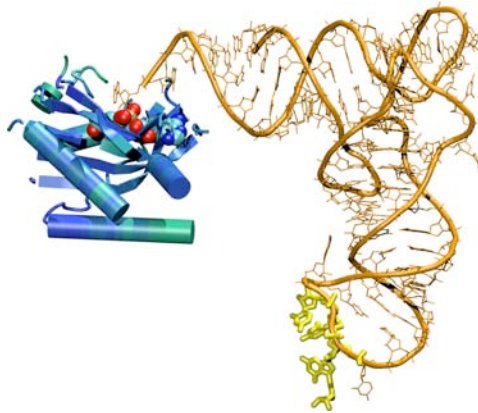


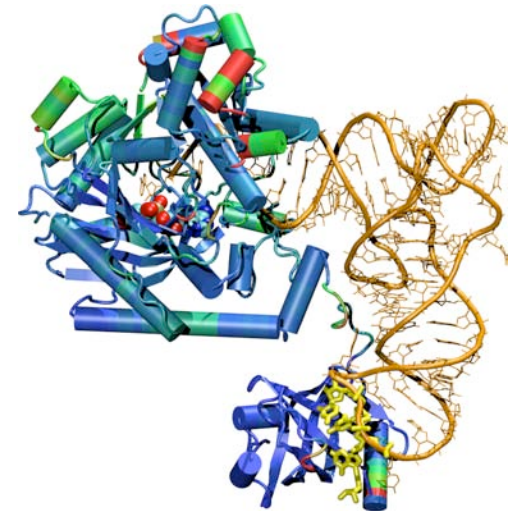
# MULTISEQ in VMD -

Revealing How Nature Designs Proteins and RNAs



		Second position						
		U	C	A	G			
U	UUU	Phe	UCU	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	CAU	His	CGU	Arg	U
	CUC		CCC	CAC		CGC		C
	CUA	Pro	CCA	CAA	Gln	CGA	G	A
	CUG		CCG	CAG		CGG		G
A	AUU	Ile	ACU	AAU	Asn	AGU	Ser	U
	AUC		ACC	AAC		AGC		C
	AUA	Thr	ACA	AAA	Lys	AGA	Arg	A
	AUG		Met/start	ACG		AAG		AGG
G	GUU	Val	GCU	GAU	Asp	GGU	Gly	U
	GUC		GCC	GAC		GGC		C
	GUA	Ala	GCA	GAA	Glu	GGA	A	A
	GUG		GCG	GAG		GGG		G

Copyright © 2004 Pearson Education, Inc., publishing as Benjamin Cummings.

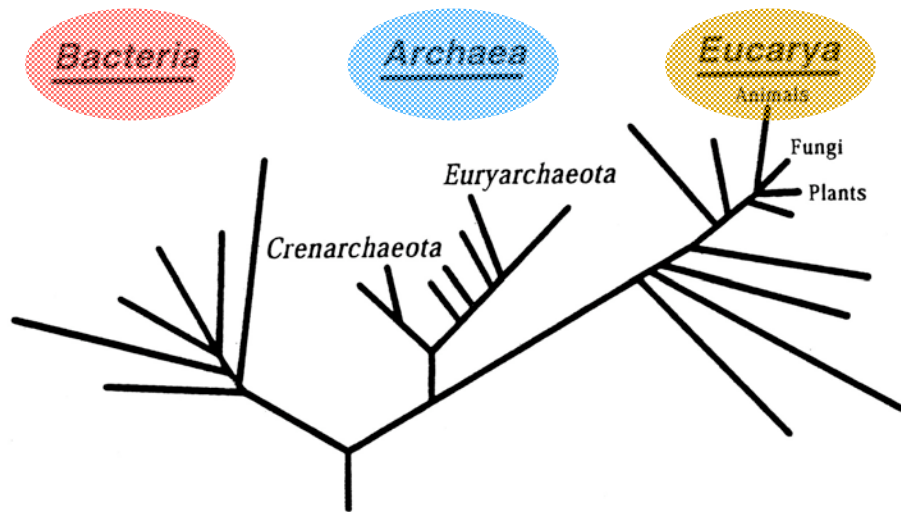


Luthey-Schulten Group

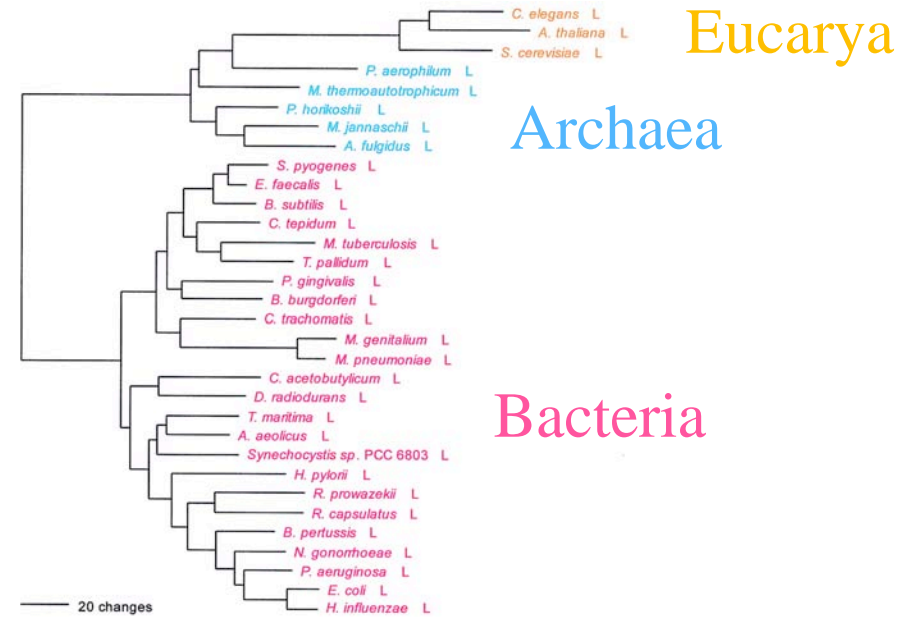
Department of Chemistry, Biophysics, and Beckman Institute  
University of Illinois at Urbana-Champaign

# Universal Phylogenetic Tree

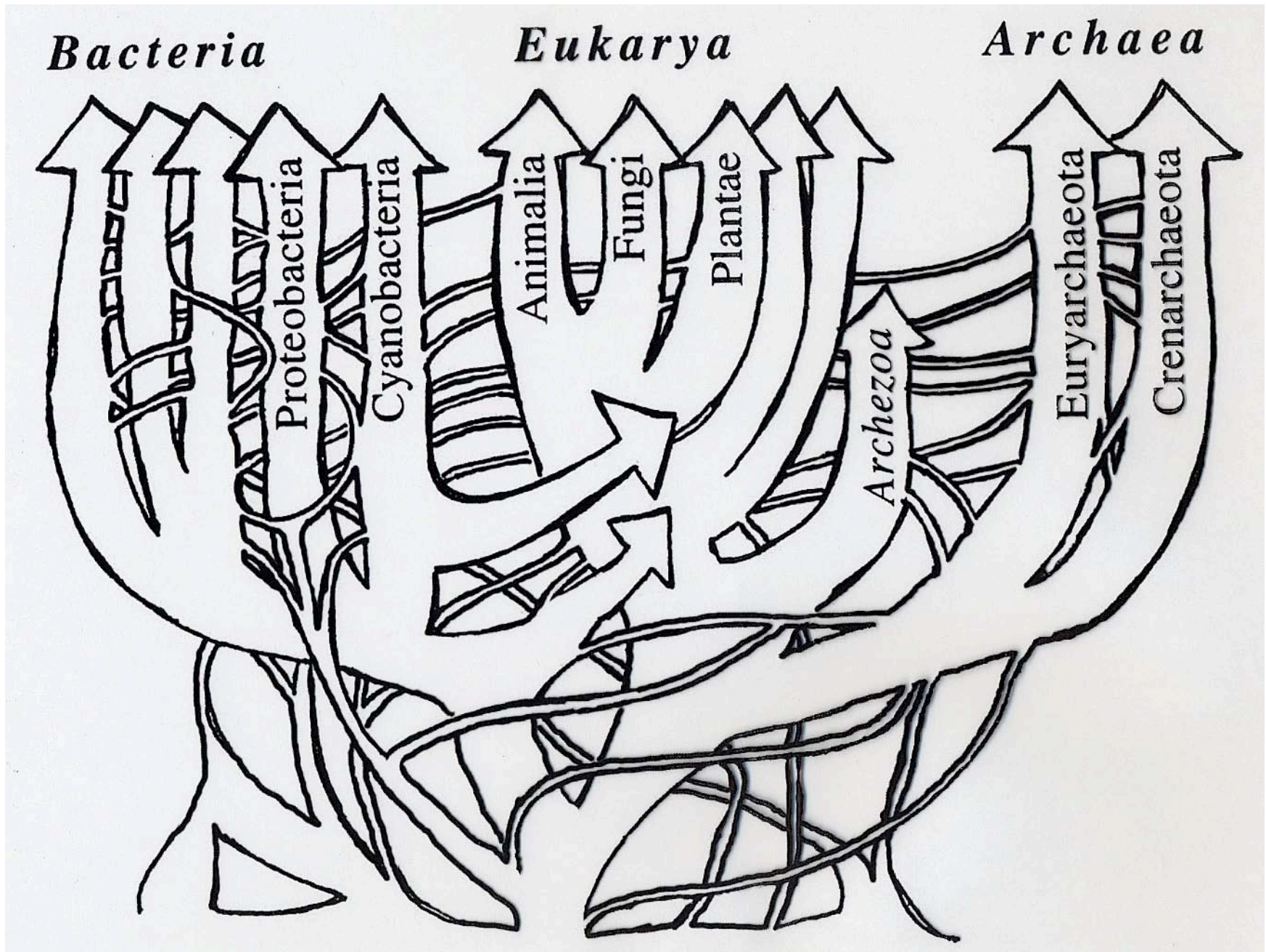
## three domains of life



Based on 16S rRNA



Leucyl-tRNA synthetase displays the full canonical phylogenetic distribution.



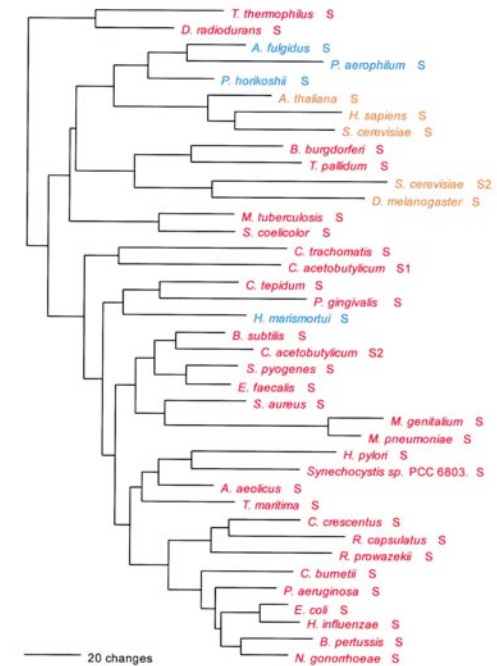
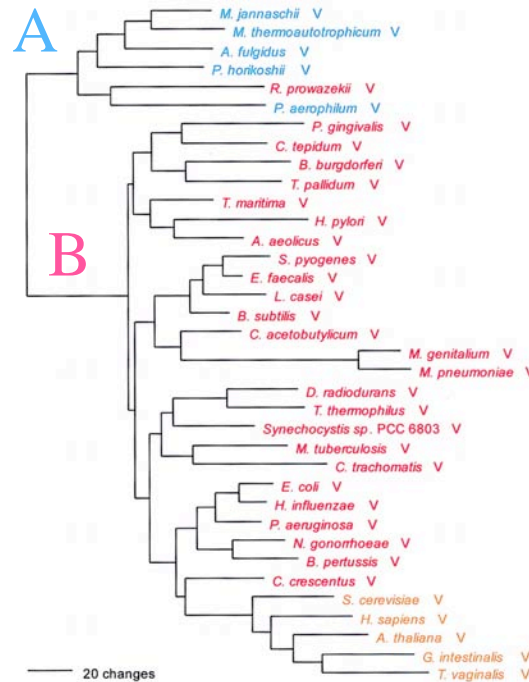
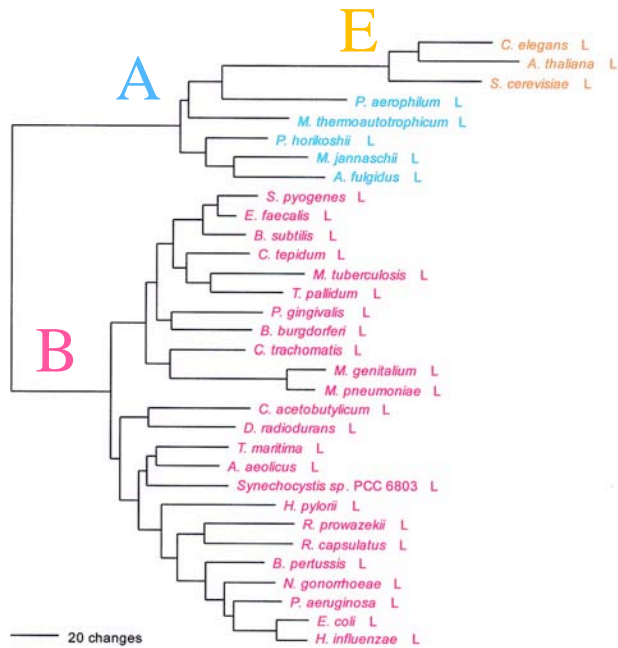
After W. Doolittle, modified by G. Olsen

# 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

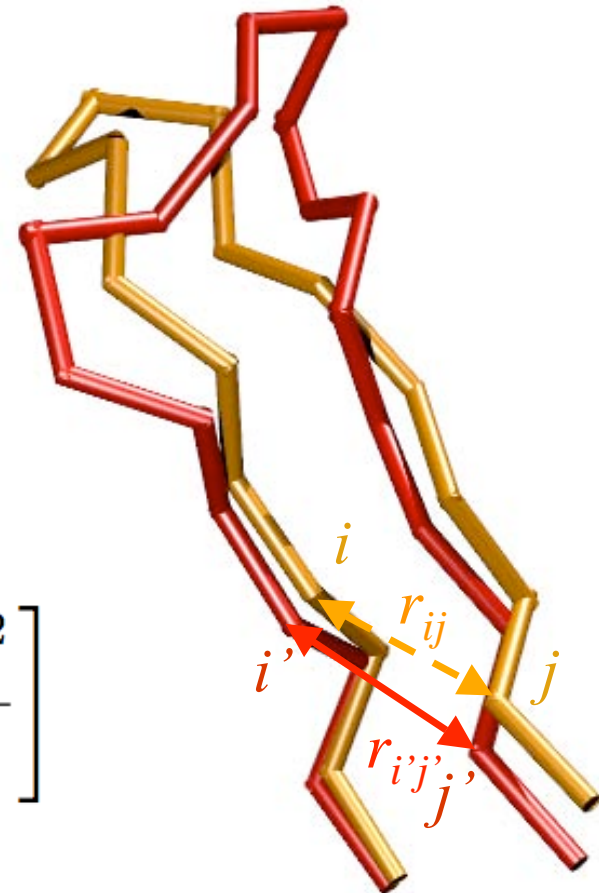
# Protein Structure Similarity Measure

## $Q_H$ Structural Homology

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

$$Q_H = \mathcal{N} [q_{aln} + q_{gap}]$$

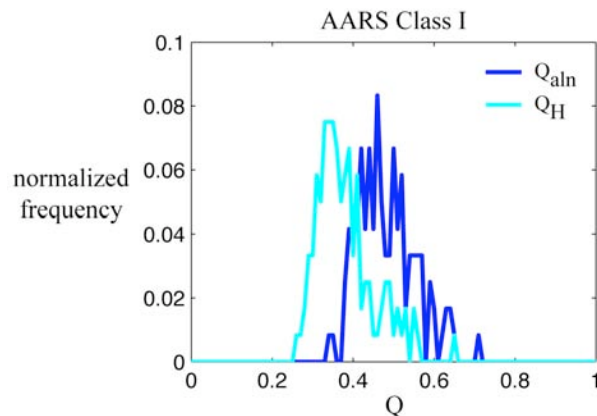
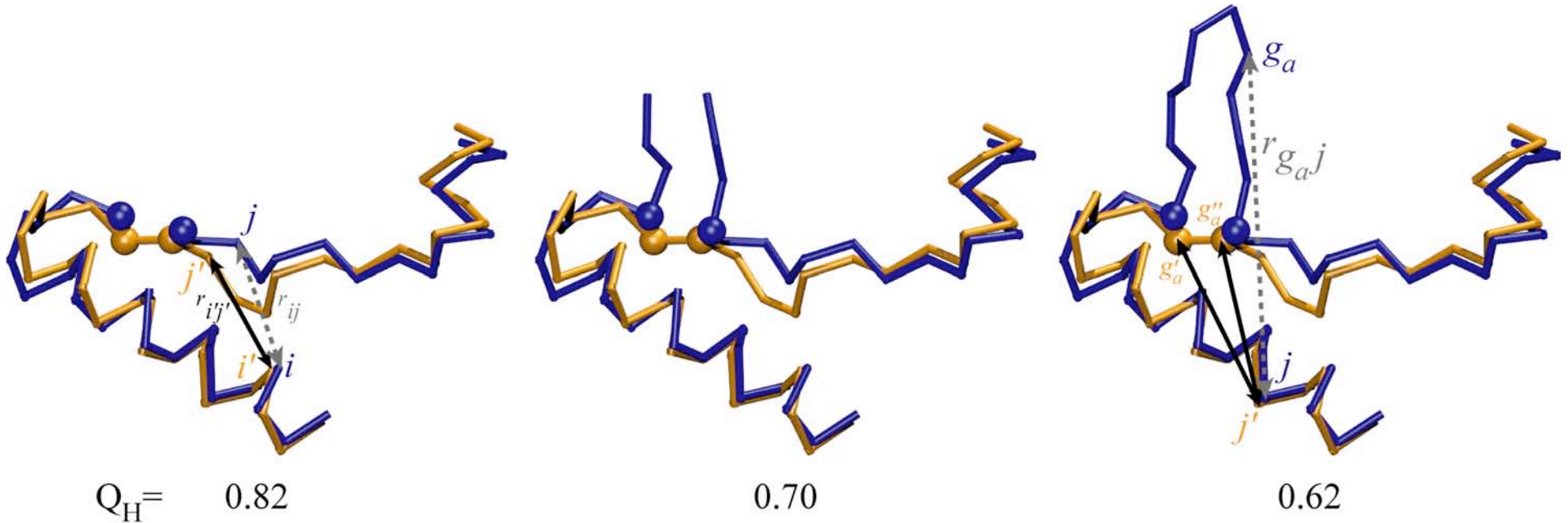
$$q_{aln} = \sum_{i < j-2} \exp \left[ -\frac{(r_{ij} - r_{i'j'})^2}{2\sigma_{ij}^2} \right]$$



# Structural Similarity Measure

## the effect of insertions

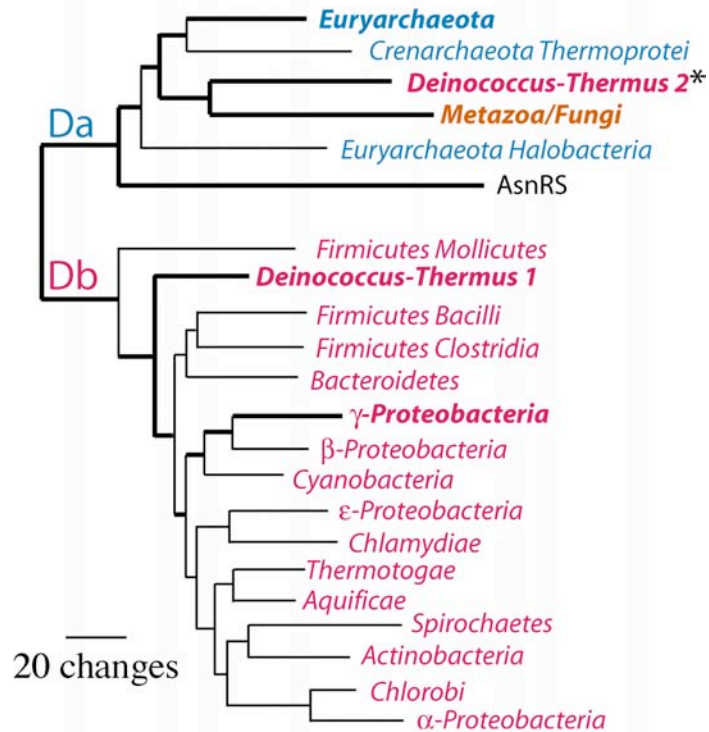
“Gaps should count as a character but not dominate” C. Woese



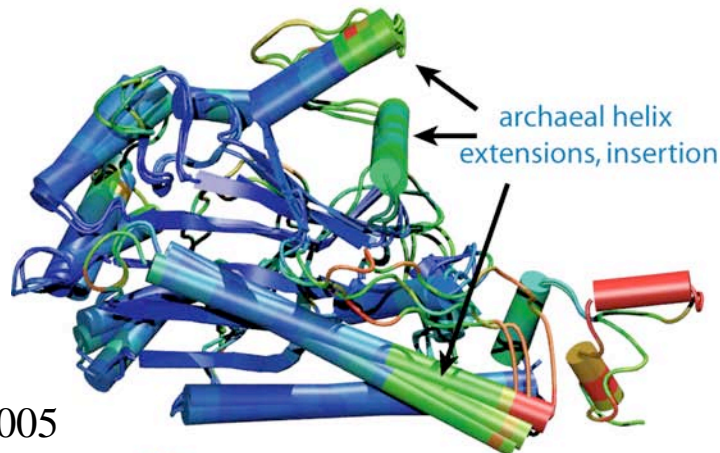
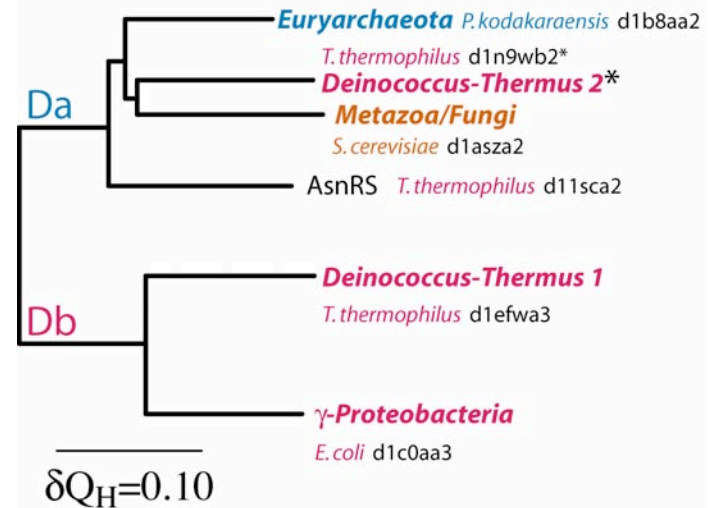
$$\begin{aligned}
 q_{gap} = & \sum_{g_a} \sum_j^{N_{aln}} \max \left\{ \exp \left[ -\frac{(r_{g_a j} - r_{g'_a j'})^2}{2\sigma_{g_a j}^2} \right], \exp \left[ -\frac{(r_{g_a j} - r_{g''_a j'})^2}{2\sigma_{g_a j}^2} \right] \right\} \\
 & + \sum_{g_b} \sum_j^{N_{aln}} \max \left\{ \exp \left[ -\frac{(r_{g_b j} - r_{g'_b j'})^2}{2\sigma_{g_b j}^2} \right], \exp \left[ -\frac{(r_{g_b j} - r_{g''_b j'})^2}{2\sigma_{g_b j}^2} \right] \right\}
 \end{aligned}$$

# Protein structure encodes evolutionary information

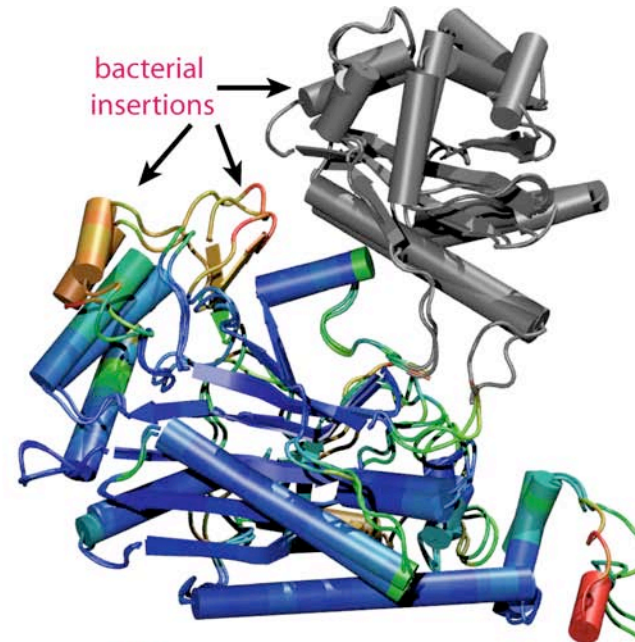
sequence-based phylogeny



structure-based phylogeny



Da - AspRS archaeal genre

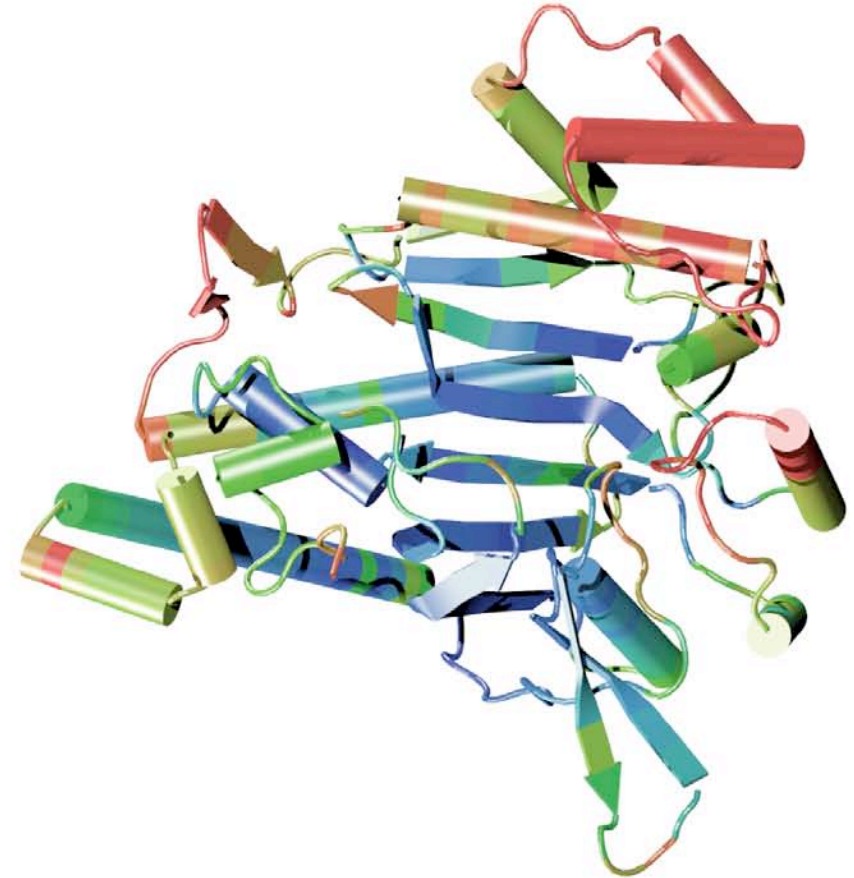
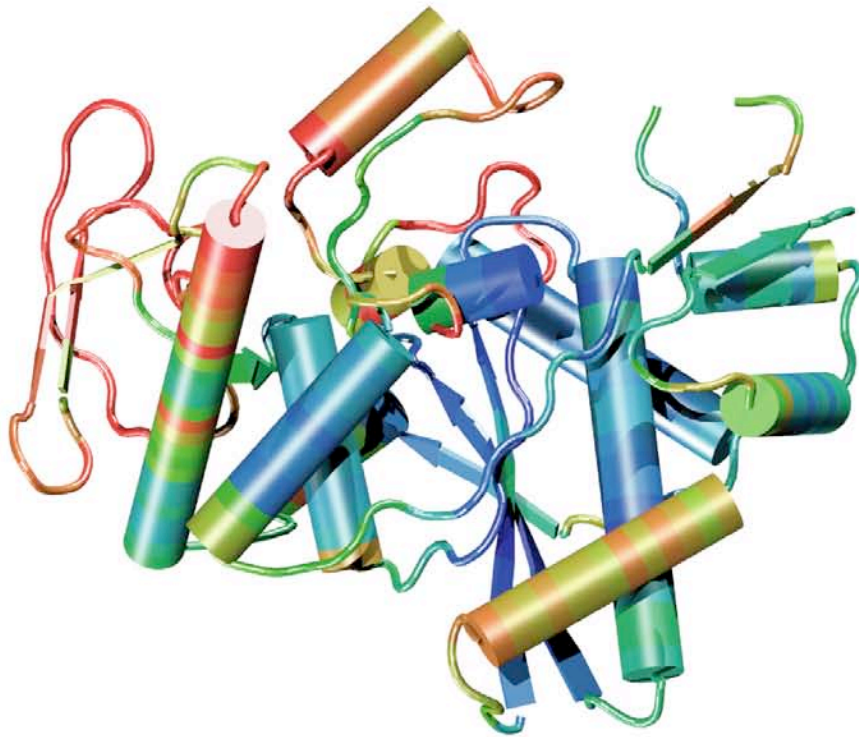
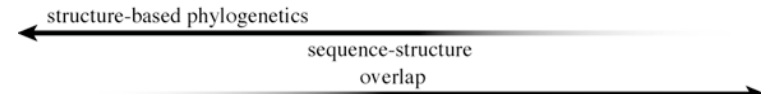
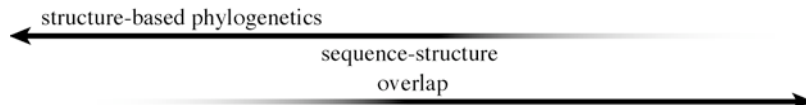


Db - AspRS bacterial genre

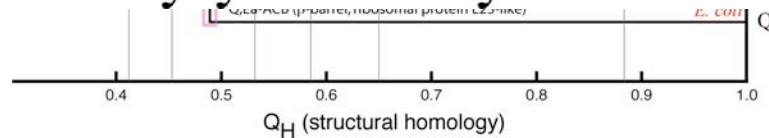
# Protein structure reveals distant evolutionary events

## Class I AARSs

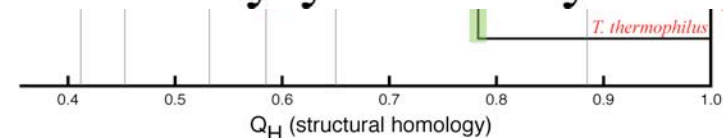
## Class II AARSs



## Class I Lysyl-tRNA Synthetase

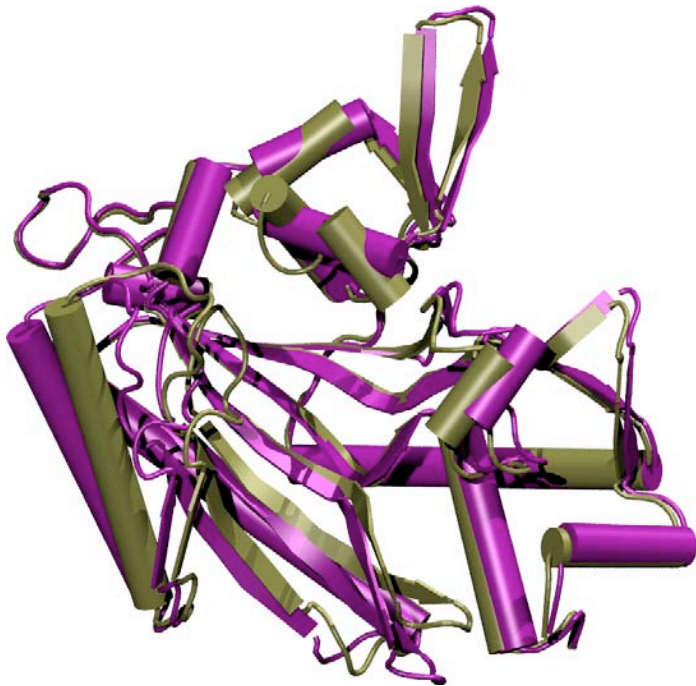


## Class II Lysyl-tRNA Synthetase





# Sequences define more recent evolutionary events



Conformational changes  
in the same protein.

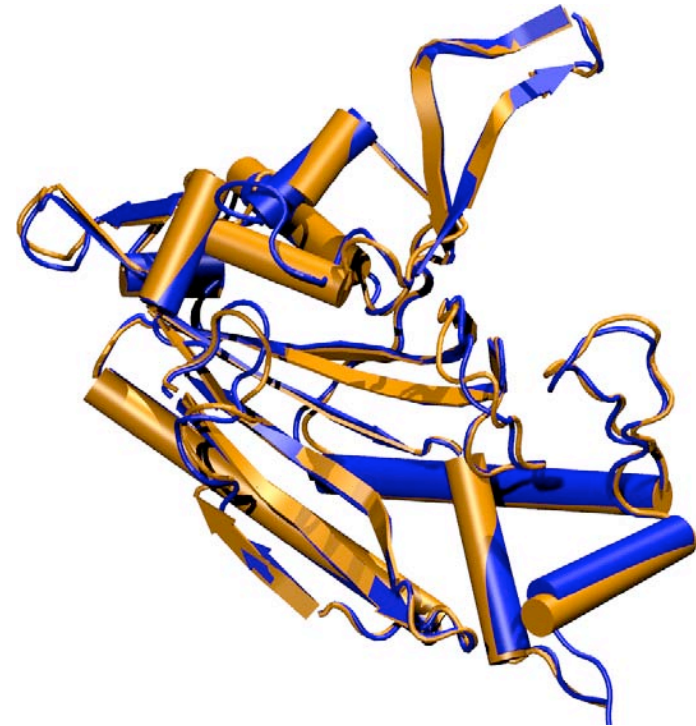
ThrRS

T-AMP analog, 1.55 Å.

T, 2.00 Å.

$Q_H = 0.80$

Sequence identity = 1.00



Structures for two  
different species.

ProRS

*M. jannaschii*, 2.55 Å.

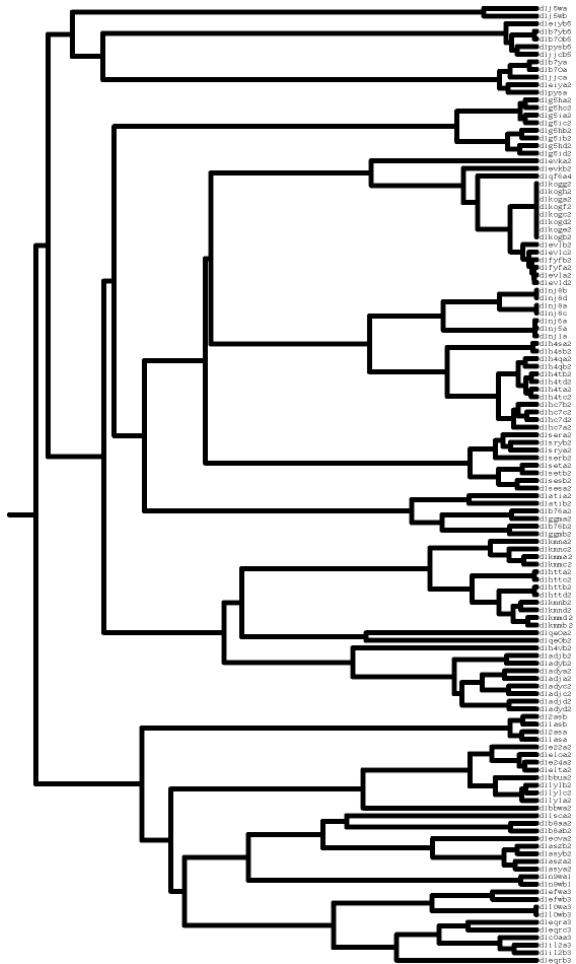
*M. thermoautotrophicus*, 3.20 Å.

$Q_H = 0.89$

Sequence identity = 0.69

# Non-redundant Representative Sets

Too much information  
129 Structures

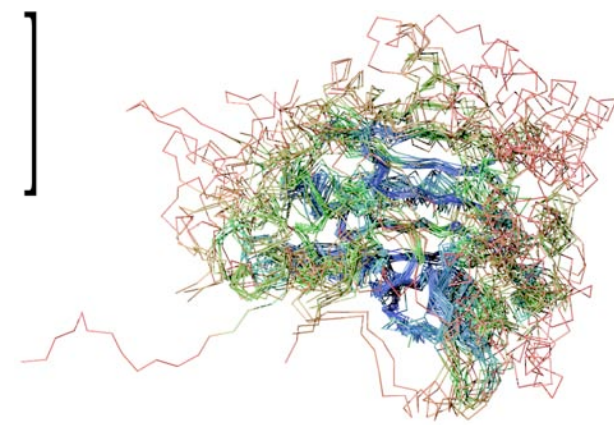
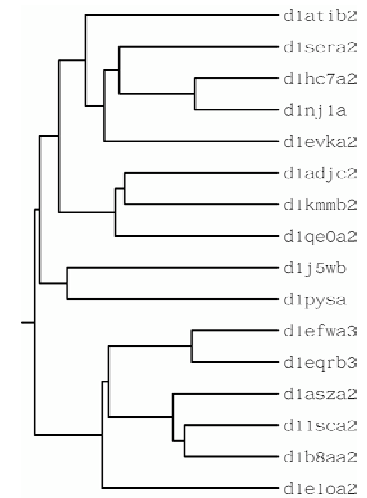


Multidimensional QR  
factorization  
of alignment matrix,  $A$ .

$$A = \left[ \begin{array}{c} \text{X} \\ \text{Y} \\ \text{Z} \\ \text{G} \end{array} \right]$$

$l_{aln}$  (vertical axis),  $k_{proteins}$  (horizontal axis),  $d=4$  (diagonal axis)

Economy of information  
16 representatives



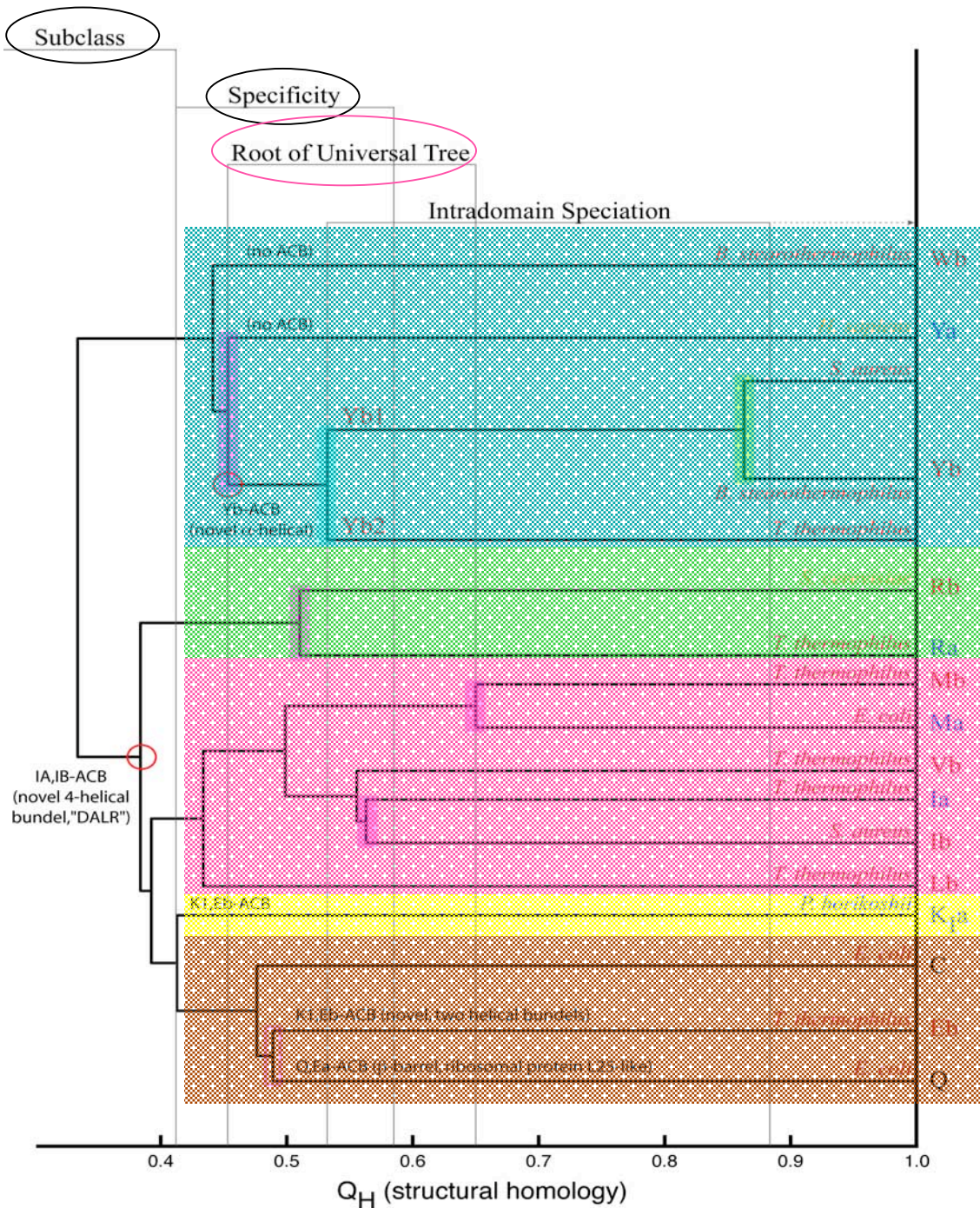
QR computes a set of maximal linearly independent structures.

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.



# Class I AARSs evolutionary events

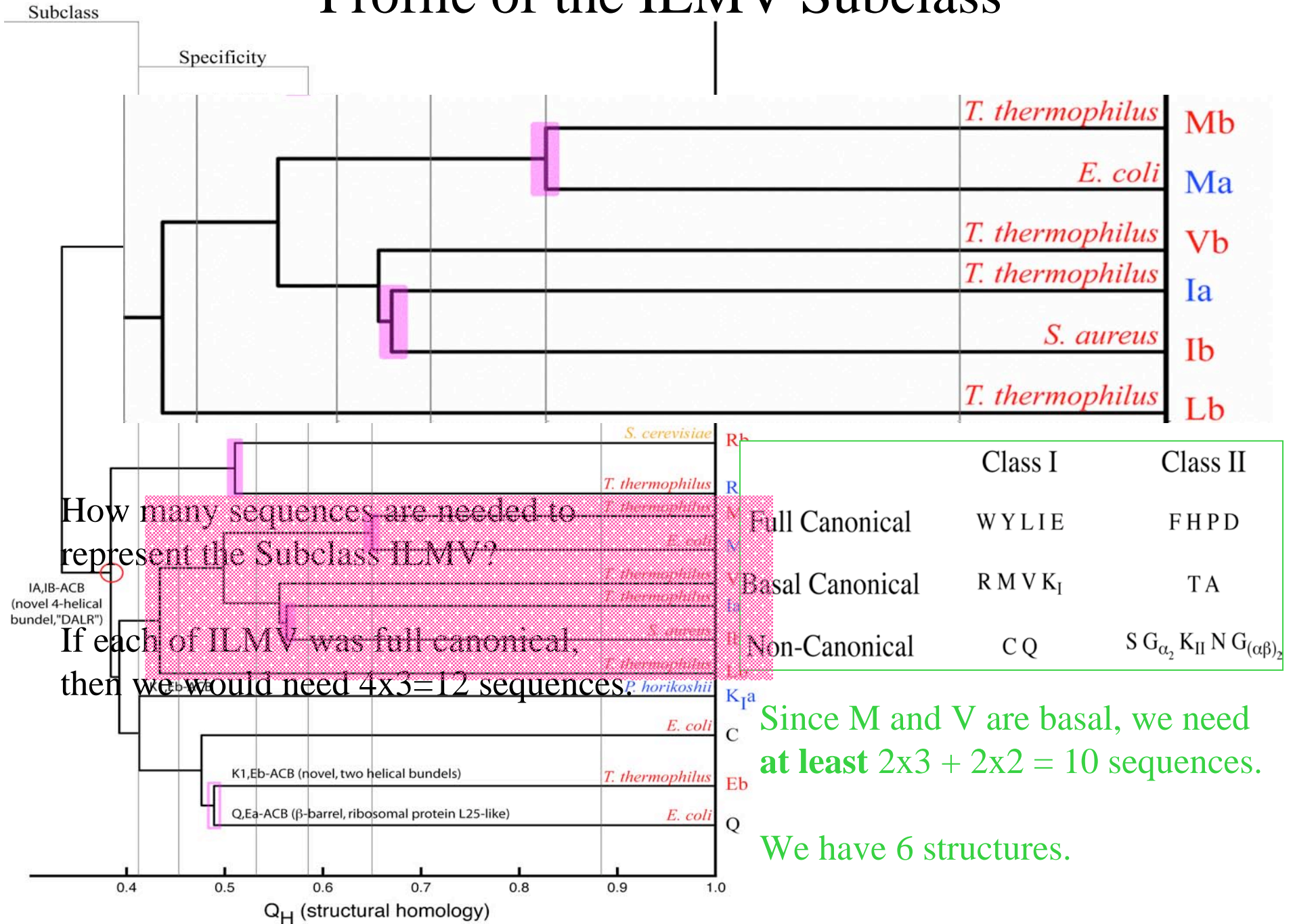


5 Subclasses

Specificity – 11 Amino acids

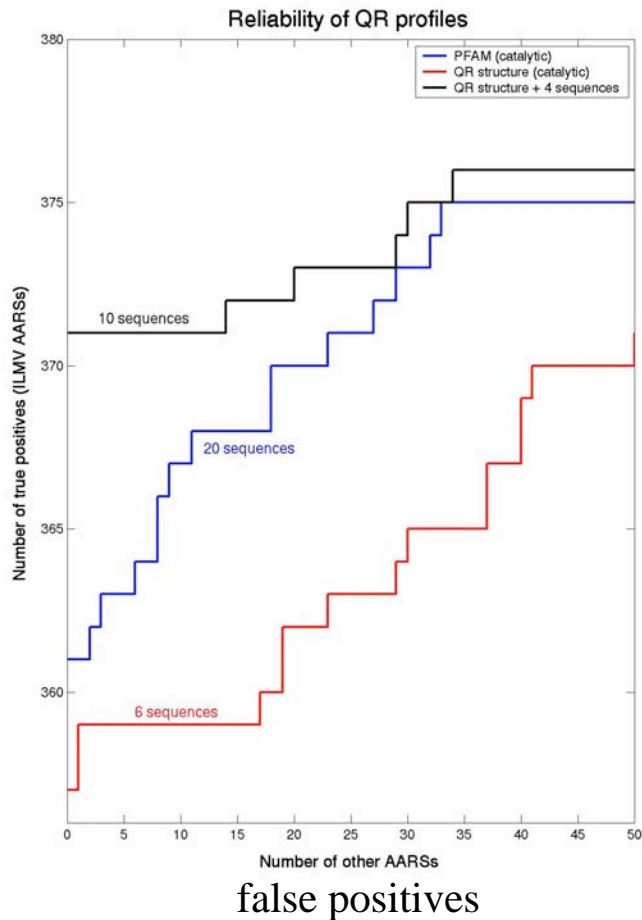
Domain of life A,B,E

# Profile of the ILMV Subclass

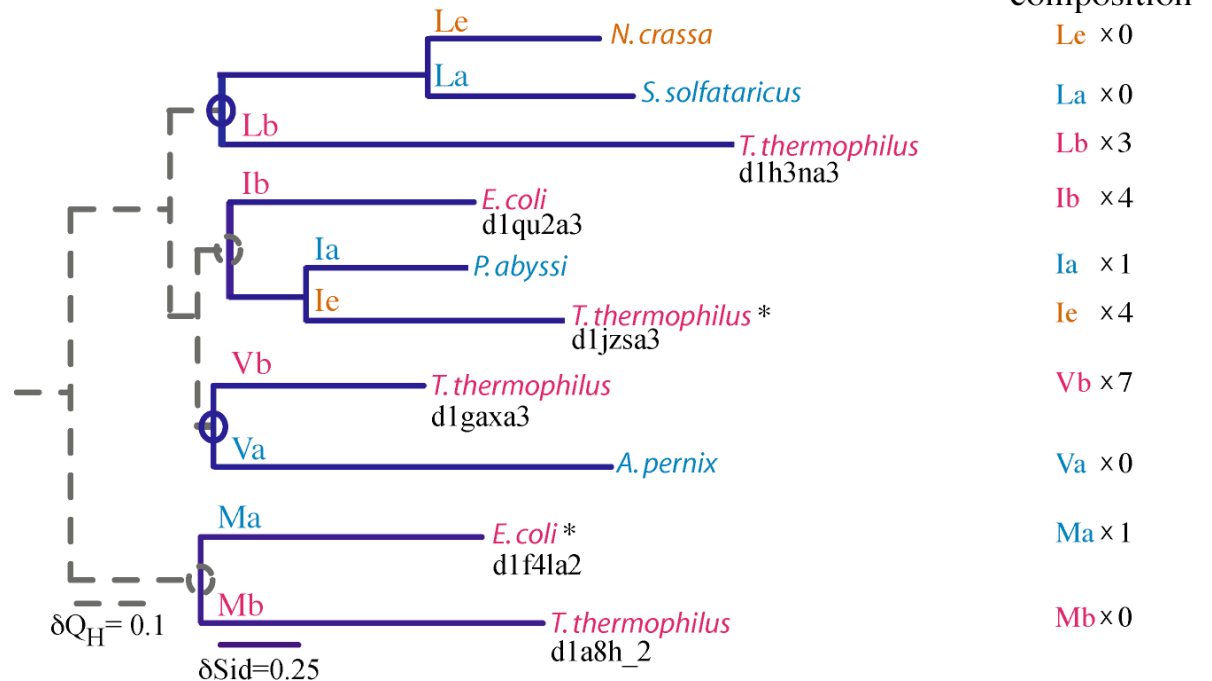


# Evolutionary Profiles for Homology Recognition

## AARS Subclass ILMV



Combined Structure-Sequence Phylogeny  
an evolutionary profile of the AARS subclass IA



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

# Genome Annotation

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

AsnRS }  
GlnRS } Indirect Mechanism  
LysRS Class I AARS  
CysRS ?

CysteinyI-tRNA(Cys) formation in *Methanocaldococcus jannaschii*: the mechanism is still unknown. *J. Bacteriology*, Jan. 2004, **186**:8-14.

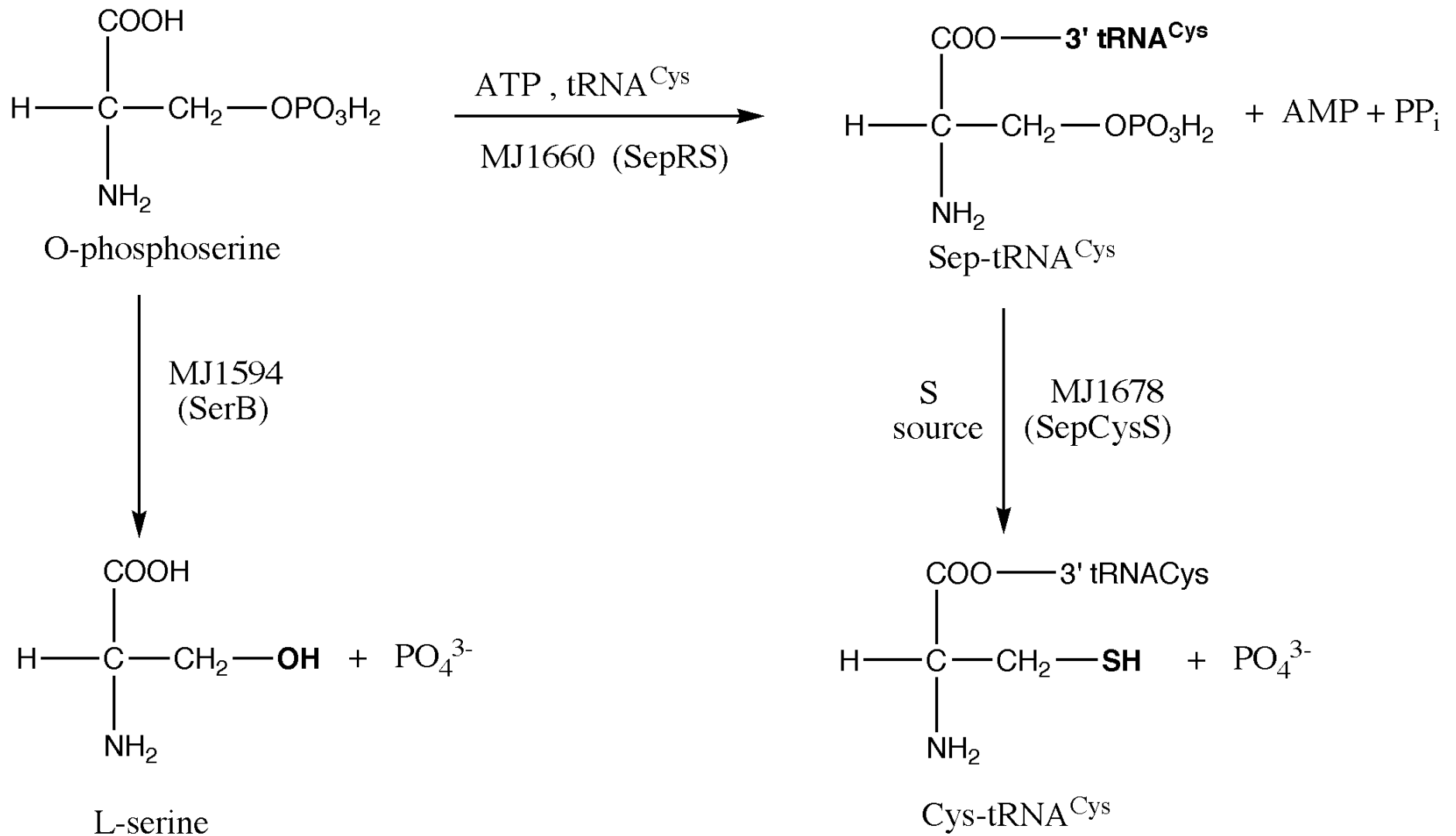
Ruan B, Nakano H, Tanaka M, Mills JA, DeVito JA, Min B, Low KB, Battista JR, and Söll D.

*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

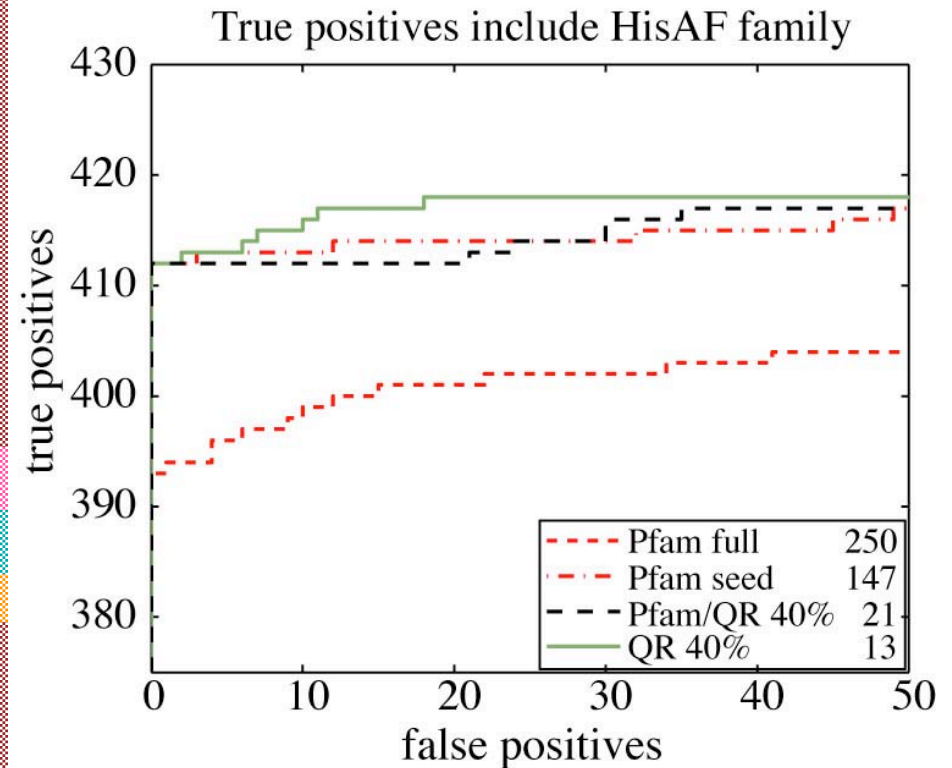
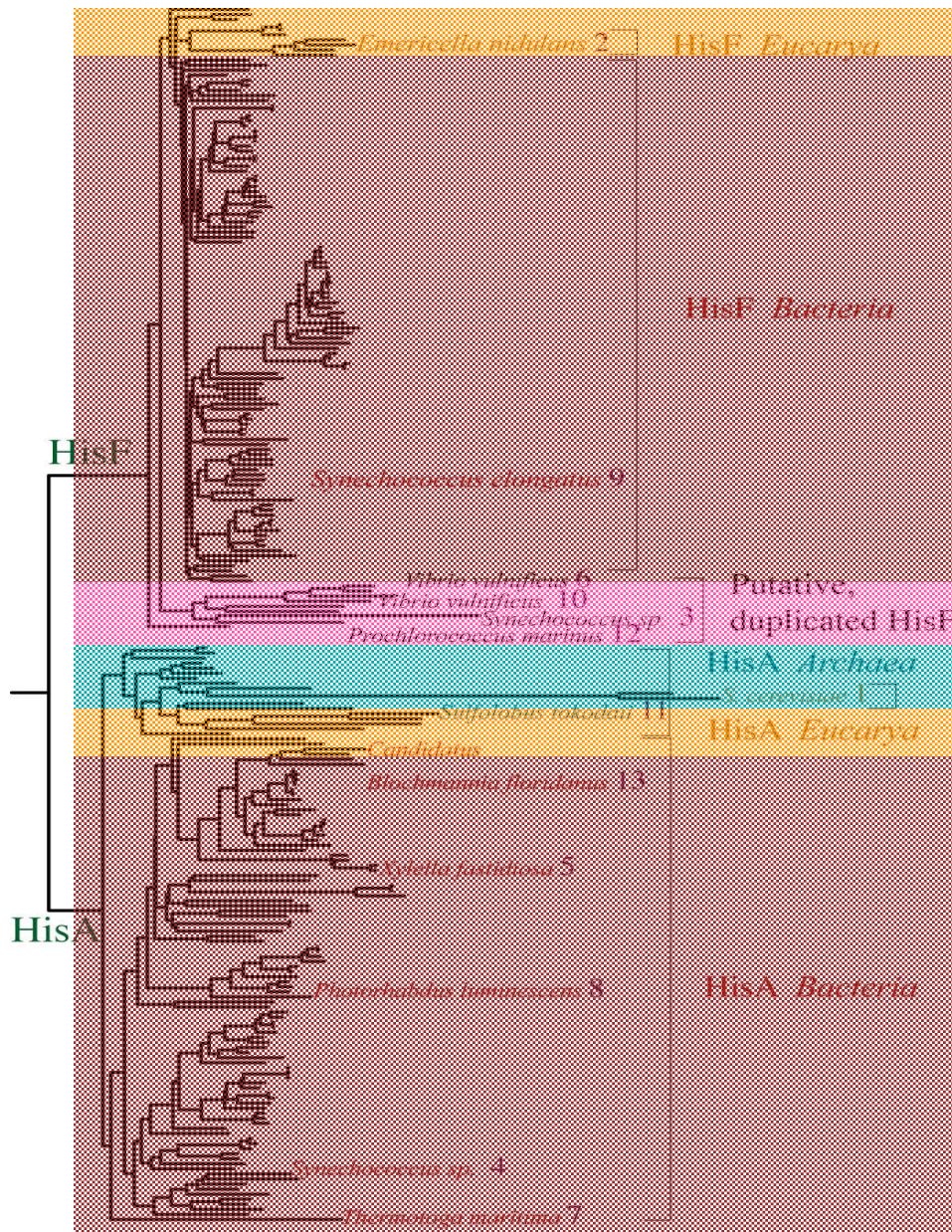
## Cysteine Biosynthesis in *Methanocaldococcus jannaschii*



Sauerwald et al. Science 2005



# Evolutionary profile for HisA-HisF family



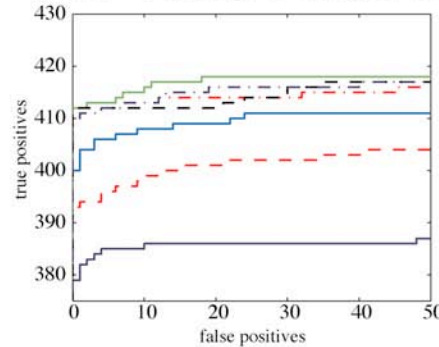
EP outperforms popular profile methods with an economy of information.

# Economy of Information

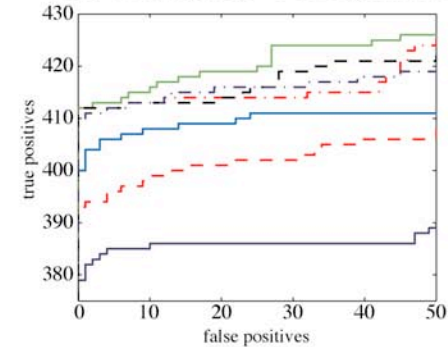
How many sequences are needed for profiles?

Profile	Nseq
--- Pfam full	250
-.- Pfam seed	147
- - - Pfam/QR 40%	21
— QR 15%	2
— QR 30%	4
— QR 40%	13
-.- QR 100%	238

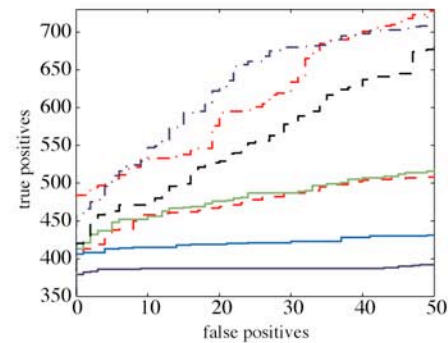
HisAF family recognition



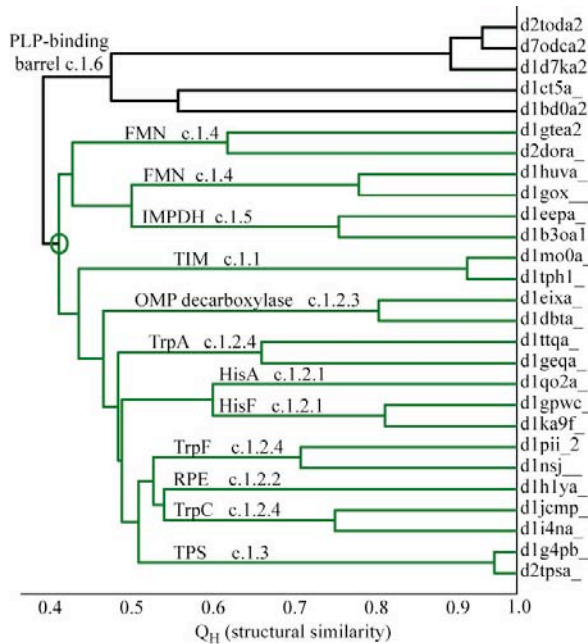
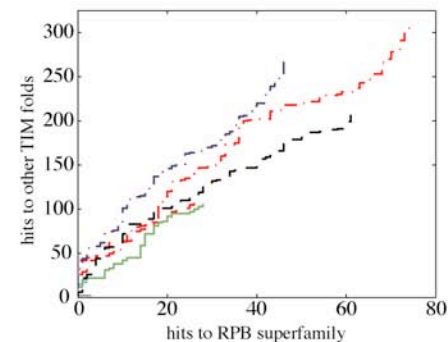
Superfamily recognition



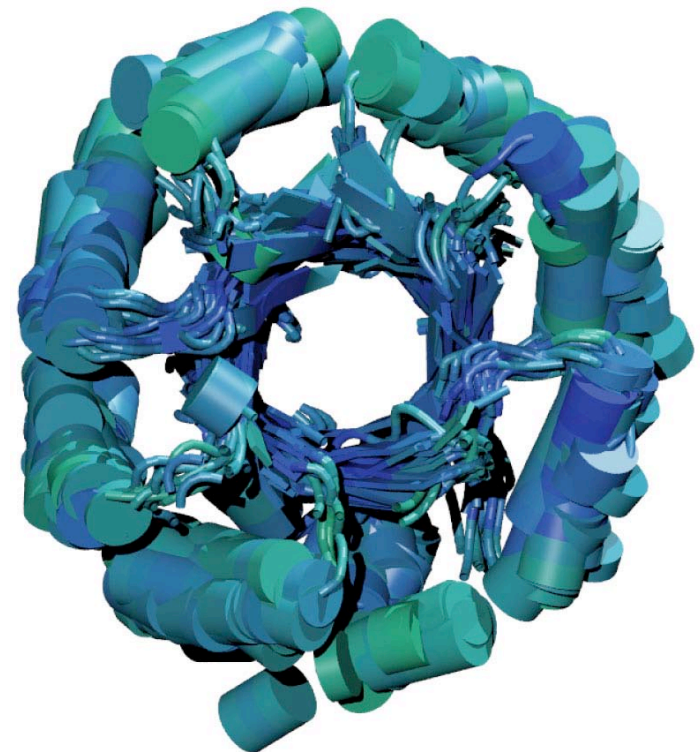
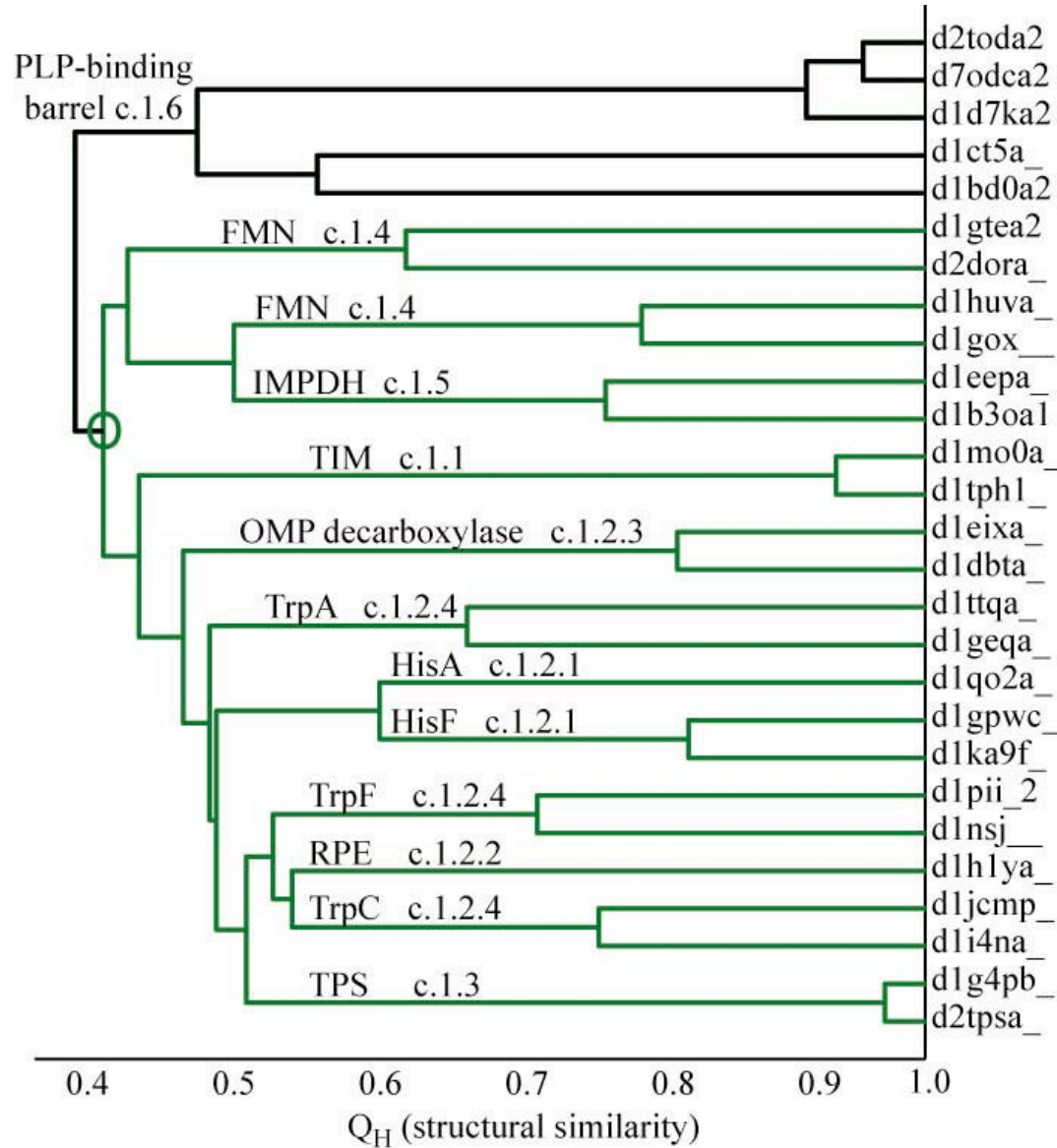
Fold recognition



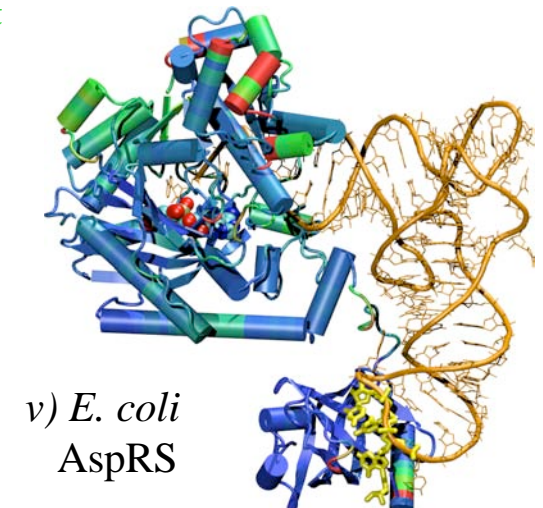
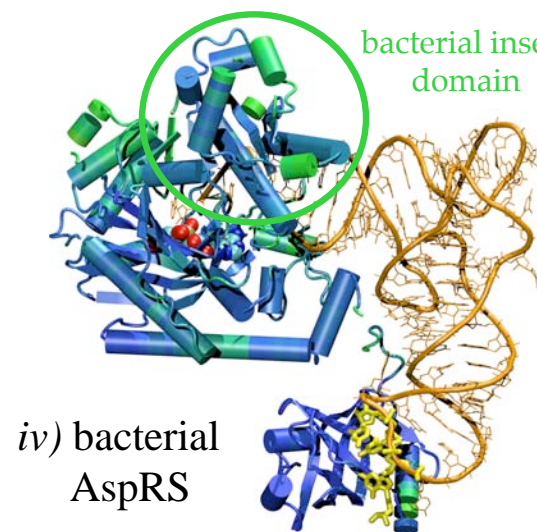
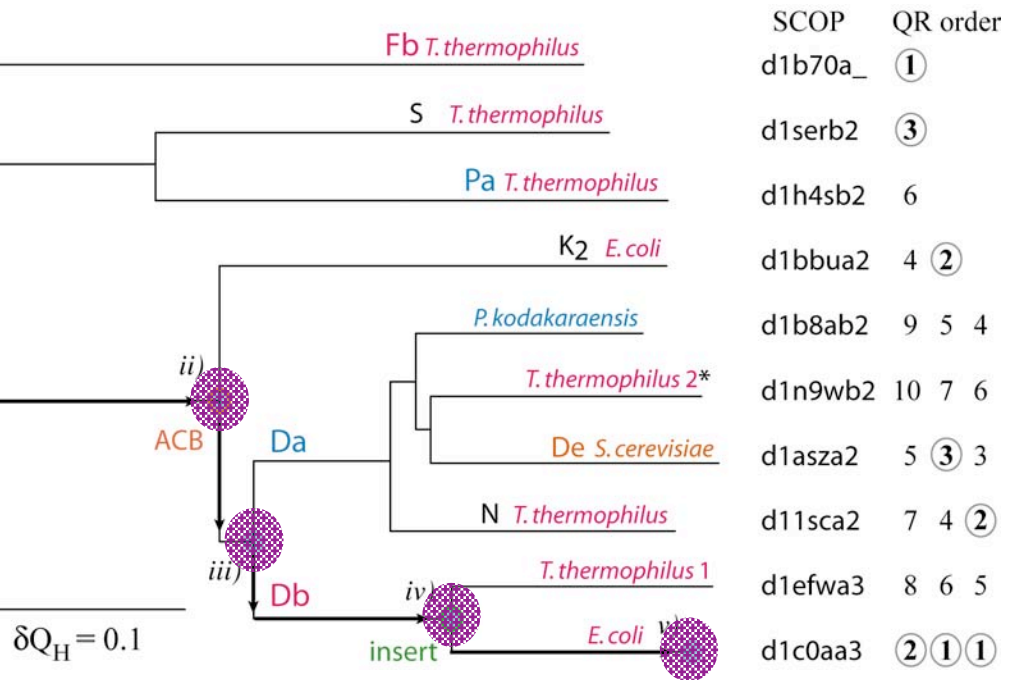
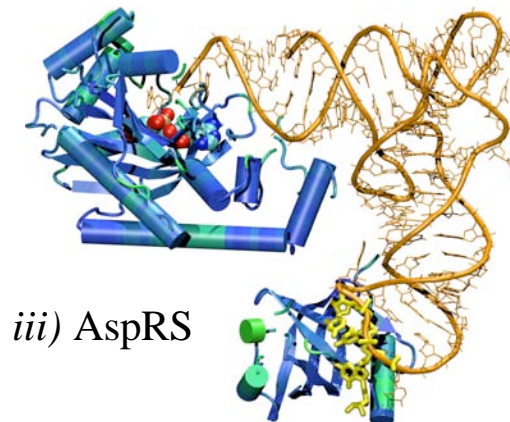
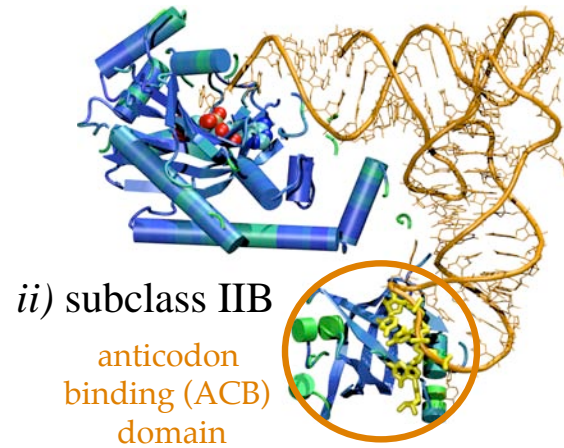
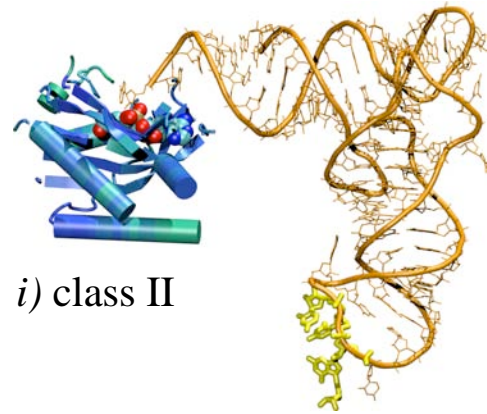
Fold versus Superfamily hits



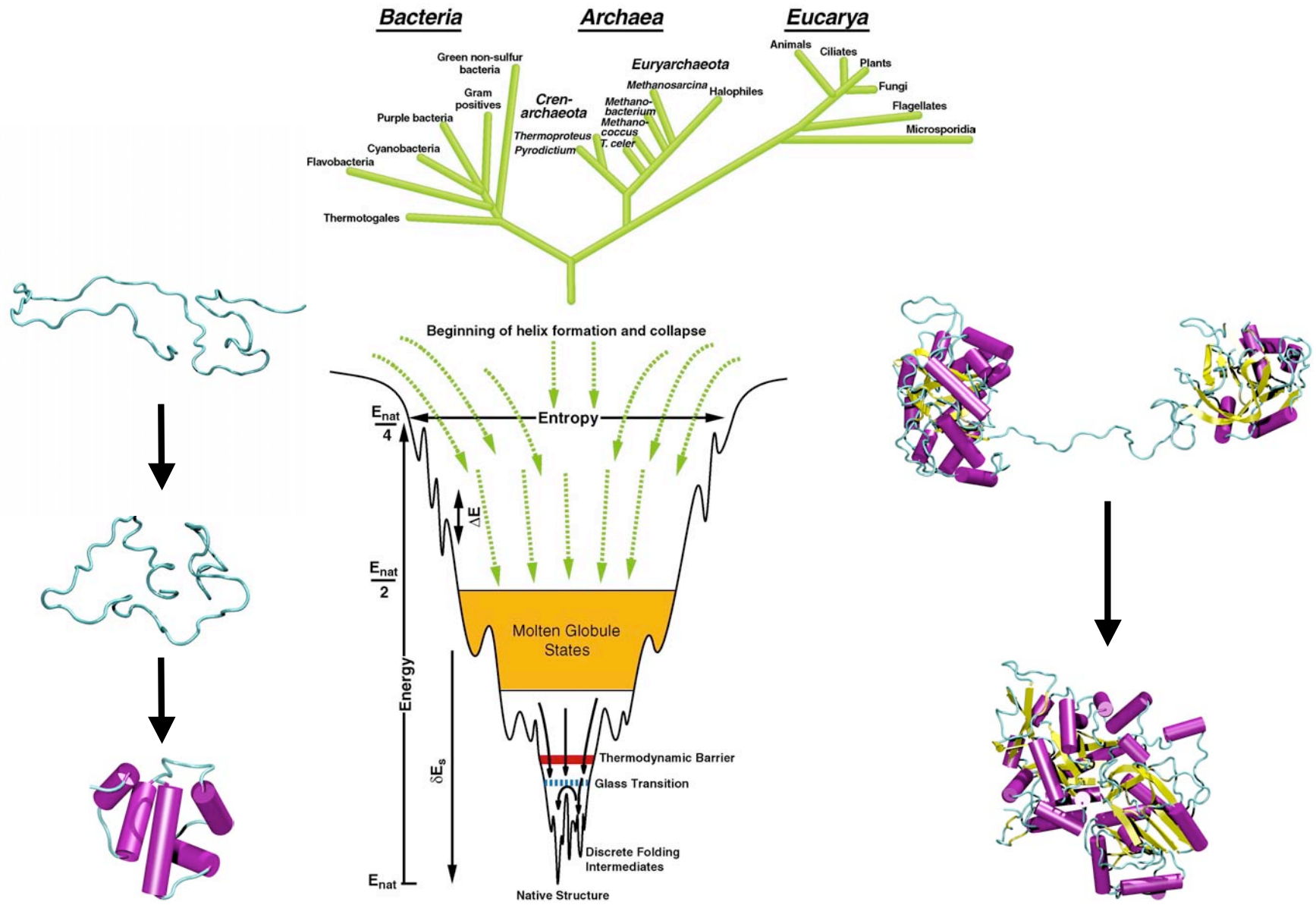
# Phylogenetic relationship between TIM barrels Found in database search with HisA-HisF profile



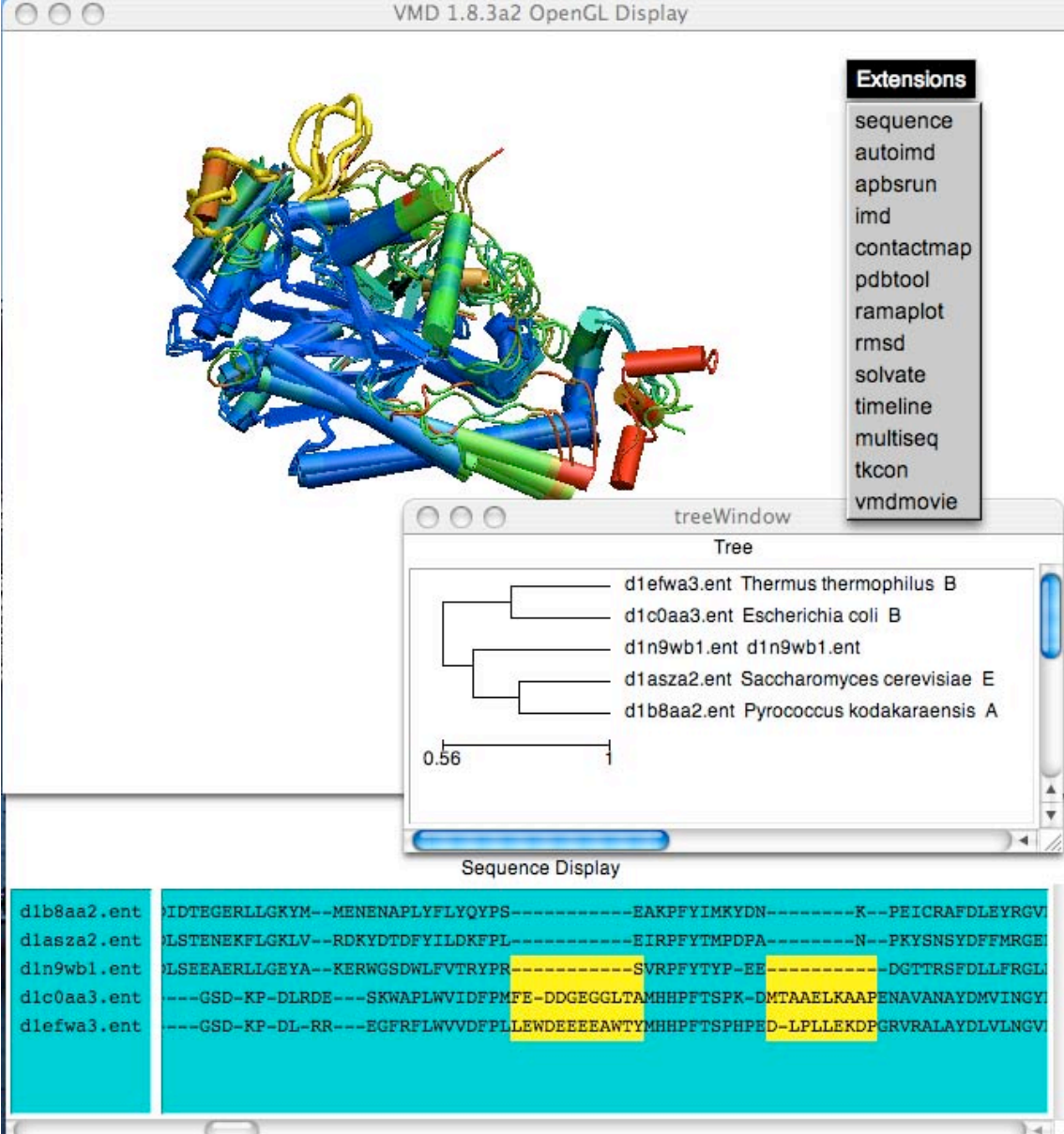
# Evolution of Structure and Function in AspRS



# Unifying the Worlds of Sequence and Structure



# Multiseq in VMD : Merging the sequence and structure worlds



The screenshot shows the VMD 1.8.3a2 OpenGL Display window. The main view displays a protein structure with a blue ribbon and a yellow stick representation. A 'treeWindow' is open, showing a phylogenetic tree with a scale bar of 0.56. The tree lists the following entries:

- d1efwa3.ent Thermus thermophilus B
- d1c0aa3.ent Escherichia coli B
- d1n9wb1.ent d1n9wb1.ent
- d1asza2.ent Saccharomyces cerevisiae E
- d1b8aa2.ent Pyrococcus kodakaraensis A

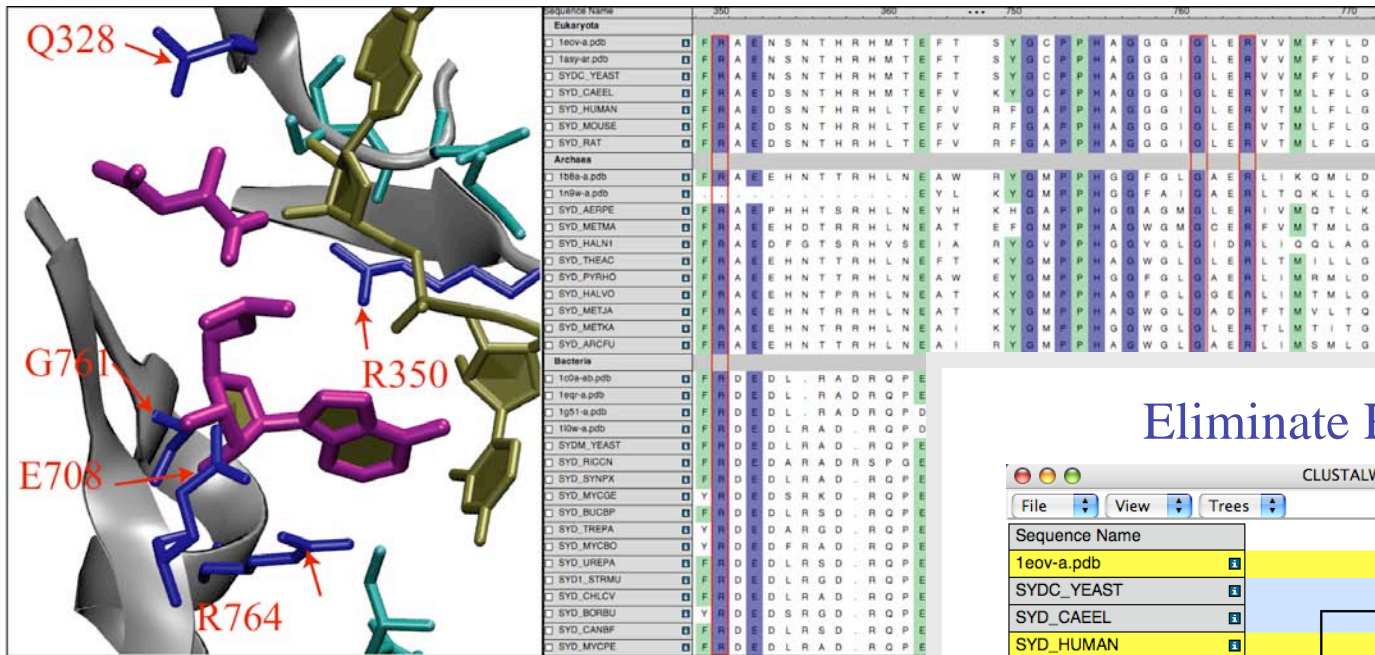
Below the tree is a 'Sequence Display' window showing a sequence alignment. The sequences are highlighted in yellow:

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

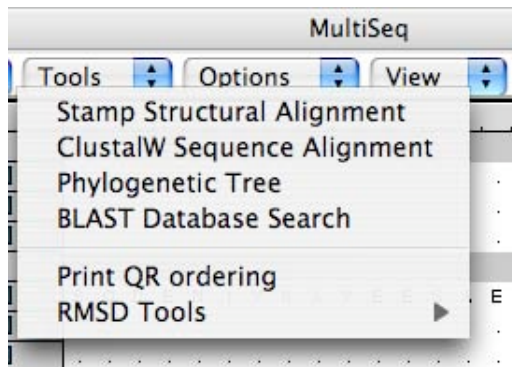
Version 1.83

# 2006 MultiSeq: New Features

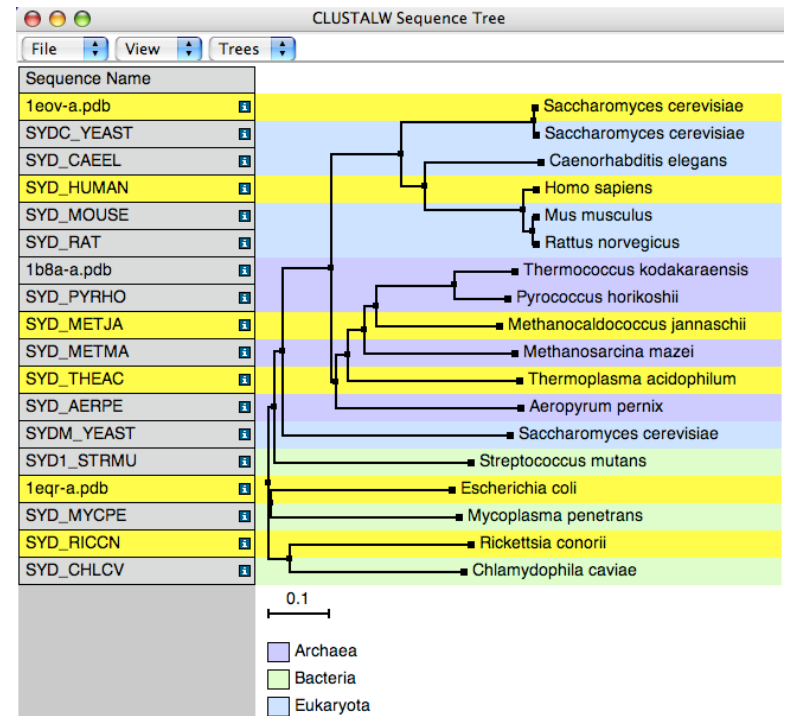
Analyze the Evolution of Sequence and Structure



## Plus More Functions



## Eliminate Redundancy



# 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**

Funding

NSF, NIH

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)

# Demonstration of New Multiseq Features

1. AspRS structures: STAMP multiple structure alignment. Color by structure (Qpair) 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