24] Tomas Vinar,[8] Charles Addo-Quaye,[11] Jeremiah Degenhardt,[8] Alexandra Denby,[8] Melissa J. Hubisz,[25] Amit Indap,[8] Carolin Kosiol,[8] Bruce T. Lahn,[25,26] Heather A. Lawson,[11] Alison Marklein,[8] Rasmus Nielsen,[27] Eric J. Vallender,[25,26] Andrew G. Clark,[28] Betsy Ferguson,[29] Ryan D. Hernandez,[8] Kashif Hirani,[1,2] Hildegard Kehrer-Sawatzki,[30] Jessica Kolb,[30] Shobha Patil,[1,2] Ling-Ling Pu,[1,2] Yanru Ren,[1,2] David Glenn Smith,[3] David A. Wheeler,[1,2] Ian Schenck,[11] Edward V. Ball,[31] Rui Chen,[1,2] David N. Cooper,[31] Belinda Giardine,[11] Fan Hsu,[22] W. James Kent,[22] Arthur Lesk,[11] David L. Nelson,[2] William E. O'Brien,[2] Kay Prüfer,[32] Peter D. Stenson,[31] James C. Wallace,[4] Hui Ke,[33] Xiao-Ming Liu,[34] Peng Wang,[33] Andy Peng Xiang,[33] Fan Yang,[33] Galt P. Barber,[22] David Haussler,[35,16] Donna Karolchik,[22] Andy D. Kern,[22] Robert M. Kuhn,[22] Kayla E. Smith,[22] Ann S. Zwieg[22]

Rhesus macaque (*Macaca mullata*) are old world monkeys found from Afghanistan to the Chinese shore of the Pacific Ocean.

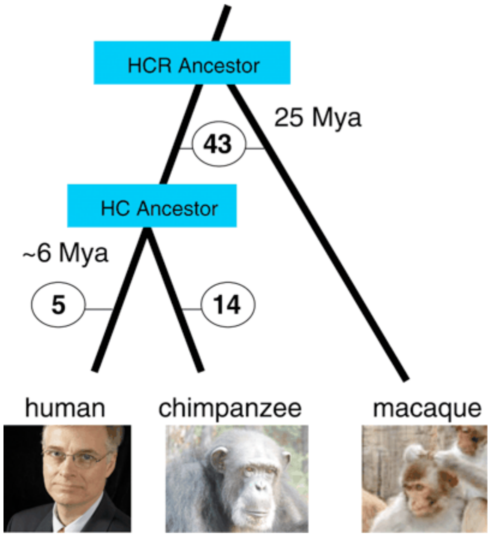


“differences that are found between humans and chimpanzees are difficult to assign as specific to either the chimpanzee or the human…

the chimpanzee analyses have on their own provided relatively few answers to the fundamental question of the nature of the specific molecular changes that make us human.”

human-macaque sequence identity 90-93%.
human-chimp sequence identity 98-99%.

Macaca Genome & Human Evolution

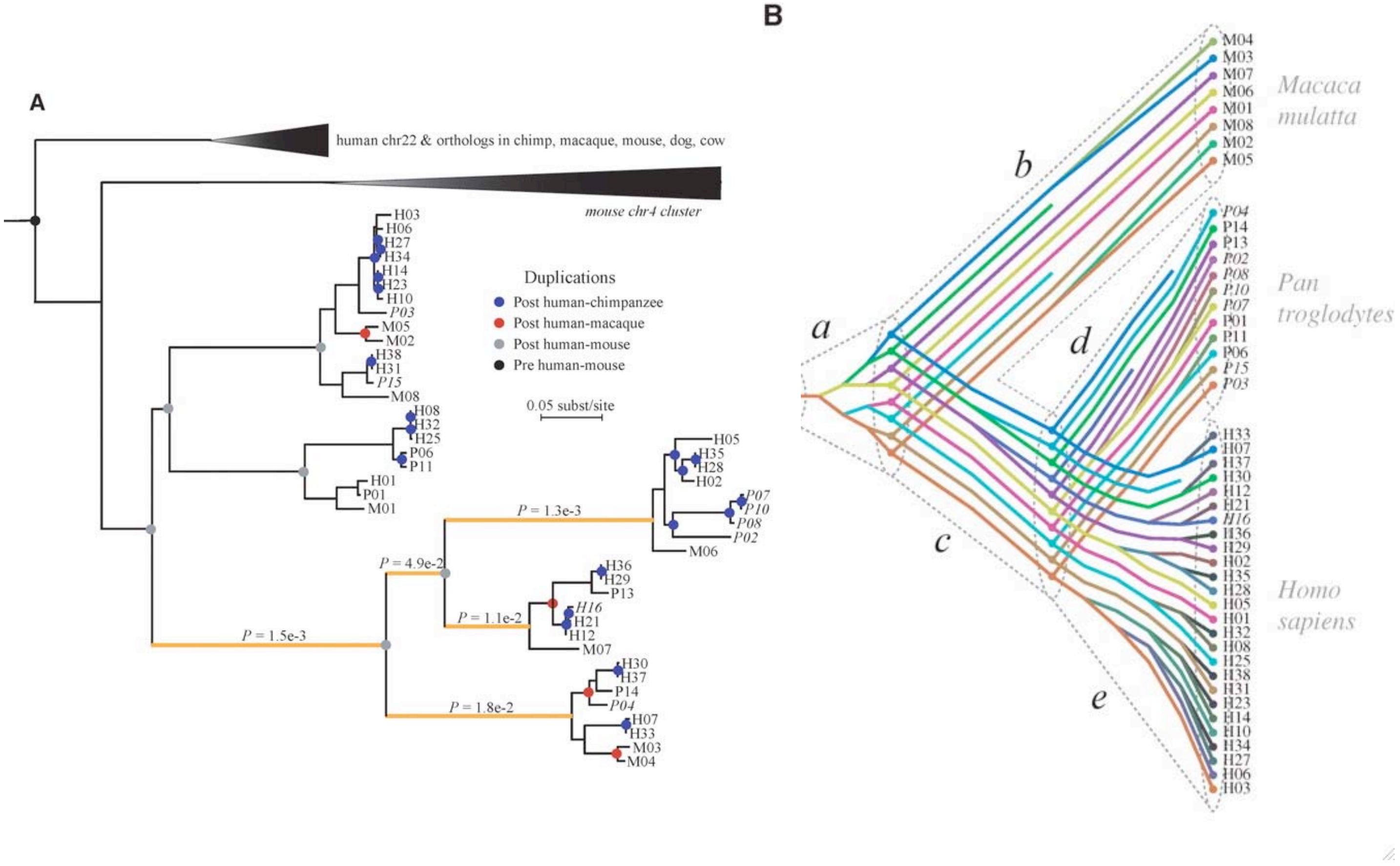


Repeat elements account for 50% of the genomes of all sequenced primates.

Similar to the human, the rhesus macaque contains about 320,000 recognizable copies from more than 100 different families of DNA transposons and more than half a million recognizable copies of endogenous retroviruses (ERVs).

ERVs demonstrate a complex phylogeny and many examples of new and expanded family members, some resulting from **horizontal transmission**.

Gene trees to phylogeny



The preferentially expressed antigen of melanoma (PRAME) gene family consists of a single gene on chromosome 22q11.22 and a cluster of several dozen genes on chromosome 1p36.21.

PRAME and PRAME-like genes are actively expressed in cancers but normally manifest testis-specific expression and may thus have a role in spermatogenesis.

There is extensive duplication early in primate evolution (a), in recent chimpanzee evolution (d), and in recent human evolution (e).

The PRAME gene cluster appears to have been much less dynamic on the macaque lineage (b) and in early hominins (c).

Assessing Phylogenomics

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.

Define genome dynamics that underlie speciation events.

Not all genes have the same history.

Incongruent gene histories result from various sources:

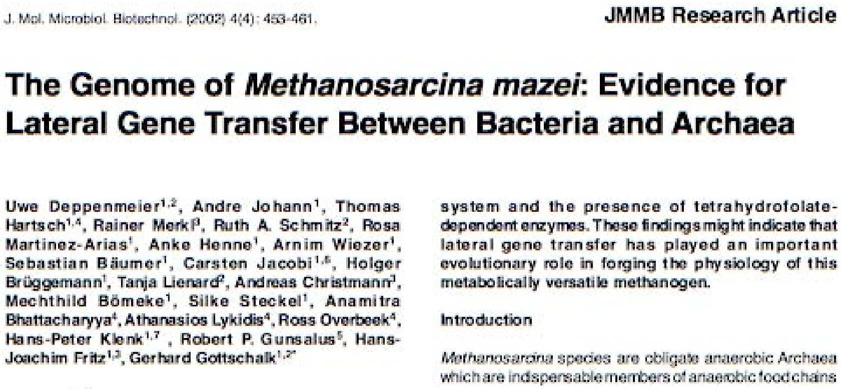
1. loss of close orthologs
2. horizontal gene transfer
3. lineage specific evolutionary rate acceleration.

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**

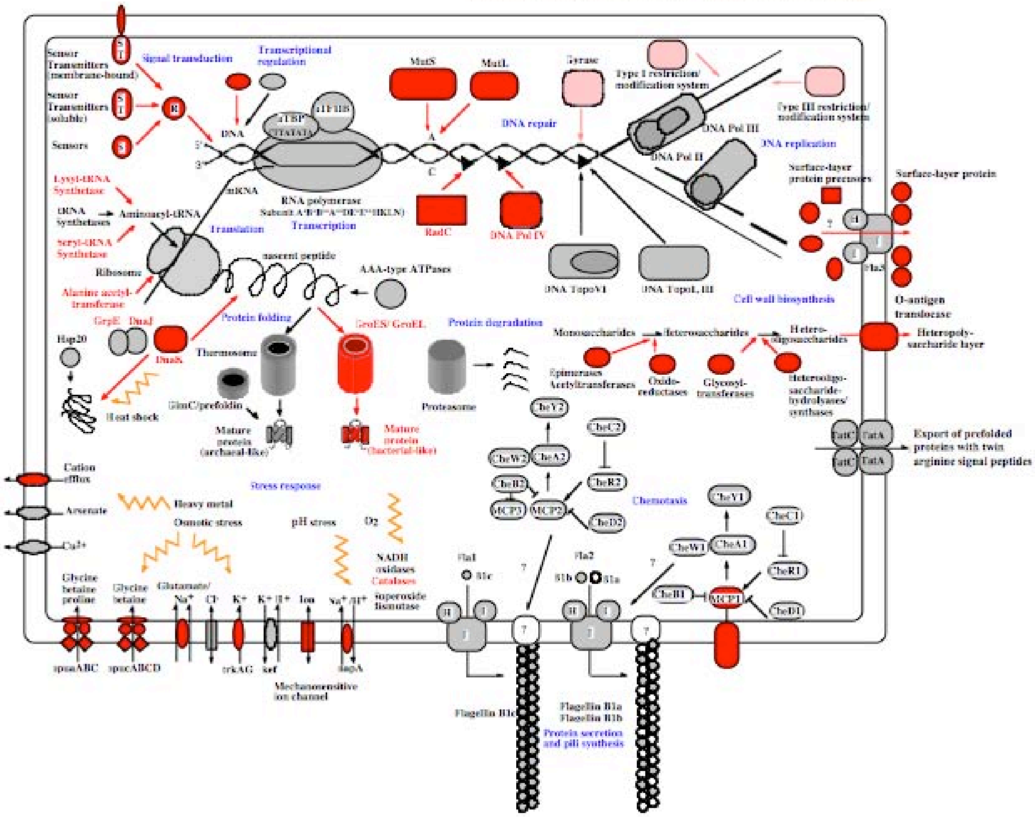
The Genome of *Methanosarcina mazei*: Evidence for Lateral Gene Transfer Between Bacteria and Archaea

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[1,6], 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

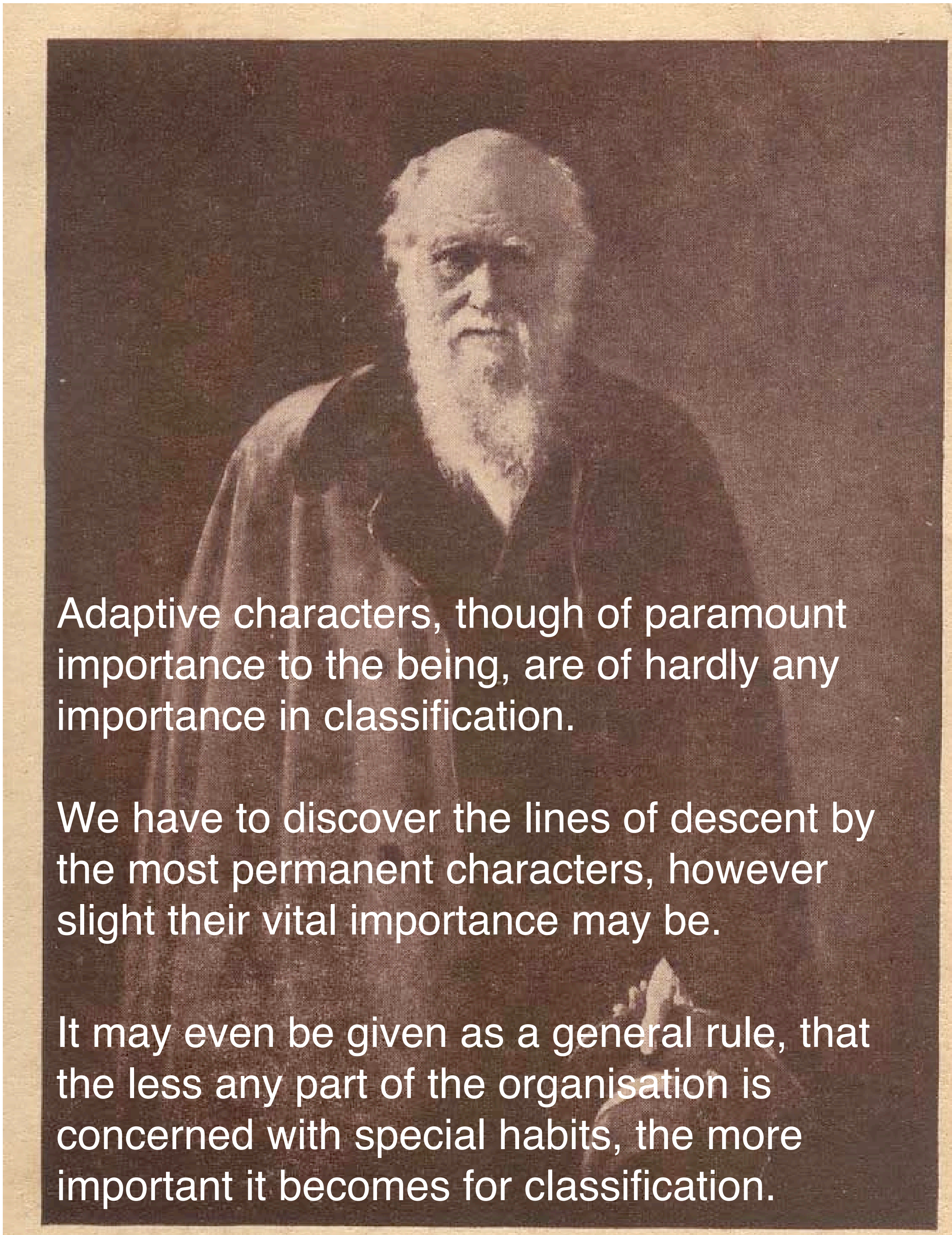
Methanosarcina species are obligate anaerobic Archaea which are indispensable members of anaerobic food chains



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