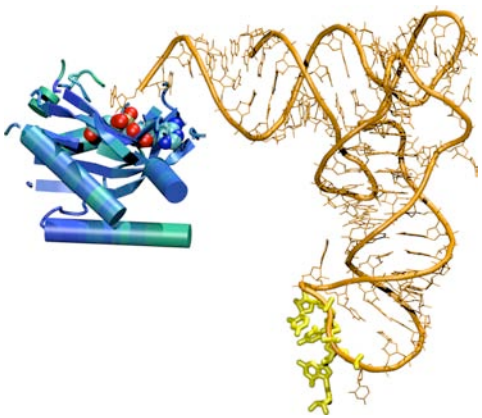
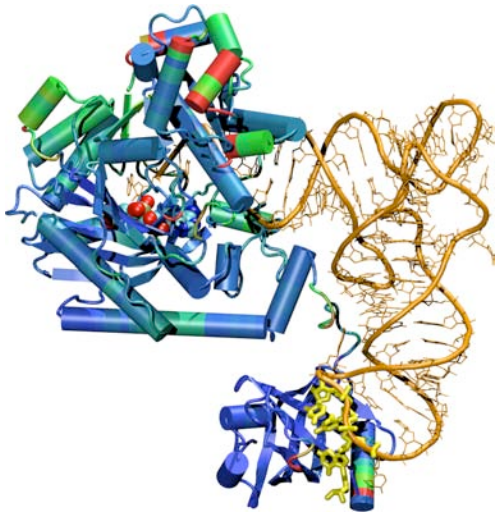
MULTISEQ in VMD -

Revealing How Nature Designs Proteins and RNAs



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

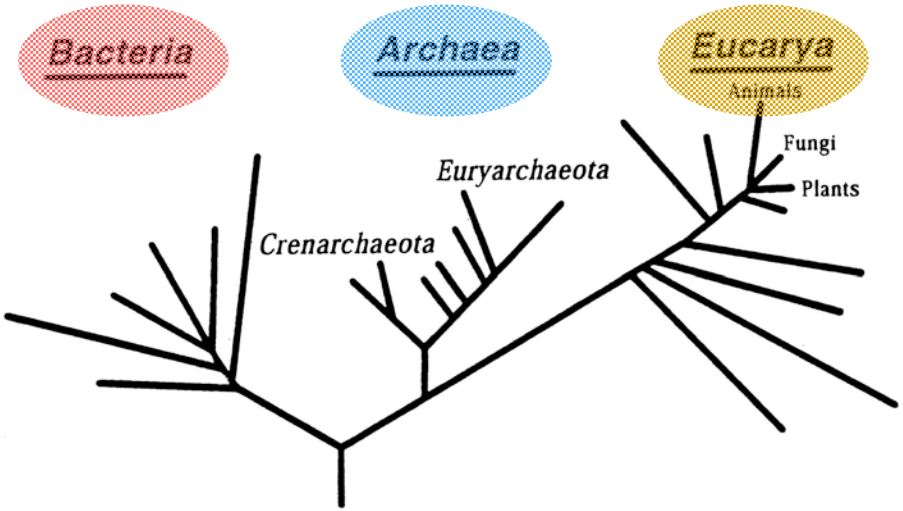


Luthey-Schulten Group

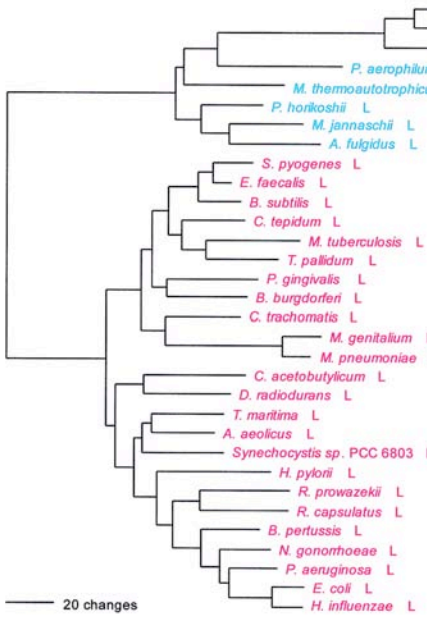
Department of Chemistry, Biophysics, and Beckman Institute
University of Illinois at Urbana-Champaign

Universal Phylogenetic Tree

three domains of life



Based on 16S rRNA



Eucarya

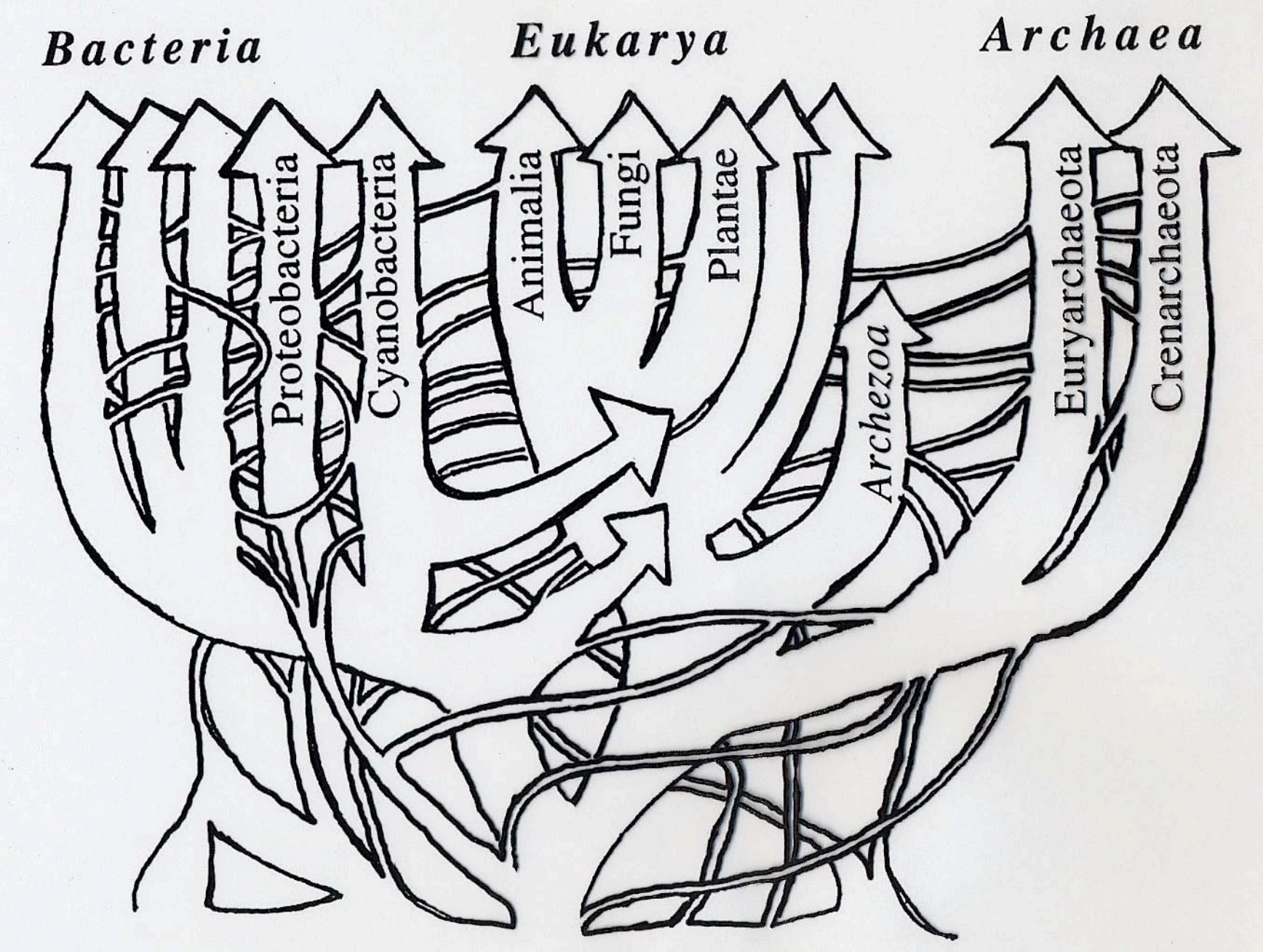
Archaea

Bacteria

Leucyl-tRNA synthetase displays the full canonical phylogenetic distribution.

for review see Woese *PNAS* 2000

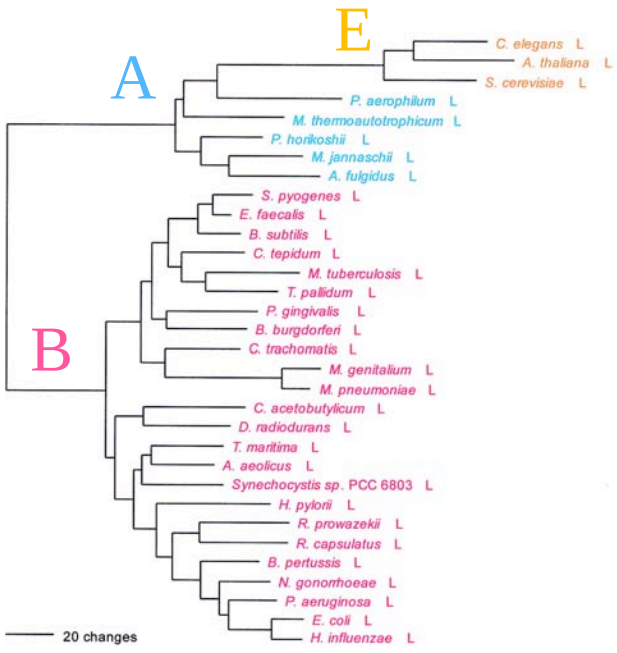
Woese, Olsen, Ibba, Soll *MMBR* 2000



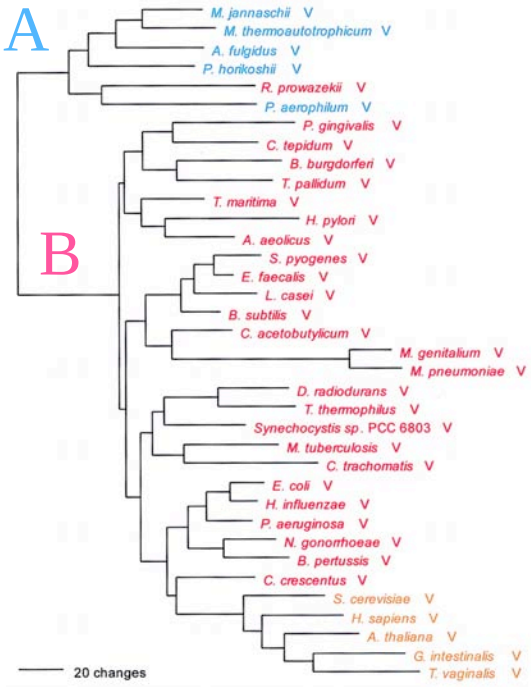
After W. Doolittle, modified by G. Olsen

Phylogenetic Distributions

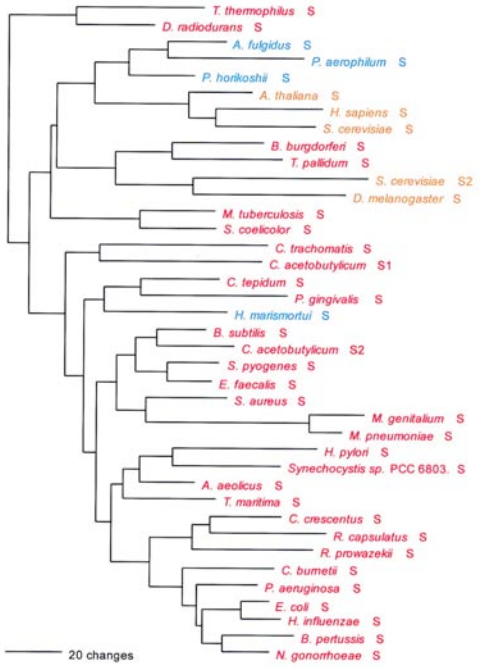
Full Canonical



Basal Canonical



Non-canonical



increasing inter-domain of life Horizontal Gene Transfer

"HGT erodes the historical trace, but does not completely erase it…." G. Olsen

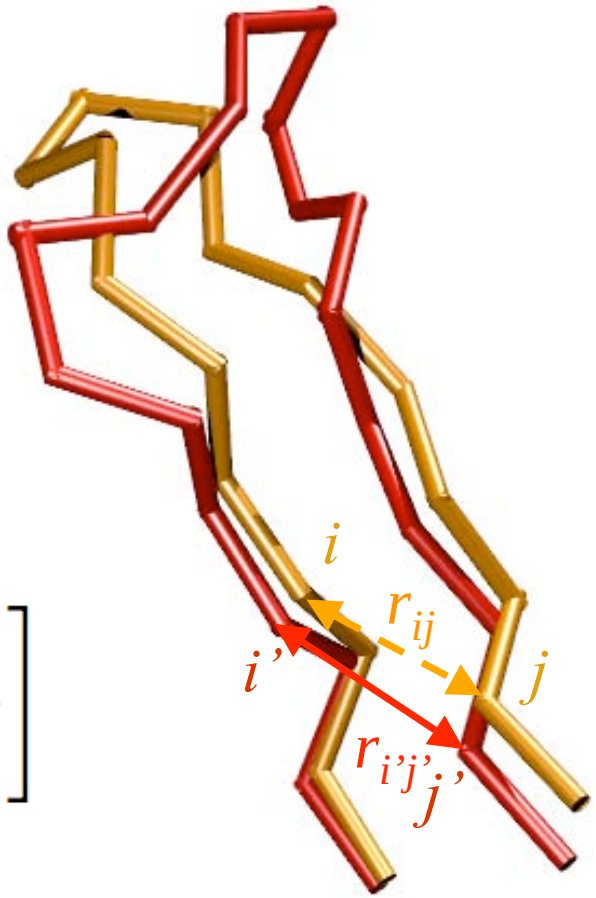
Protein Structure Similarity Measure

Q_H Structural Homology

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

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

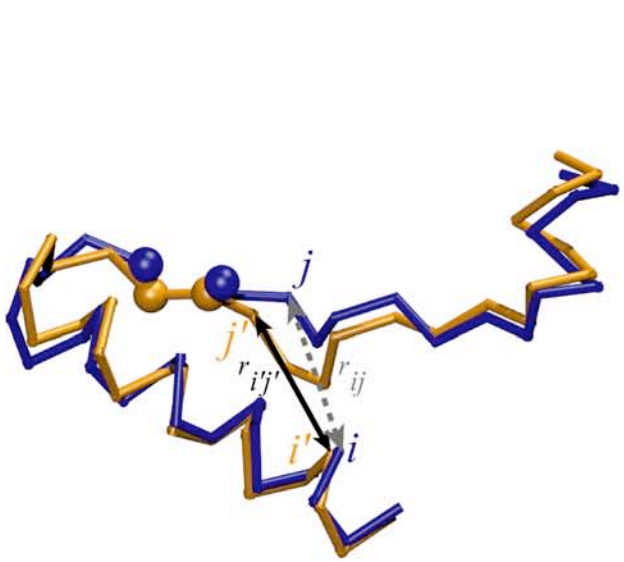
$$q_{aln} = \sum_{i<j-2} \exp\left[-\frac{(r_{ij} - r_{i'j'})^2}{2\sigma_{ij}^2}\right]$$

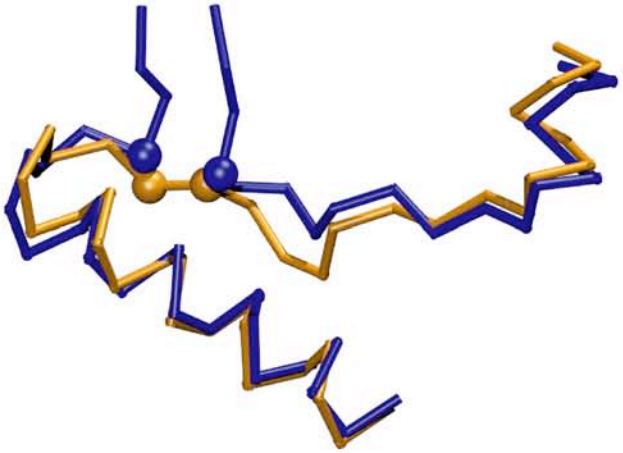


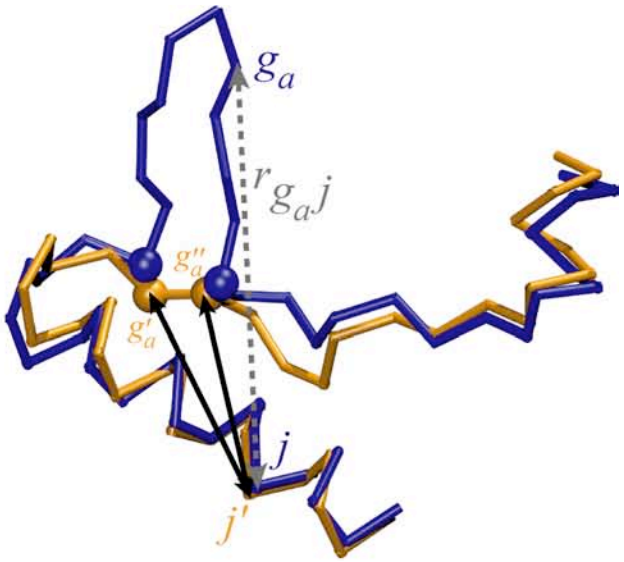
O'Donoghue & Luthey-Schulten *MMBR* 2003.

Structural Similarity Measure the effect of insertions

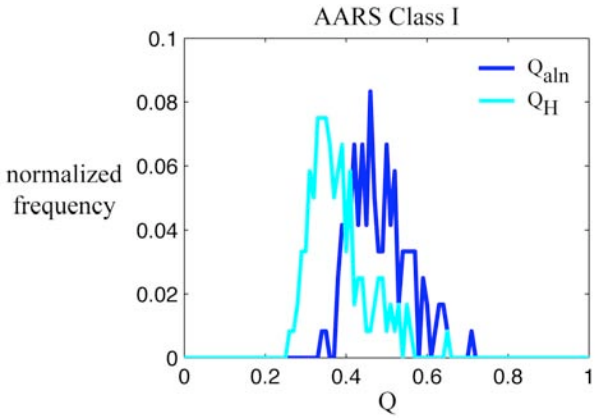
"Gaps should count as a character but not dominate" C. Woese







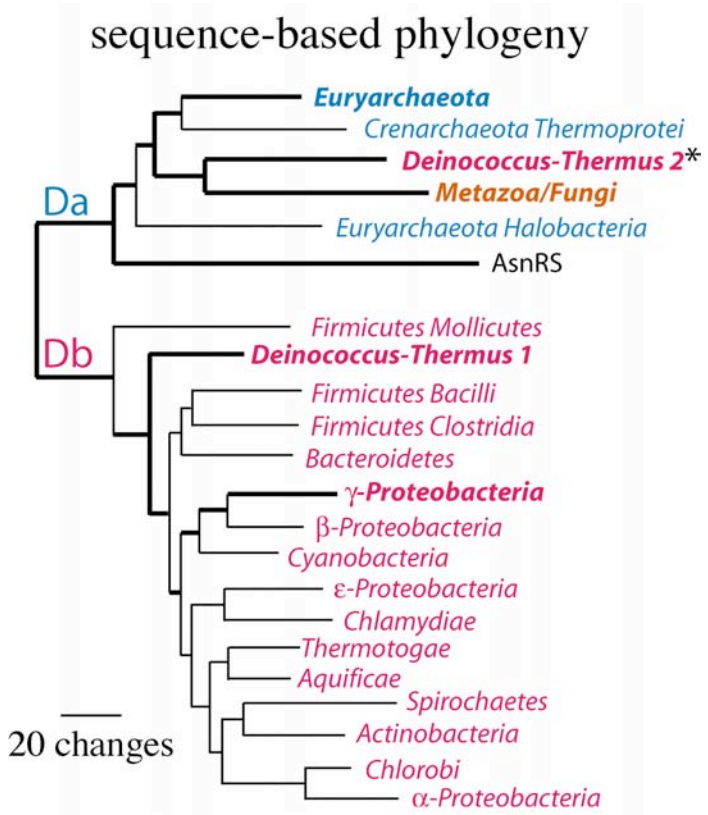
Q_H= 0.82 0.70 0.62

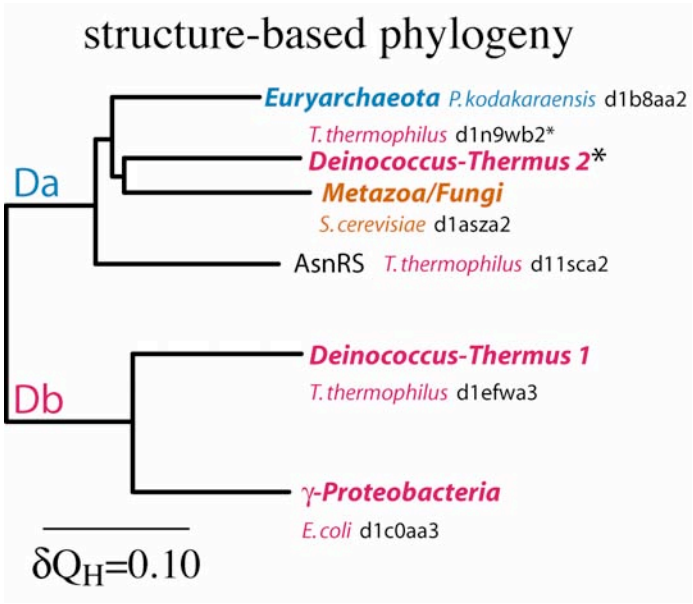


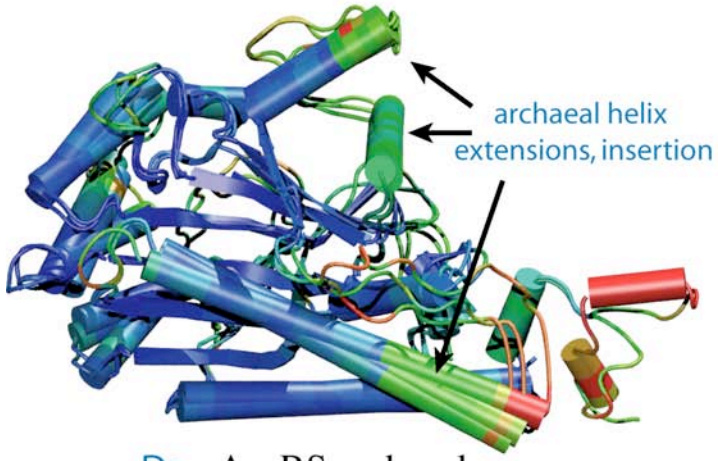
$$q_{gap} = \sum_{g_a}\sum_{j}^{N_{aln}} \max\left\{\exp\left[-\frac{\left(r_{g_aj} - r_{g'_aj'}\right)^2}{2\sigma^2_{g_aj}}\right], \exp\left[-\frac{\left(r_{g_aj} - r_{g''_aj'}\right)^2}{2\sigma^2_{g_aj}}\right]\right\}$$

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

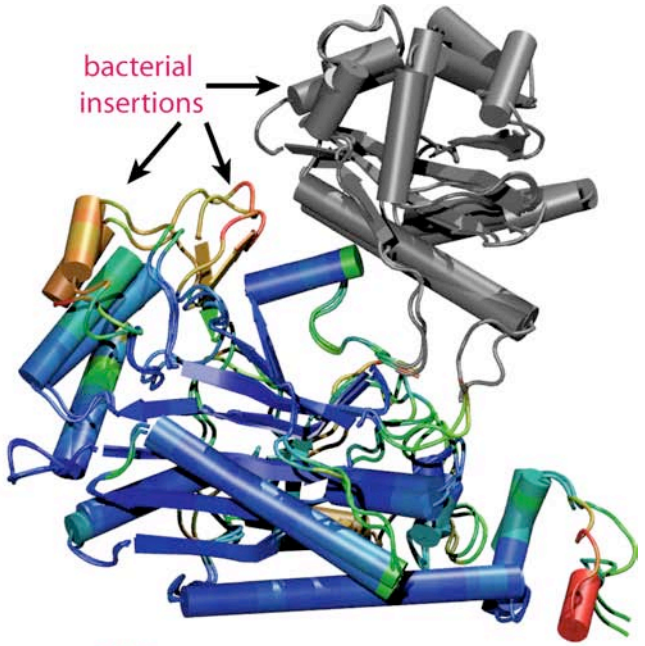
Protein structure encodes evolutionary information







Da - AspRS archaeal genre



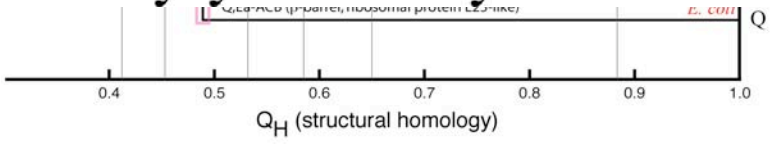
Db - AspRS bacterial genre

JMB 2005
MMBR 2003

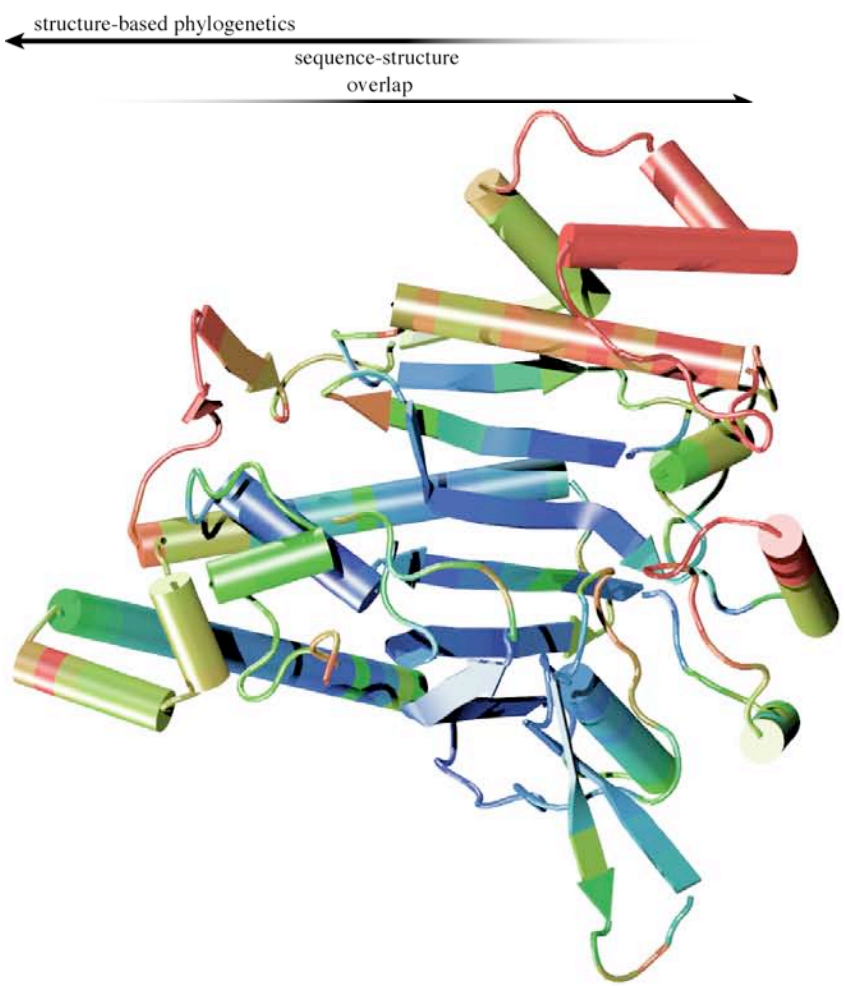
Protein structure reveals distant evolutionary events

Class I AARSs

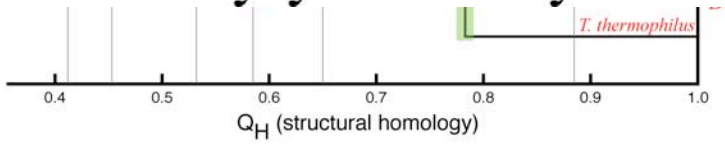
Class I Lysyl-tRNA Synthetase



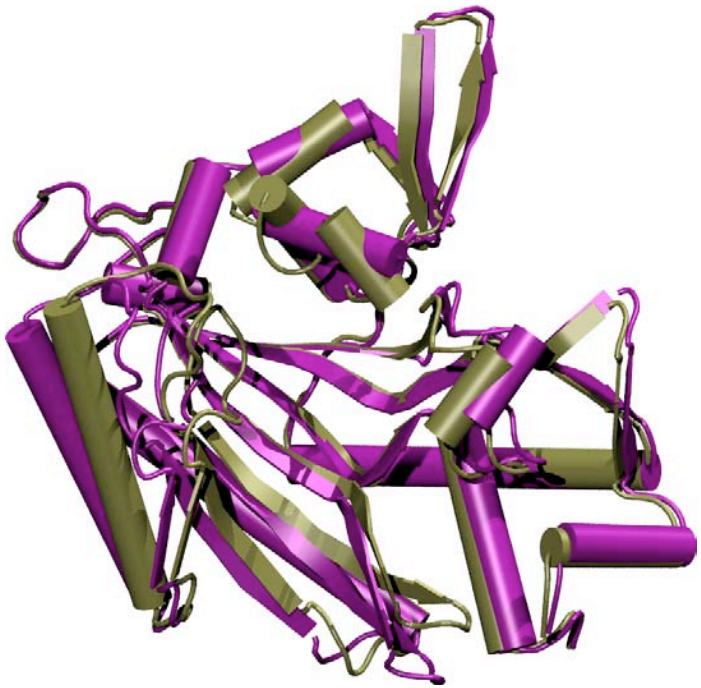
Class II AARSs



Class II Lysyl-tRNA Synthetase



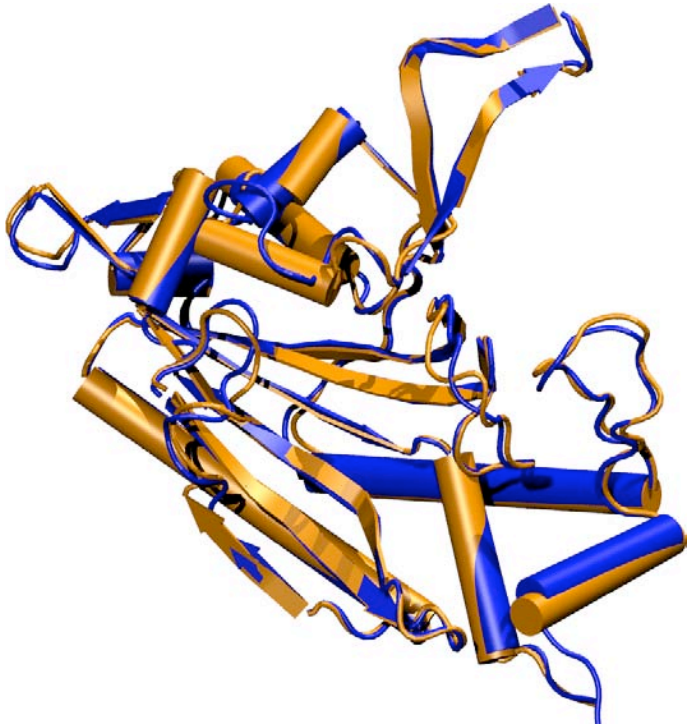
Sequences define more recent evolutionary events



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



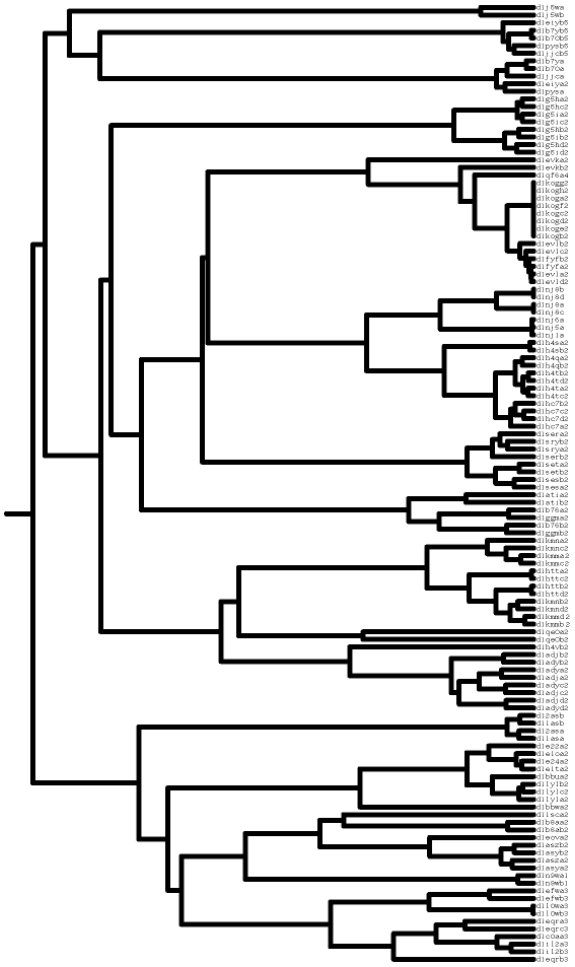
Structures for two
different species.

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

Q_H = 0.89
Sequence identity = 0.69

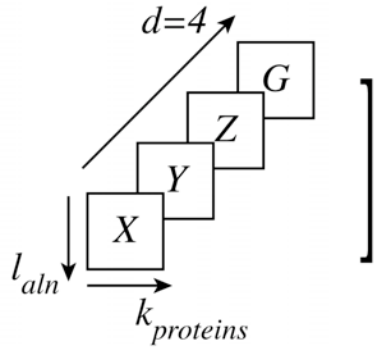
Non-redundant Representative Sets

Too much information
129 Structures

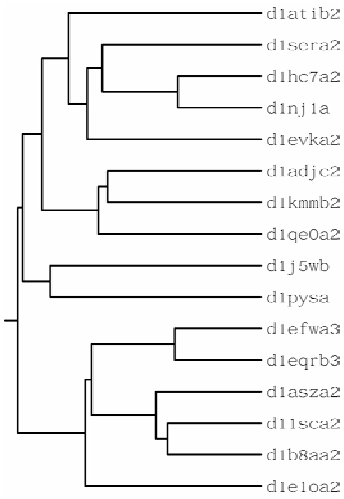


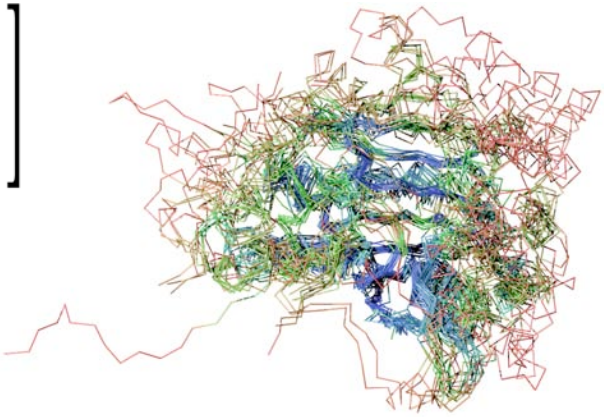
Multidimensional QR factorization of alignment matrix, *A*.

$$A = \left[\quad \right]$$



Economy of information
16 representatives





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 $g = \eta \dfrac{\|X\|_{F_4} + \|Y\|_{F_4} + \|Z\|_{F_4}}{\|G\|_{F_4}}$

adjustable parameter

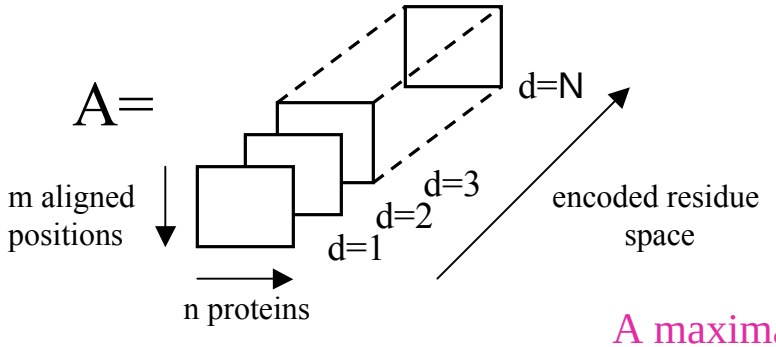
Sequence Space

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)

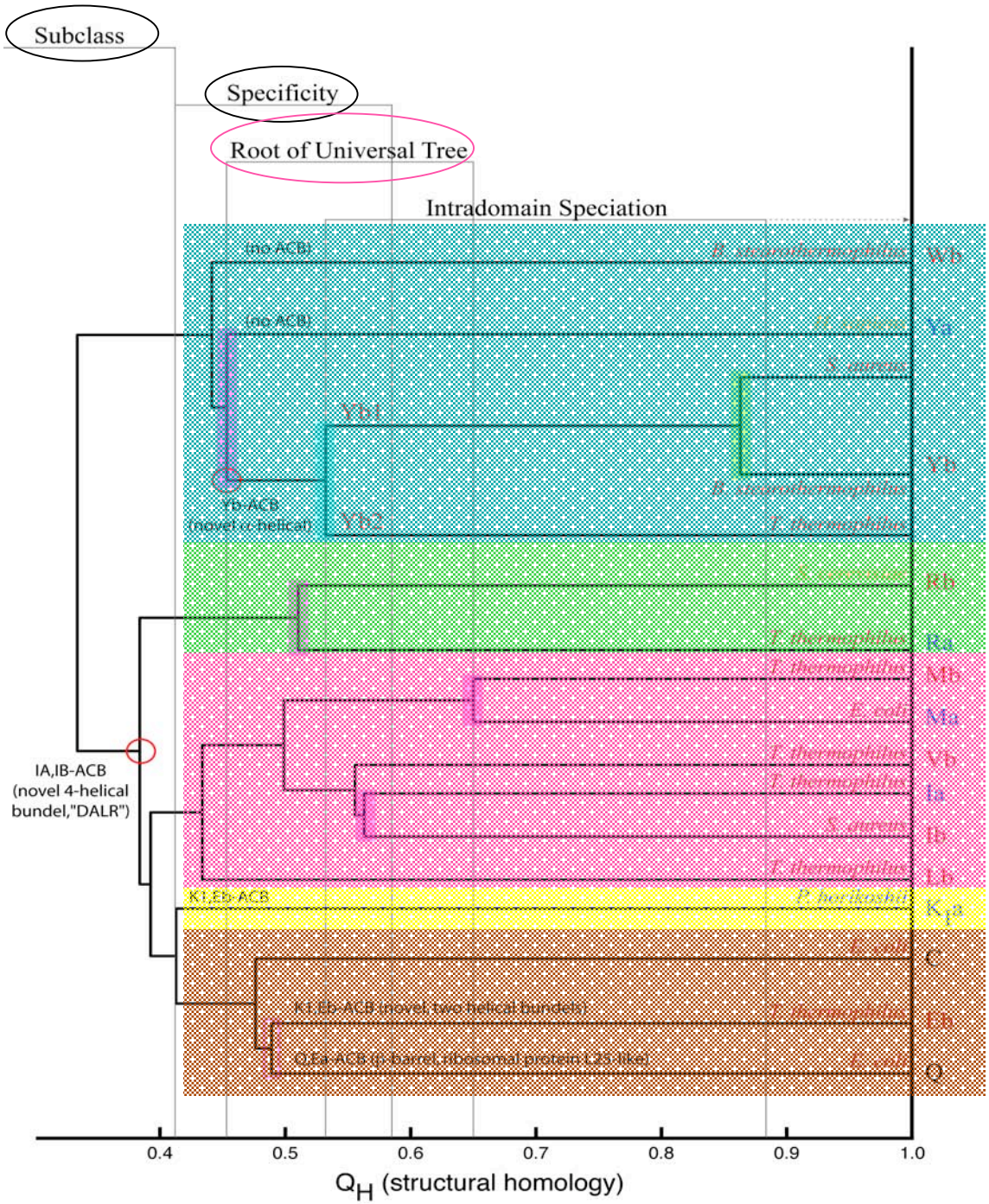
Alignment is a Matrix with Linearly Dependent Columns



$$Q_{(d)}^T A_{(d)} P = Q_{(d)}^T \begin{bmatrix} & & & G \\ & & & Z \\ & & & Y \\ & & & X \\ m_{aln} & & & n_{proteins} \end{bmatrix} P = \tilde{R}_{(d)}$$

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

Class I AARSs evolutionary events

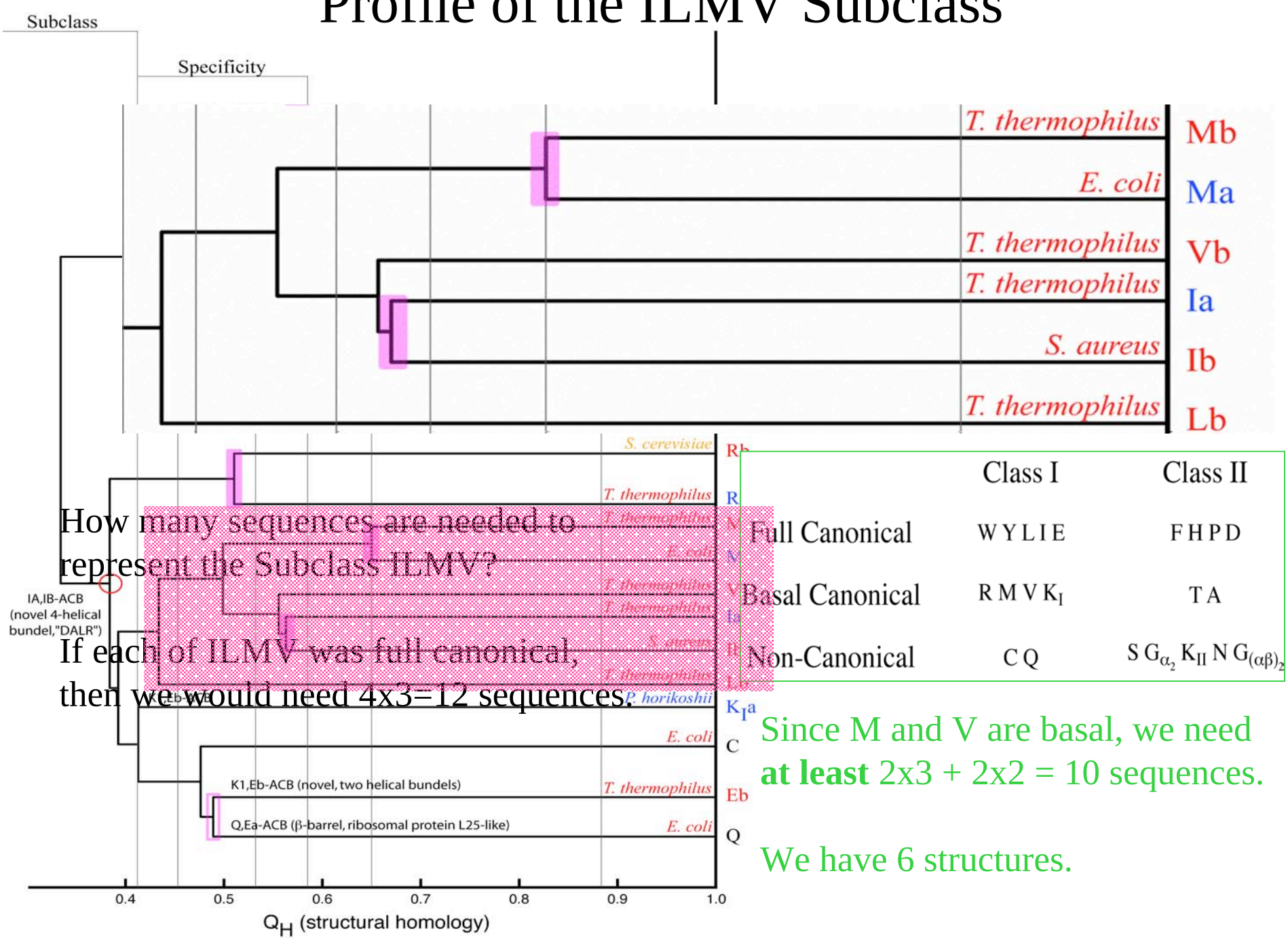


5 Subclasses

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?

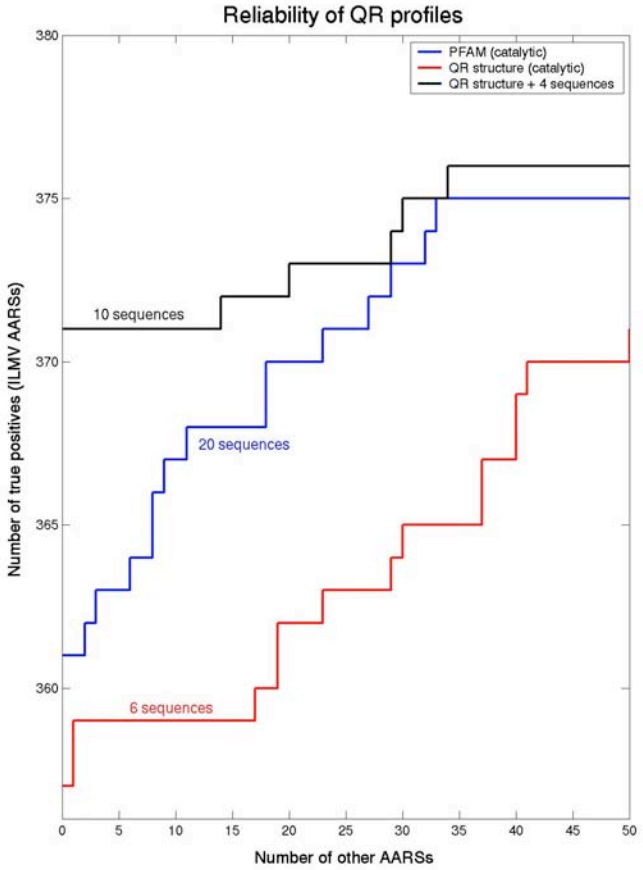
If each of ILMV was full canonical, then we would need 4x3=12 sequences.

	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 $G_{(\alpha\beta)_2}$

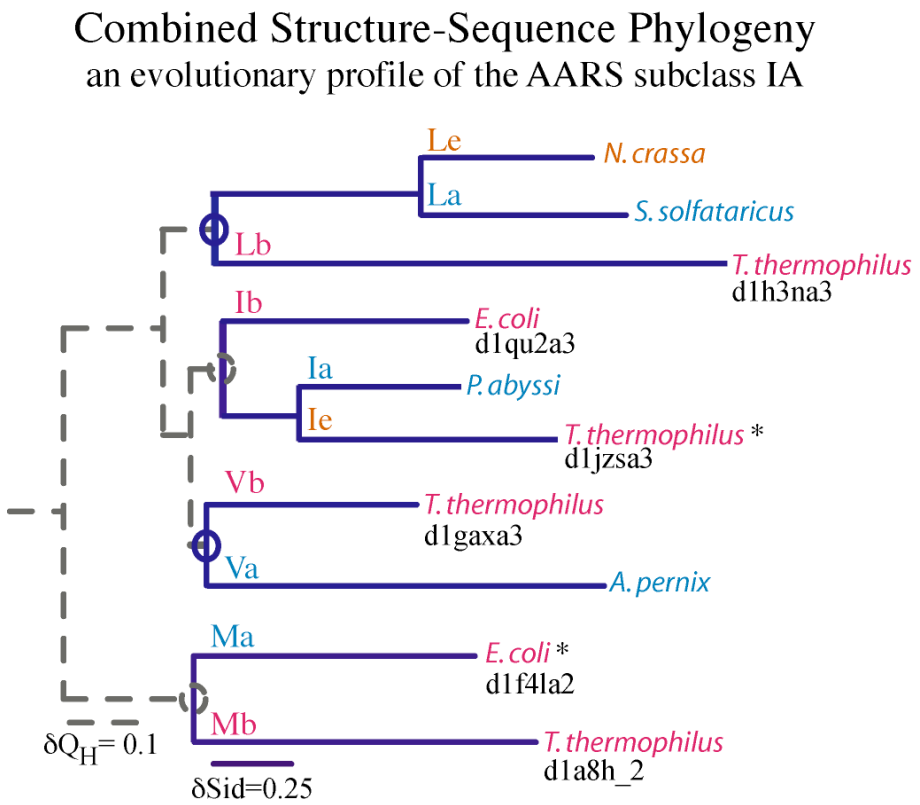
Since M and V are basal, we need **at least** 2x3 + 2x2 = 10 sequences.

We have 6 structures.

Evolutionary Profiles for Homology Recognition
AARS Subclass ILMV



false positives



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 ?

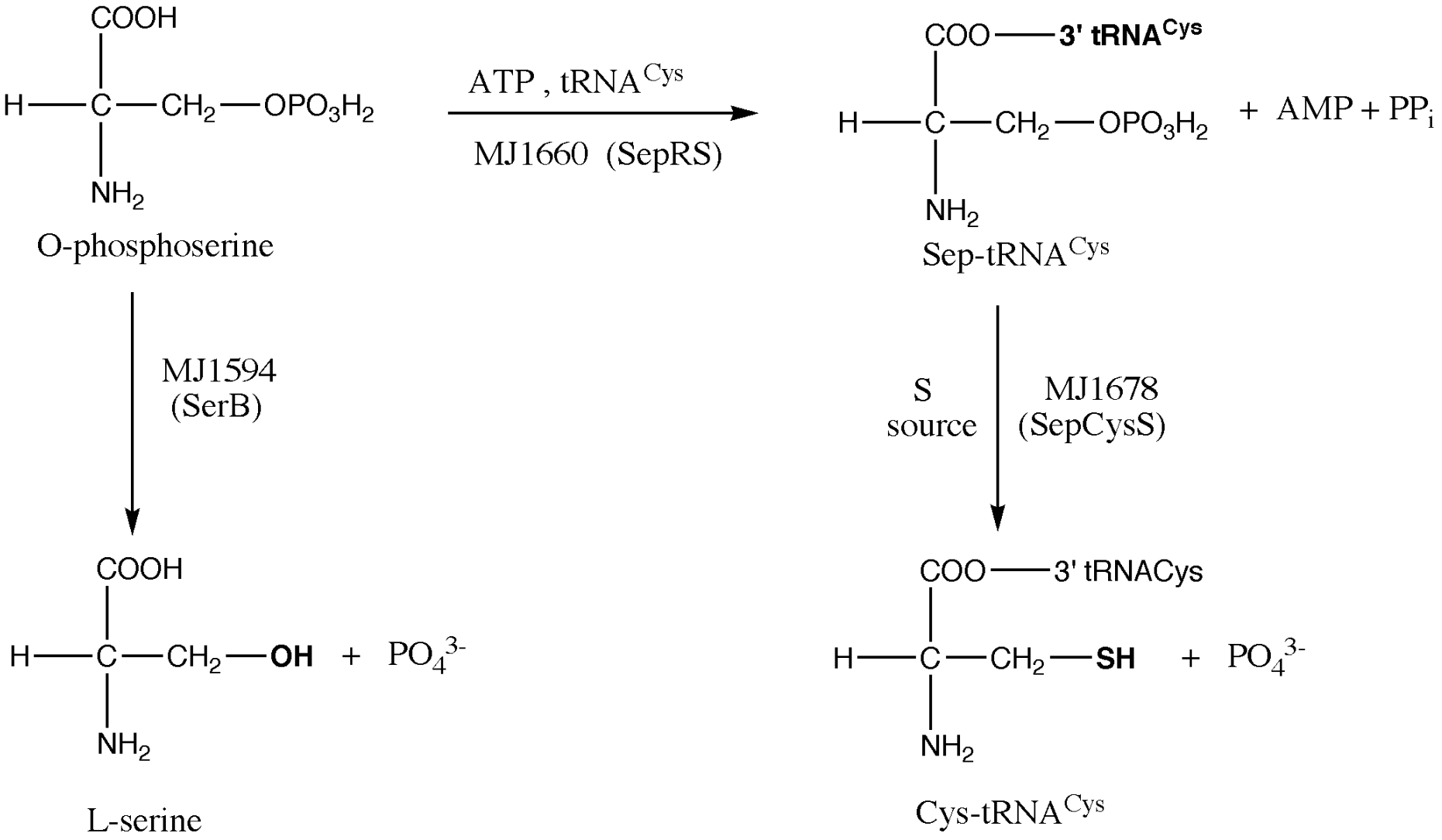
Cysteinyl-tRNA(Cys) formation in *Methanocaldococcus jannaschii*: the mechanism is still unknown. *J. Bacteriology*, Jan. 2004, **186:**8-14.
Ruan B, Nakano H, Tanaka M, Mills JA, DeVito JA, Min B, Low KB, Battista JR, and Söll D.

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	← MJ1660
AlaRS	5.1e-02	
GlyRS	0.12	
PheRS β-chain	0.15	
DNA repair protein	7.5	

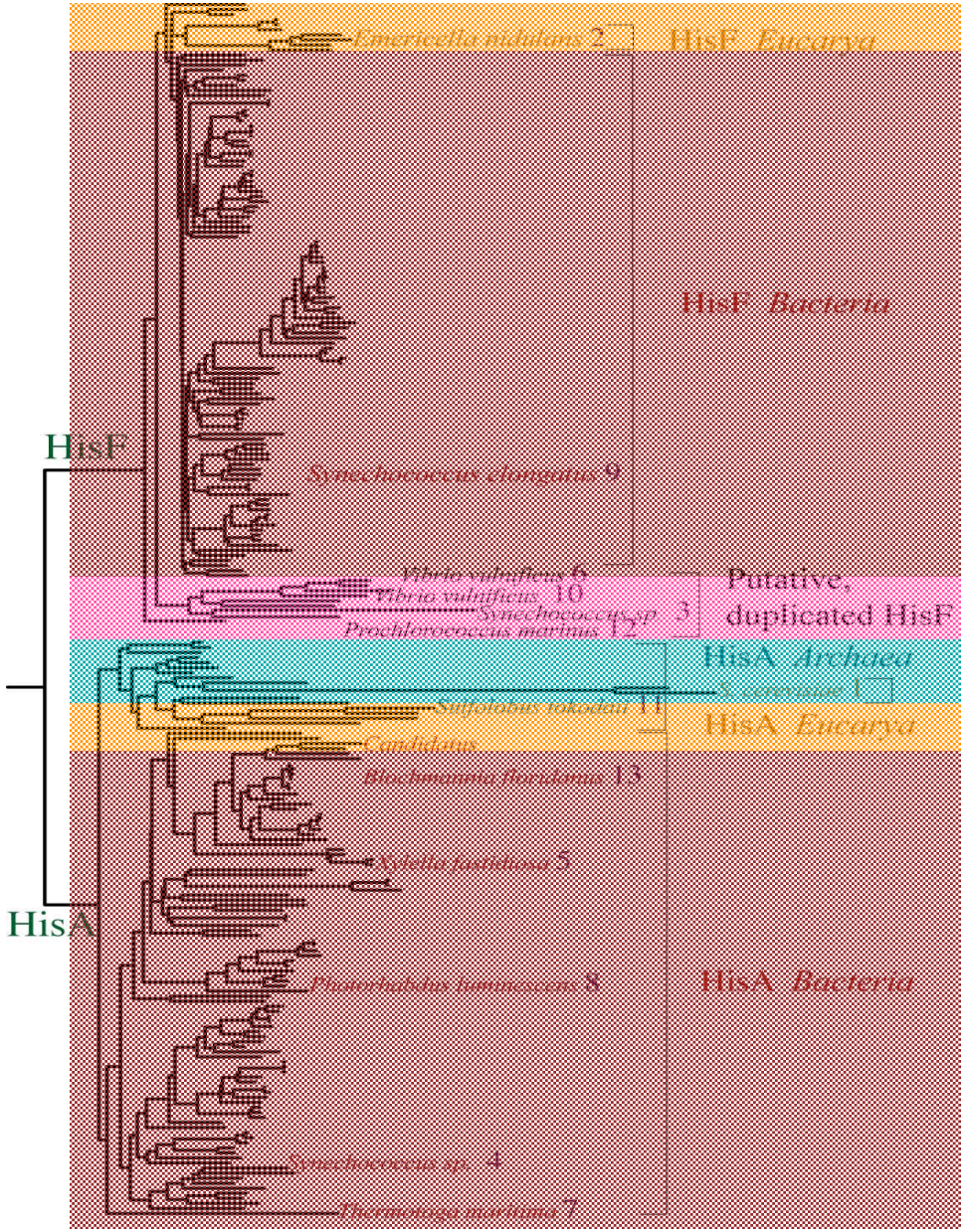
Sethi, et. al., PNAS, **102** ,2005

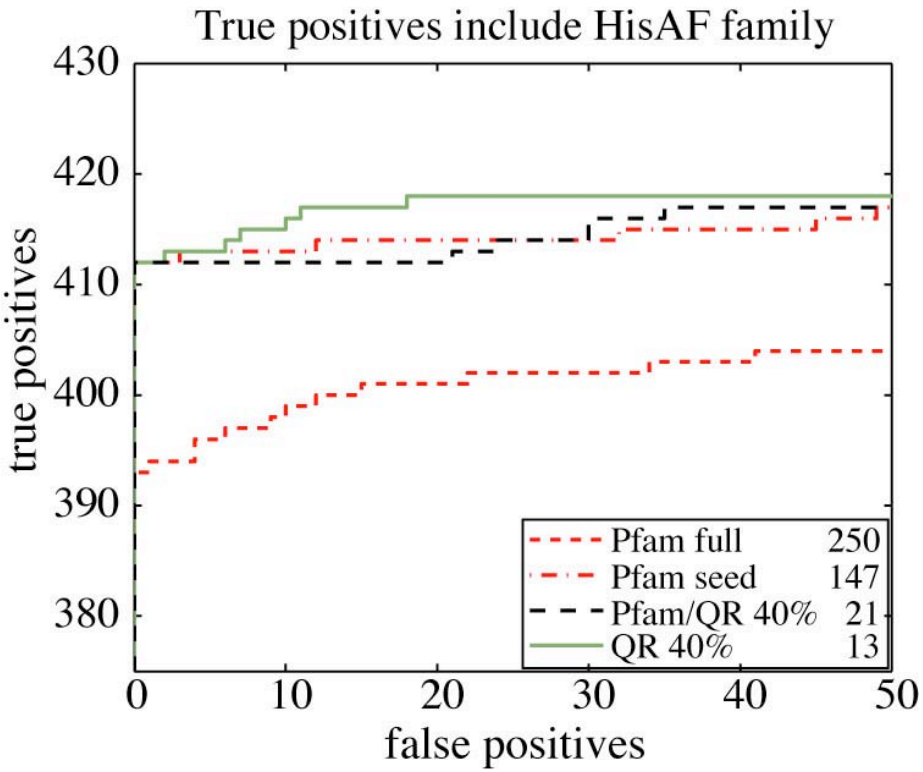
Cysteine Biosynthesis in *Methanocaldococcus jannaschii*



Sauerwald et al. Science 2005

Evolutionary profile for HisA-HisF family





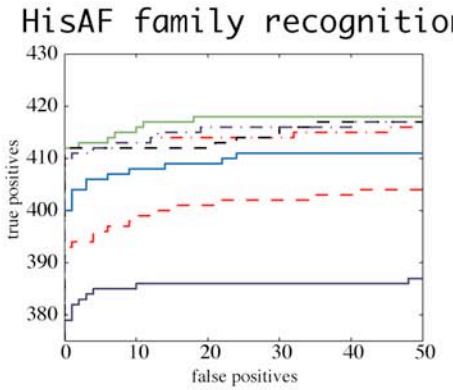
EP outperforms popular profile methods with an economy of information.

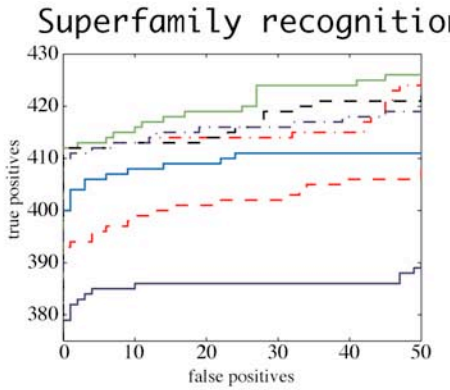
Sethi, et. al., PNAS, 2005.

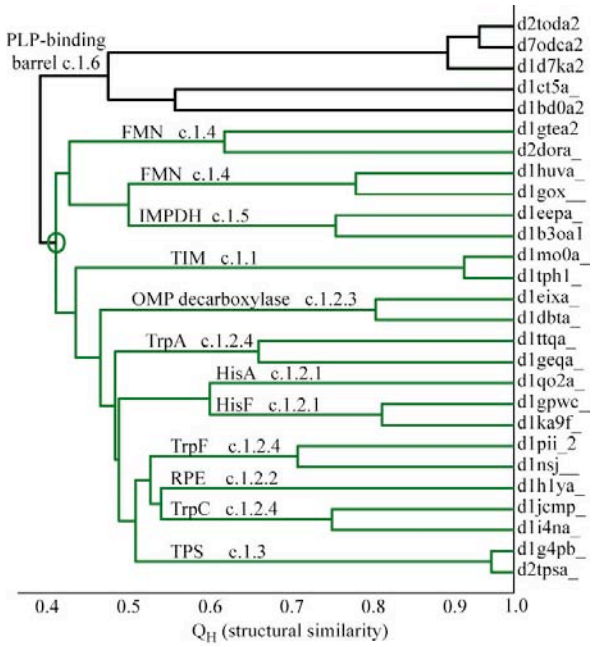
Economy of Information

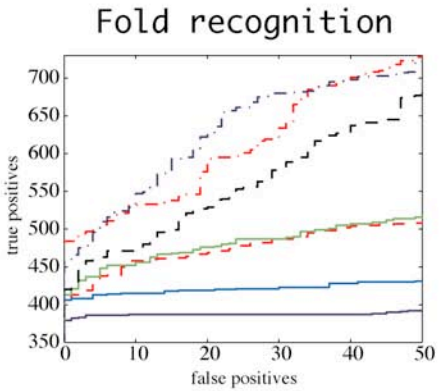
How many sequences are needed for profiles?

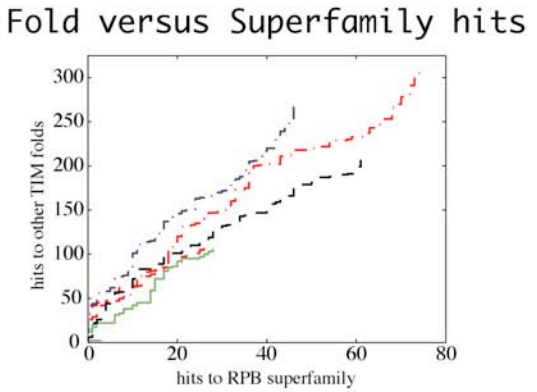
Profile	N_{seq}
Pfam full	250
Pfam seed	147
Pfam/QR 40%	21
QR 15%	2
QR 30%	4
QR 40%	13
QR 100%	238





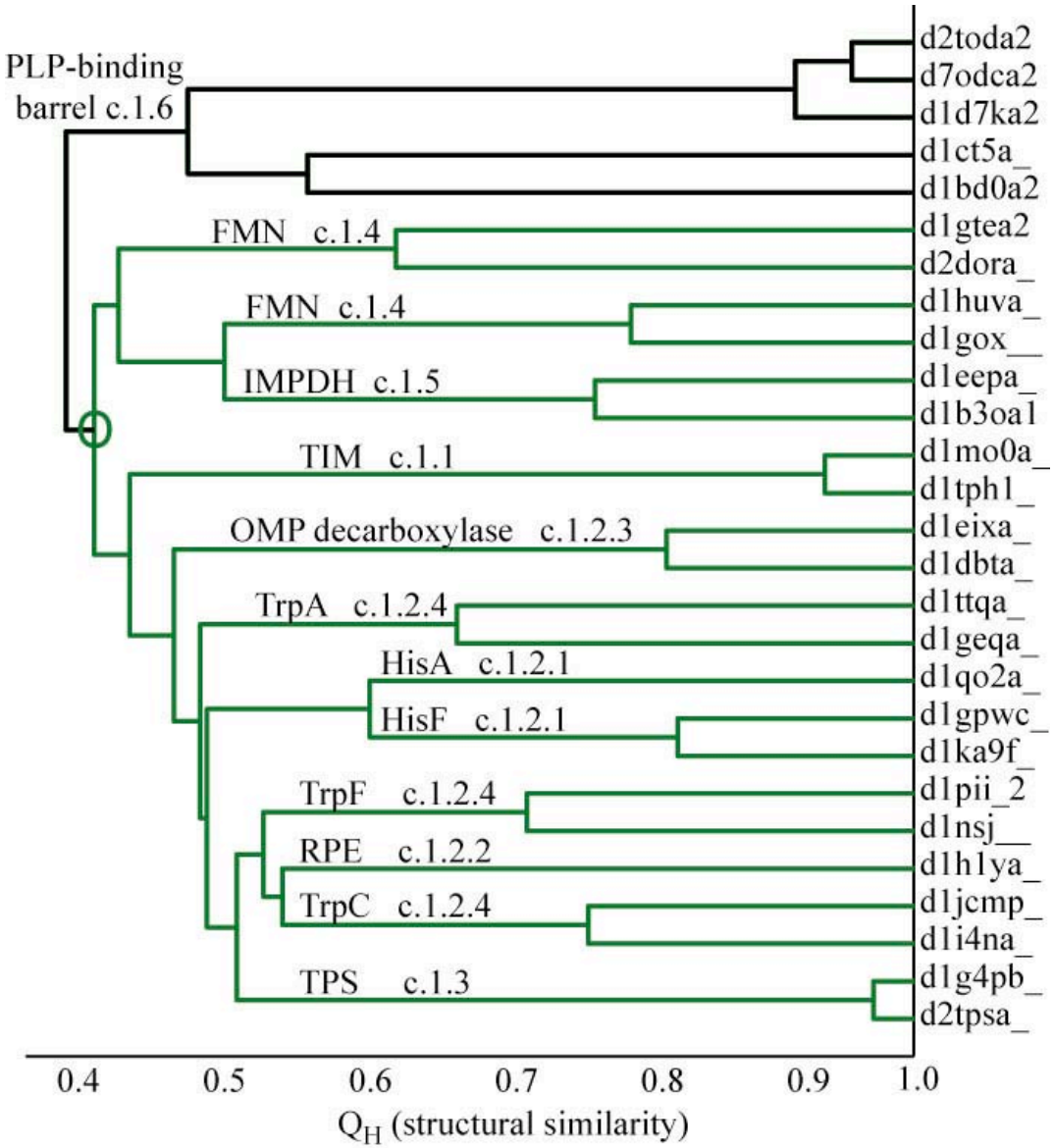


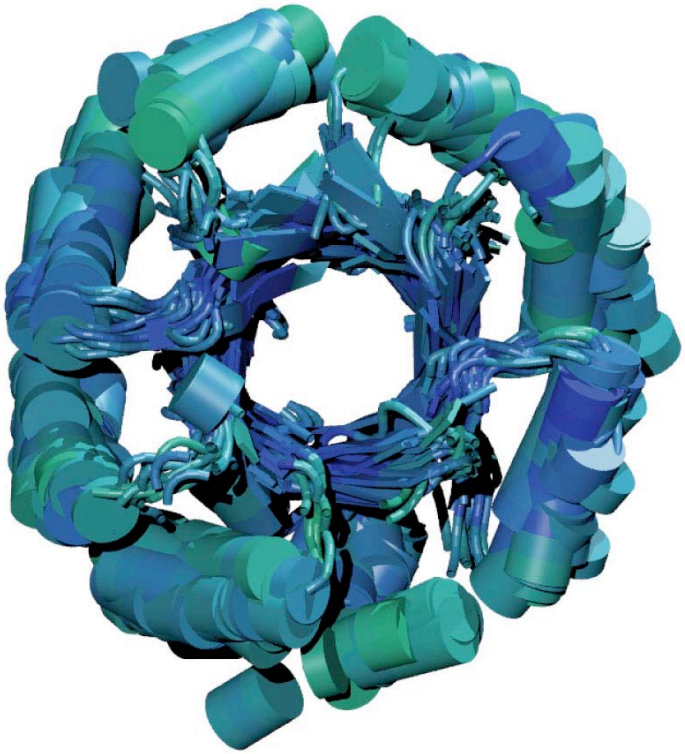




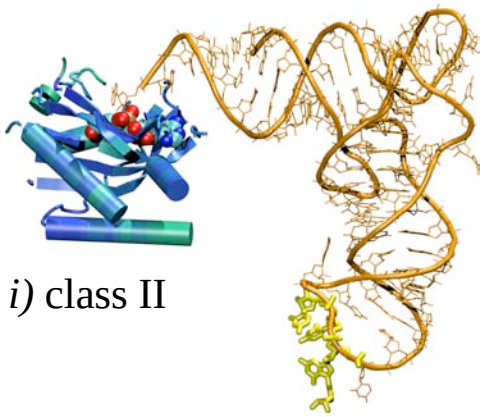
A. Sethi, P. O'Donoghue, ZLS, PNAS **102**, 2005

Phylogenetic relationship between TIM barrels Found in database search with HisA-HisF profile

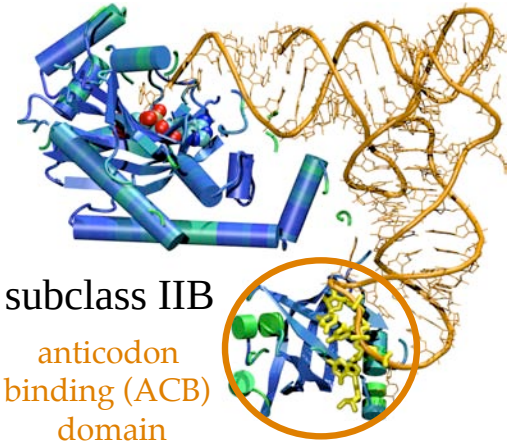




Evolution of Structure and Function in AspRS

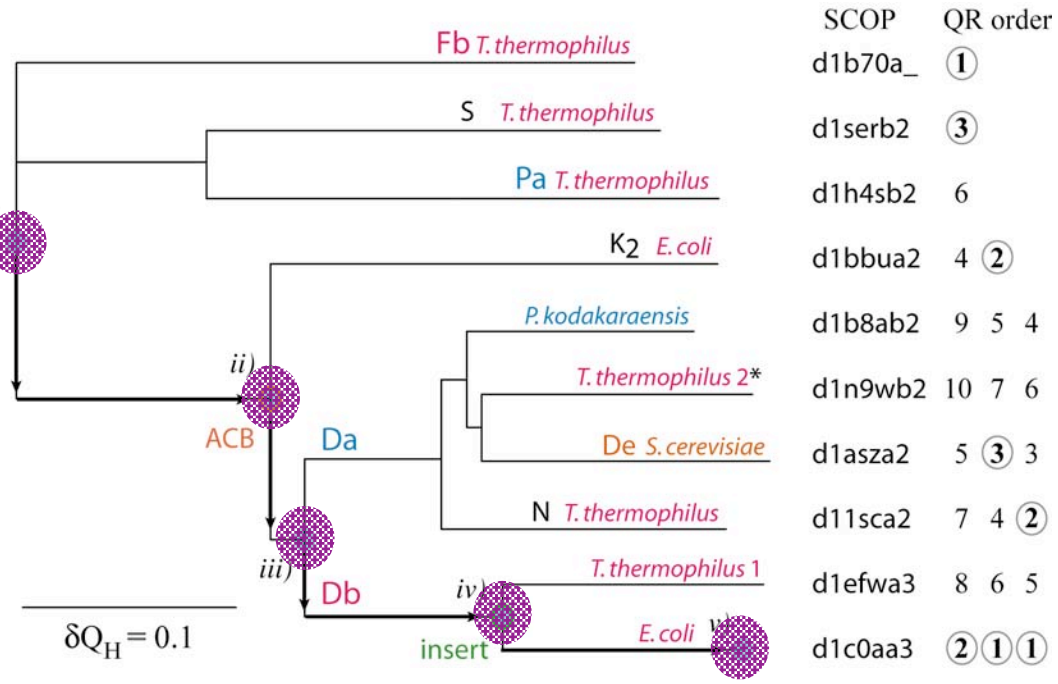


i) class II

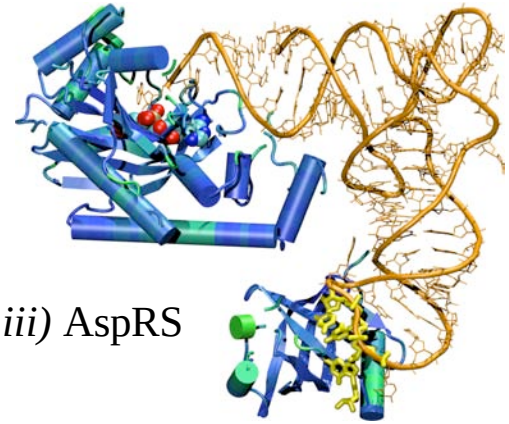


ii) subclass IIB

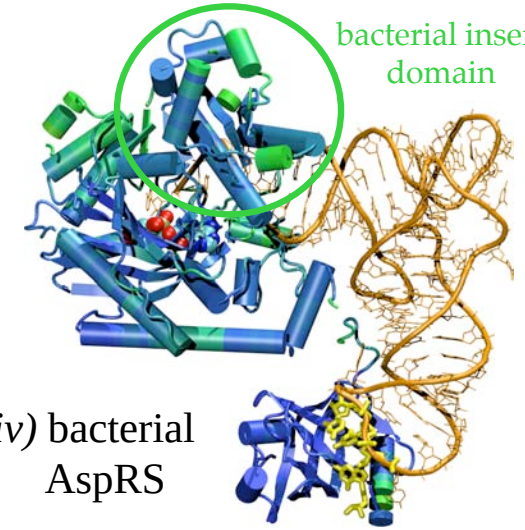
anticodon binding (ACB) domain



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)

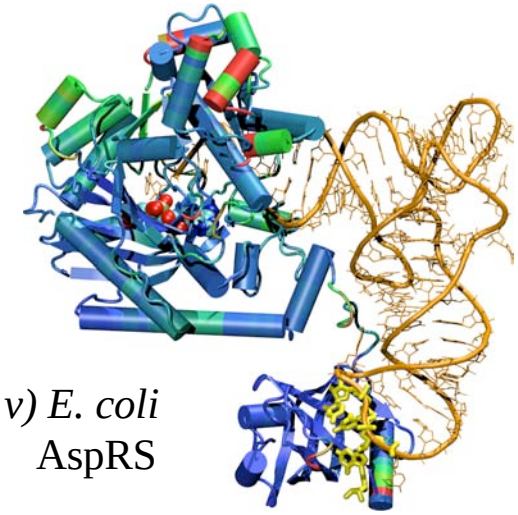


iii) AspRS



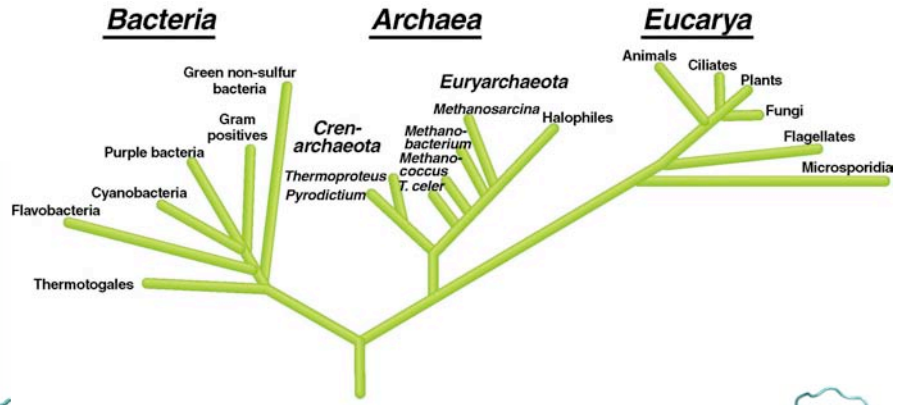
bacterial insert domain

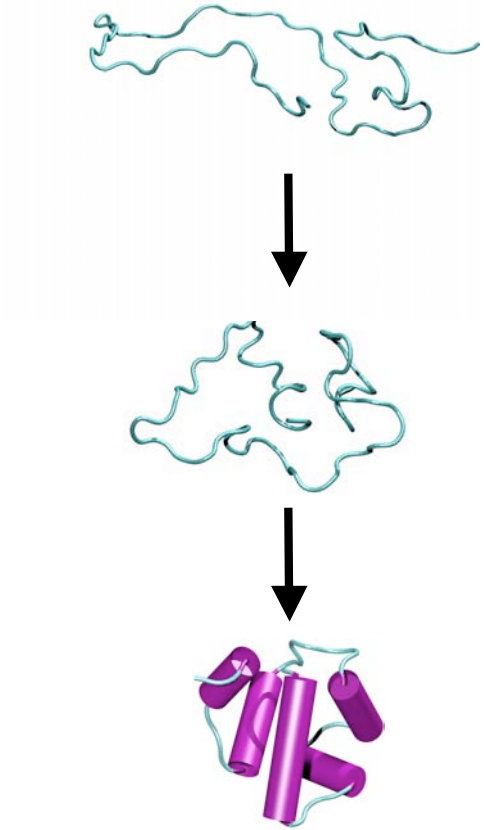
iv) bacterial AspRS

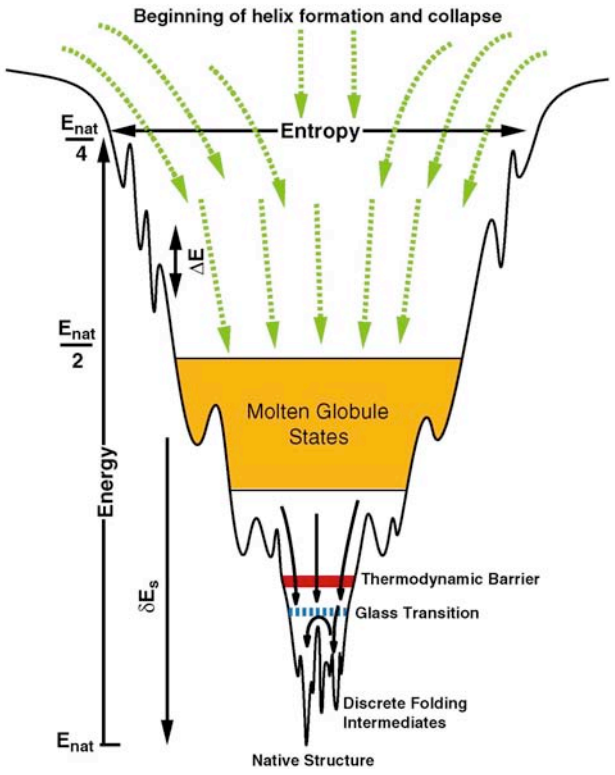


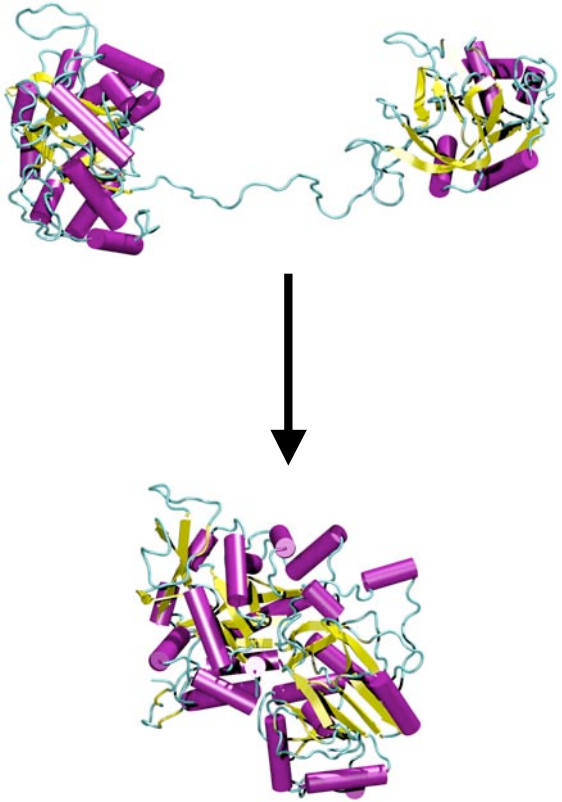
v) E. coli AspRS

Unifying the Worlds of Sequence and Structure



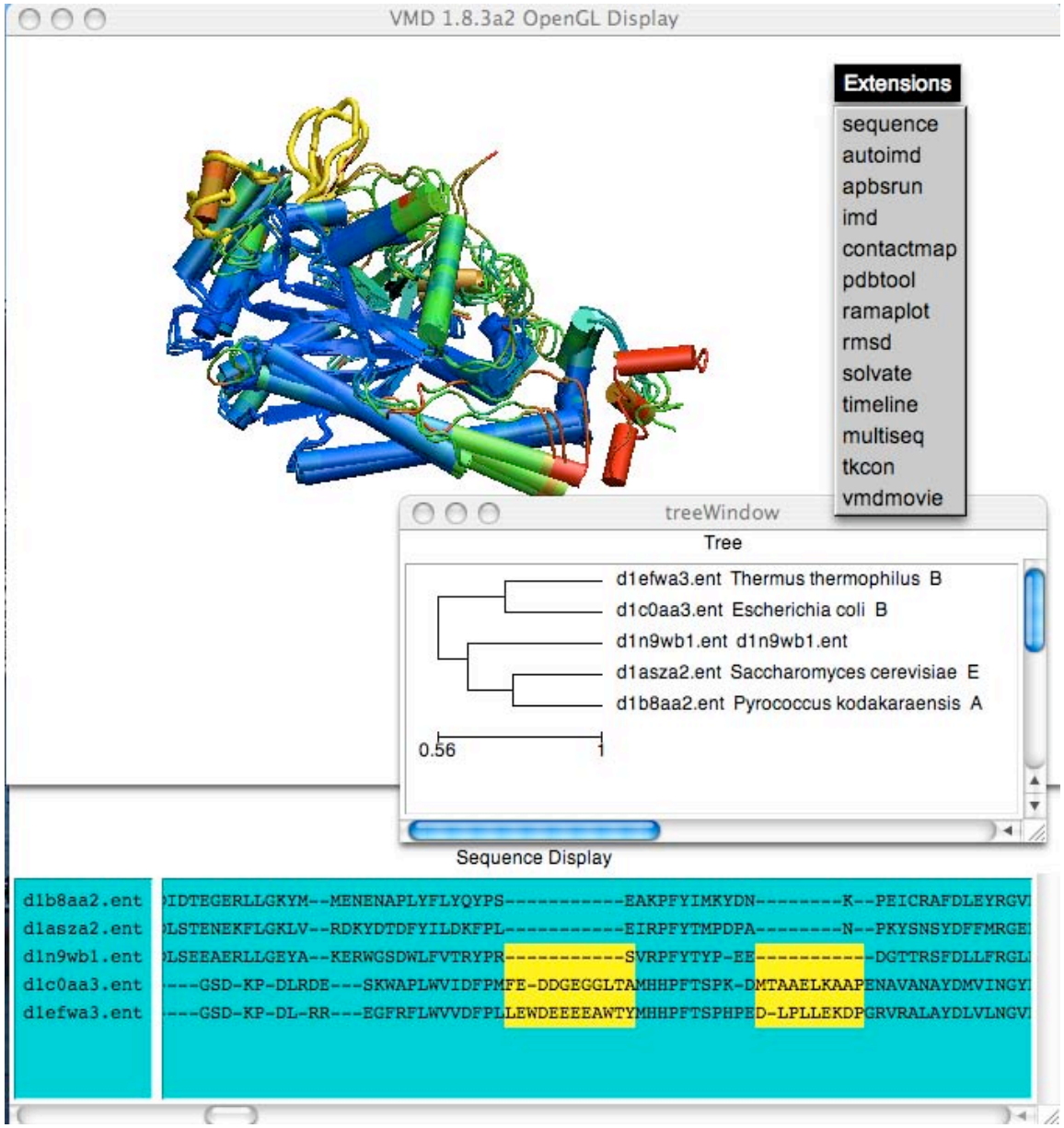






Chicago 2005

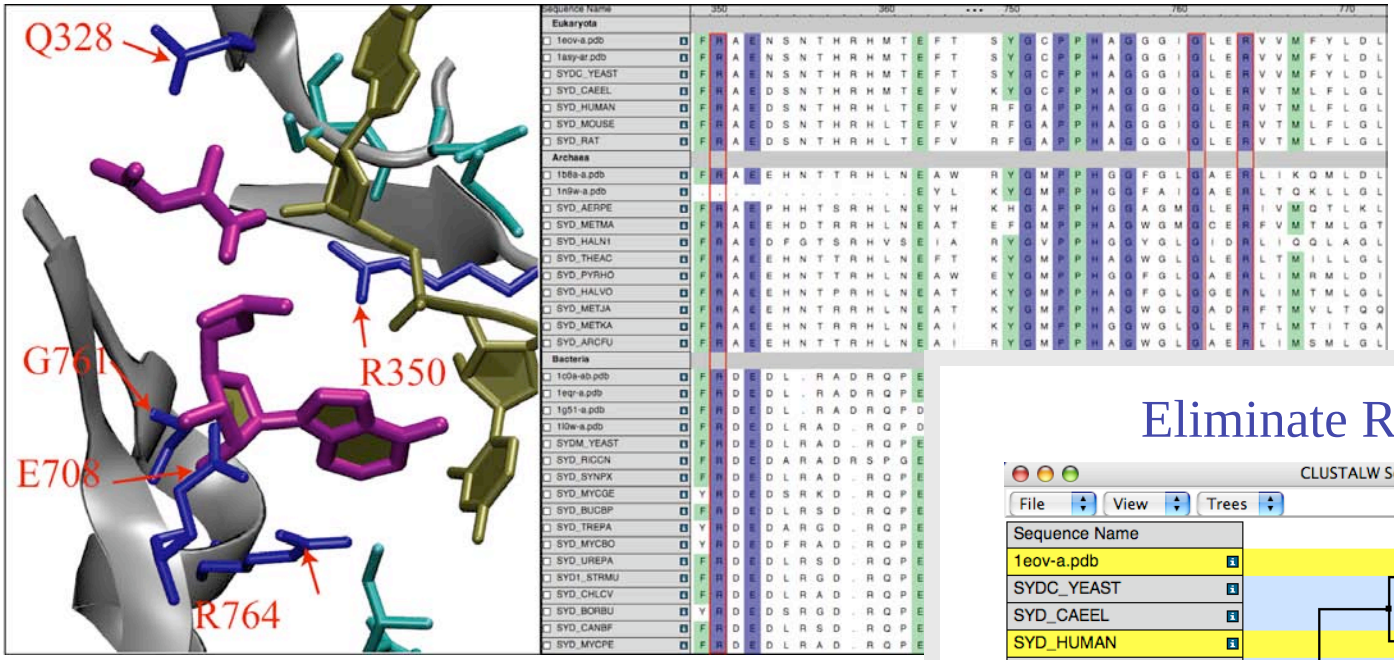
Multiseq in VMD : Merging the sequence and structure worlds



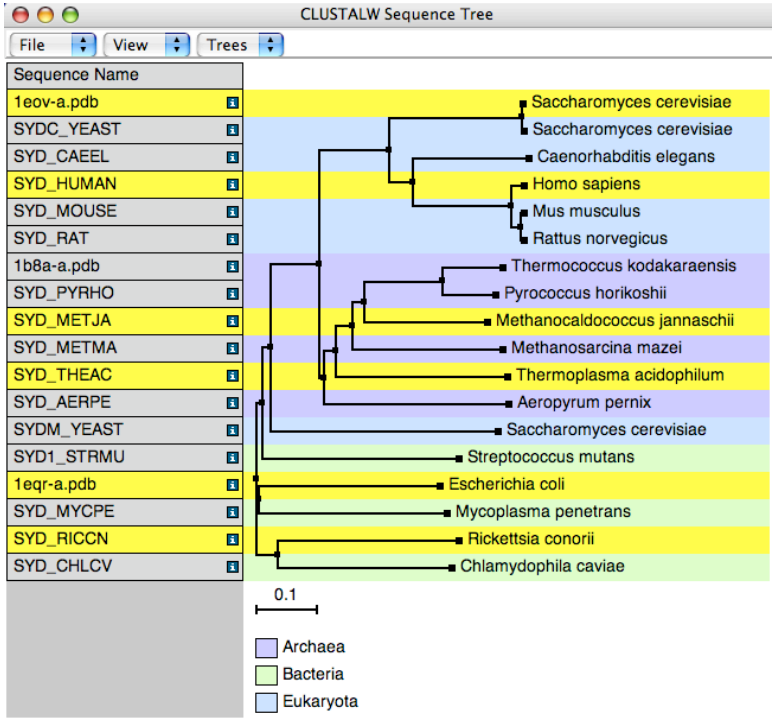
Version 1.83

2006 MultiSeq: New Features

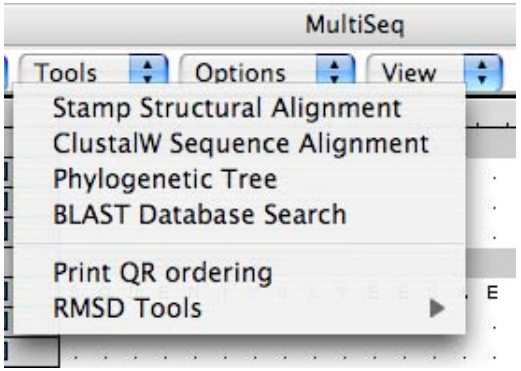
Analyze the Evolution of Sequence and Structure



Eliminate Redundancy



Plus More Functions



List of New Features in Multiseq

1. INPUT: Sequences and structures of proteins and nucleic acids from file or Blast searches of specialized databases:

   Structural (PDB, SCOP, ASTRAL, NDB, VIPER..)
   Sequence (NCBI, ASTRAL, modified tRNA, Viral)
   Sequence Editor and Electronic Notebook

2. TOOLS:
   Alignments (STAMP, CLUSTAL, TCoffee)
   Database Searches - BLAST and VMD/Multiple DB searches
   QR reduction, Phylogenetic tree - UPGMA, NJ
   Conservation Mappings, RMSD plots
   Covariance and Coordination Analysis

Acknowledgements

Patrick O'Donoghue

Anurag Sethi

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

Graphics Programmers VMD

Elijah Roberts, Dan Wright, John Eargle

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)

Funding

NSF, NIH

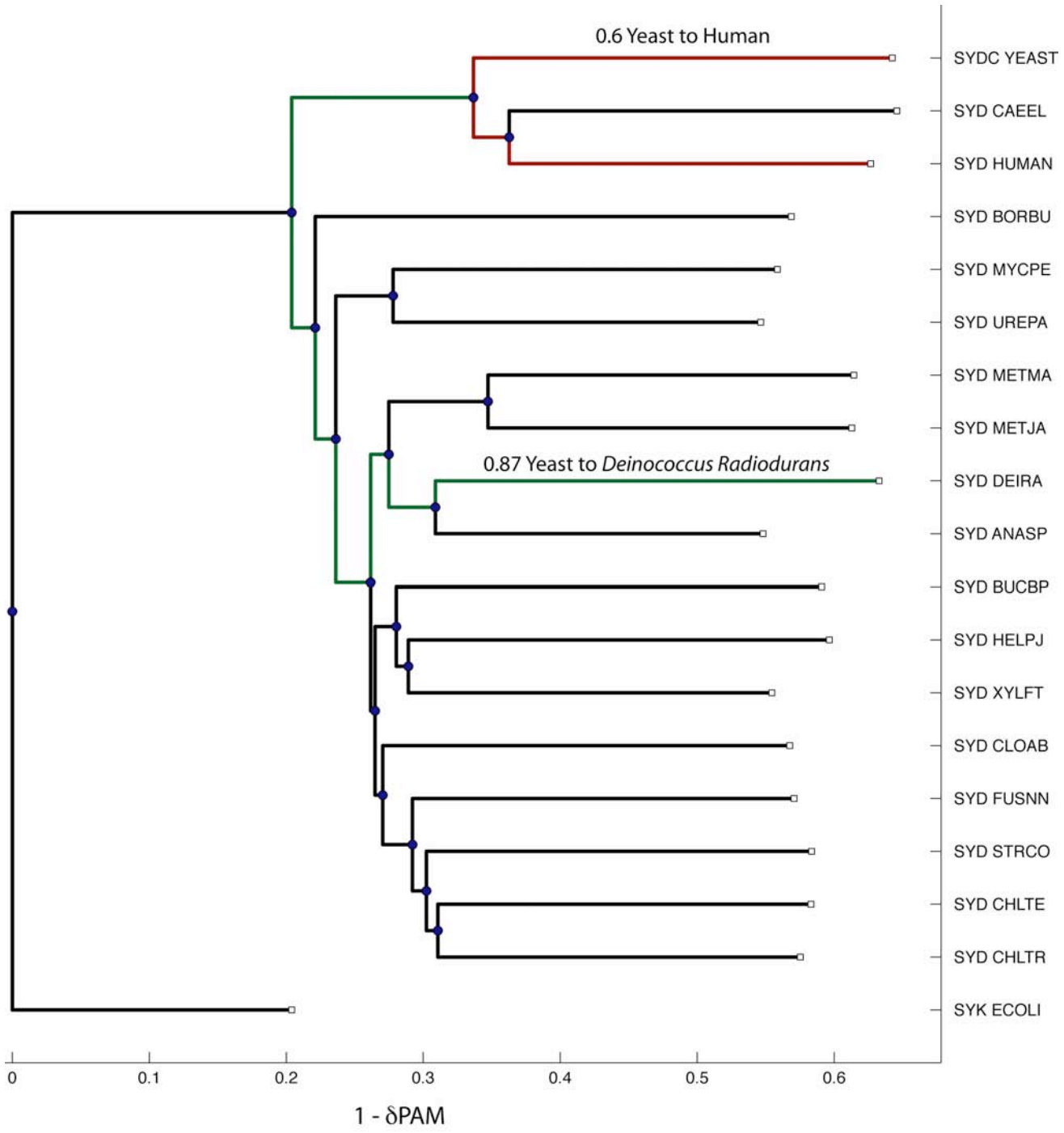
Demonstration of New Multiseq Features

1. AspRS structures: STAMP multiple structure alignment. Color by structure (Qpair) and and sequence conservation. Tcl script - seq ID and Sec. Str. Information in beta field.

2. Sequence Editor and Electronic Notebook

3. AspRS Sequences (from BLAST database search): Automated grouping by domains of life. Sequence conservation by domain of life. Mapping of sequence and structure information onto structures. CLUSTAL alignment to structural profile.

4. Phylogenetic trees of structure and sequences: HGT and QR algorithm for sequences. Evolutionary profiles


0.6 Yeast to Human
0.87 Yeast to *Deinococcus Radiodurans*
SYDC YEAST
SYD CAEEL
SYD HUMAN
SYD BORBU
SYD MYCPE
SYD UREPA
SYD METMA
SYD METJA
SYD DEIRA
SYD ANASP
SYD BUCBP
SYD HELPJ
SYD XYLFT
SYD CLOAB
SYD FUSNN
SYD STRCO
SYD CHLTE
SYD CHLTR
SYK ECOLI
0
0.1
0.2
0.3
0.4
0.5
0.6
1 - δPAM

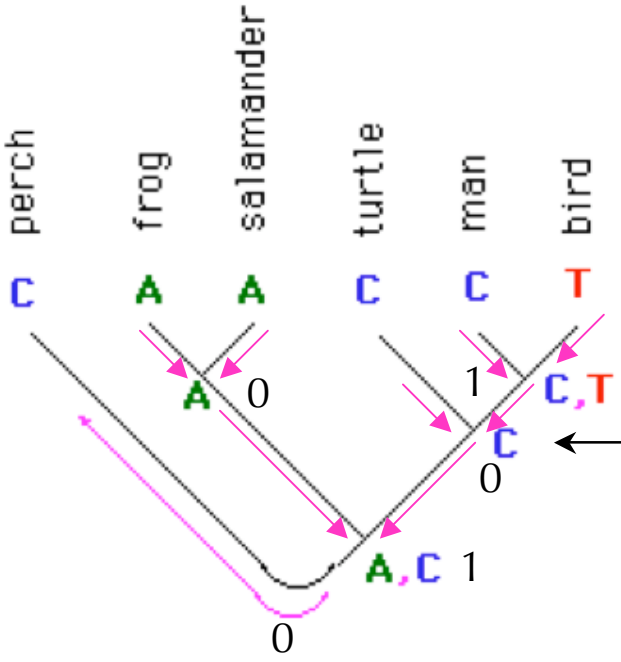
1. Show distance matrix for NJ/UPGMA for small number 3-4 sequences.Give algebraic equations needed for NJ.

2. MP/ML trees: Animate through several tree topologies generated by paup to describe the search through tree space.

Maximum Parsimony
Fitch optimization

Assign characters to the ancestral nodes and calculate the number of steps (sequence changes) required by a data set on a given tree.

"Downpass" algorithm traces back through the tree from leaves to root.



If decendent characters intersect
add 0 to total length.

If descendent characters do not intersect,
their union set is assigned to the
node add 1 to total length.

The intersection of C and (C,T) is C.
This ancestral node is assigned the "state" C.
The total length is unchanged.

The length on this tree for this site is 2.
The length of this toplogy for the sequences in the alignment is the sum of length over all sites gives.

Adapted from Mark E. Siddall, http://research.amnh.org/~siddall/methods/day3.html