MULTISEQ in VMD -
Revealing How Nature Designs Proteins and RNAs

Luthey-Schulten Group

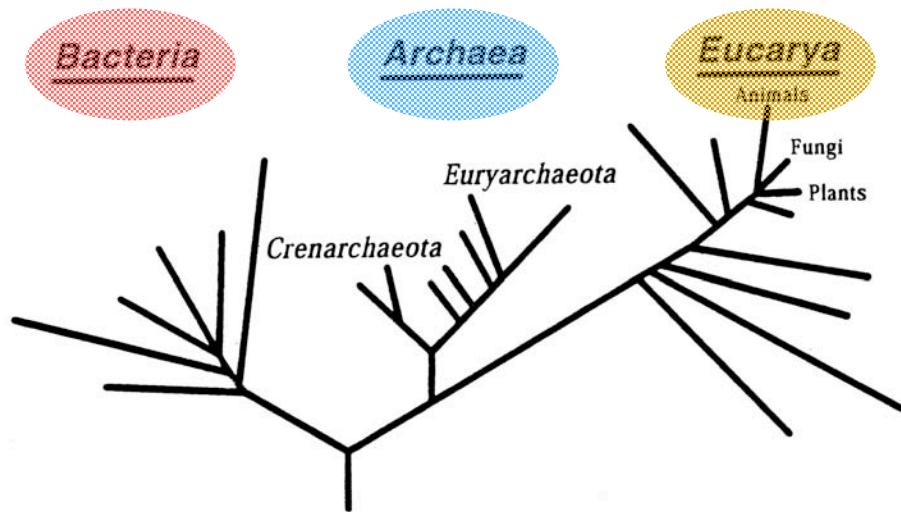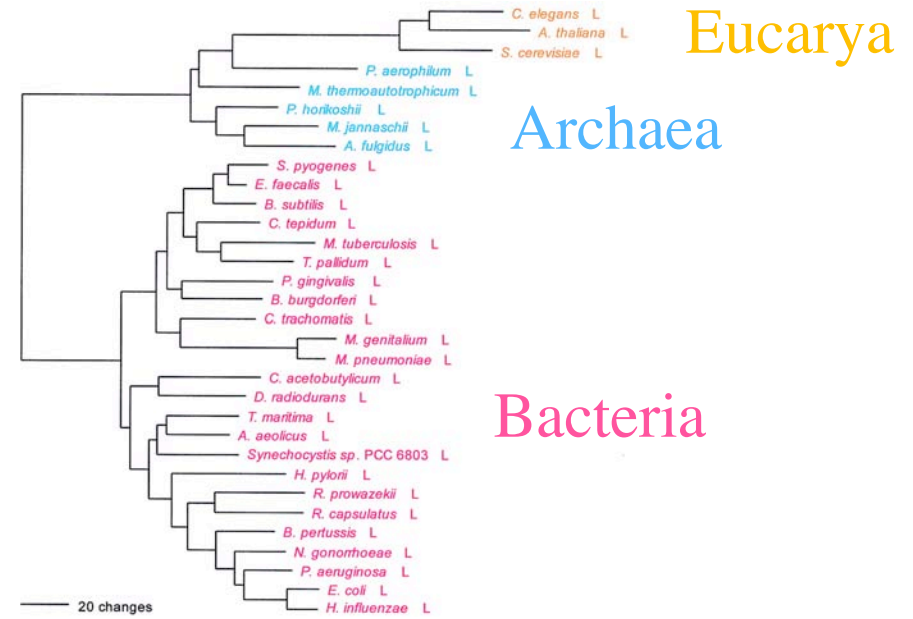Department of Chemistry, Biophysics, and Beckman Institute
University of Illinois at Urbana-Champaign
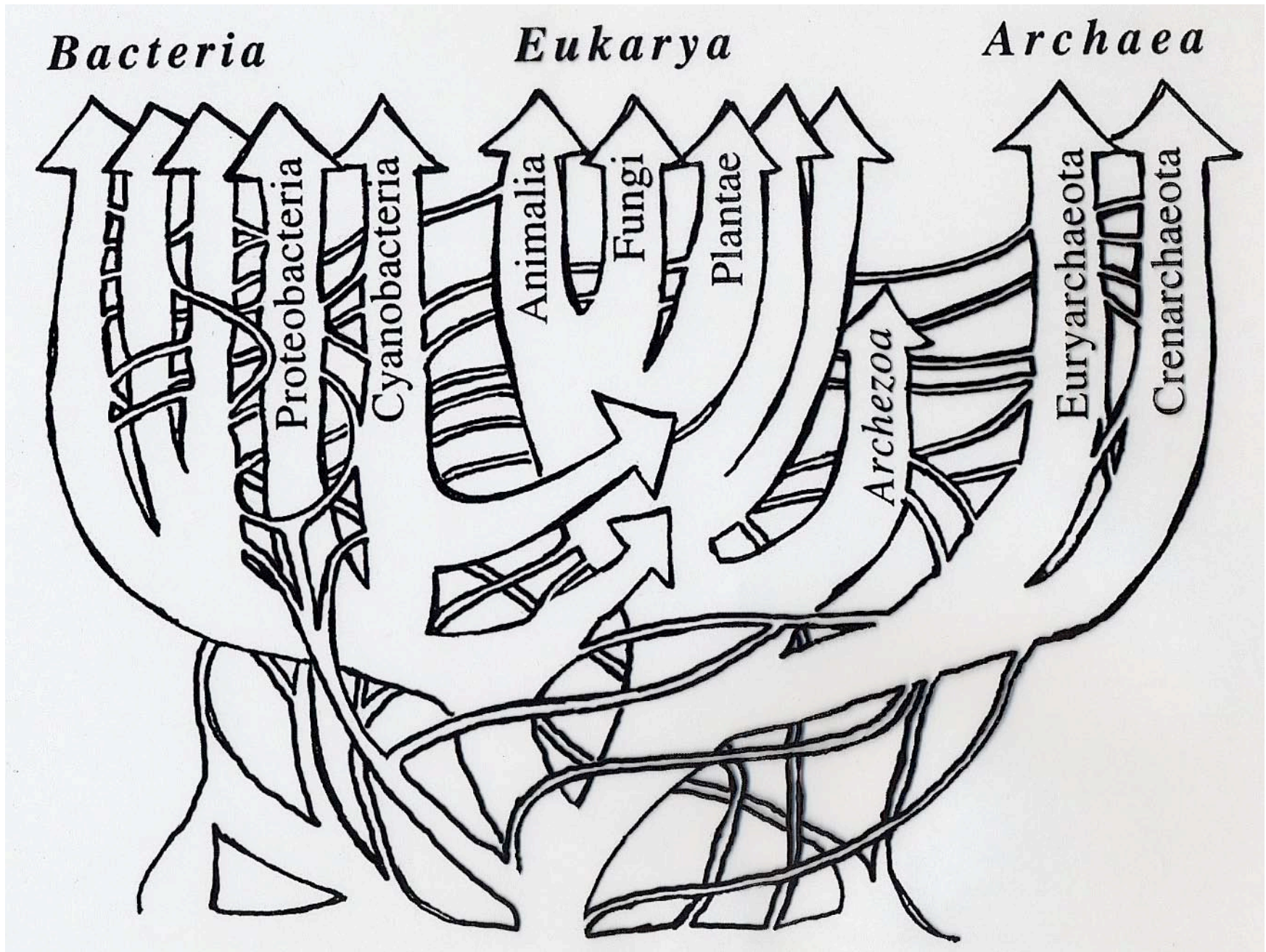
Universal Phylogenetic Tree
three domains of life

Bacteria   Archaea   Eucarya
Animals

Euryarchaeota
Crenarchaeota
Fungi
Plants

Based on 16S rRNA

C. elegans L
A. thaliana L
S. cerevisiae L
P. aerophilum L
M. thermoautotrophicum L
P. horikoshii L
M. jannaschii L
A. fulgidus L
S. pyogenes L
E. faecalis L
B. subtilis L
C. tepidum L
M. tuberculosis L
T. pallidum L
P. gingivalis L
B. burgdorferi L
C. trachomatis L
M. genitalium L
M. pneumoniae L
C. acetobutylicum L
D. radiodurans L
T. maritima L
A. aeolicus L
Synechocystis sp. PCC 6803 L
H. pylorii L
R. prowazekii L
R. capsulatus L
B. pertussis L
N. gonorrhoeae L
P. aeruginosa L
E. coli L
H. influenzae L

20 changes

Eucarya
Archaea
Bacteria

Leucyl-tRNA synthetase displays the
full canonical phylogenetic distribution.

for review see Woese *PNAS* 2000

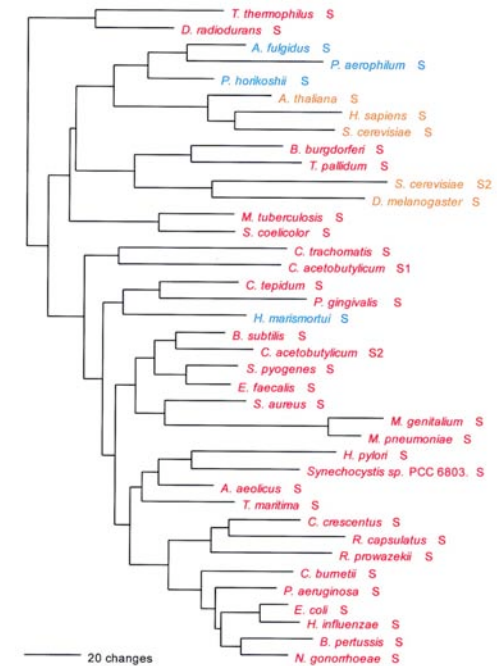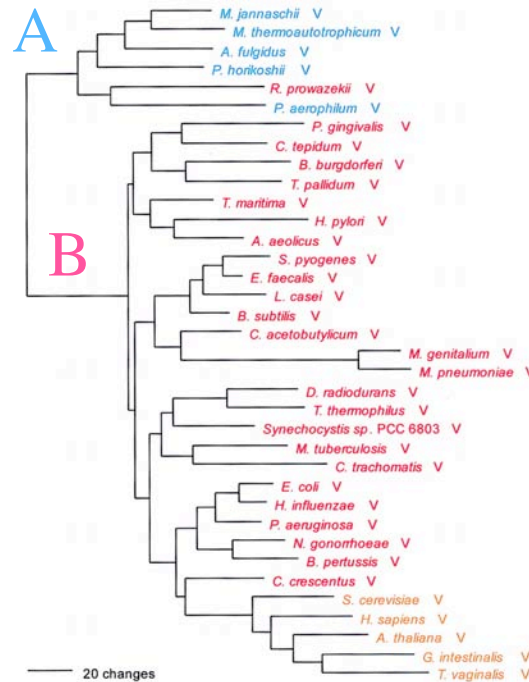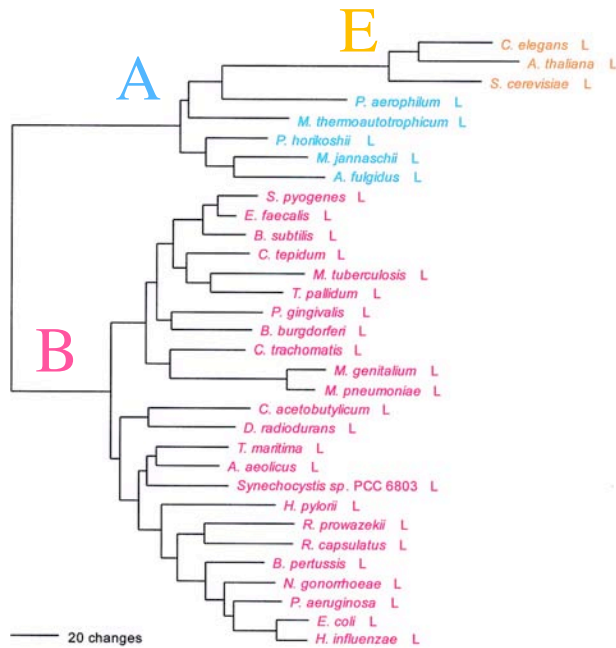Woese, Olsen, Ibba, Soll *MMBR* 2000

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
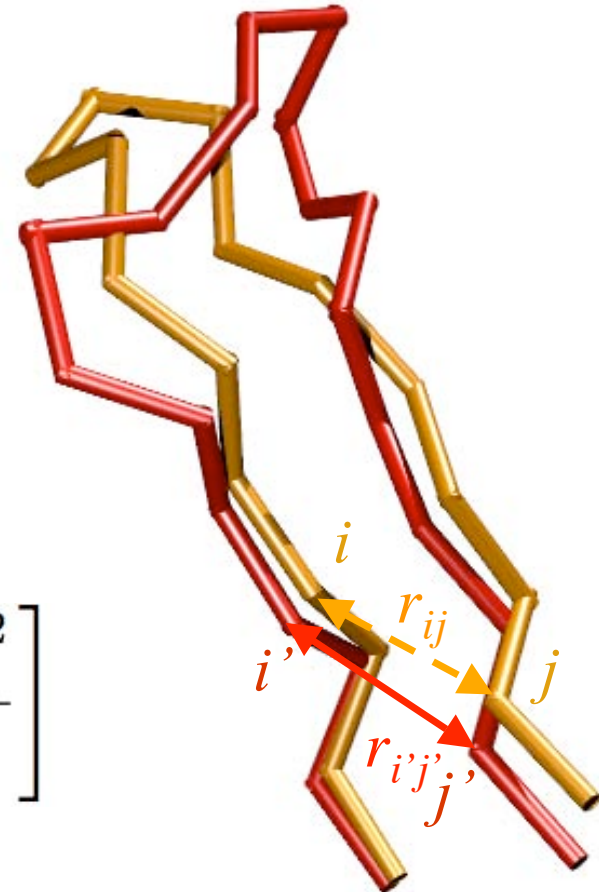
Protein Structure Similarity Measure

Q_H Structural Homology

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

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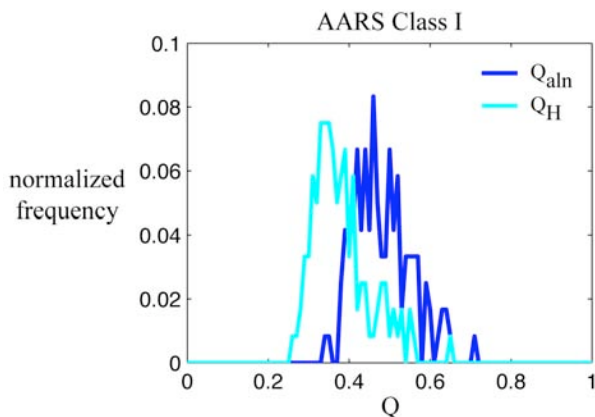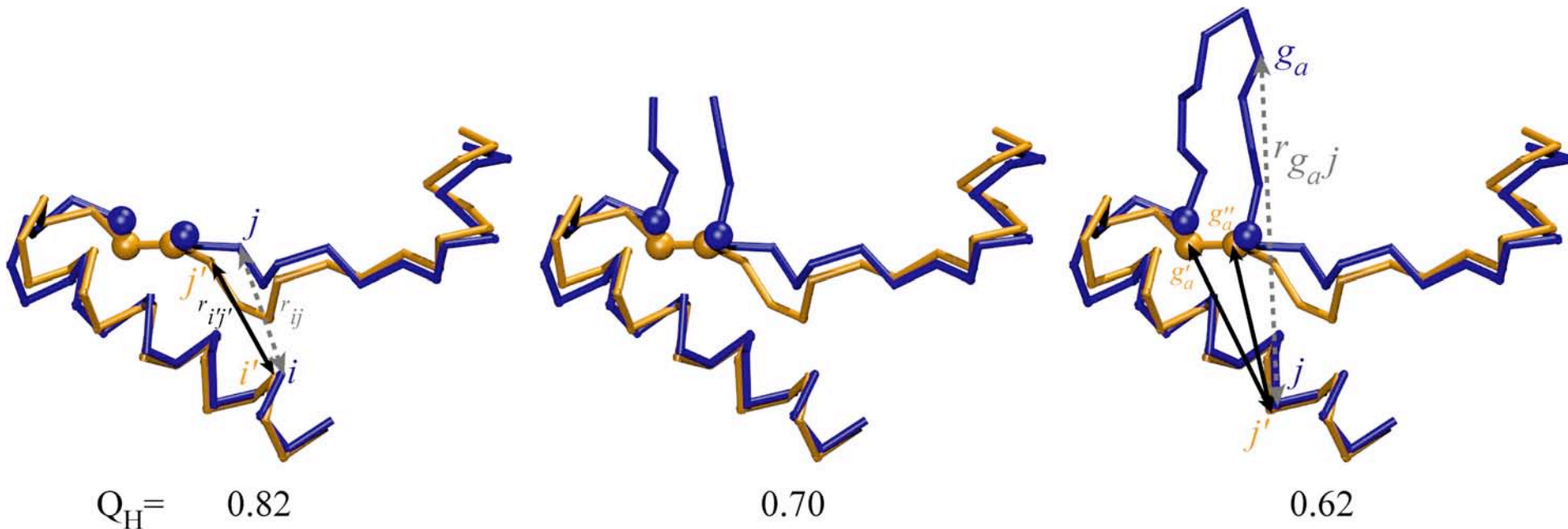
$$q_{aln} = \sum_{i<j-2} \exp \left[-\frac{(r_{ij} - r_{i'j'})^2}{2\sigma_{ij}^2} \right]$$

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

Structural Similarity Measure
the effect of insertions

$Q_H = \quad 0.82 \qquad\qquad 0.70 \qquad\qquad 0.62$

AARS Class I

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

$$+ \sum_{g_b}\sum_{j}^{N_{aln}} \max\left\{\exp\left[-\frac{\left(r_{g_b j} - r_{g_b' j'}\right)^2}{2\sigma_{g_b j}^2}\right], \exp\left[-\frac{\left(r_{g_b j} - r_{g_b'' j'}\right)^2}{2\sigma_{g_b j}^2}\right]\right\}$$
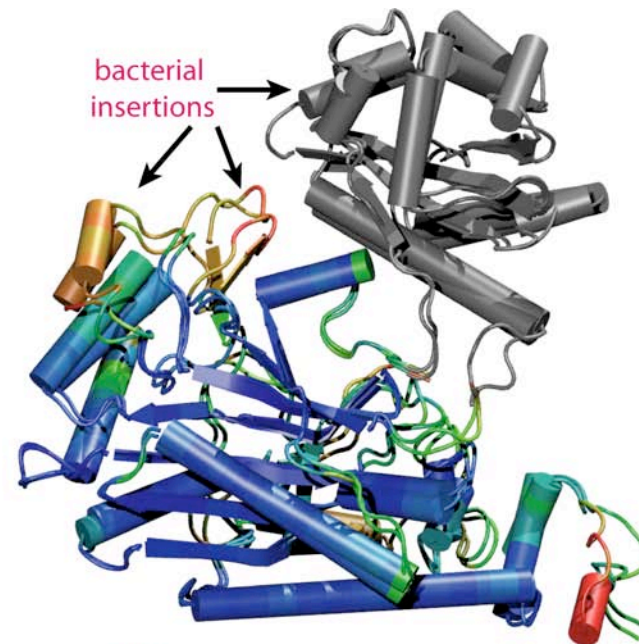
Protein structure encodes evolutionary information

sequence-based phylogeny
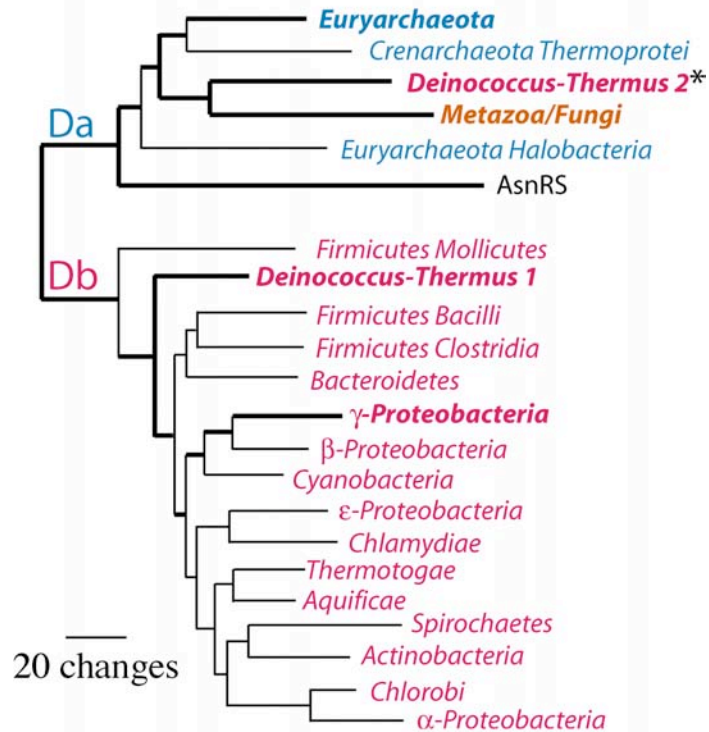
structure-based phylogeny

Da
Db

20 changes

Euryarchaeota
Crenarchaeota Thermoprotei
Deinococcus-Thermus 2*
Metazoa/Fungi
Euryarchaeota Halobacteria
AsnRS

Firmicutes Mollicutes
Deinococcus-Thermus 1
Firmicutes Bacilli
Firmicutes Clostridia
Bacteroidetes
γ-*Proteobacteria*
β-*Proteobacteria*
Cyanobacteria
ε-*Proteobacteria*
Chlamydiae
Thermotogae
Aquificae
Spirochaetes
Actinobacteria
Chlorobi
α-*Proteobacteria*

Da
Db

Euryarchaeota *P. kodakaraensis* d1b8aa2
T. thermophilus d1n9wb2*
Deinococcus-Thermus 2*
Metazoa/Fungi
S. cerevisiae d1asza2
AsnRS *T. thermophilus* d11sca2

Deinococcus-Thermus 1
T. thermophilus d1efwa3

γ-*Proteobacteria*
E. coli d1c0aa3

$\delta Q_H = 0.10$

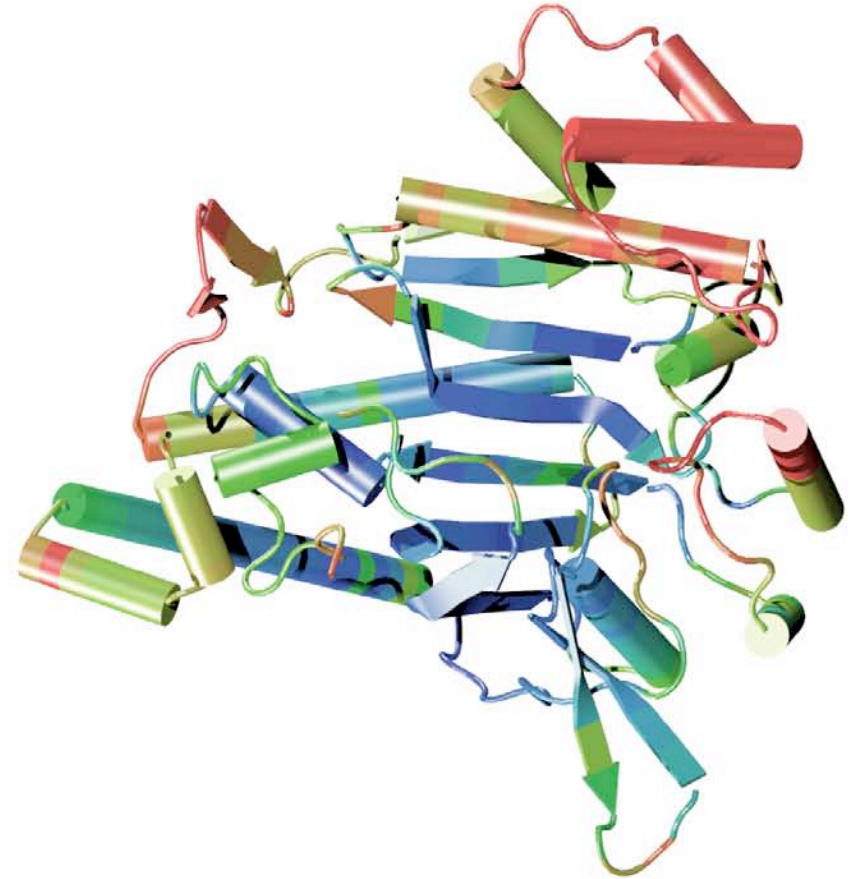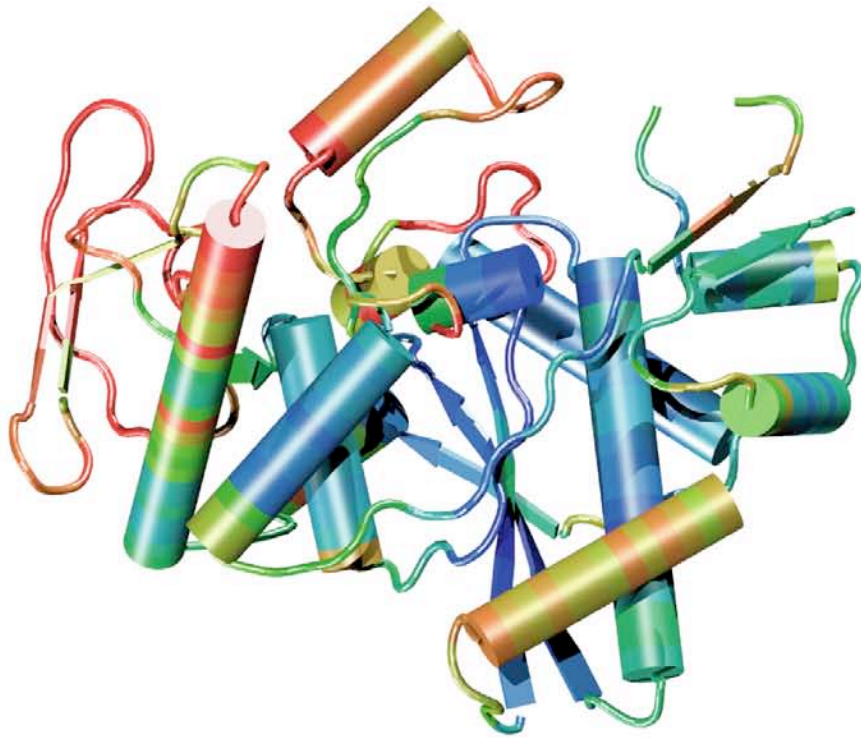archaeal helix extensions, insertion

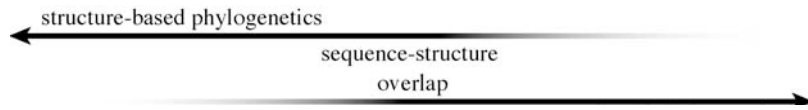bacterial insertions

Da - AspRS archaeal genre

Db - AspRS bacterial genre

JMB 2003, 2005

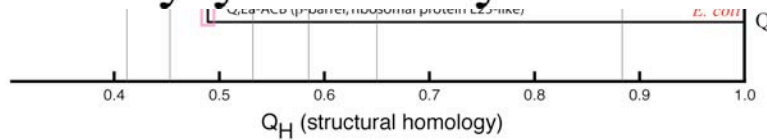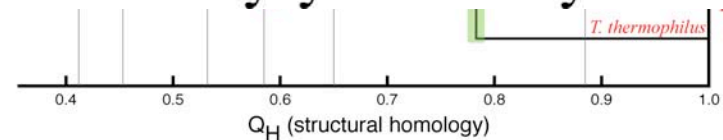# Protein structure reveals distant evolutionary events

Class I AARSs

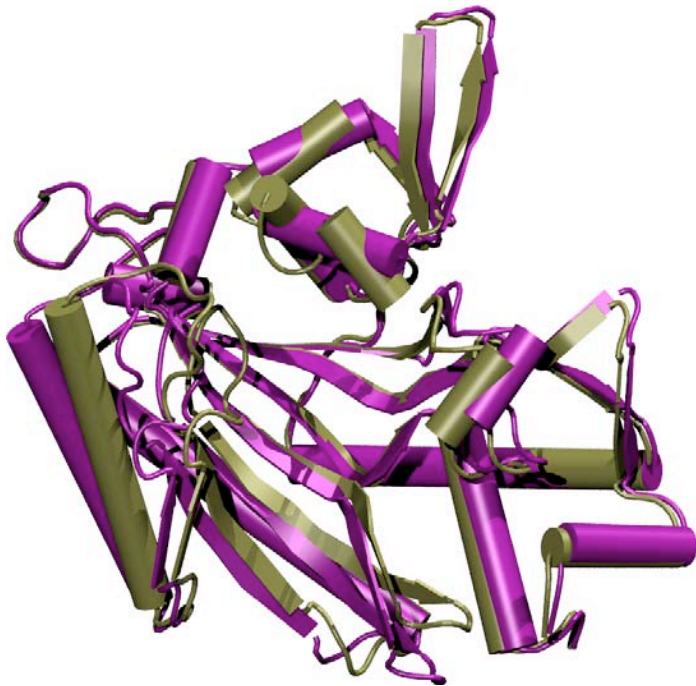## Class II AARSs
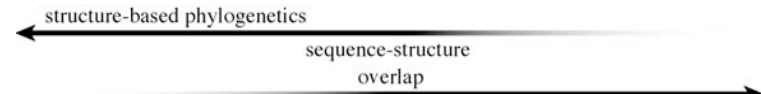
Class I Lysyl-tRNA Synthetase

Class II Lysyl-tRNA Synthetase

Sequences define more recent evolutionary events

Conformational changes
in the same protein.
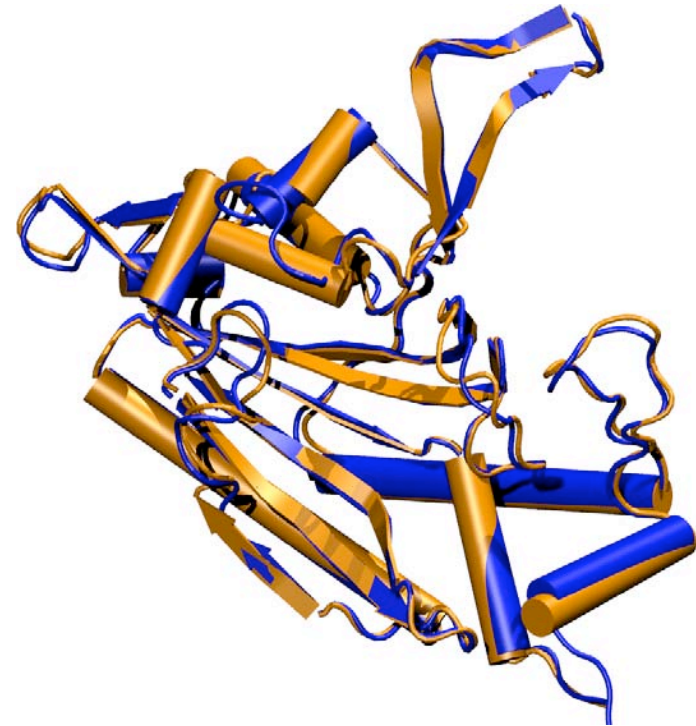
ThrRS
T-AMP analog, 1.55 A.
T, 2.00 A.

$Q_H = 0.80$
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
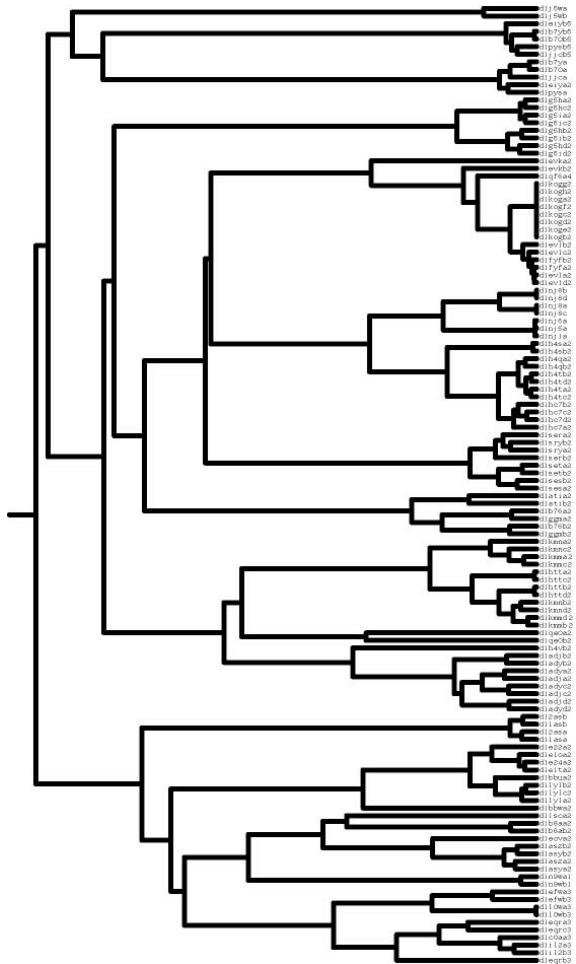
Multidimensional QR
factorization
of alignment matrix, A.

$$A = \begin{bmatrix} & & & G \\ & & Z & \\ & Y & & \\ X & & & \end{bmatrix}$$

$d=4$

l_{aln}

$k_{proteins}$
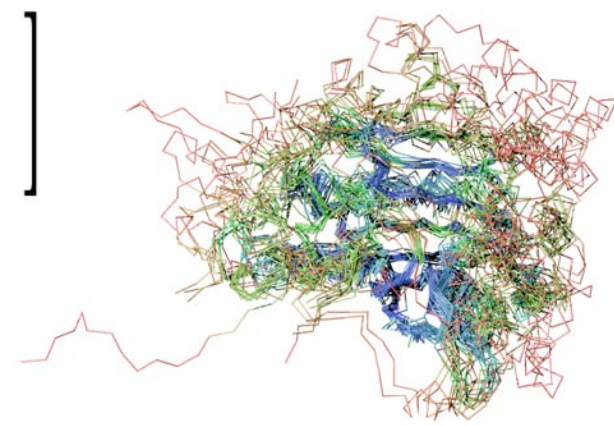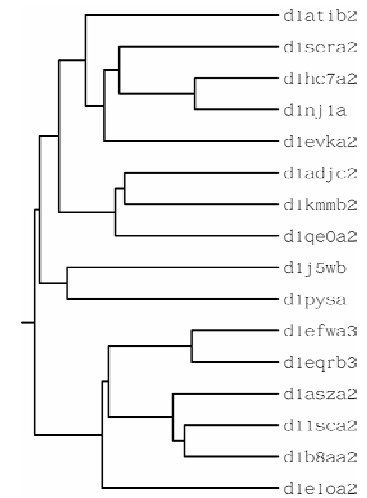
QR computes a set of maximal linearly independent structures.

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
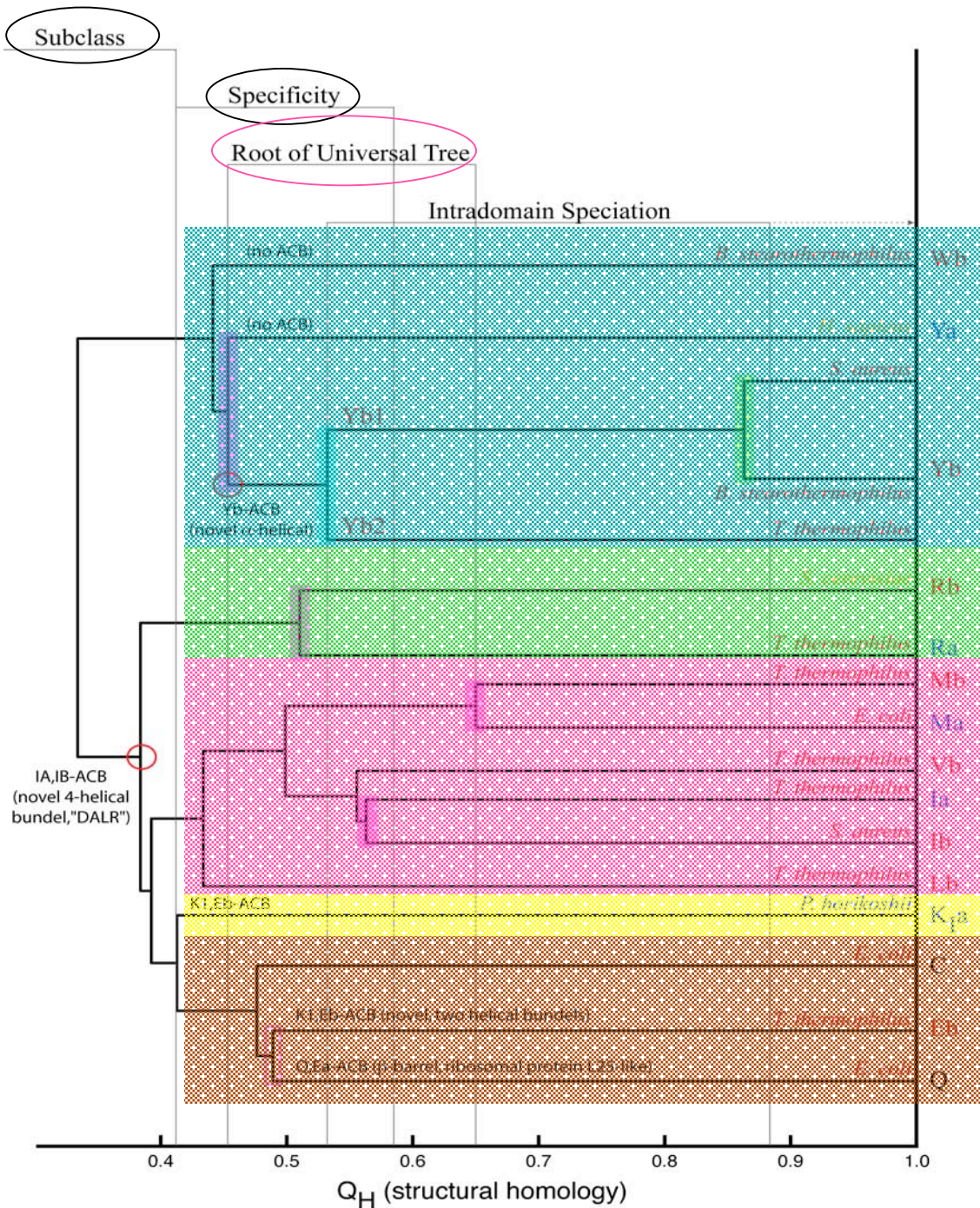
Encoding Structure
Rotated Cartesian + Gap = 4-space

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

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

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

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

$$Q^T_{(d)} A_{(d)} P = Q^T_{(d)} \begin{bmatrix} & & \end{bmatrix} \quad P = \tilde{R}_{(d)}$$

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

Subclass

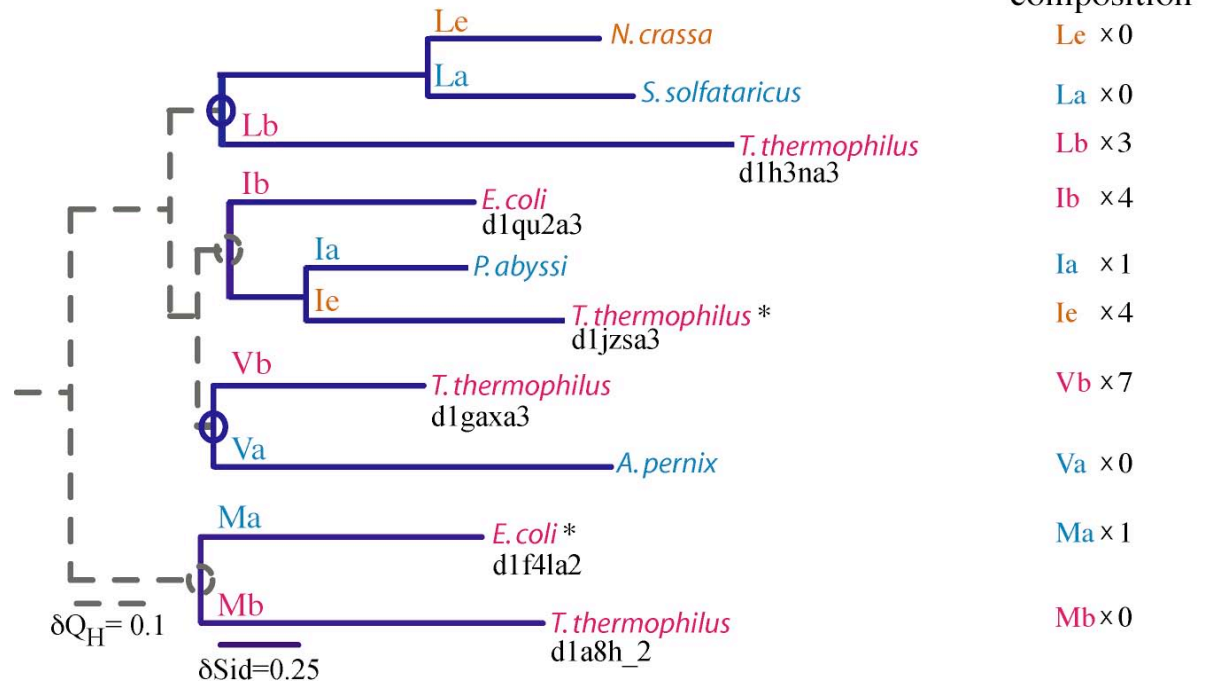Specificity

	T. thermophilus	Mb
	E. coli	Ma
	T. thermophilus	Vb
	T. thermophilus	Ia
	S. aureus	Ib
	T. thermophilus	Lb

IA,IB-ACB
(novel 4-helical
bundel,"DALR")

K1,Eb-ACB (novel, two helical bundels)

Q,Ea-ACB (β-barrel, ribosomal protein L25-like)

S. cerevisiae R?

T. thermophilus R

T. thermophilus M

E. coli M

T. thermophilus V

T. thermophilus L?

S. aureus I?

T. thermophilus L?

P. horikoshii K_Ia

E. coli C

T. thermophilus Eb

E. coli Q

Q_H (structural homology)

0.4   0.5   0.6   0.7   0.8   0.9   1.0

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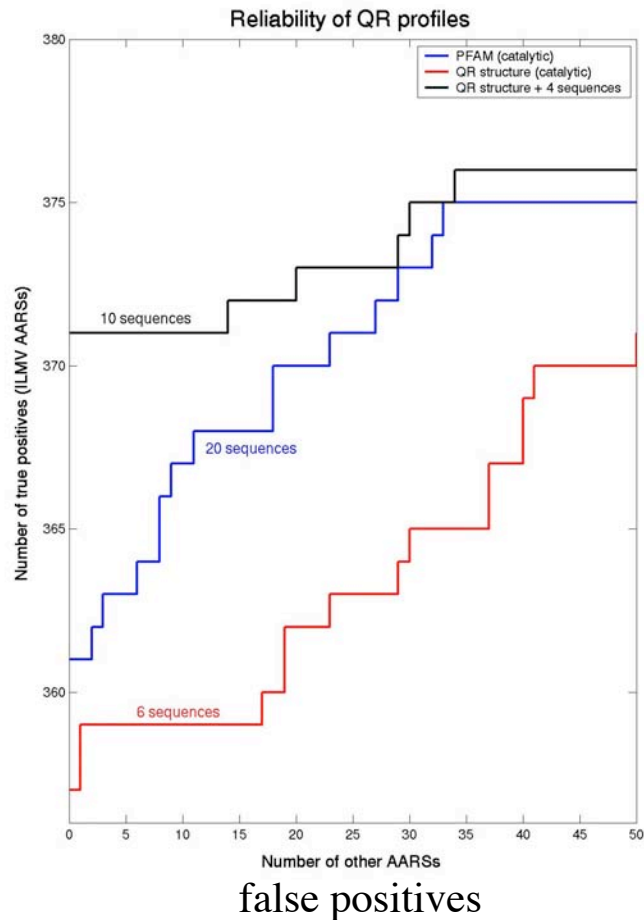
The composition of the profile matters.
Choosing the right 10 sequence makes all the difference.

Genome Annotation

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

AsnRS
GlnRS } Indirect Mechanism

LysRS   Class I AARS

CysRS   ?

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

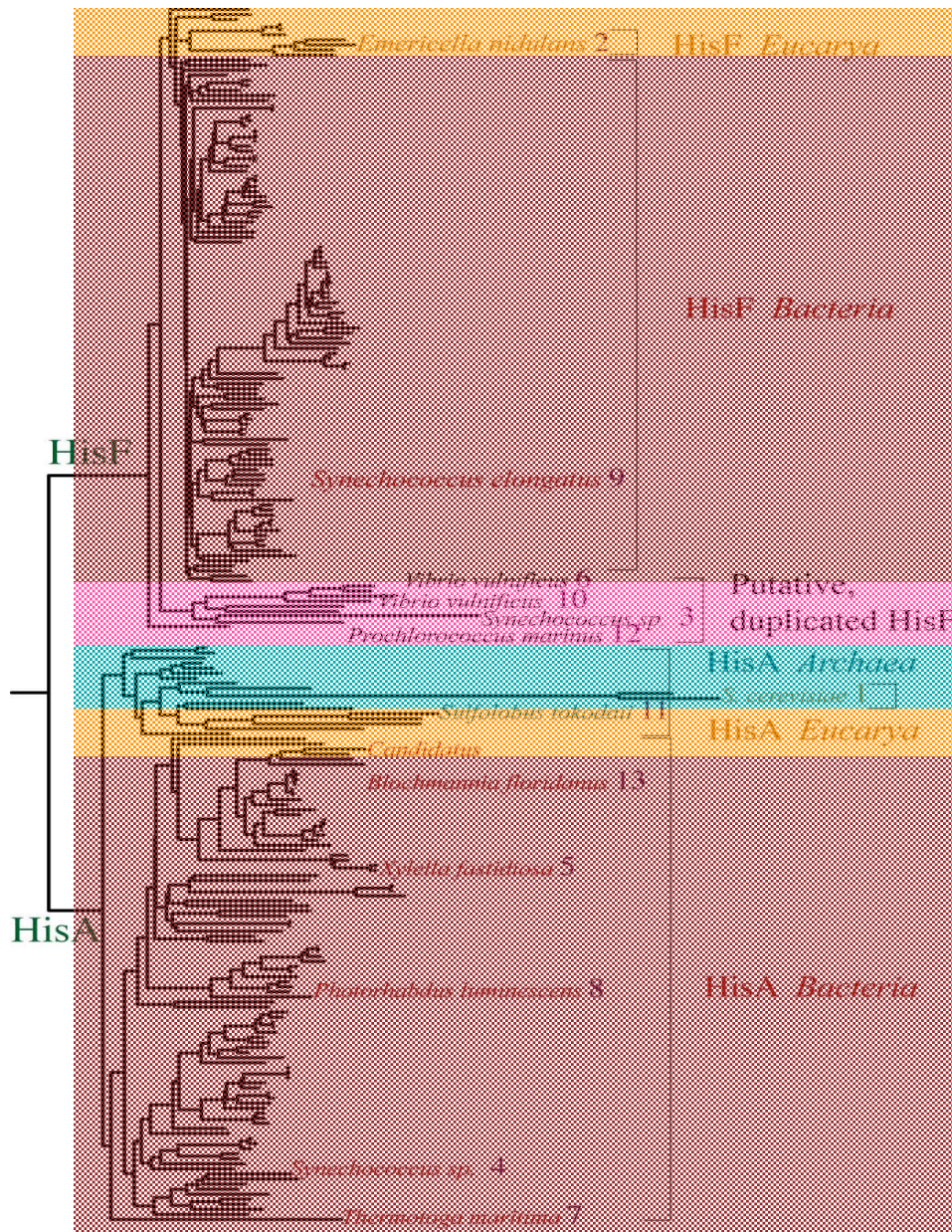| Protein	E-value
HisRS	1.1e-10
AspRS	1.9e-10
PheRS α-chain	9.5e-10
ThrRS	6.6e-04
ProRS	9.1e-03
SerRS	9.2e-03
putative CysRS	1.6e-02 ← MJ1660
AlaRS	5.1e-02
GlyRS	0.12
PheRS β-chain	0.15
DNA repair protein	7.5

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

Cysteine Biosynthesis in *Methanocaldococcus jannaschii*
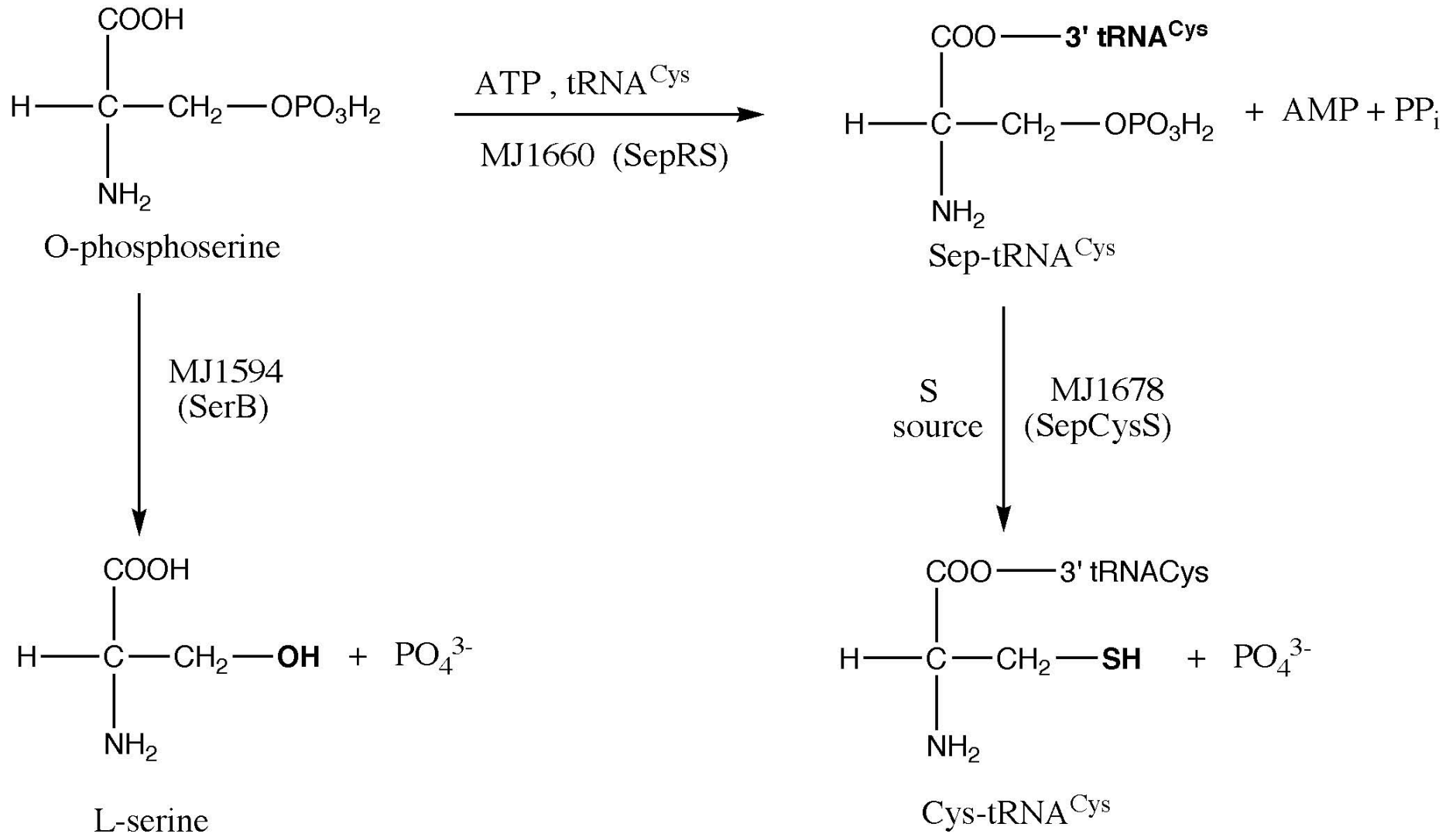
Sauerwald et al. Science 2005
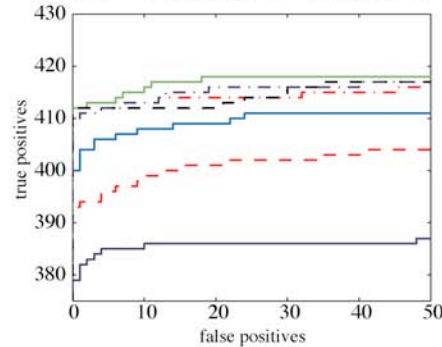
Evolutionary profile for HisA-HisF family
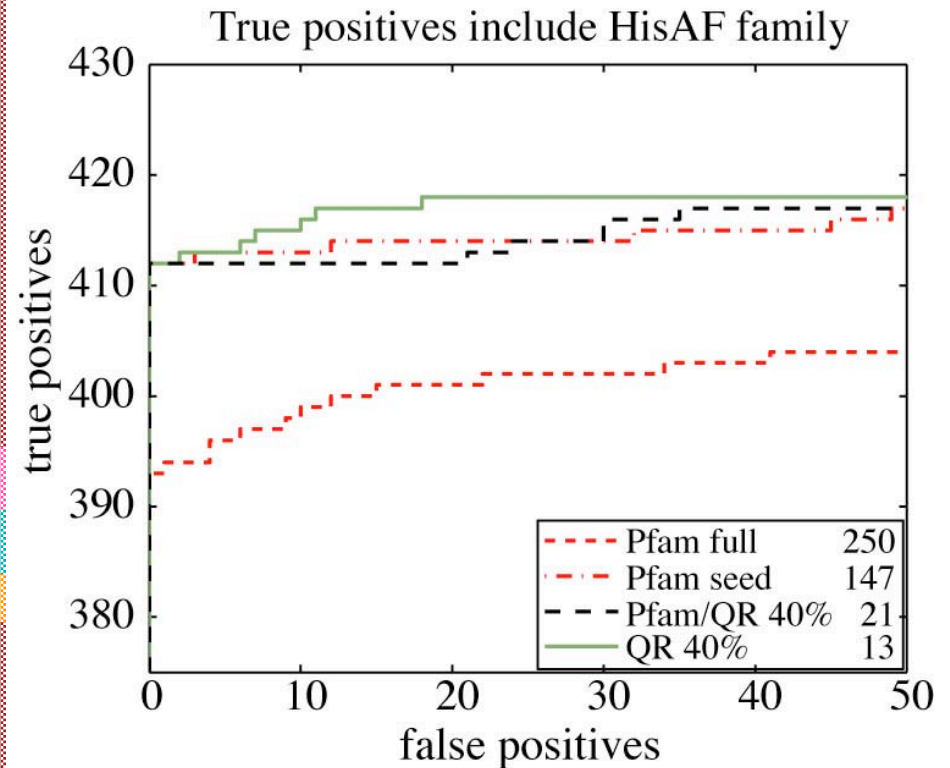
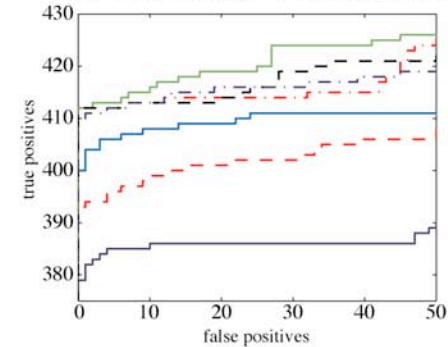EP outperforms popular profile methods with an economy of information.

Sethi, et. al., PNAS, 2005.

Economy of Information

How many sequences are needed for profiles?
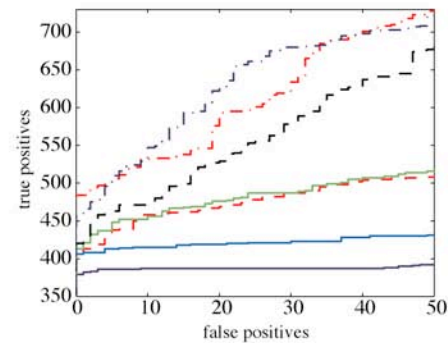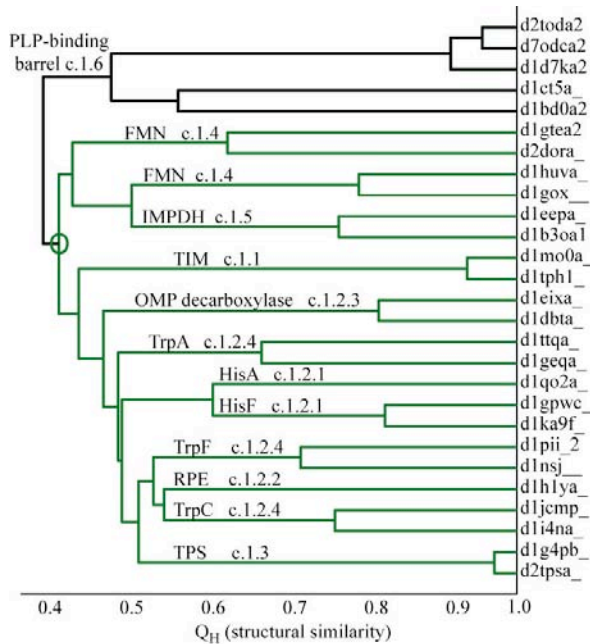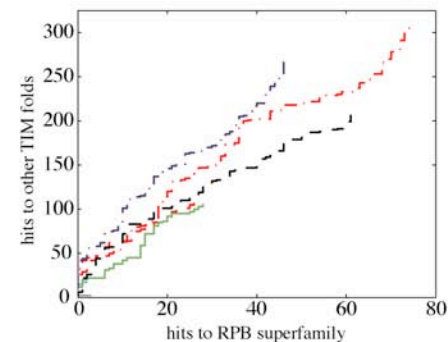
Phylogenetic relationship between TIM barrels
Found in database search with HisA-HisF profile

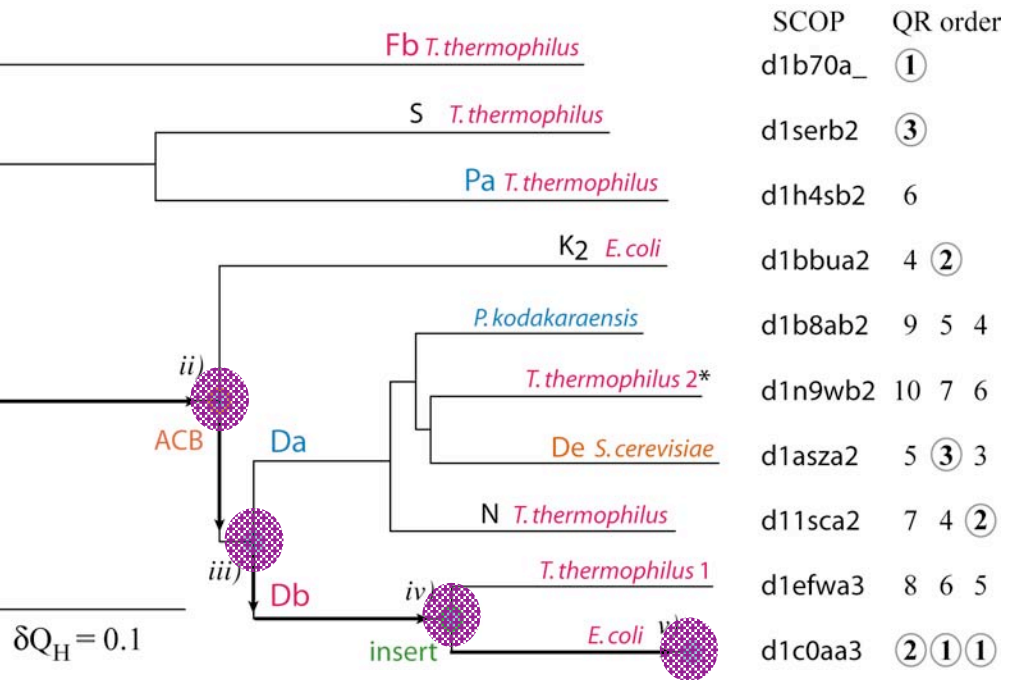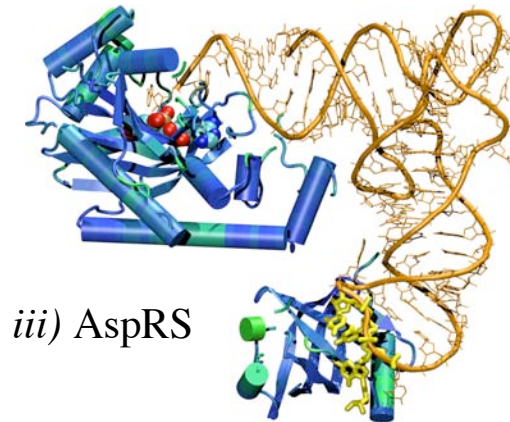|	
PLP-binding barrel c.1.6	d2toda2
	d7odca2
	d1d7ka2
	d1ct5a_
	d1bd0a2
FMN c.1.4	d1gtea2
	d2dora_
FMN c.1.4	d1huva_
	d1gox__
IMPDH c.1.5	d1eepa_
	d1b3oa1
TIM c.1.1	d1mo0a_
	d1tph1_
OMP decarboxylase c.1.2.3	d1eixa_
	d1dbta_
TrpA c.1.2.4	d1ttqa_
	d1geqa_
HisA c.1.2.1	d1qo2a_
HisF c.1.2.1	d1gpwc_
	d1ka9f_
TrpF c.1.2.4	d1pii_2
	d1nsj__
RPE c.1.2.2	d1h1ya_
TrpC c.1.2.4	d1jcmp_
	d1i4na_
TPS c.1.3	d1g4pb_
	d2tpsa_

0.4   0.5   0.6   0.7   0.8   0.9   1.0

Q_H (structural similarity)

Evolution of Structure and Function in AspRS

i) class II

ii) subclass IIB

anticodon
binding (ACB)
domain

	SCOP	QR order
Fb *T. thermophilus*	d1b70a_	①
S *T. thermophilus*	d1serb2	③
Pa *T. thermophilus*	d1h4sb2	6
K₂ *E. coli*	d1bbua2	4 ②
P. kodakaraensis	d1b8ab2	9 5 4
T. thermophilus 2*	d1n9wb2	10 7 6
De *S. cerevisiae*	d1asza2	5 ③ 3
N *T. thermophilus*	d1lsca2	7 4 ②
T. thermophilus 1	d1efwa3	8 6 5
E. coli	d1c0aa3	② ① ①

$\delta Q_H = 0.1$

iii) AspRS

iv) bacterial
AspRS

bacterial insert
domain

v) E. coli
AspRS

Unifying the Worlds of Sequence and Structure
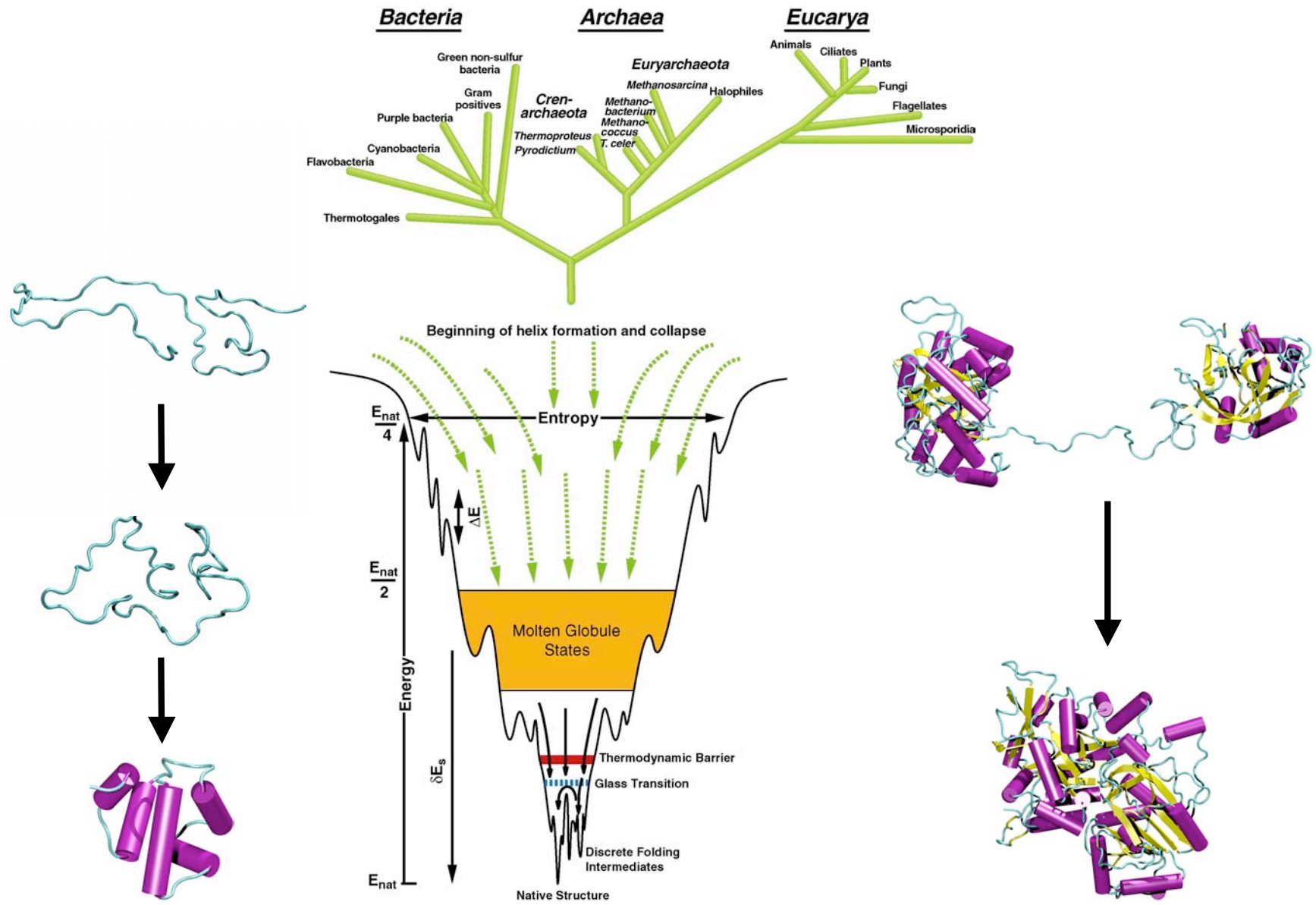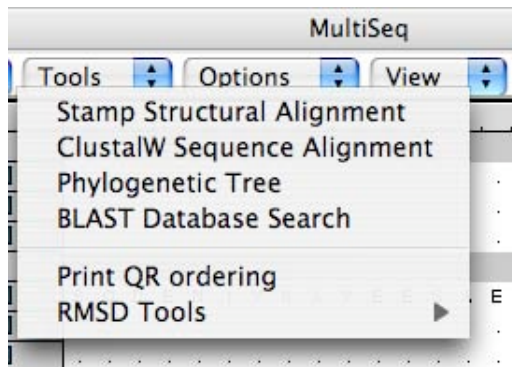
Multiseq in VMD : Merging the sequence and structure worlds

Version 1.83

2006 MultiSeq: New Features

Analyze the Evolution of Sequence and Structure

Eliminate Redundancy

Plus More Functions
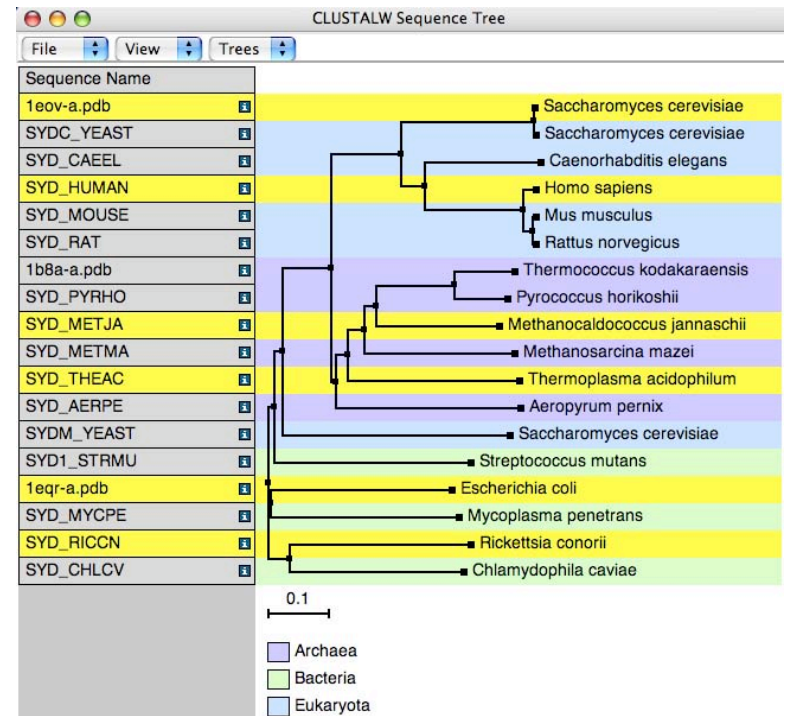
List of New Features in Multiseq

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

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)