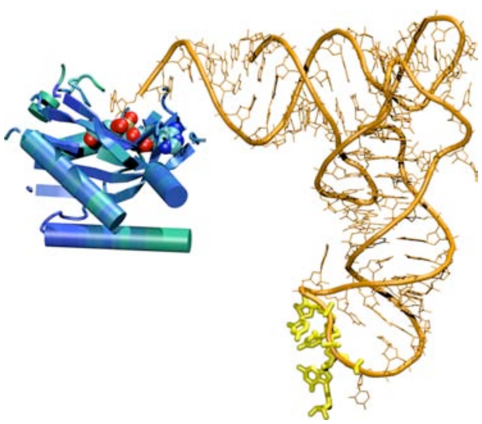
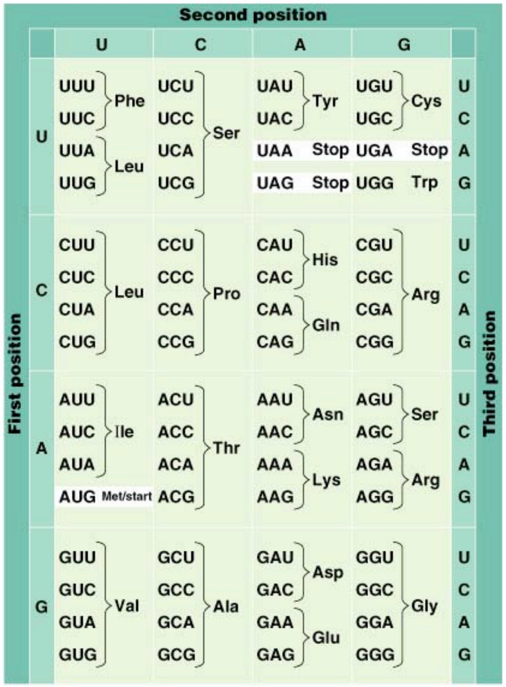
Evolution of Protein Structure





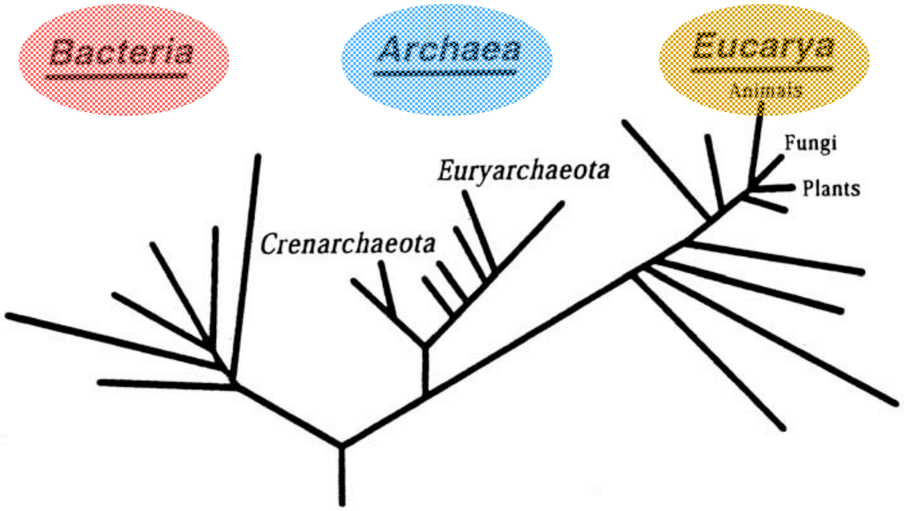
	Second position								
First position	U		C		A		G		Third position
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met/start	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

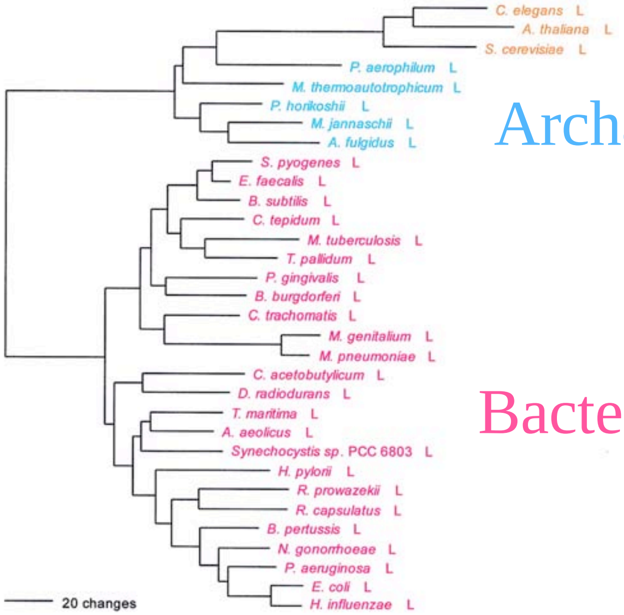
Luthey-Schulten Group

Department of Chemistry, Biophysics, and Beckman Institute
University of Illinois at Urbana-Champaign

Universal Phylogenetic Tree

three domains of life





Eucarya

Archaea

Bacteria

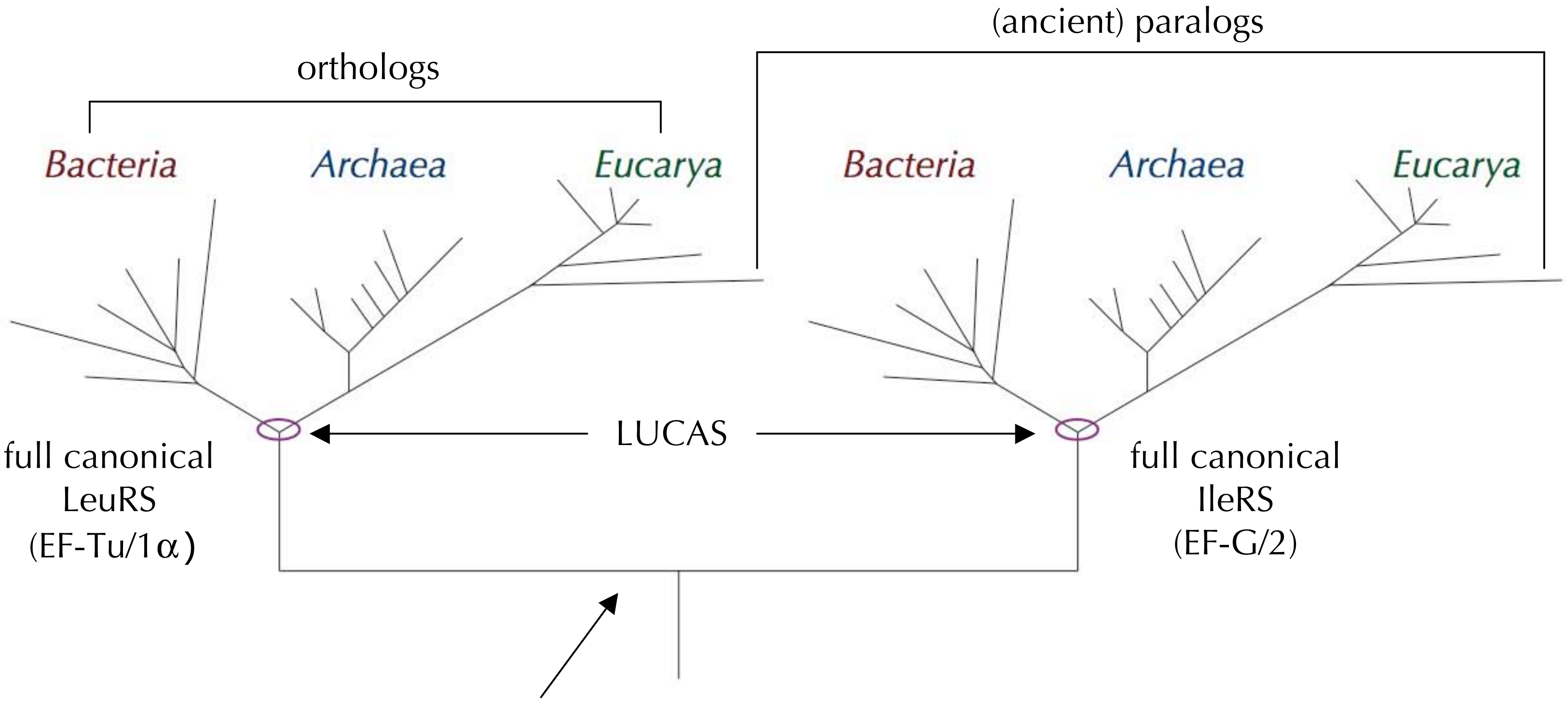
Based on 16S rRNA

Leucyl-tRNA synthetase displays the full canonical phylogenetic distribution.

for review see Woese *PNAS* 2000

Woese, Olsen, Ibba, Soll *MMBR* 2000

Evolutionary Theory: Gene Duplication Prior to LUCAS



Represents gene duplication prior to LUCAS, divergence of function from ancestral gene.

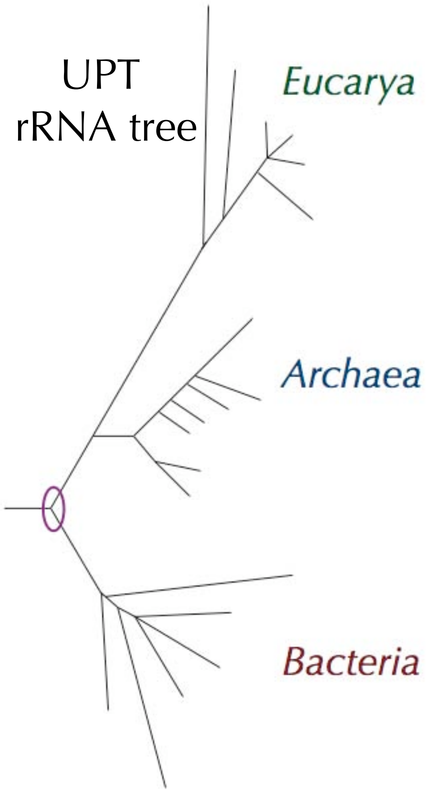
What kind of organisms existed at this time?

What genes were available to LUCAS and its ancestors?

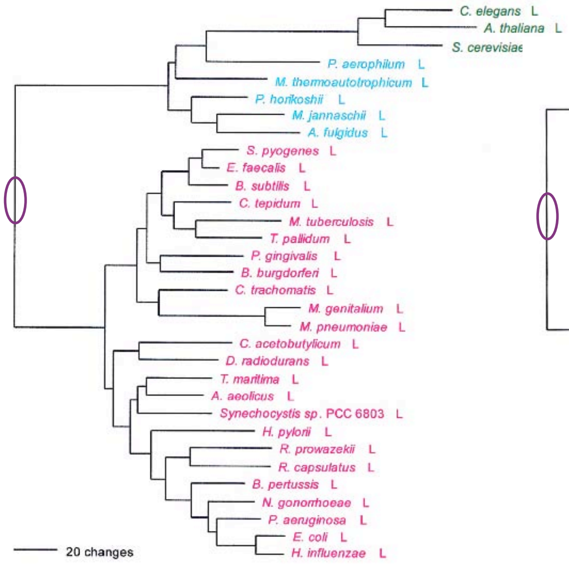
N. Iwabe et al. (1989) Proc Natl Acad Sci. 86: 9355-9.

Evolutionary Theory: Phylogenetic Patterns

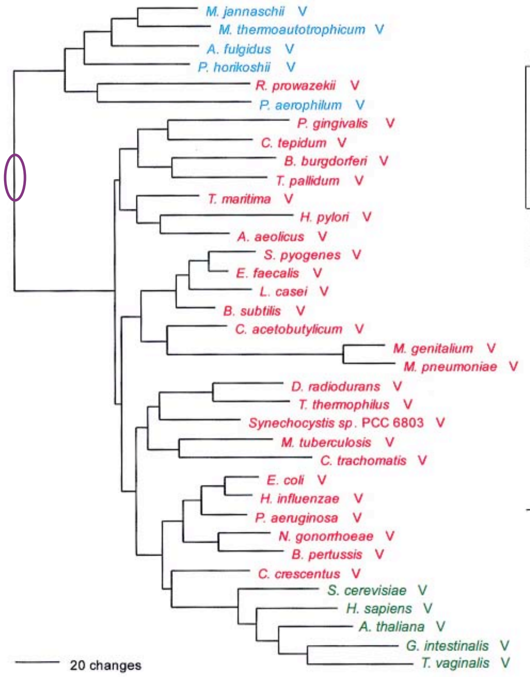
UPT
rRNA tree



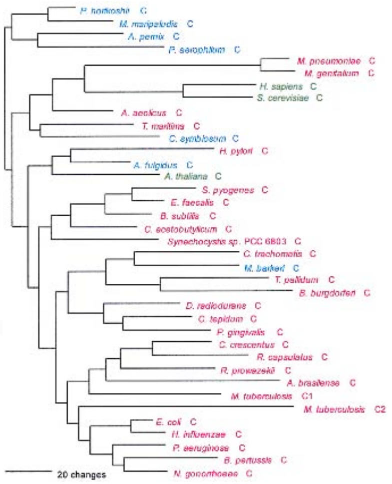
full canonical
LeuRS



basal canonical
ValRS



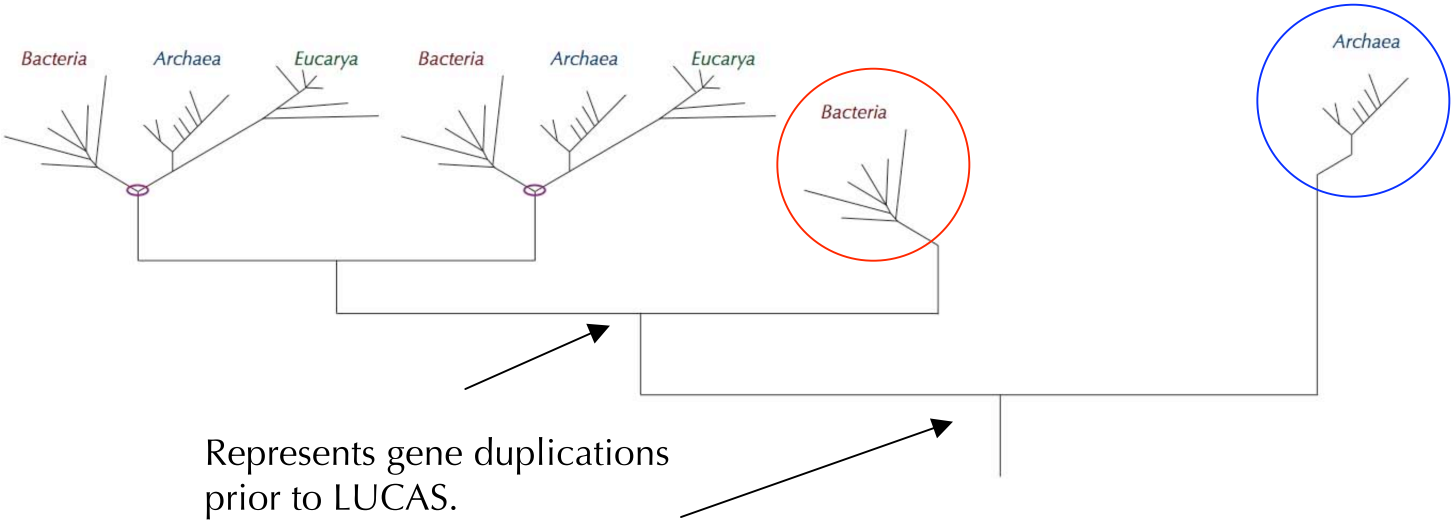
non-canonical
CysRS



Horizontal Gene Transfer Events →

C. R. Woese, G. J. Olsen, M. Ibba & D. Söll (2000) *MMBR*. 64, 202-236.

Evolutionary Theory: Deep phylogeny of protein families

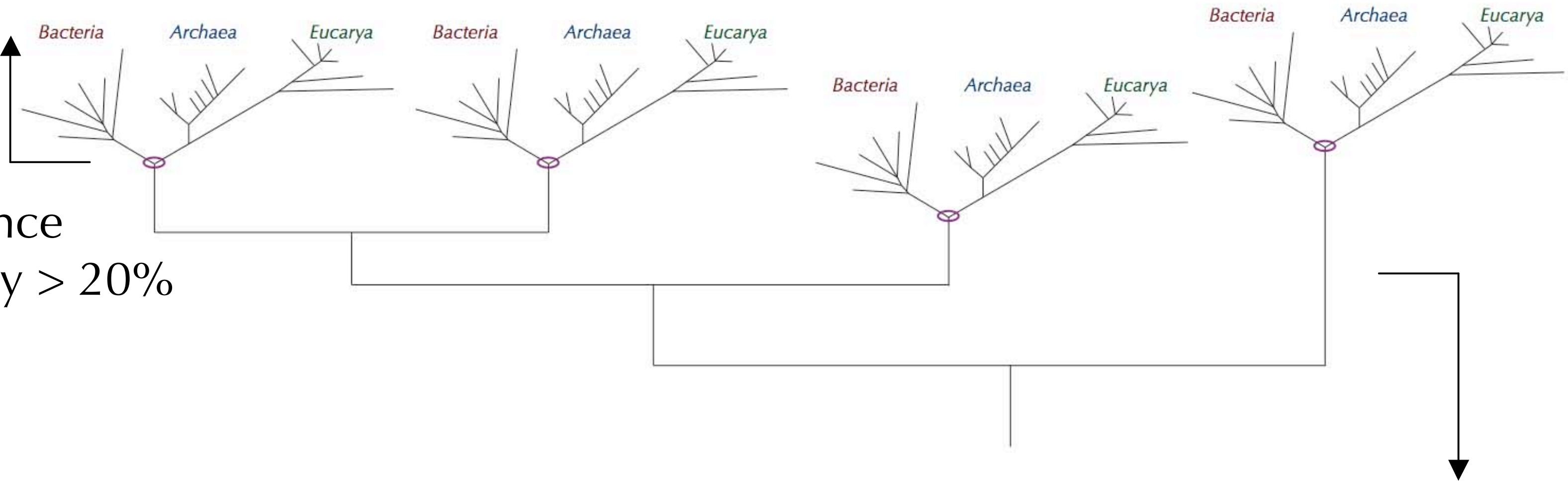


Represents gene duplications
prior to LUCAS.

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

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

Evolutionary Theory: Sequence Signal Decays



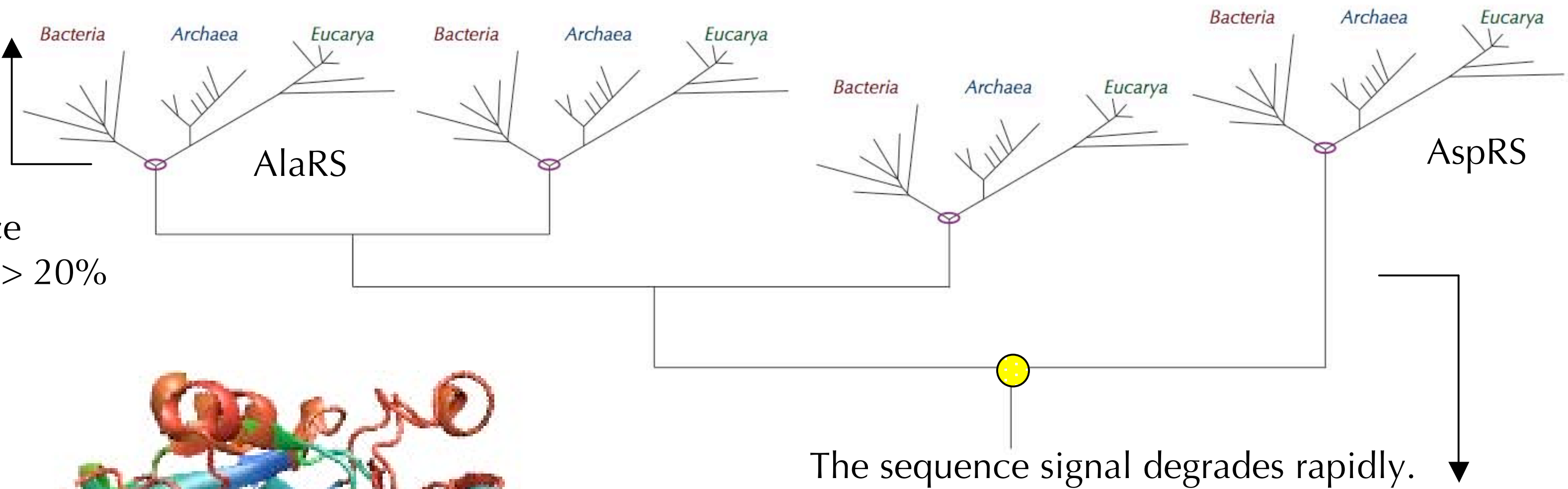
sequence
identity > 20%

The sequence signal degrades rapidly.
sequence identity < 10%.

As sequence similarity degrades alignment and phylogeny become unreliable.

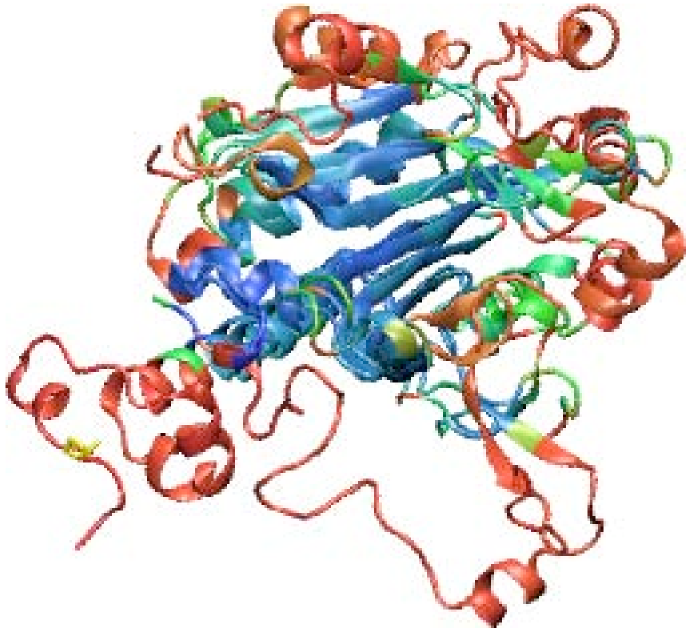
How can we probe the molecular evolution of these ancient events?

The Relationship Between Sequence & Structure



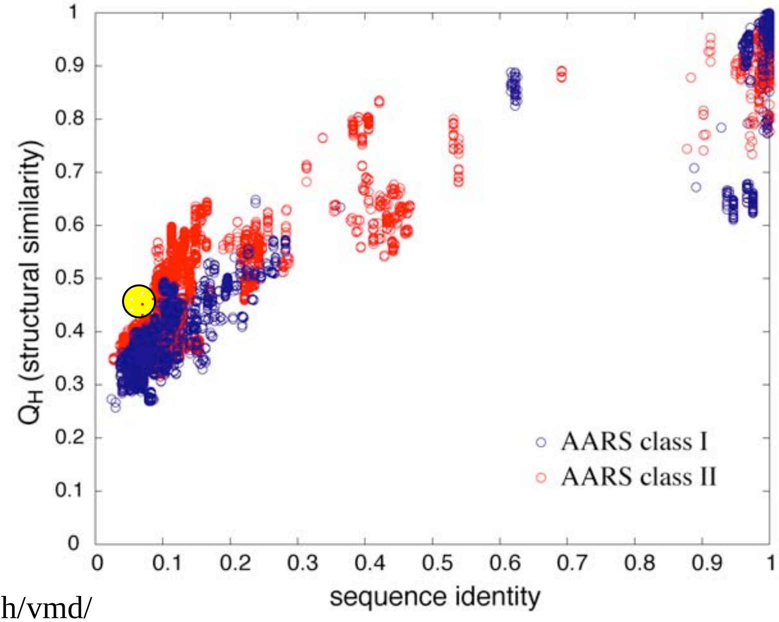
sequence
identity > 20%

The sequence signal degrades rapidly.
sequence identity < 10%



Structural superposition of AlaRS & AspRS.

● Sequence id = 0.055, Q_H= 0.48



O’Donoghue & Luthey-Schulten (2003) *MMBR* 67: 550–73.
Structural alignment & visualization software @ http://www.ks.uiuc.edu/Research/vmd/

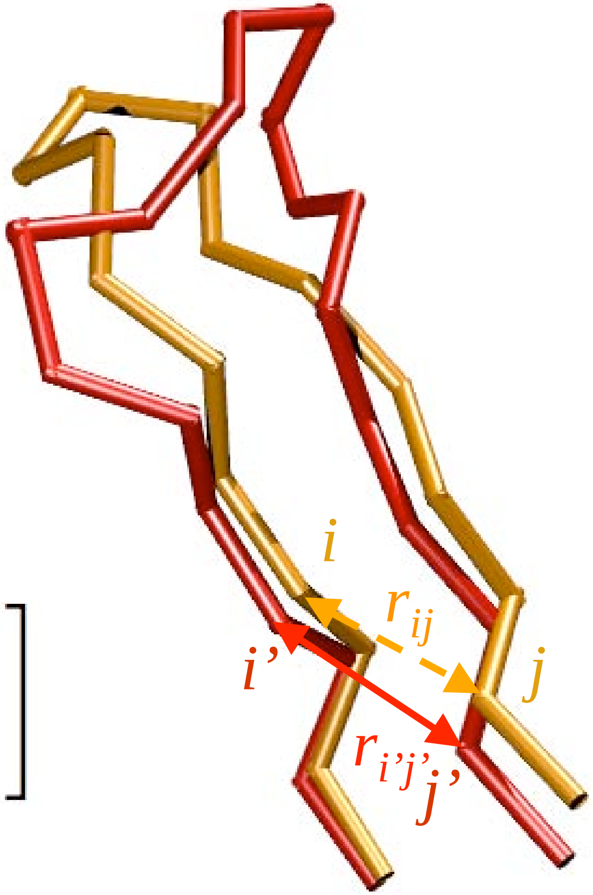
Protein Structure Similarity Measure

Q_H Structural Homology

fraction of native contacts for aligned residues + presence and perturbation of gaps

$$Q_H = \aleph \left[q_{aln} + q_{gap} \right]$$

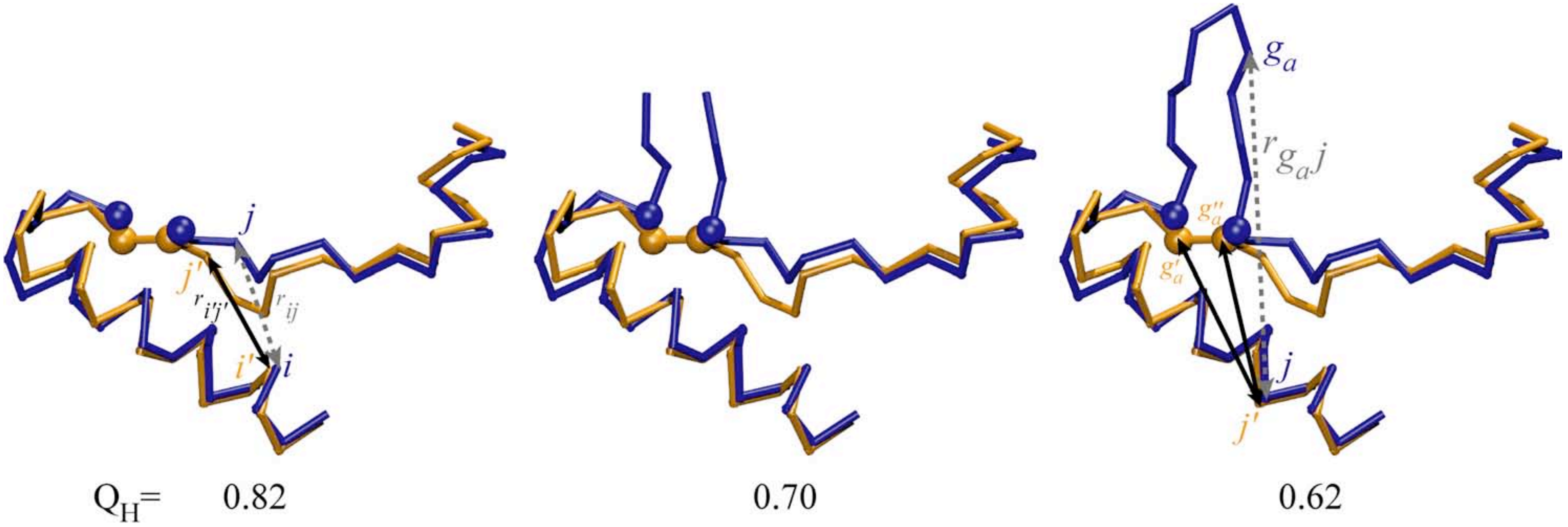
$$q_{aln} = \sum_{i<j-2} \exp\left[-\frac{(r_{ij} - r_{i'j'})^2}{2\sigma_{ij}^2} \right]$$

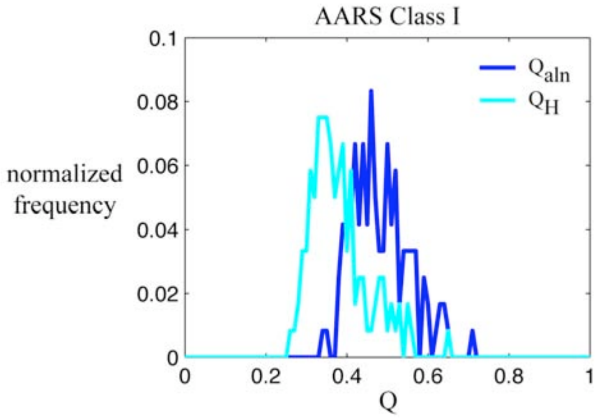


O’Donoghue & Luthey-Schulten *MMBR* 2003.

Structural Similarity Measure the effect of insertions

"Gaps should count as a character but not dominate" C. Woese

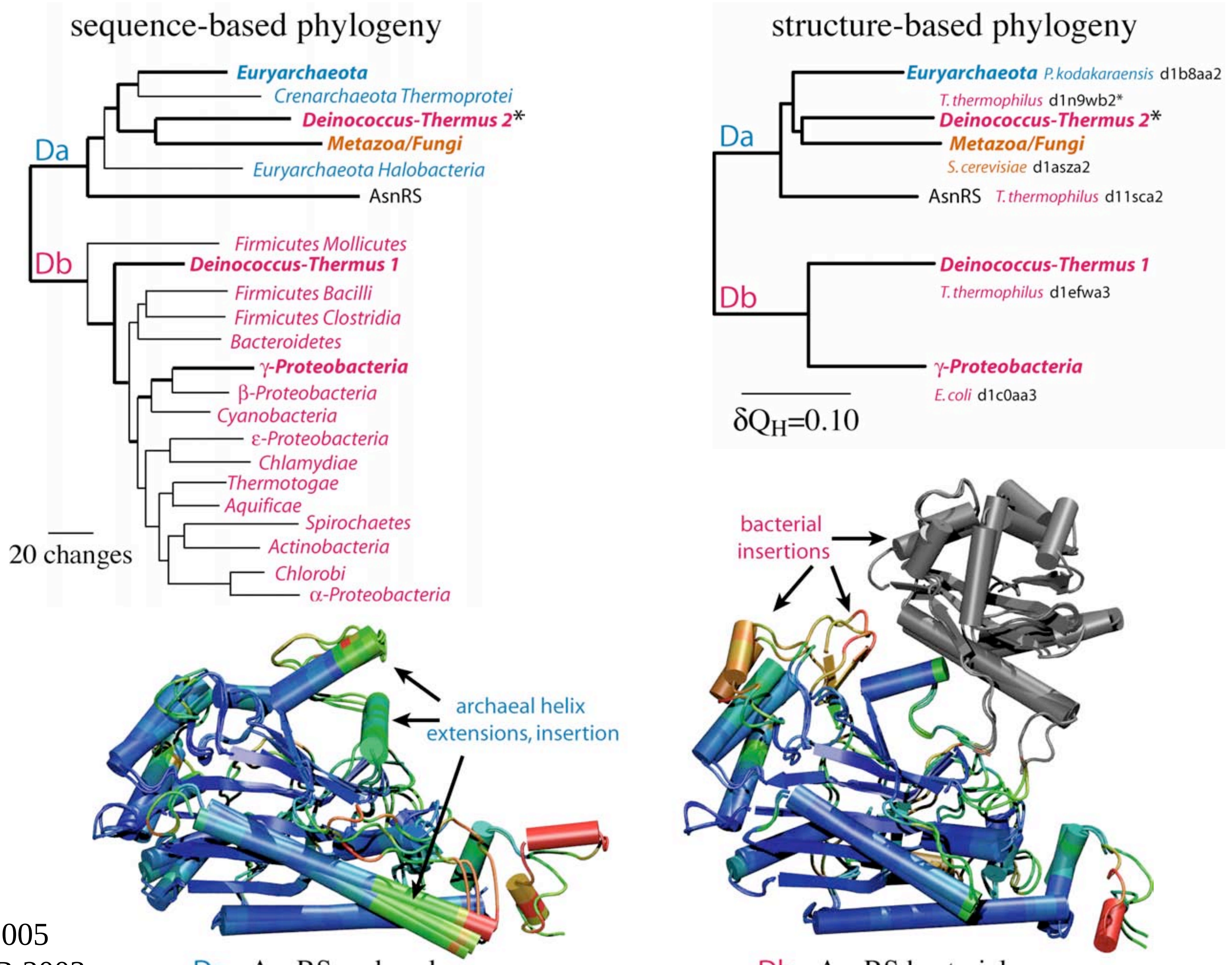




$$
q_{gap} = \sum_{g_a} \sum_{j}^{N_{aln}} \max \left\{ \exp\left[-\frac{\left(r_{g_a j} - r_{g'_a j'}\right)^2}{2\sigma^2_{g_a j}} \right], \exp\left[-\frac{\left(r_{g_a j} - r_{g''_a j'}\right)^2}{2\sigma^2_{g_a j}} \right] \right\}
$$

$$
+ \sum_{g_b} \sum_{j}^{N_{aln}} \max \left\{ \exp\left[-\frac{\left(r_{g_b j} - r_{g'_b j'}\right)^2}{2\sigma^2_{g_b j}} \right], \exp\left[-\frac{\left(r_{g_b j} - r_{g''_b j'}\right)^2}{2\sigma^2_{g_b j}} \right] \right\}
$$

Protein structure encodes evolutionary information



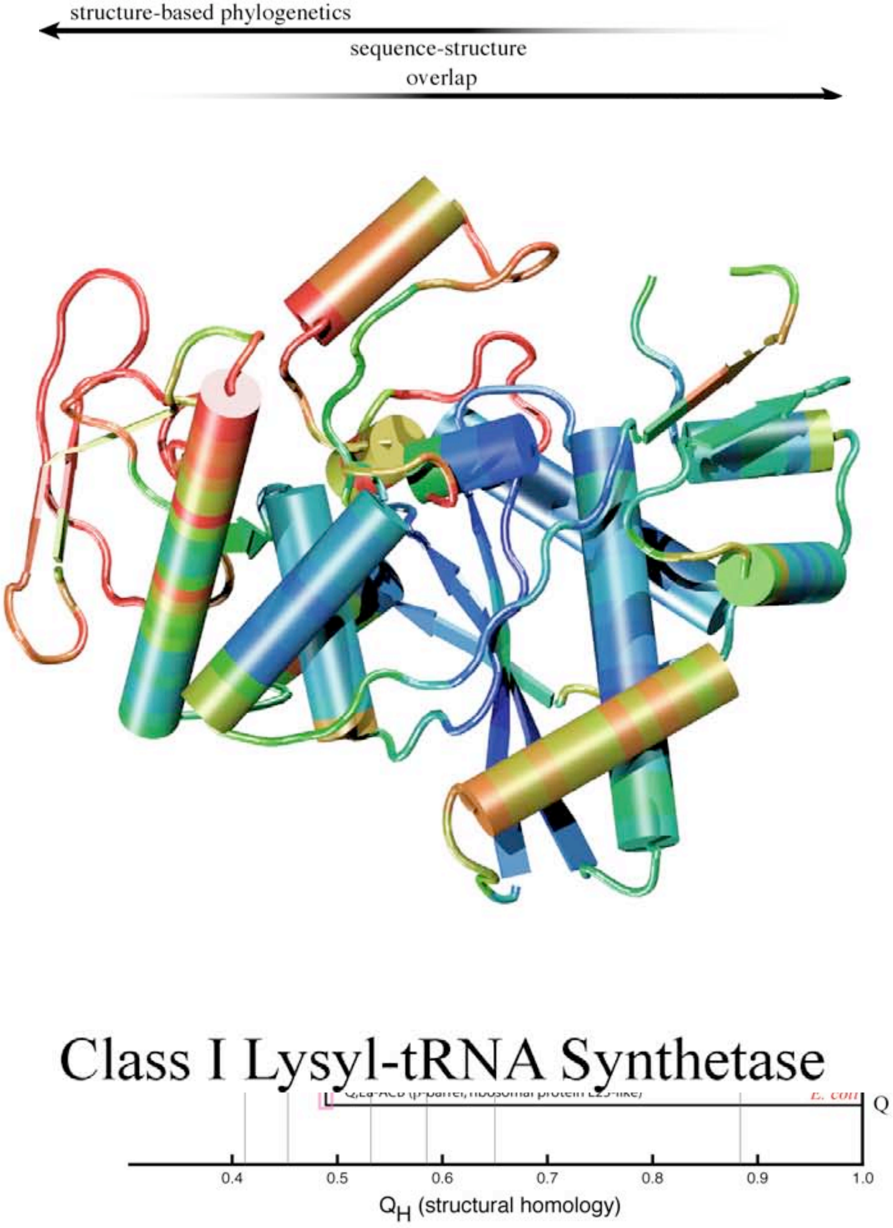
Da - AspRS archaeal genre

Db - AspRS bacterial genre

JMB 2005
MMBR 2003

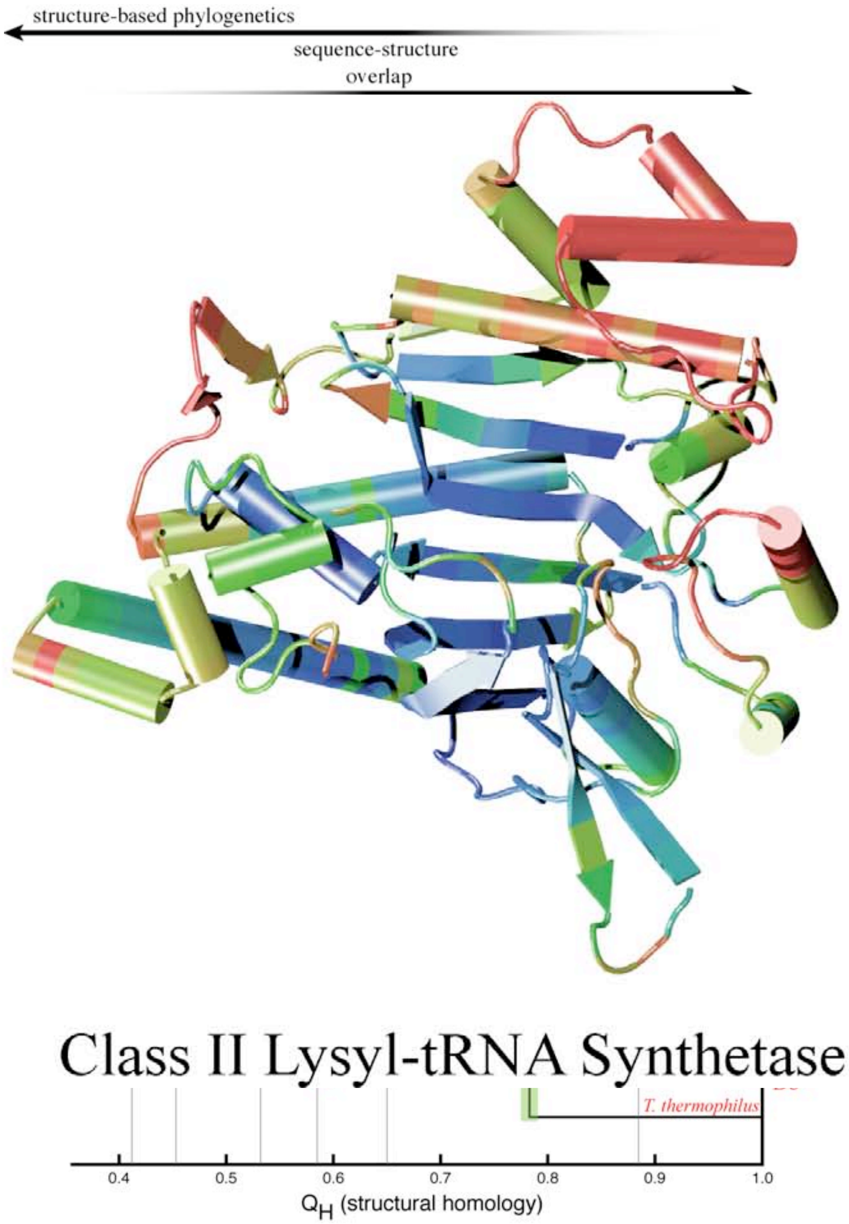
Protein structure reveals distant evolutionary events

Class I AARSs


structure-based phylogenetics
sequence-structure overlap
Q,La-ACB (β-barrel, ribosomal protein L25-like)
E. coli
Q
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Q_H (structural homology)

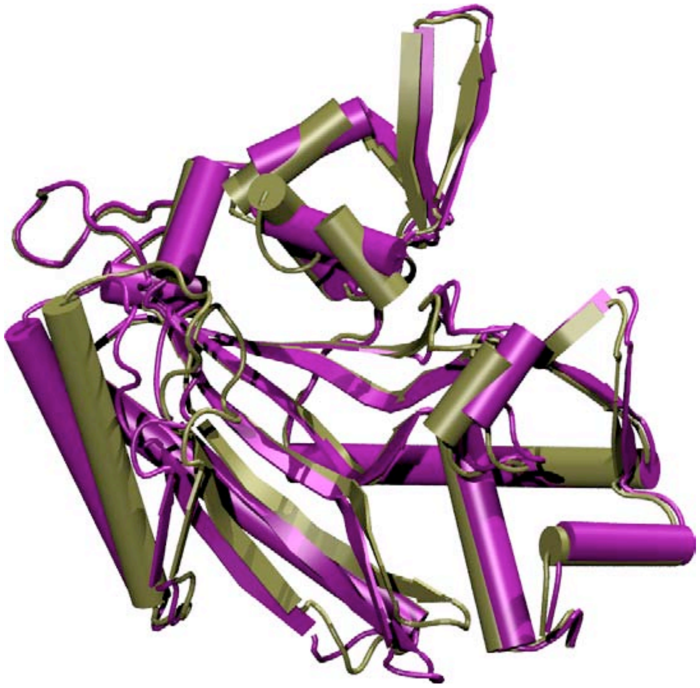
Class I Lysyl-tRNA Synthetase

Class II AARSs


structure-based phylogenetics
sequence-structure overlap
T. thermophilus
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Q_H (structural homology)

Class II Lysyl-tRNA Synthetase

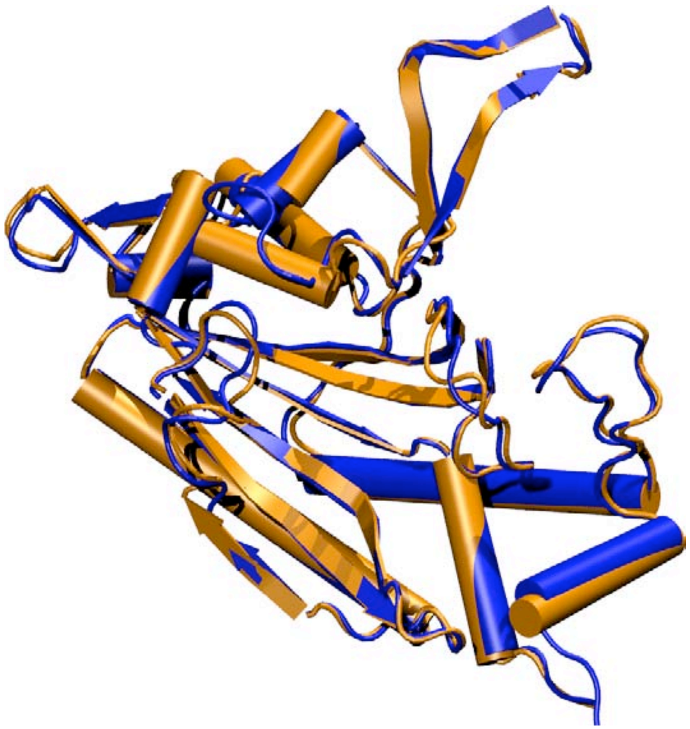
Sequences define more recent evolutionary events



Conformational changes in the same protein.

ThrRS
T-AMP analog, 1.55 A.
T, 2.00 A.

$Q_H = 0.80$
Sequence identity = 1.00



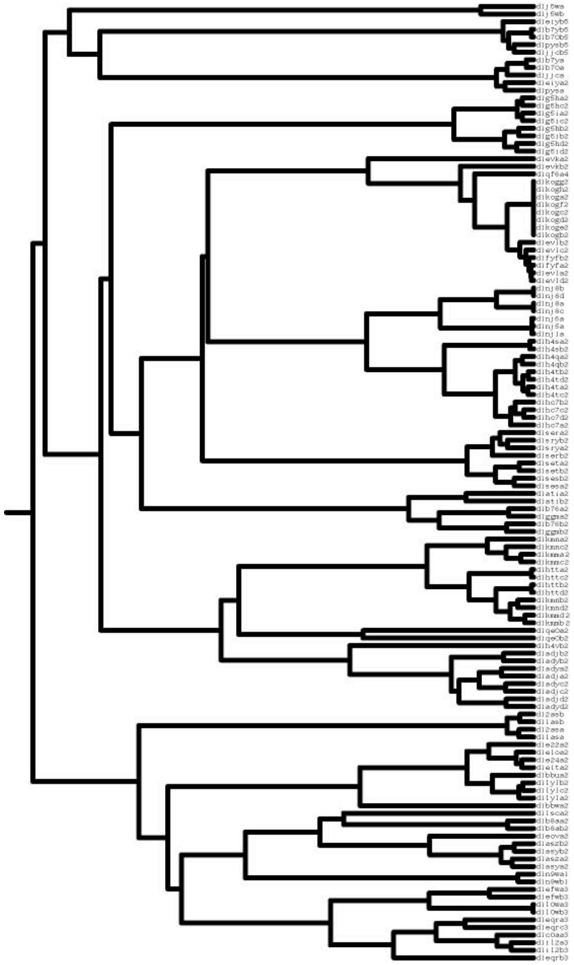
Structures for two different species.

ProRS
M. jannaschii, 2.55 A.
M. thermoautotrophicus, 3.20 A.

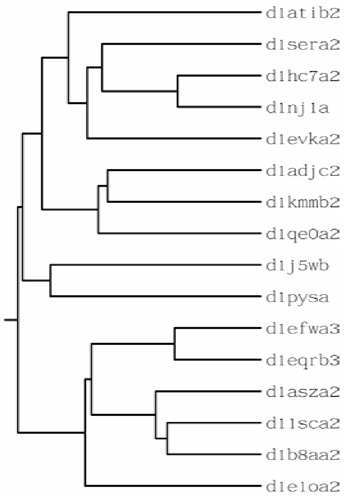
$Q_H = 0.89$
Sequence identity = 0.69

Non-redundant Representative Sets

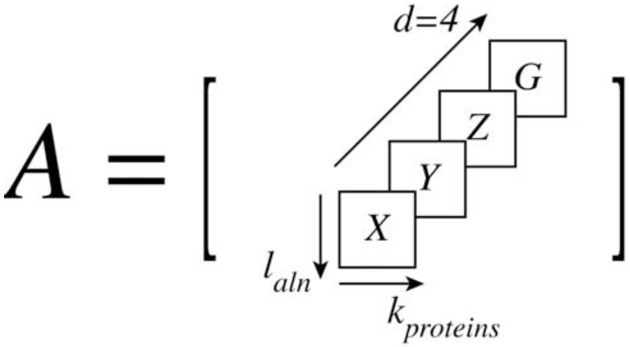
Too much information
129 Structures

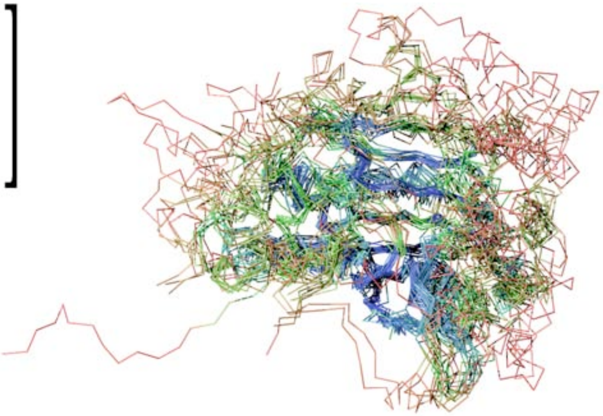


Economy of information
16 representatives



Multidimensional QR factorization of alignment matrix, ***A***.





QR computes a set of maximal linearly independent structures.

P. O'Donoghue and Z. Luthey-Schulten (2003) *MMBR* **67**:550-571.

P. O'Donoghue and Z. Luthey-Schulten (2005) *J. Mol. Biol.*, **346**, 875-894.

Numerical Encoding of Proteins in a Multiple Alignment

Encoding Structure

Rotated Cartesian + Gap = 4-space

Aligned position $(x_{C_\alpha}, y_{C_\alpha}, z_{C_\alpha}, 0)$

Gapped position $(0, 0, 0, g)$

Gap Scaling $g = \gamma \dfrac{\|X\|_{F_4} + \|Y\|_{F_4} + \|Z\|_{F_4}}{\|G\|_{F_4}}$

adjustable parameter

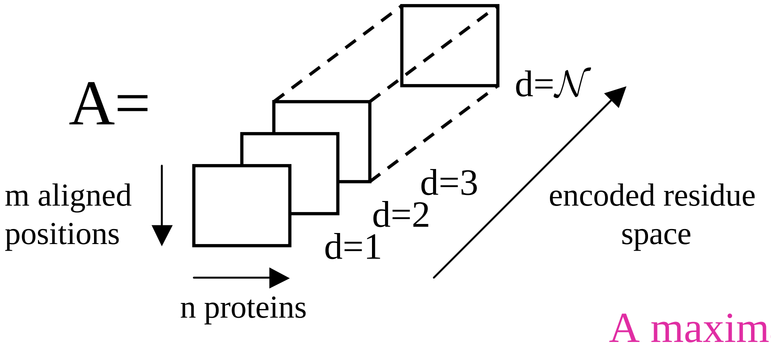
Sequence Space

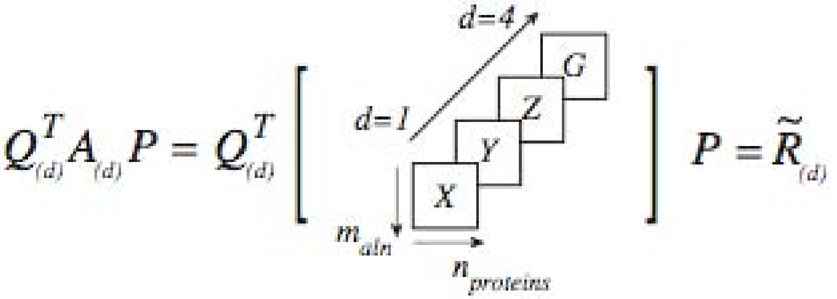
Orthogonal Encoding = 24-space

23 amino acids (20 + B, X, Z) + gap

A = (1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
B = (0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
C = (0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
…
GAP = (0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1)

Alignment is a Matrix with Linearly Dependent Columns





A maximal linearly independent subset can be determined with respect to a threshold, e.g., similarity measure threshold.

Applications of Evolutionary Profiles

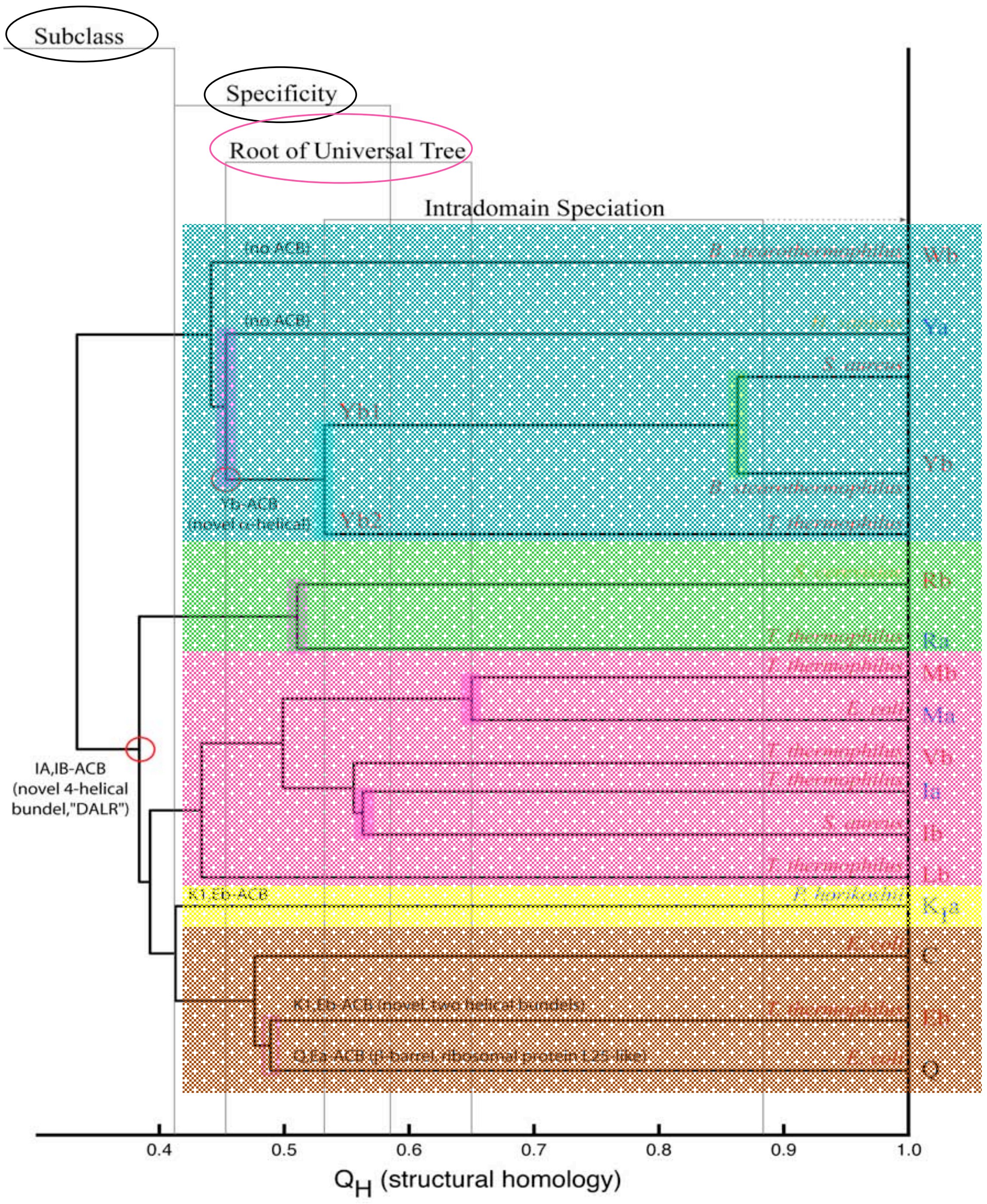
I. **Genome Annotation** AARS - MJ1660

II. **Conserved Core -- Folding Nuclei? HD Exchange?**

III. **Functional Ancestor ?**

IV. **Classification of Protein Structures -** Superfamilies

Class I AARSs evolutionary events

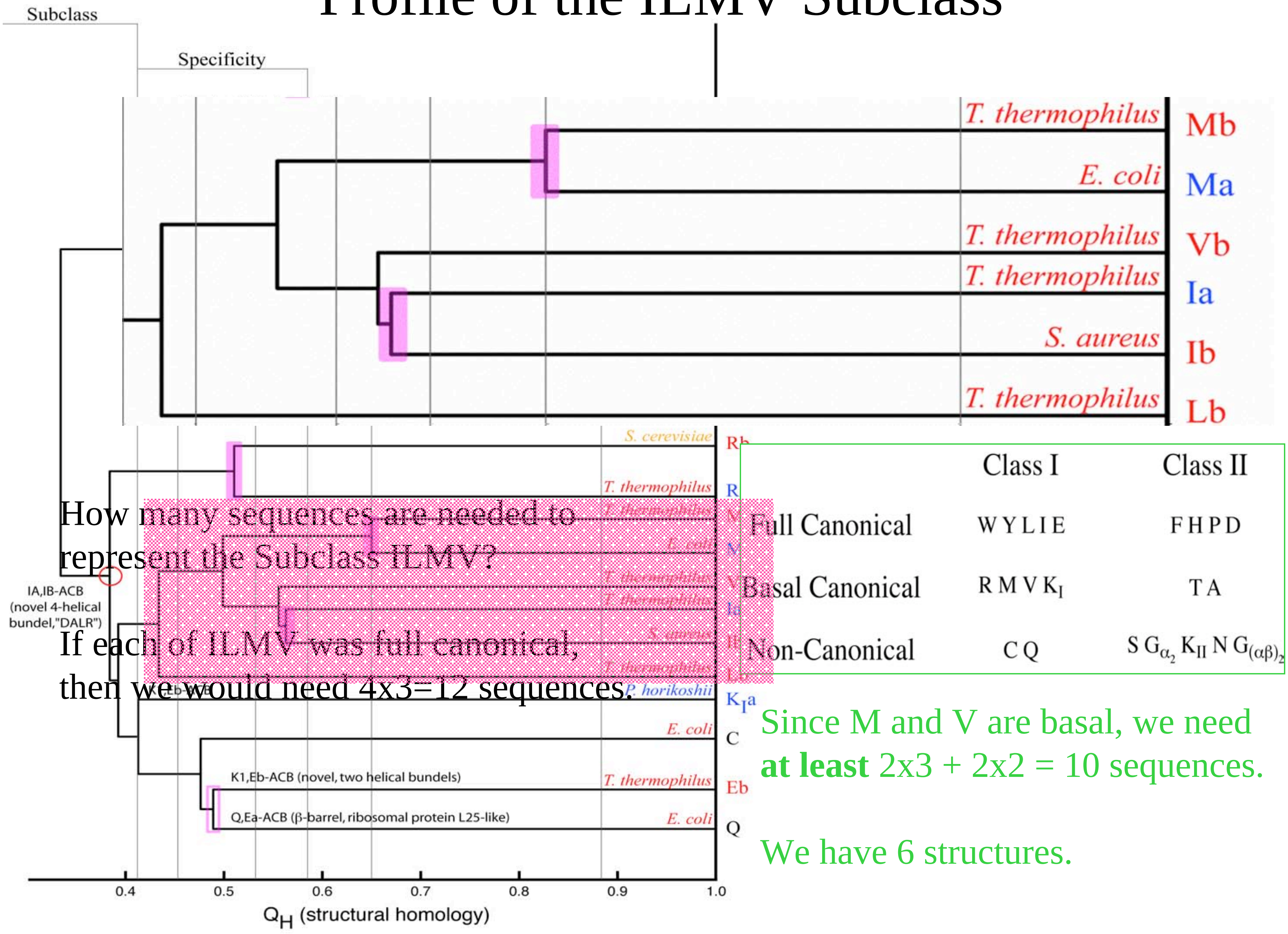


5 Subclasses

Specificity – 11 Amino acids

Domain of life A,B,E

Profile of the ILMV Subclass



How many sequences are needed to represent the Subclass ILMV?

If each of ILMV was full canonical, then we would need 4x3=12 sequences.

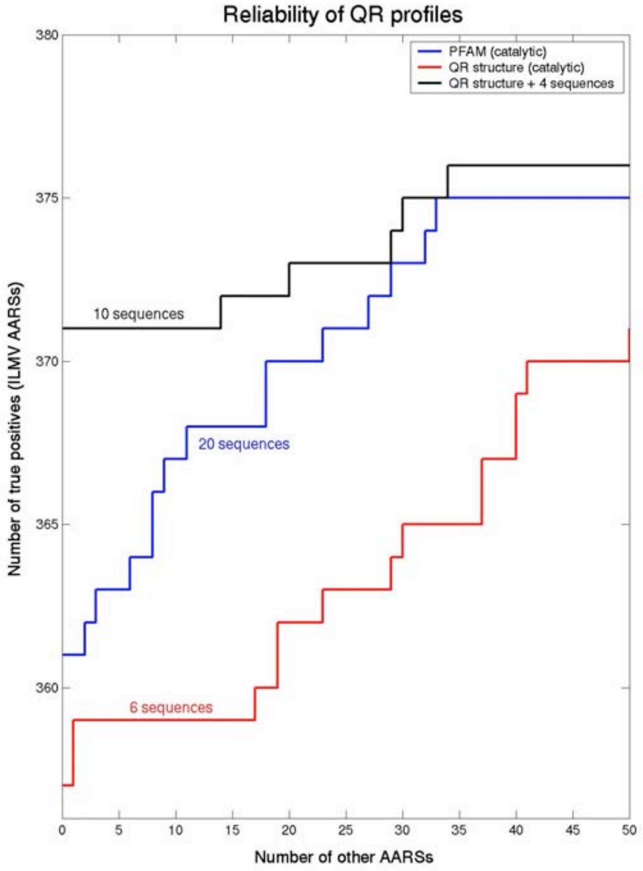
	Class I	Class II
Full Canonical	W Y L I E	F H P D
Basal Canonical	R M V K_I	T A
Non-Canonical	C Q	S G_{α_2} K_{II} N $G_{(\alpha\beta)_2}$

Since M and V are basal, we need **at least** 2x3 + 2x2 = 10 sequences.

We have 6 structures.

Evolutionary Profiles for Homology Recognition

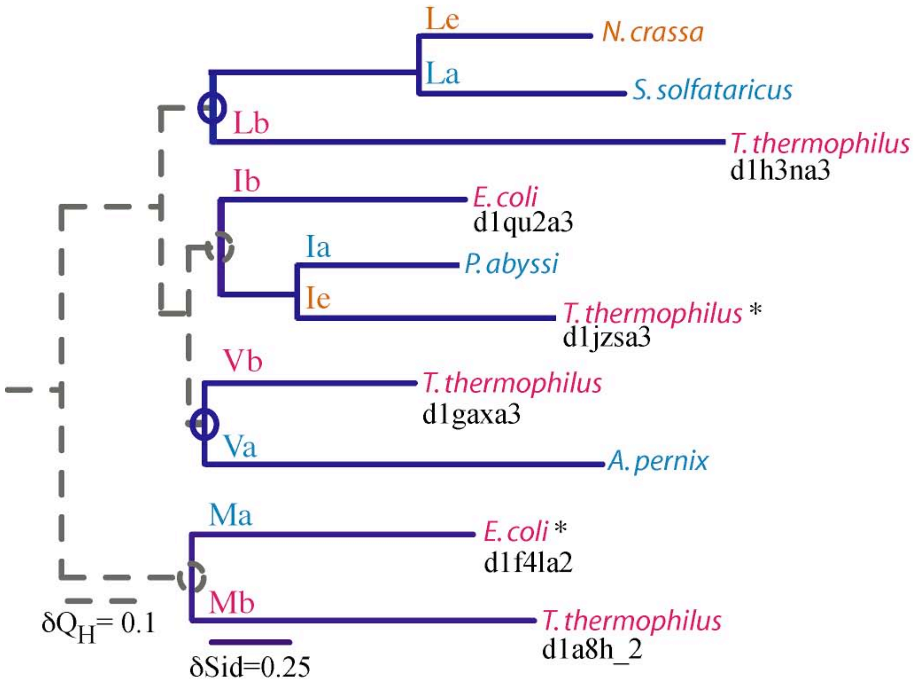
AARS Subclass ILMV



false positives

Combined Structure-Sequence Phylogeny

an evolutionary profile of the AARS subclass IA



Pfam profile composition

Le	×0
La	×0
Lb	×3
Ib	×4
Ia	×1
Ie	×4
Vb	×7
Va	×0
Ma	×1
Mb	×0

The composition of the profile matters.
Choosing the right 10 sequence makes all the difference.

A. Sethi, P. O’Donoghue, Z. Luthey-Schulten (2005) *JMB, PNAS*

Genome Annotation

M.jannaschii genome was completely sequenced in 1996.
Genome had four missing AARSs:

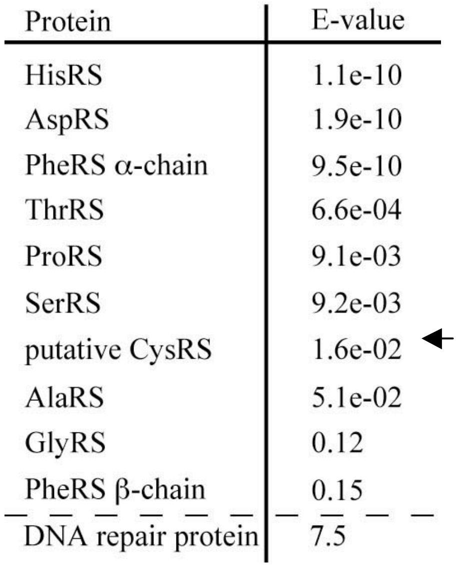
AsnRS
GlnRS } Indirect Mechanism

LysRS Class I AARS

CysRS ?

Cysteinyl-tRNA(Cys) formation in *Methanocaldococcus jannaschii*: the mechanism is still unknown. *J. Bacteriology*, Jan. 2004, **186:**8-14.
Ruan B, Nakano H, Tanaka M, Mills JA, DeVito JA, Min B, Low KB, Battista JR, and Söll D.

M. jannaschii genome database search using EP of class II AARS with HMMER



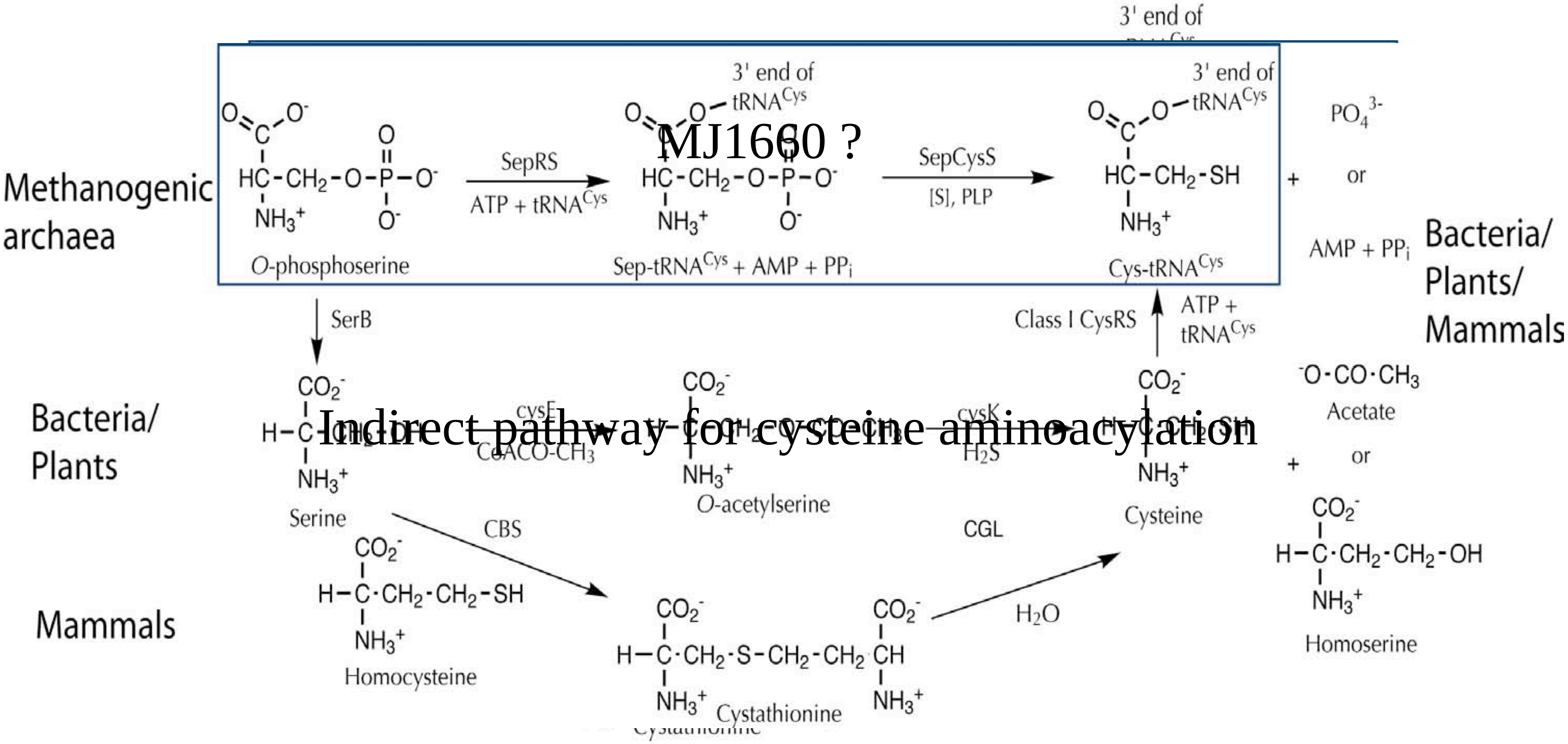
Protein	E-value
HisRS	1.1e-10
AspRS	1.9e-10
PheRS α-chain	9.5e-10
ThrRS	6.6e-04
ProRS	9.1e-03
SerRS	9.2e-03
putative CysRS	1.6e-02
AlaRS	5.1e-02
GlyRS	0.12
PheRS β-chain	0.15
DNA repair protein	7.5

← MJ1660

A. Sethi, P. O'Donoghue and
Z. Luthey-Sculten. PNAS, **102.** 2005

Pathways for cysteine biosynthesis

Direct pathway for cysteine aminoacylation

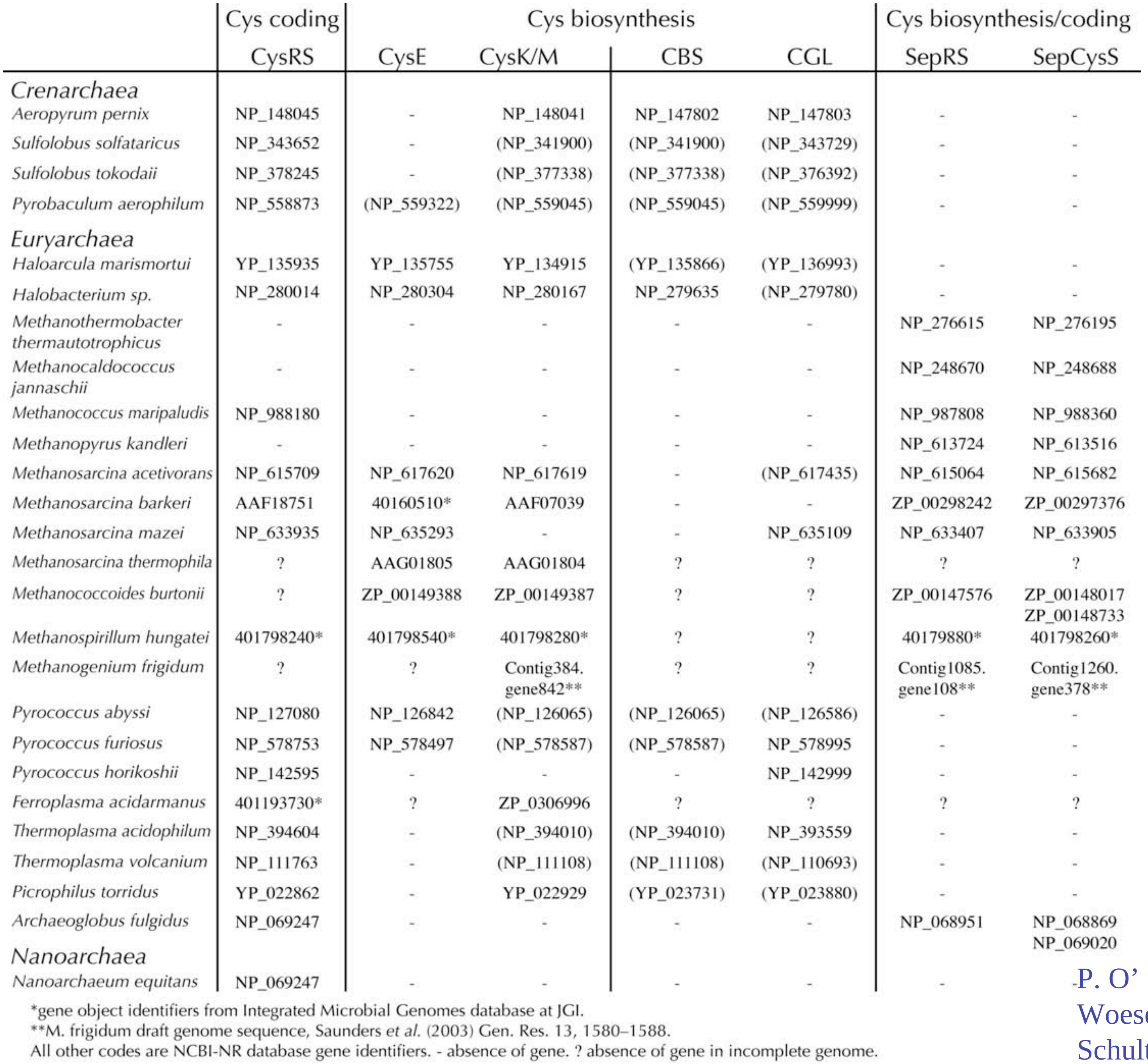


MJ1660 ?

Indirect pathway for cysteine aminoacylation

Sauerwald et al., Science, 307, 2005, 1969-1972.

Genes for Cysteine Biosynthesis and Aminoacylation

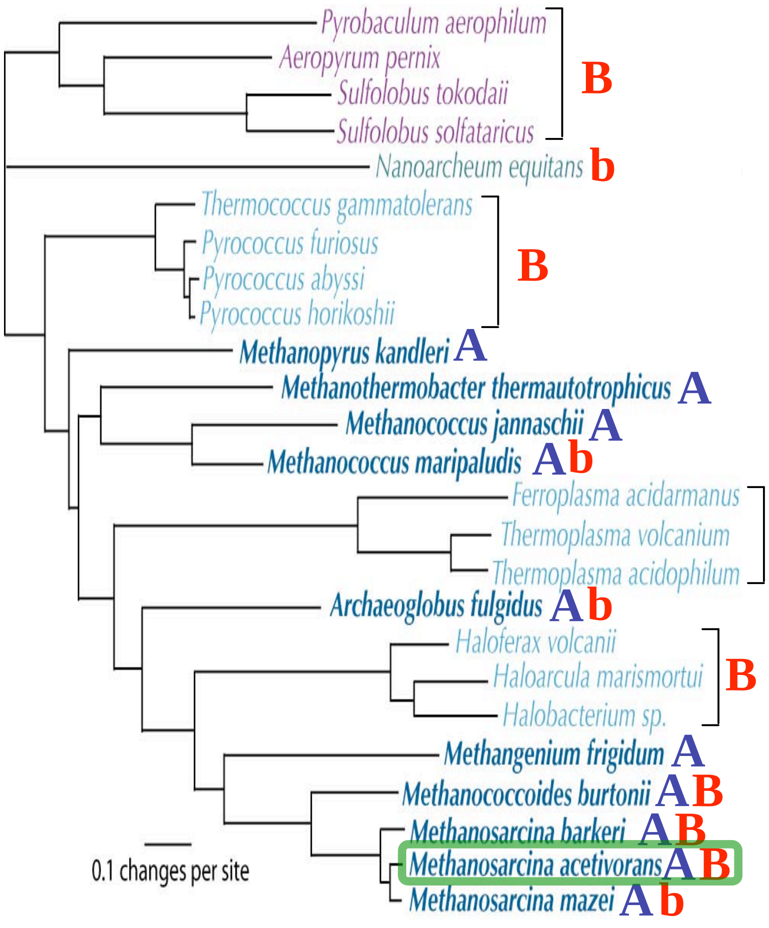


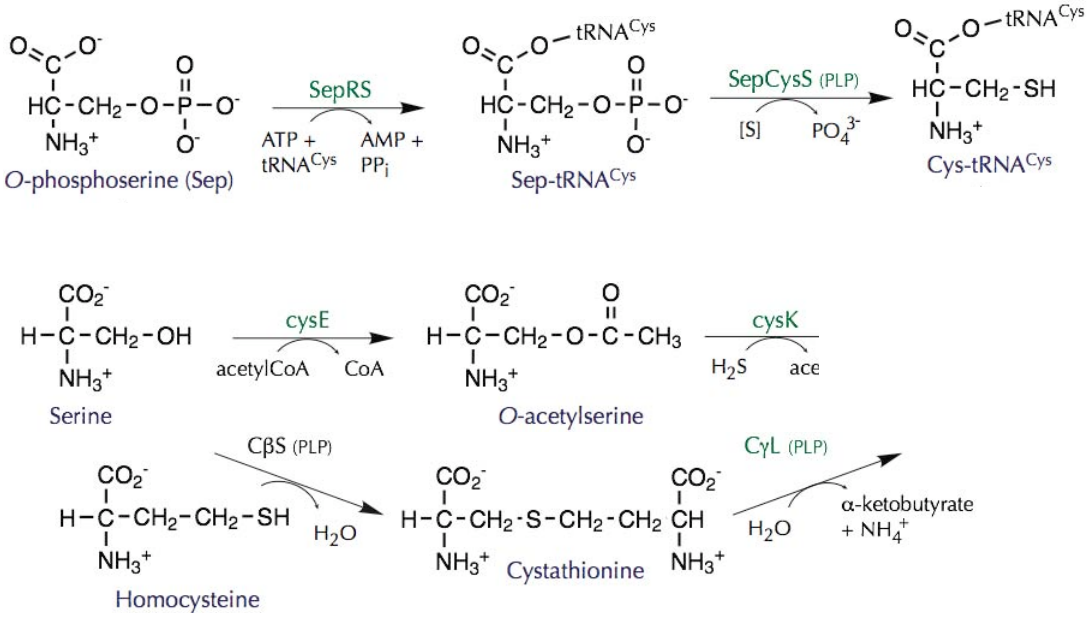
	Cys coding	Cys biosynthesis				Cys biosynthesis/coding	
	CysRS	CysE	CysK/M	CBS	CGL	SepRS	SepCysS
Crenarchaea							
Aeropyrum pernix	NP_148045	-	NP_148041	NP_147802	NP_147803	-	-
Sulfolobus solfataricus	NP_343652	-	(NP_341900)	(NP_341900)	(NP_343729)	-	-
Sulfolobus tokodaii	NP_378245	-	(NP_377338)	(NP_377338)	(NP_376392)	-	-
Pyrobaculum aerophilum	NP_558873	(NP_559322)	(NP_559045)	(NP_559045)	(NP_559999)	-	-
Euryarchaea							
Haloarcula marismortui	YP_135935	YP_135755	YP_134915	(YP_135866)	(YP_136993)	-	-
Halobacterium sp.	NP_280014	NP_280304	NP_280167	NP_279635	(NP_279780)	-	-
Methanothermobacter thermautotrophicus	-	-	-	-	-	NP_276615	NP_276195
Methanocaldococcus jannaschii	-	-	-	-	-	NP_248670	NP_248688
Methanococcus maripaludis	NP_988180	-	-	-	-	NP_987808	NP_988360
Methanopyrus kandleri	-	-	-	-	-	NP_613724	NP_613516
Methanosarcina acetivorans	NP_615709	NP_617620	NP_617619	-	(NP_617435)	NP_615064	NP_615682
Methanosarcina barkeri	AAF18751	40160510*	AAF07039	-	-	ZP_00298242	ZP_00297376
Methanosarcina mazei	NP_633935	NP_635293	-	-	NP_635109	NP_633407	NP_633905
Methanosarcina thermophila	?	AAG01805	AAG01804	?	?	?	?
Methanococcoides burtonii	?	ZP_00149388	ZP_00149387	?	?	ZP_00147576	ZP_00148017 ZP_00148733
Methanospirillum hungatei	401798240*	401798540*	401798280*	?	?	40179880*	401798260*
Methanogenium frigidum	?	?	Contig384. gene842**	?	?	Contig1085. gene108**	Contig1260. gene378**
Pyrococcus abyssi	NP_127080	NP_126842	(NP_126065)	(NP_126065)	(NP_126586)	-	-
Pyrococcus furiosus	NP_578753	NP_578497	(NP_578587)	(NP_578587)	NP_578995	-	-
Pyrococcus horikoshii	NP_142595	-	-	-	NP_142999	-	-
Ferroplasma acidarmanus	401193730*	?	ZP_0306996	?	?	?	?
Thermoplasma acidophilum	NP_394604	-	(NP_394010)	(NP_394010)	NP_393559	-	-
Thermoplasma volcanium	NP_111763	-	(NP_111108)	(NP_111108)	(NP_110693)	-	-
Picrophilus torridus	YP_022862	-	YP_022929	(YP_023731)	(YP_023880)	-	-
Archaeoglobus fulgidus	NP_069247	-	-	-	-	NP_068951	NP_068869 NP_069020
Nanoarchaea							
Nanoarchaeum equitans	NP_069247	-	-	-	-	-	-

*gene object identifiers from Integrated Microbial Genomes database at JGI.
**M. frigidum draft genome sequence, Saunders *et al.* (2003) Gen. Res. 13, 1580–1588.
All other codes are NCBI-NR database gene identifiers. - absence of gene. ? absence of gene in incomplete genome.

P. O’ Donoghue, A. Sethi, C. Woese, and Z. Luthey-Schulten, PNAS, 2005.

Cysteine Coding & Biosynthesis





A - SepRS/SepCysS (VGT)

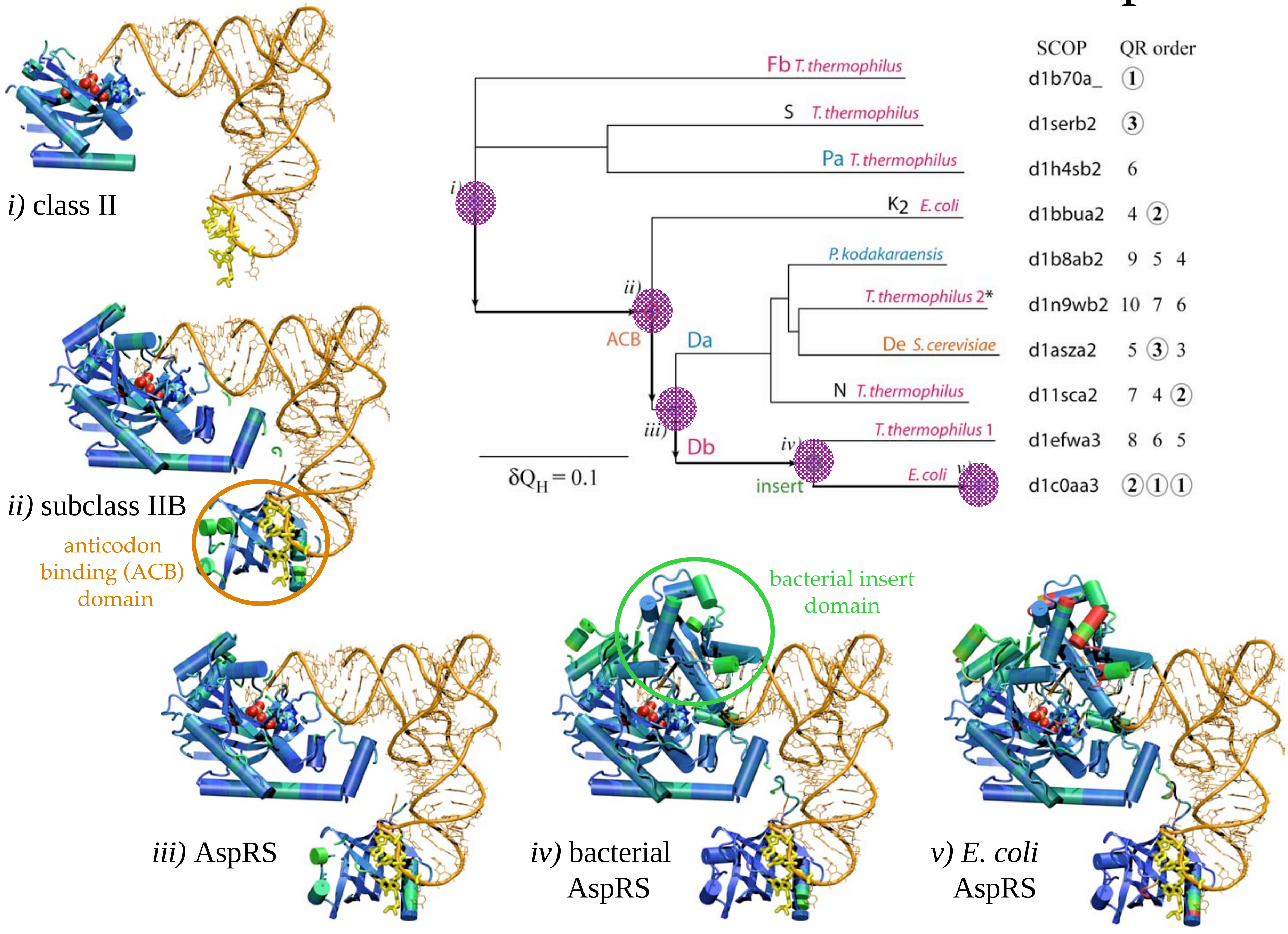
b - CysRS (HGT)

B - Biosynthesis genes (HGT)

Genetic tools will help uncover the role of the native versus the acquired pathways.

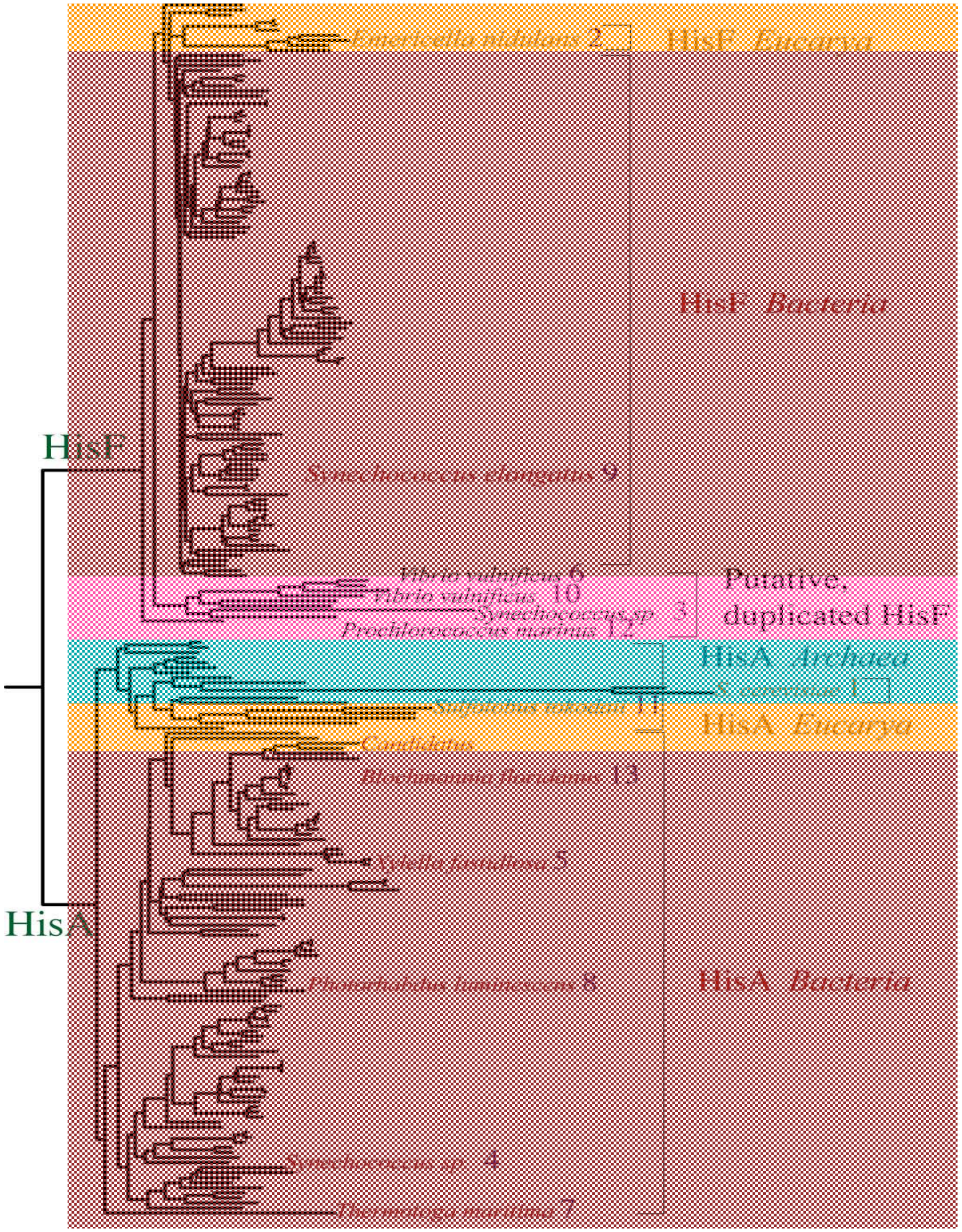
Experiments: Patrick O'Donoghue (Chemistry, UIUC,Yale), Bill Metcalf (Microbiology, IGB, UIUC) , Claudia Reich (Microbiology, UIUC), Michael Hohn & Dieter Söll (Yale University).

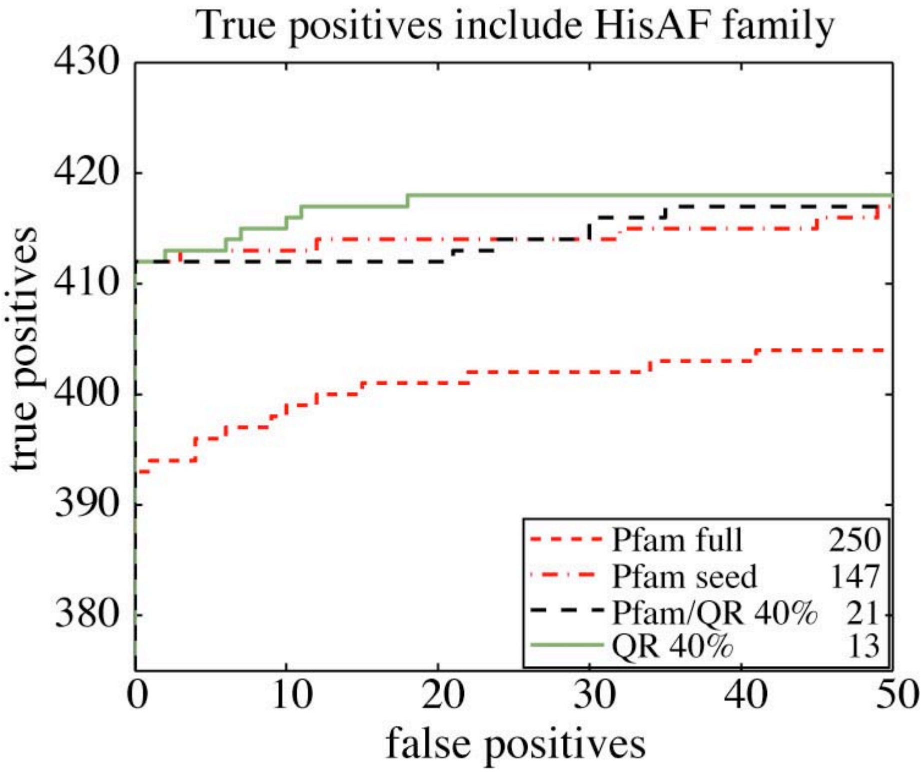
Evolution of Structure and Function in AspRS



SCOP	QR order		
d1b70a_	①		
d1serb2	③		
d1h4sb2	6		
d1bbua2	4	②	
d1b8ab2	9	5	4
d1n9wb2	10	7	6
d1asza2	5	③	3
d11sca2	7	4	②
d1efwa3	8	6	5
d1c0aa3	②	①	①

Evolutionary profile for HisA-HisF family

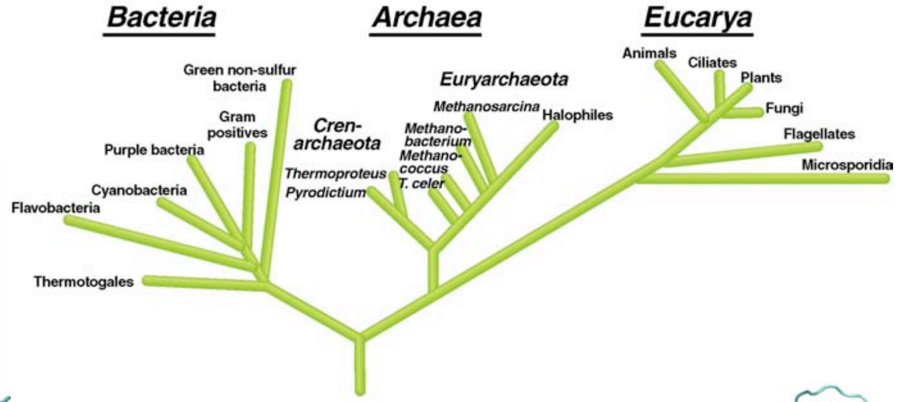


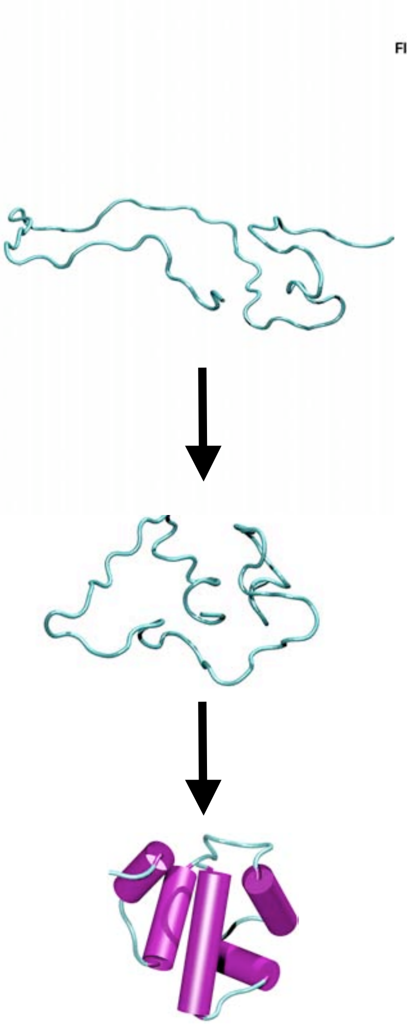


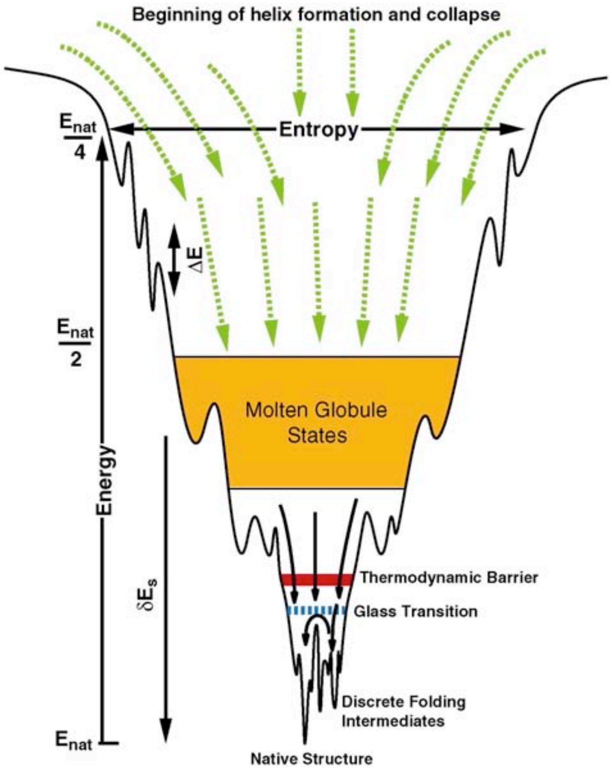
EP outperforms popular profile methods with an economy of information.

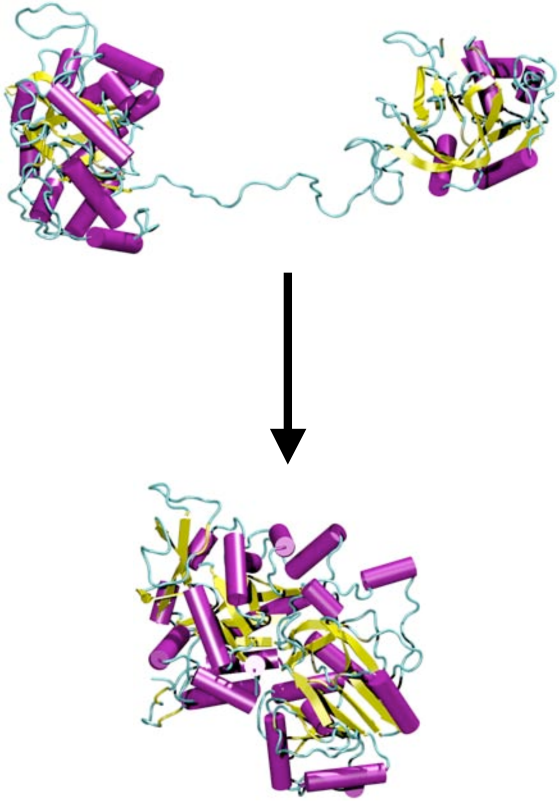
Sethi, et. al., PNAS, 2005.

Unifying the Worlds of Sequence and Structure

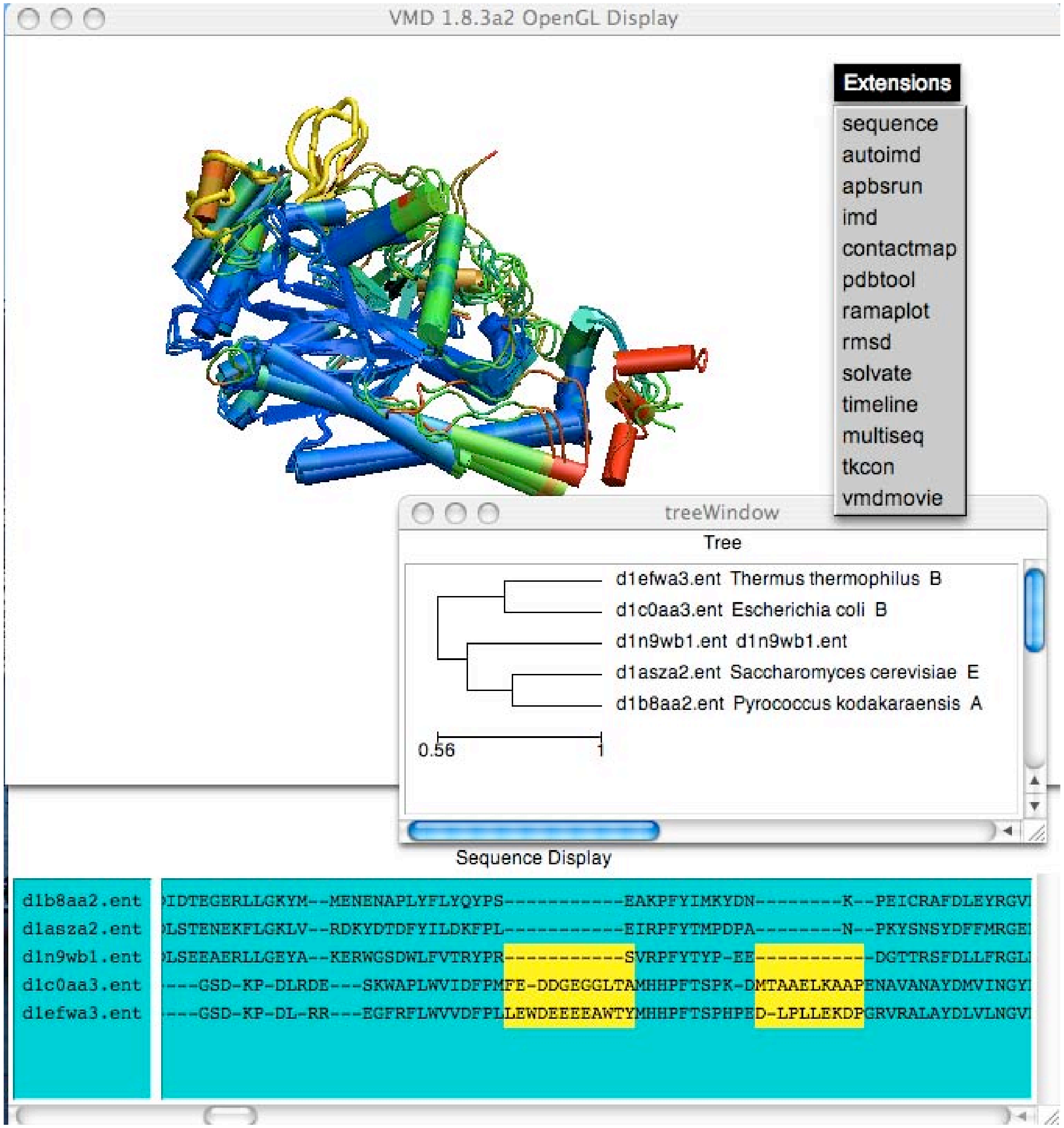








Multiseq in VMD : Merging the sequence and structure worlds

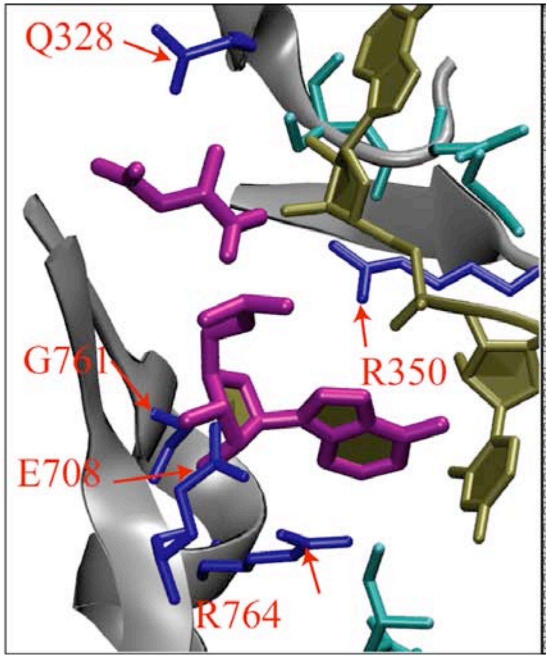

VMD 1.8.3a2 OpenGL Display
Extensions
sequence
autoimd
apbsrun
imd
contactmap
pdbtool
ramaplot
rmsd
solvate
timeline
multiseq
tkcon
vmdmovie
treeWindow
Tree
d1efwa3.ent Thermus thermophilus B
d1c0aa3.ent Escherichia coli B
d1n9wb1.ent d1n9wb1.ent
d1asza2.ent Saccharomyces cerevisiae E
d1b8aa2.ent Pyrococcus kodakaraensis A
0.56
1
Sequence Display

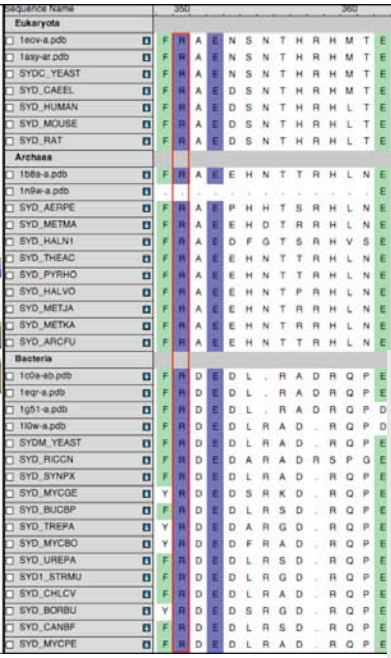
```
d1b8aa2.ent  IDTEGERLLGKYM--MENENAPLYFLYQYPS-----------EAKPFYIMKYDN--------K--PEICRAFDLEYRGV
d1asza2.ent  LSTENEKFLGKLV--RDKYDTDFYILDKFPL-----------EIRPFYTMPDPA--------N--PKYSNSYDFFMRGE
d1n9wb1.ent  LSEEAERLLGEYA--KERWGSDWLFVTRYPR-----------SVRPFYTYP-EE-----------DGTTRSFDLLFRGL
d1c0aa3.ent  ----GSD-KP-DLRDE---SKWAPLWVIDFPMFE-DDGEGGLTAMHHPFTSPK-DMTAAELKAAPENAVANAYDMVINGY
d1efwa3.ent  ----GSD-KP-DL-RR---EGFRFLWVVDFPLLEWDEEEEAWTYMHHPFTSPHPED-LPLLEKDPGRVRALAYDLVLNGV
```

Version 1.83

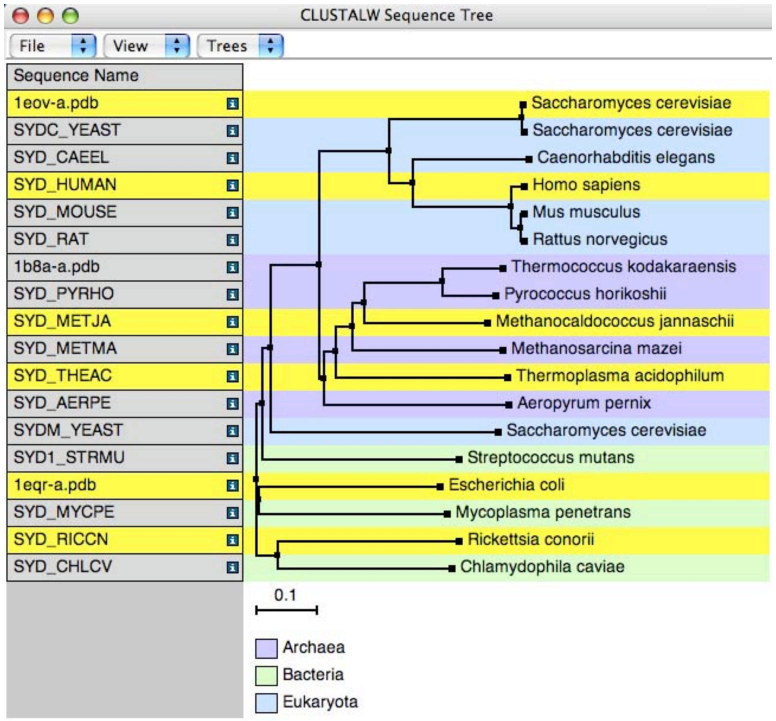
2006 MultiSeq: New Features

Analyze the Evolution of Sequence and Structure

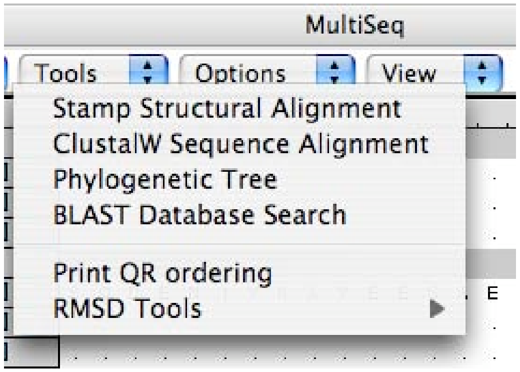




Eliminate Redundancy



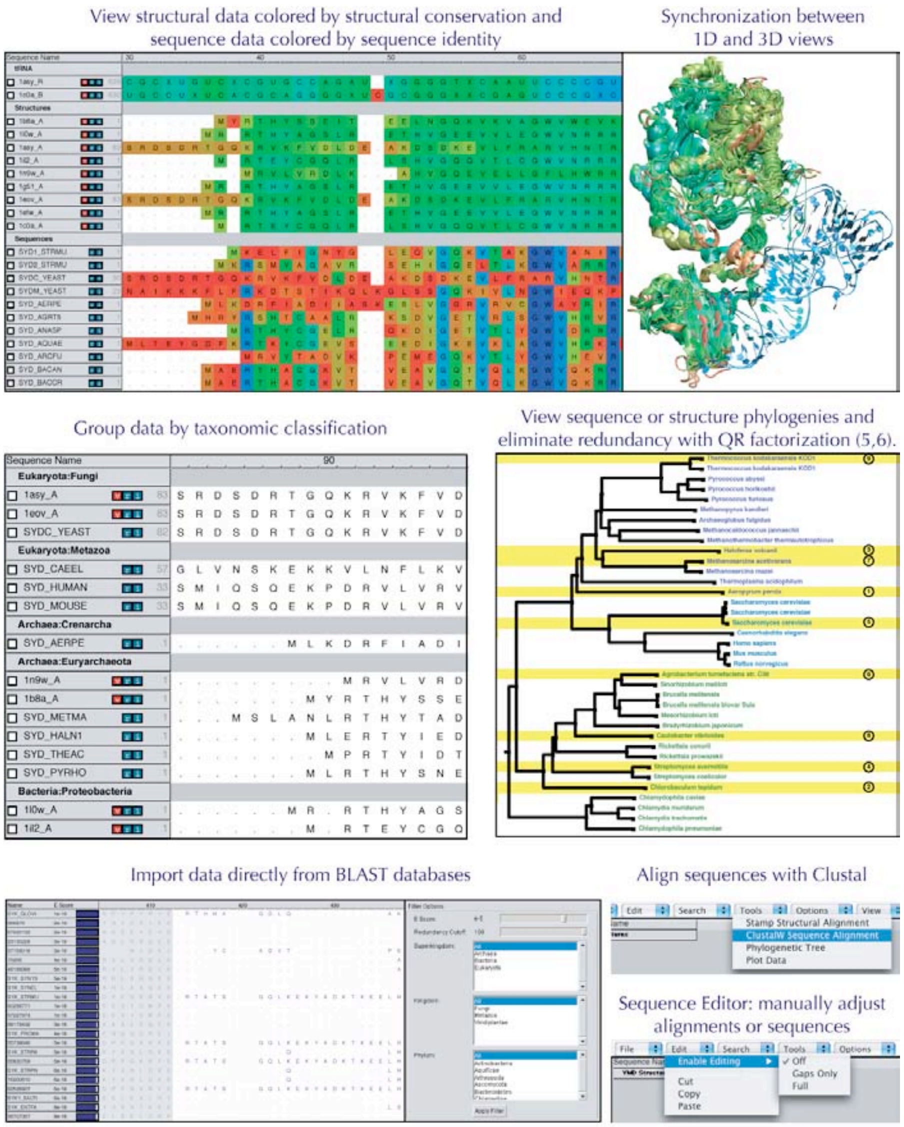
Plus More Functions



Multiseq in VMD 1.8.5

J. Eargle, D. Wright, ZLS Bioinformatics. 2006 Feb 15;22(4):504-6.

E. Roberts, J. Eargle, D. Wright, ZLS in BMC Bioinformatics. Sept. 2006 .



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

“Evolution of Structures in Biomolecules”
Lectures and Tutorials
Frankfurt, 2006

John Eargle, **Elijah Roberts**, Dan Wright, and ZLS.

Acknowledgements

Patrick O’Donoghue

Anurag Sethi

Rommie Amaro
Felix Autenrieth
Alexis Black
John Eargle
Taras Pogorelov
Elijah Roberts
Dan Wright

Funding

NSF, NIH, DOE

Graphics Programmers VMD

Elijah Roberts, Dan Wright, John Eargle

Mike Bach, John Stone

Collaborators

Evolutionary Studies

Gary Olsen, Carl Woese (UIUC)

QR Algorithms

Mike Heath (UIUC)

Protein Structure Prediction

Peter Wolynes, Jose Onuchic (UCSD)

Ken Suslick (UIUC)