

*Introduction to evolutionary concepts and
VMD/MultiSeq - Part I*

Characterizing molecular systems

Zaida (Zan) Luthey-Schulten

Dept. Chemistry, Physics, Beckman Institute, Institute of
Genomics Biology, & Center for Biophysics

Workshop June 2015, Pittsburgh

NIH Center Macromolecular Modeling and Bioinformatics

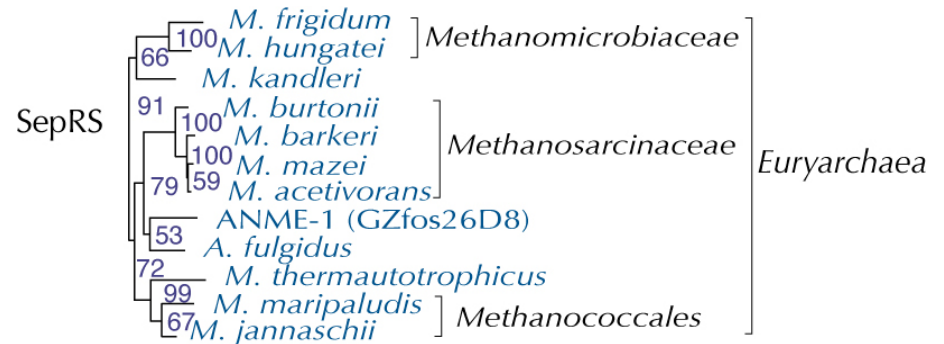
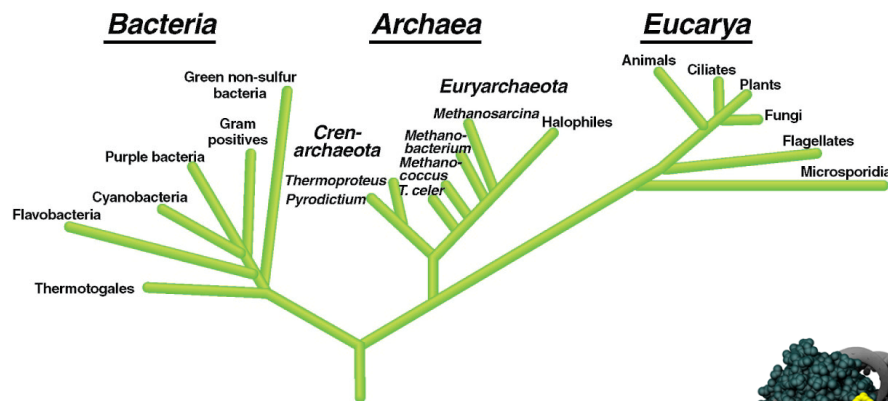


I L L I N O I S

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

VMD/MultiSeq - “A Tool to Think”

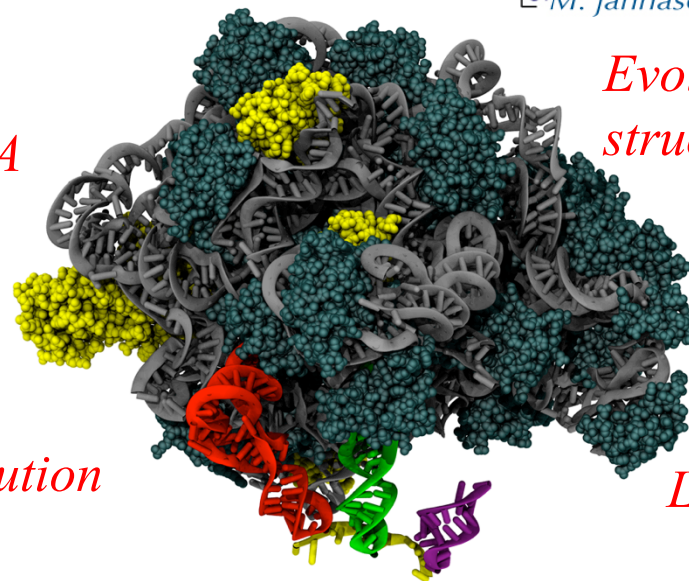
Carl Woese - “*VMD is far from a simple visualization tool for a biologist, it is a true thinking tool. Without it a whole class of biological hypotheses would simply not exist.*”



UPT - Woese 16S rRNA

Evolutionary profiles for protein structure & function prediction

Signatures ribosomal evolution



LSU (23S rRNA + rproteins)

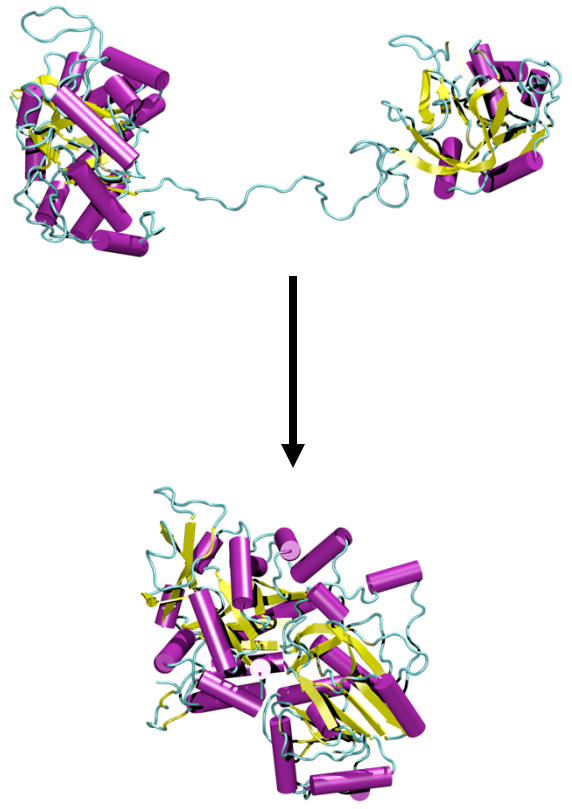
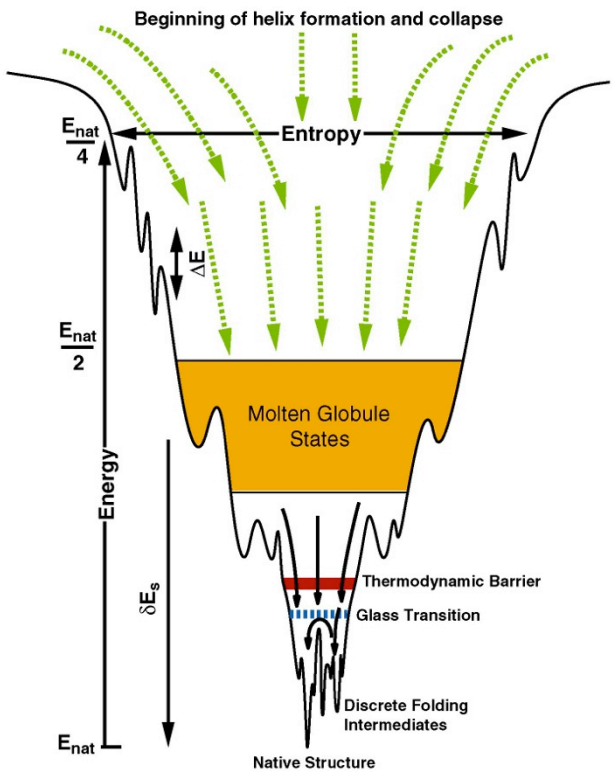
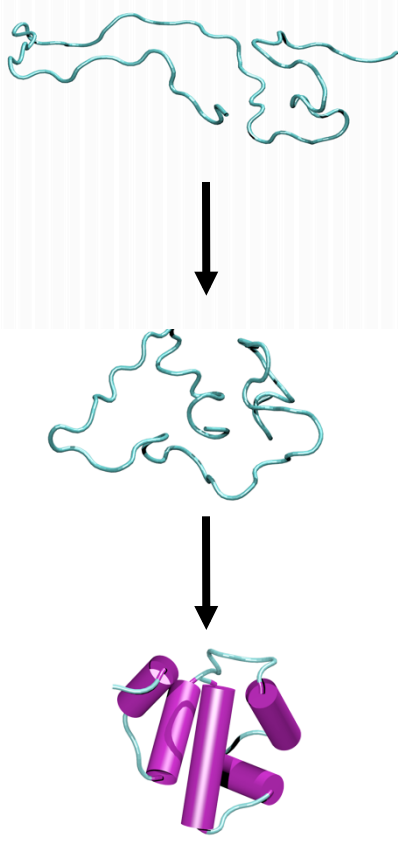
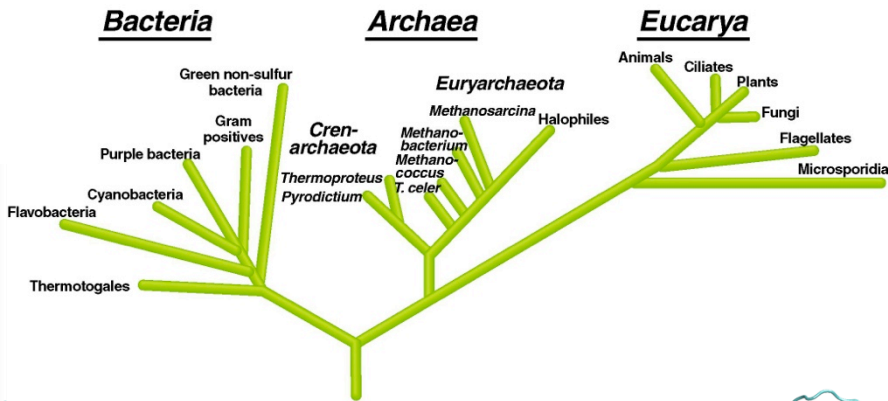
Why Look at More Than One Sequence?

1. Multiple Sequence Alignment shows patterns of conservation

Sequence Name	800	810	820	830																																						
<input checked="" type="checkbox"/> SYN_THEAC 357	S	Q	R	I	W	N	Y	D	E	L	M	Q	R	I	R	E	A	N	L	D	E	S	.	A	Y	Y	W	Y	V
<input checked="" type="checkbox"/> SYNC_CAEEL 473	S	M	R	I	W	K	E	D	Q	L	L	A	A	F	E	K	G	G	L	D	S	K	N	.	.	Y	Y	W	Y	M	
<input checked="" type="checkbox"/> SYNC_MOUSE 475	S	M	R	S	W	D	S	E	E	I	L	E	G	Y	K	R	E	G	I	D	P	A	P	.	.	Y	Y	W	Y	T		
<input checked="" type="checkbox"/> SYNC_DEBHA 480	S	M	R	T	Y	D	N	D	E	L	V	A	A	I	K	R	E	G	L	D	L	D	S	.	.	Y	Y	W	F	T			
<input checked="" type="checkbox"/> SYNC_YEAST 482	S	M	R	I	D	D	M	D	E	L	M	A	G	F	K	R	E	G	I	D	T	D	A	.	.	Y	Y	W	F	I		
<input checked="" type="checkbox"/> SYNC_HUMAN 476	S	M	R	I	F	D	S	E	E	I	L	A	G	Y	K	R	E	G	I	D	P	T	P	.	.	Y	Y	W	Y	T			
<input checked="" type="checkbox"/> SYK2_METMA 433	Y	S	E	L	N	D	P	L	E	Q	E	K	R	F	E	E	Q	D	K	K	R	K	L	G	D	L	E	A	Q	T	V	D	Y	D	F	I	
<input checked="" type="checkbox"/> SYK_HUMAN 499	Y	T	E	L	N	D	P	M	R	Q	R	Q	L	F	E	E	Q	A	K	A	K	A	A	G	D	D	E	A	M	F	I	D	E	N	F	C	
<input checked="" type="checkbox"/> SYK2_METAC 433	Y	S	E	L	N	D	P	L	E	Q	E	K	R	F	E	E	Q	D	K	K	R	K	L	G	D	L	E	A	Q	T	V	D	Y	D	F	I	
<input checked="" type="checkbox"/> SYK_MOUSE 497	Y	T	E	L	N	D	P	V	R	Q	R	Q	L	F	E	E	Q	A	K	A	K	A	A	G	D	D	E	A	M	F	I	D	E	N	F	C	
<input checked="" type="checkbox"/> SYK_CRIGR 499	Y	T	E	L	N	D	P	M	R	Q	R	Q	L	F	E	E	Q	A	K	A	K	A	A	G	D	D	E	A	M	F	I	D	E	N	F	C	
<input checked="" type="checkbox"/> SYK_ORYSA 524	Y	T	E	L	N	D	P	V	V	Q	R	Q	R	F	E	E	Q	L	K	D	R	Q	S	G	D	D	E	A	M	A	L	D	E	T	F	C	

2. Are these positions functionally important? Active sites, folding,..
3. What and how many sequences should be included?
4. Where do I find the sequences and structures for MS alignment?
5. How to generate pairwise and multiple sequence alignments?

Protein (RNA) Folding, Structure, & Function



New Tools in VMD/MultiSeq

Protein / RNA
Sequence Data

SwissProt DB (400K),
Greengenes RNA (100K)
Signatures, Zoom

Metadata Information,
Clustal, MAFFT &
Phylogenetic Trees

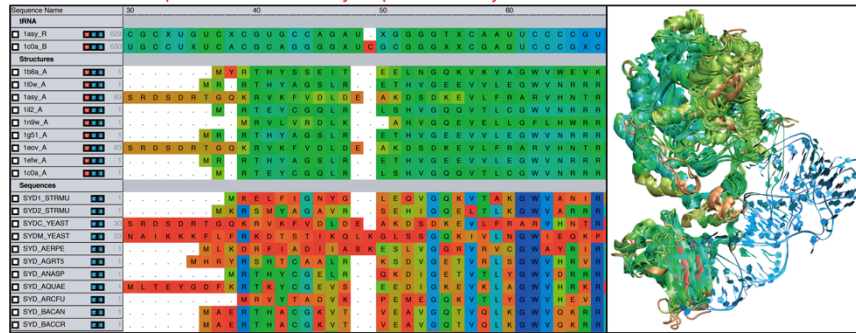
RAXml Trees,
Genomic Content,
Temperature DB

Blast & PsiBlast

Sequence Editor

View structural data colored by structural conservation and
sequence data colored by sequence identity

Synchronization between
1D and 3D views



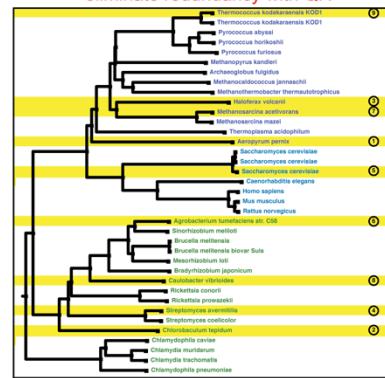
Sequence /Structure
Alignment

Protein & RNA
secondary structure

Group data by taxonomic classification

View sequence / structure phylogenies and
eliminate redundancy with QR

Sequence Name		90
Eukaryota:Fungi		
1asy_A	83	S R D S D R T G Q K R V K F V D
1teov_A	83	S R D S D R T G Q K R V K F V D
SYDC_YEAST	82	S R D S D R T G Q K R V K F V D
Eukaryota:Metazoa		
SYD_GAEL	57	G L V N S K E K K V L N F L K V
SYD_HUMAN	35	S M I Q S Q E K P D R V L V R V
SYD_MOUSE	35	S M I Q S Q E K P D R V L V R V
Archaea:Crenarcha		
SYD_AERPE	1 M L K D R F I A D I
Archaea:Euryarchaeota		
1n9w_A	1 M R V L V R D
1b8a_A	1 M Y R T H Y S S E
SYD_METMA	1	. . . M S L A N L R T H Y T A D
SYD_HALN1	1 M L E R T Y I E D
SYD_THEAC	1 M P R T Y I D T
SYD_PVRHO	1 M L R T H Y S N E
Bacteria:Proteobacteria		
110w_A	1 M R . R T H Y A G S
112_A	1 M . R T E Y C G Q

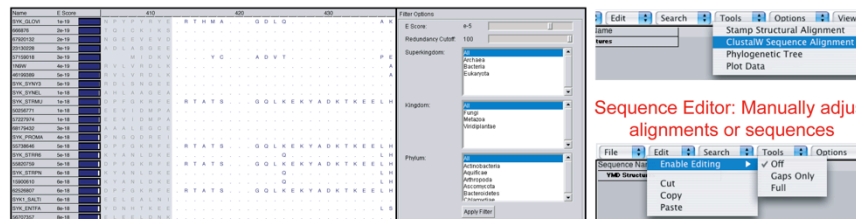


QR non-redundant
seq / str sets

Cluster
analysis /
Bioinformatics
scripting
Tutorials MultiSeq/
AARS
EF-Tu/Ribosome

Import data directly from BLAST databases

Align sequences with Clustal

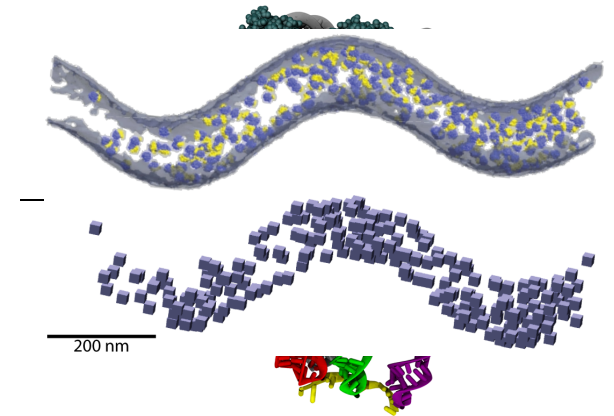
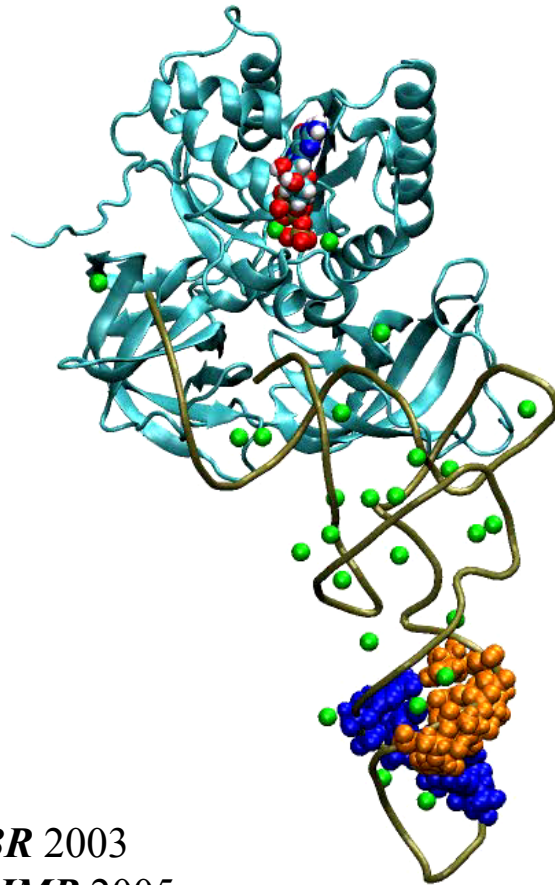
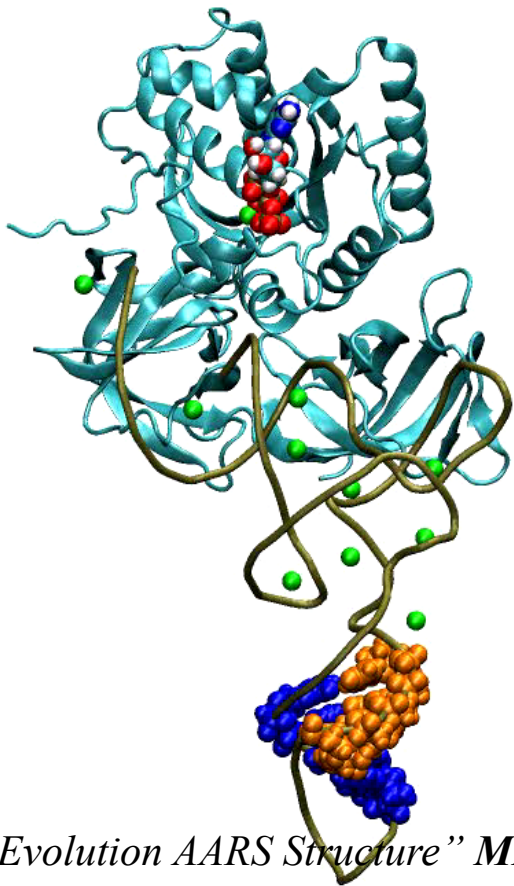


J. Eargle, D. Wright, Z. Luthey-Schulten, *Bioinformatics*, 22:504 (2006)

E. Roberts, J. Eargle, D. Wright, Z. Luthey-Schulten, *BMC Bioinformatics*, 7:382 (2006)

Protein:RNA Complexes in Translation

Evolutionary Analysis & Dynamics



“Evolution AARS Structure” *MMBR* 2003
 “Evol. Profiles Class I&II AARS” *JMB* 2005
 “Evolution SepRS/CysRS” *PNAS* 2005
 “Dynamic Signaling Network” *PNAS* 2009
 “Exit Strategy Charged tRNA” *JMB* 2010
 “Mistransl. in Mycoplasma” *PNAS* 2011
 “Capture & Selection of ATP” *JACS* 2013

“Dynamical Recognition & tRNA
 Dynamics” *JMB* 2008, *FEBS* 2010
Network Viewer, Bioinf., JCTC 2012

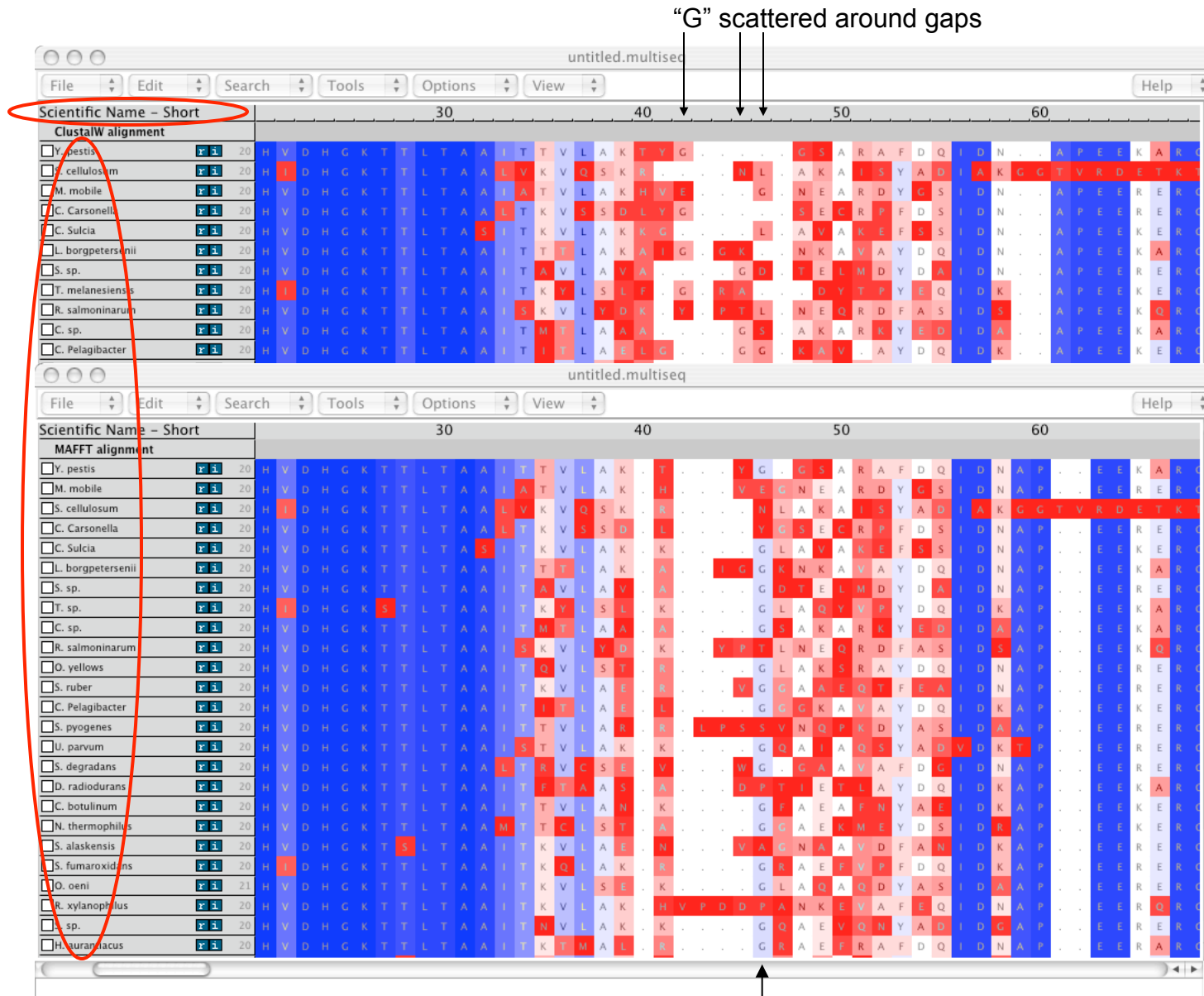
r-Proteins/r-RNA
Ribosome LSU
 “Signatures ribosomal evolution”
PNAS 2008, *BMC* 2009, *BJ* 2010
 “Motion L1 Stalk:tRNA” *JMB* 2010,
 “Ribosome Biogenesis” *JPC* 2012,3
 “Whole cell simulations on GPUs”
IEEE 2009, *Plos CB* 2011, *PRL* 2011,
JCC 2013, *PNAS* 2013,
PRL 2013, *CSB* 2013
Nature 2014

Basic principles of evolutionary analysis for proteins & RNAs

- Comparative analysis of sequences and **structures**
- Multiple sequence alignments (**gaps and editing**)
- Sequence and **structure** phylogenetic trees*
- Reference to 16S rRNA tree
- Horizontal or lateral gene transfer events
- Genomic context
- Evolutionary profiles representing diversity
- Conservation analysis of evolutionary profiles

*Various models of evolutionary change

Alignment of ~200 EF-Tu sequences in VMD/MultiSeq



“Classic”
ClustalW
alignment
~5 minutes

MAFFT7*
alignment

~ 30 seconds
More sequences!

<http://www.clustal.org/clustal2/>

* MAFFT v7.221, Katoh and Standley, Mol.Biol and Evol. 2015

Sequence Alignment & Dynamic Programming

number of possible alignments:

Seq. 1: $a_1 a_2 a_3 - - a_4 a_5 \dots a_n$
 Seq. 2: $c_1 - c_2 c_3 c_4 c_5 - \dots c_m$



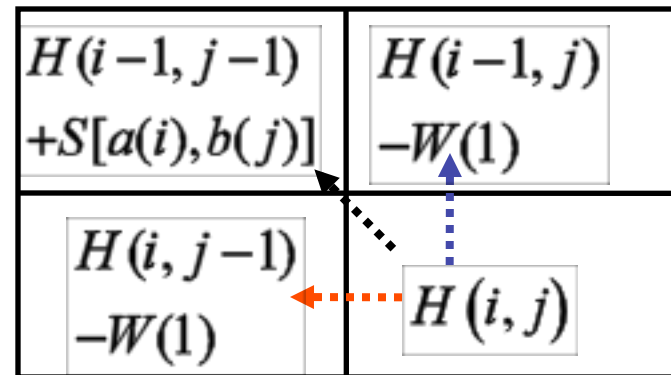
$$= \binom{2n}{n} = 2^{2n} (\sqrt{n\pi})^{-1}$$

Needleman-Wunsch alignment algorithm

$$H(i, j) = \text{MAX} \begin{cases} H(i-1, j-1) + S[a(i), b(j)] \\ H(i, j-k) - W(k), \\ H(i-m, j) - W(m) \end{cases}$$

S : substitution matrix

A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	
5	-2	-1	-1	-2	0	-1	1	-2	-1	-2	-1	-1	-3	-2	1	0	-3	-2	0	-1	-1	0	A
-2	9	0	-1	-3	2	-1	-3	0	-3	-2	3	-1	-2	-3	-1	-2	-2	-1	-2	-1	0	-1	R
-1	0	8	2	-2	1	-1	0	1	-2	-3	0	-2	-3	-2	1	0	-4	-2	-3	4	0	-1	N
-1	-1	2	9	-2	-1	2	-2	0	-4	-3	0	-3	-4	-2	0	-1	-5	-3	-3	6	1	-1	D
-2	-3	-2	-2	16	-4	-2	-3	-4	-4	-2	-3	-3	-2	-5	-1	-1	-6	-4	-2	-2	-3	-2	C
0	2	1	-1	-4	8	2	-2	0	-3	-2	1	-1	-4	-2	1	-1	-1	-3	0	4	-1	Q	
-1	-1	-1	2	-2	2	7	-3	0	-4	-2	1	-2	-3	0	0	-1	-2	-2	-3	1	5	-1	E
1	-3	0	-2	-3	-2	-3	8	-2	-4	-4	-2	-2	-3	-1	0	-2	-2	-3	-4	-1	-2	-1	G
-2	0	1	0	-4	0	0	-2	13	-3	-2	-1	1	-2	-2	-1	-2	-5	2	-4	0	0	-1	H
-1	-3	-2	-4	-4	-3	-4	-4	-3	6	2	-3	1	1	-2	-2	-1	-3	0	4	-3	-4	-1	I
-2	-2	-3	-3	-2	-2	-2	-4	-2	2	6	-2	3	2	-4	-3	-1	-1	0	2	-3	-2	-1	L
-1	3	0	0	-3	1	1	-2	-1	-3	-2	6	-1	-3	-1	0	0	-2	-1	-2	0	1	-1	K
-1	-1	-2	-3	-3	-1	-2	-2	1	1	3	-1	7	0	-2	-2	-1	-2	1	1	-3	-2	0	M
-3	-2	-3	-4	-2	-4	-3	-3	-2	1	2	-3	0	9	-4	-2	-1	1	4	0	-3	-4	-1	F
-2	-3	-2	-2	-5	-2	0	-1	-2	-2	-4	-1	-2	-4	11	-1	0	-4	-3	-2	-1	-2	0	P
1	-1	1	0	-1	1	0	0	-1	-2	-3	0	-2	-2	-1	5	2	-5	-2	-1	0	0	0	S
0	-2	0	-1	-1	-1	-2	-2	-1	-1	0	-1	-1	0	2	6	-4	-1	1	0	-1	0	-1	T
-3	-2	-4	-5	-6	-1	-2	-2	-5	-3	-1	-2	-2	1	-4	-5	4	19	3	-3	-4	-2	-2	W
-2	-1	-2	-3	-4	-1	-2	-3	2	0	0	-1	1	4	-3	-2	-1	3	9	-1	-3	-2	-1	Y
0	-2	-3	-3	-2	-3	-3	-4	-4	4	2	-2	1	0	-3	-1	1	-3	-1	5	-3	-3	-1	V
-1	-1	4	6	-2	0	1	-1	0	-3	-3	0	-3	-3	-2	0	0	-4	-3	-3	5	2	-1	B
-1	0	0	1	-3	4	5	-2	0	-4	-2	1	-2	-4	-1	0	-1	-2	-2	-3	2	5	-1	Z
0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	0	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-1	X



Score Matrix H: Traceback

gap penalty $W = -6$

Reference: "Biological Sequence Analysis - Probabilistic Models of Proteins and Nucleic Acids" R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, Cambridge U. P. London, 1998; pp. 19-22 (see also other sections)

Needleman-Wunsch Global Alignment

Similarity Values

		M	G	K	P
M		5	-3	-1	-2
G		-3	6	-2	-2
P		-2	-2	-1	7
K		-1	-2	5	-1
K		-1	-2	5	-1
P		-2	-2	-1	7

Initialization of Gap Penalties

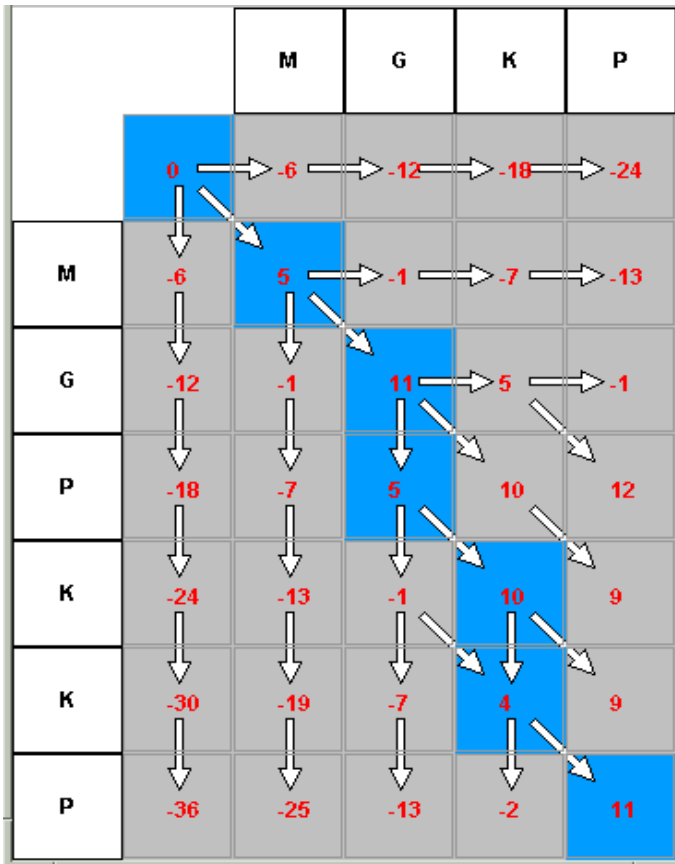
		M	G	K	P	
		0	-6	-12	-18	-24
M		-6	5	-3	-1	-2
G		-12	-3	6	-2	-2
P		-18	-2	-2	-1	7
K		-24	-1	-2	5	-1
K		-30	-1	-2	5	-1
P		-36	-2	-2	-1	7

Filling out the Score Matrix H

		M	G	K	P	
		0	-6	-12	-18	-24
M		-6	5	-1	-7	-13
G		-12	-1	11	-2	-2
P		-18	-2	-2	-1	7
K		-24	-1	-2	5	-1
K		-30	-1	-2	5	-1
P		-36	-2	-2	-1	7

		M	G	K	P	
		0	-6	-12	-18	-24
M		-6	5	-1	-7	-13
G		-12	-1	11	5	-1
P		-18	-7	5	10	12
K		-24	-13	-1	10	9
K		-30	-19	-7	4	9
P		-36	-25	-13	-2	11

Traceback and Alignment



The Alignment

M	G	-	K	-	P
:	:		:		:
M	G	P	K	K	P

Traceback (blue) from optimal score

STAMP - Multiple Structural Alignments

1. Initial Alignment Inputs

- Multiple Sequence alignment
- Ridged Body “Scan”
- Pairwise Alignments and Hierarchical Clustering

2. Refine Initial Alignment & Produce Multiple Structural Alignment

$$P_{ij} = \left\{ e^{-d_{ij}^2/2E_1} \right\} \left\{ e^{-s_{ij}^2/2E_2} \right\}$$

probability that residue i on structure A is equivalent to residue j on structure B .

d_{ij} —distance between i & j

s_{ij} —conformational similarity; function of rms between $i-1, i, i+1$ and $j-1, j, j+1$.

- Dynamic Programming (Smith-Waterman) through P matrix gives optimal set of equivalent residues.
- This set is used to re-superpose the two chains. Then iterate until alignment score is unchanged.
- This procedure is performed for all pairs with no gap penalty

Multiple Structural Alignments

STAMP – cont' d

2. Refine Initial Alignment & Produce Multiple Structural Alignment

Alignment score:

$$S_C = \frac{S_p}{L_p} \frac{L_p - i_A}{L_A} \frac{L_p - i_B}{L_B}$$

$$S_p = \sum_{\text{aln.path}} P_{ij}$$

L_p, L_A, L_B – length of alignment, sequence A, sequence B

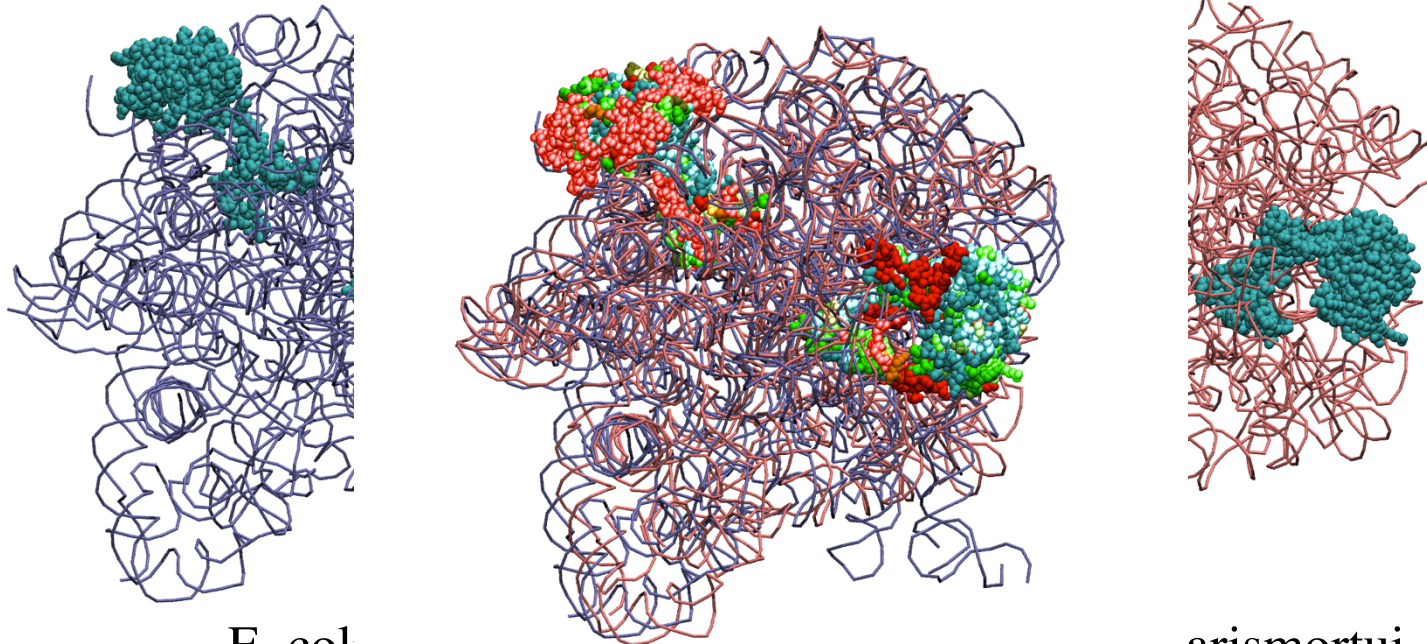
i_A, i_B – length of gaps in A and B.

Multiple Alignment:

- Create a dendrogram using the alignment score.
- Successively align groups of proteins (from branch tips to root).
- When 2 or more sequences are in a group, then average coordinates are used.

Structural Overlaps - STAMP

Ribosome large subunit showing ribosomal proteins L2 and L3
180,000 atoms in 4 rRNAs and 58 proteins



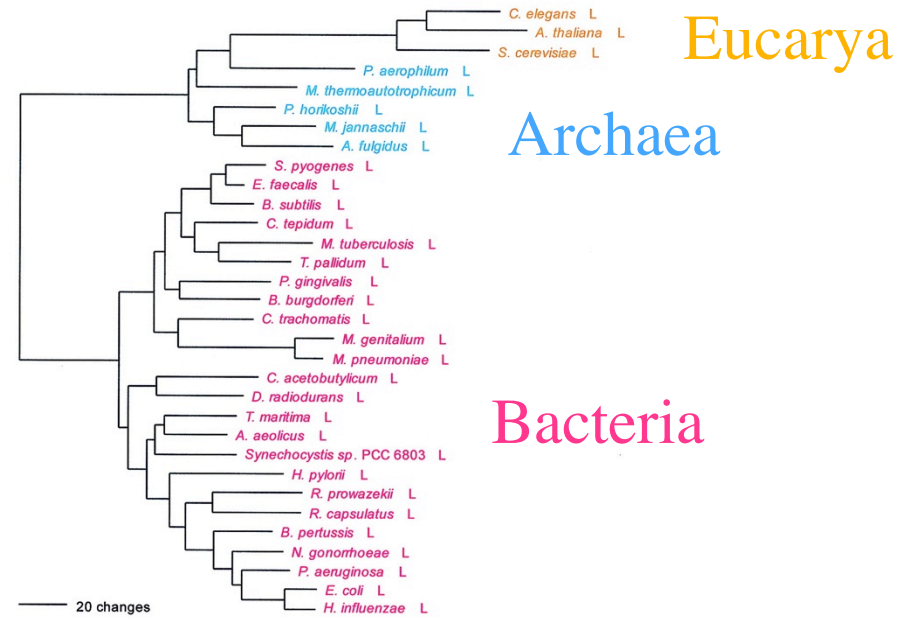
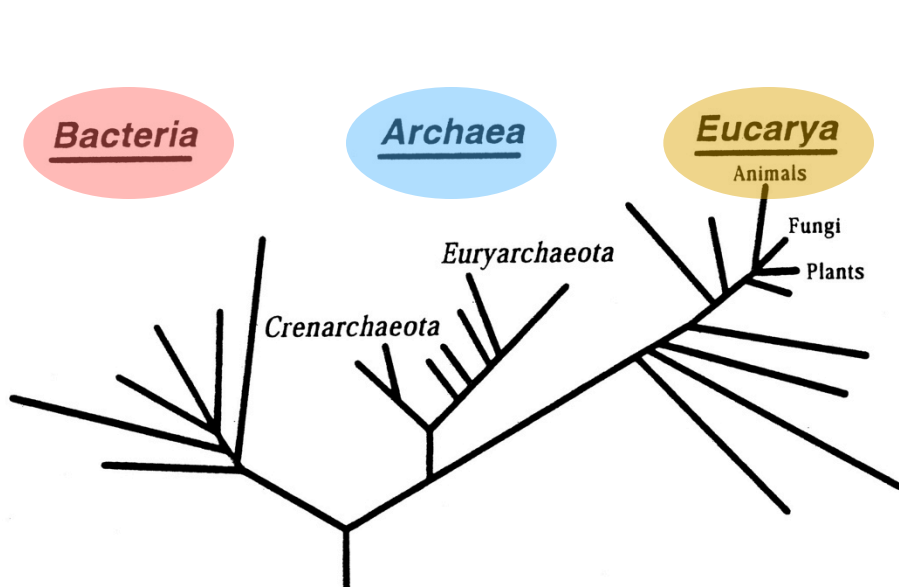
E. coli

Thermoplasma arismortui

Sequence Name	50	60	70	80	90																																						
23S rRNA																																											
<input type="checkbox"/> 2aw4_B v r i 49	A	U	G	A	A	G	G	A	C	G	U	G	C	U	A	A	U	C	U	G	C	G	A	U	A	A	G	C	G	U	C	G	G	U	A	A	G	G	U	G	A	U	
<input type="checkbox"/> 1s72_0 v r i 58	C	A	A	G	C	U	G	C	G	A	U	A	A	G	C	C	A	U	G	G	G	G	A	G	C	C	G	C	A	C	G	G	A	G	G	C	G	A	A	G	A	A	
5S rRNA																																											
<input checked="" type="checkbox"/> 2aw4_A v r i 48	U	C	A	G	A	A	G	U	G	A	A	A	C	G	C	C	G	U	A	G	C	G	C	C	G	A	U	G	G	.	.	.	U	A	G	U	G	U	G	G	G	.	
<input checked="" type="checkbox"/> 1s72_9 v r i 47	A	C	G	G	A	A	G	A	U	A	A	A	G	C	C	C	A	C	C	A	G	C	G	U	U	C	C	G	G	G	G	A	G	U	A	C	U	G	G	A	G	U	G
Ribosomal Protein L2																																											
<input type="checkbox"/> 2aw4_C v r i 41	.	.	.	G	R	N	N	N	G	R	I	T	T	R	H	I	G	G	G	H	K	Q	A	Y	R	I	V	.	D	F	K	R	N	K	.	D	.	.	G	I	P	A	
<input type="checkbox"/> 1s72_A v r i 11	R	G	T	S	T	F	R	A	.	.	P	S	H	R	Y	K	A	D	L	E	H	R	K	V	E	D	G	D	V	I	A	G		
Ribosomal Protein L3																																											
<input type="checkbox"/> 2aw4_D v r i 11	M	T	R	I	F	T	E	D	G	V	S	I	P	V	T	V	I	E	V	E	A	N	R	V	T	Q	V	K	.	.	.			
<input type="checkbox"/> 1s72_B v r i 49	T	H	V	V	L	V	N	D	E	P	N	S	P	R	E	G	M	E	E	T	.	V	P	V	T	V	I	E	T	P	P	M	R	A	V	A	L	R	A	Y	E	D	

Universal Phylogenetic Tree

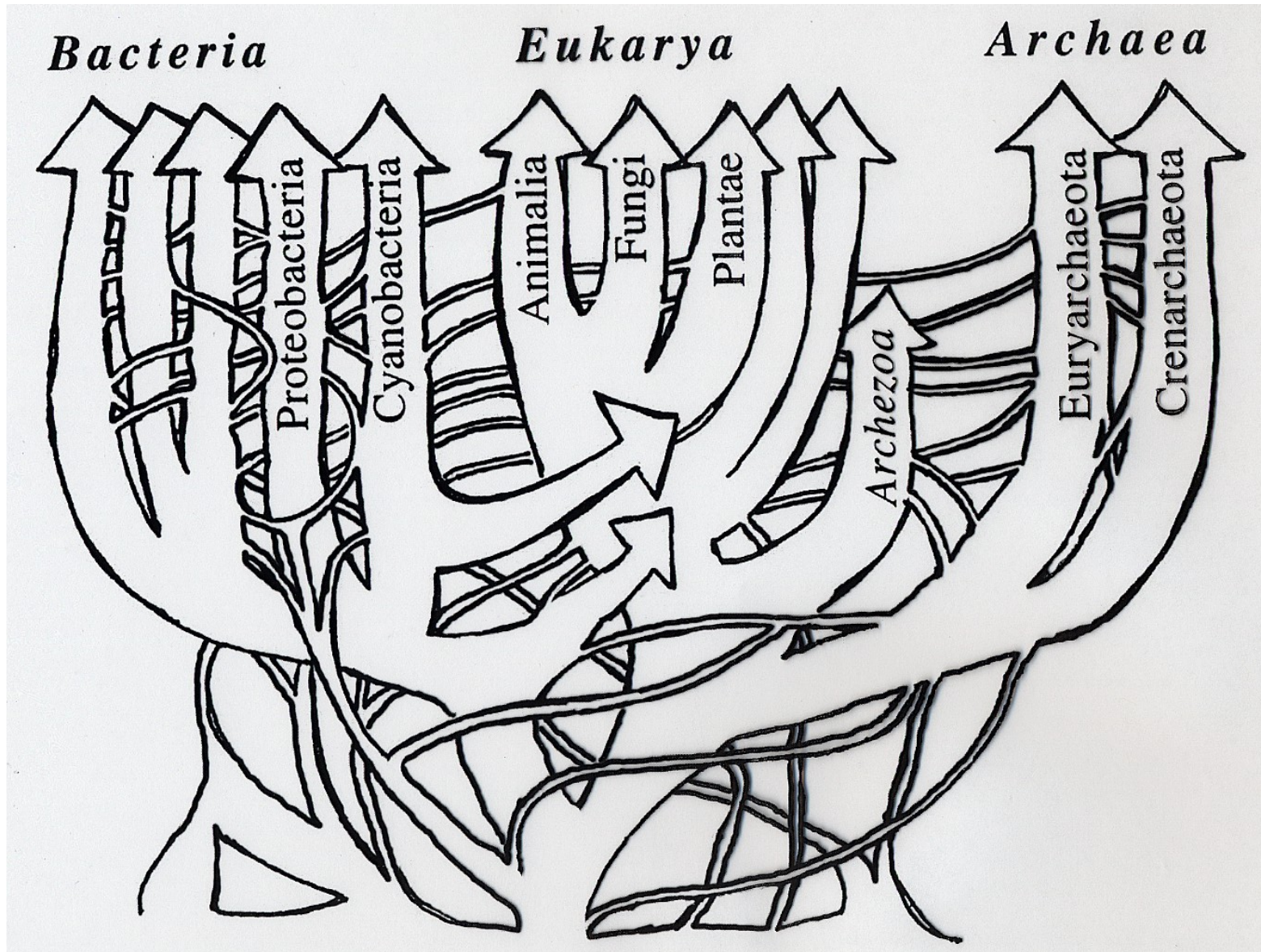
3 domains of life



Reference 16S rRNA tree

Leucyl-tRNA synthetase displays the full canonical phylogenetic distribution.

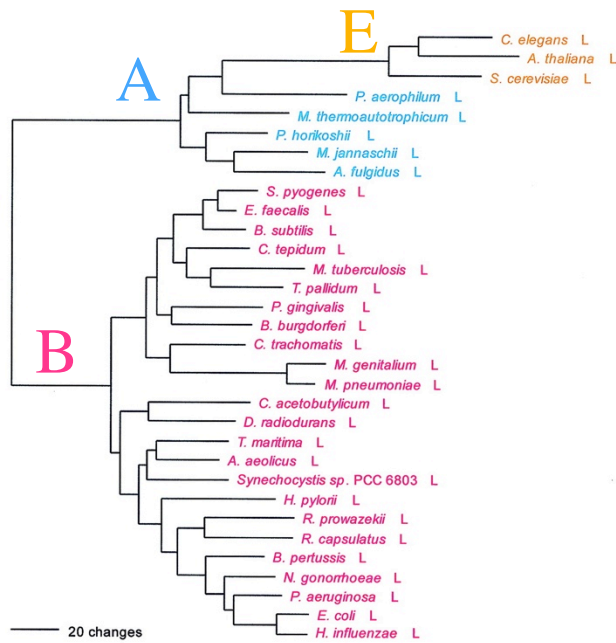
Look for horizontal gene transfer events



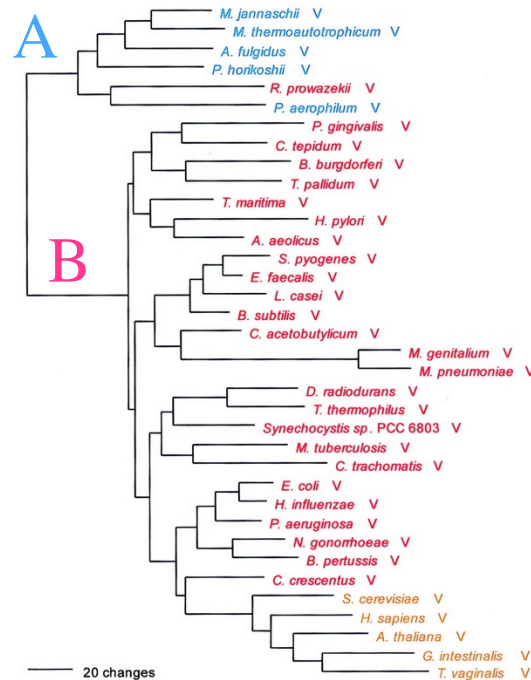
After W. Doolittle, modified by G. Olsen

Phylogenetic Distributions

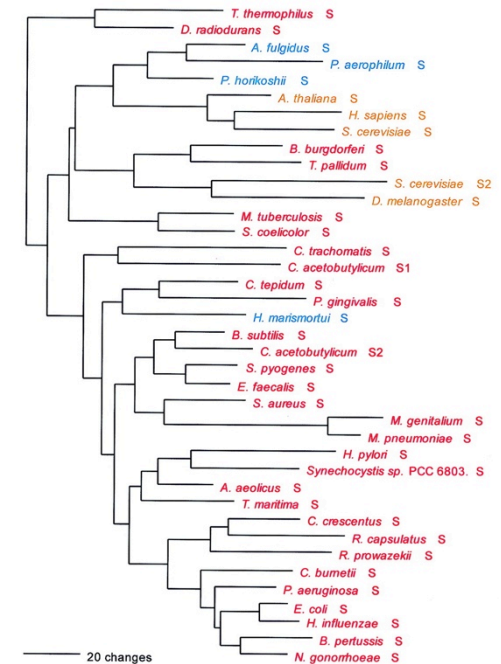
Full Canonical



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

Woese, Olsen, Ibba, Soll *MMBR* 2000

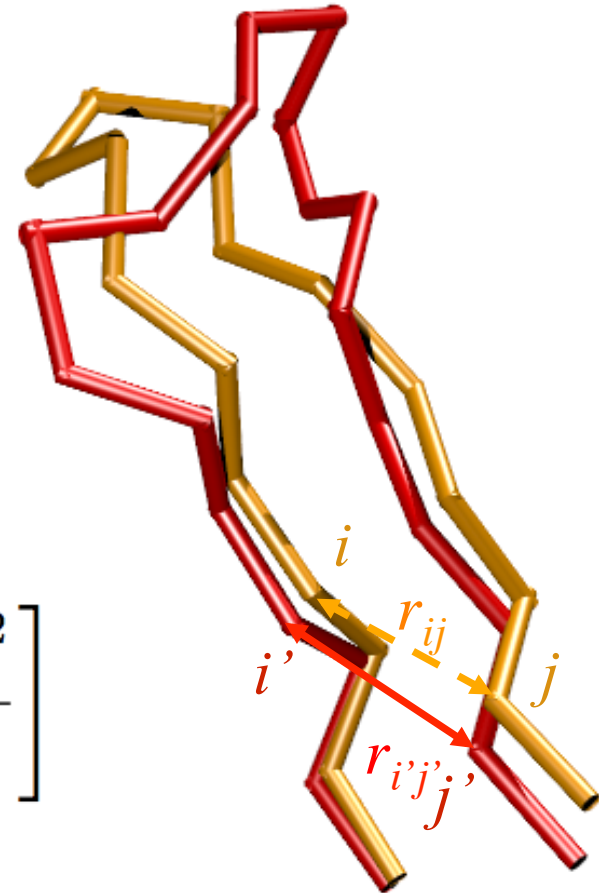
Protein Structure Similarity Measure

Q_H Structural Homology

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

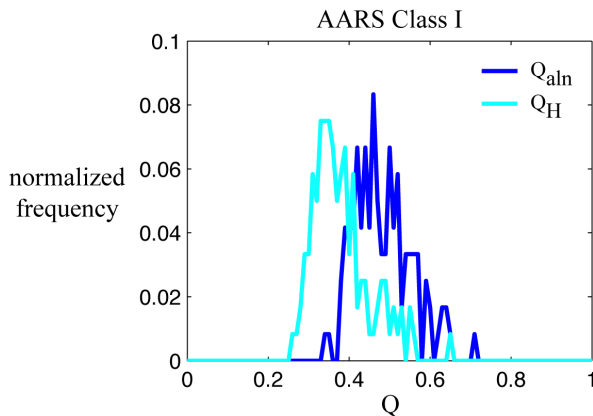
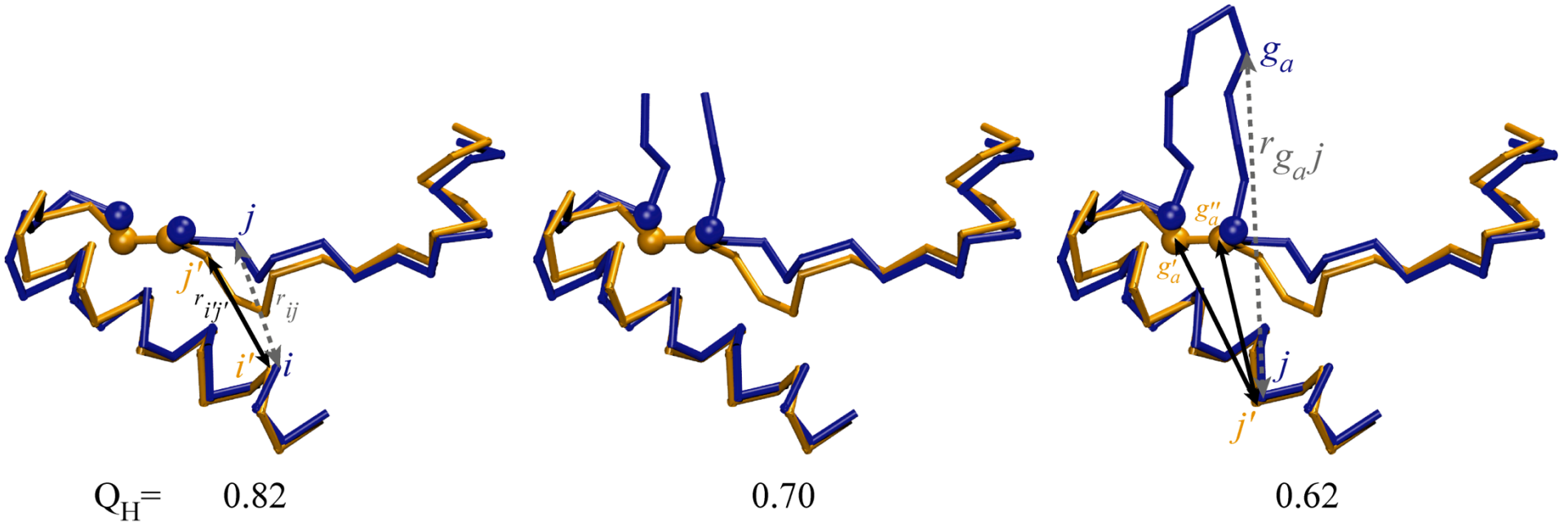
$$Q_H = \mathcal{N} [q_{aln} + q_{gap}]$$

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



Structural Similarity Measure: The effect of insertions

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

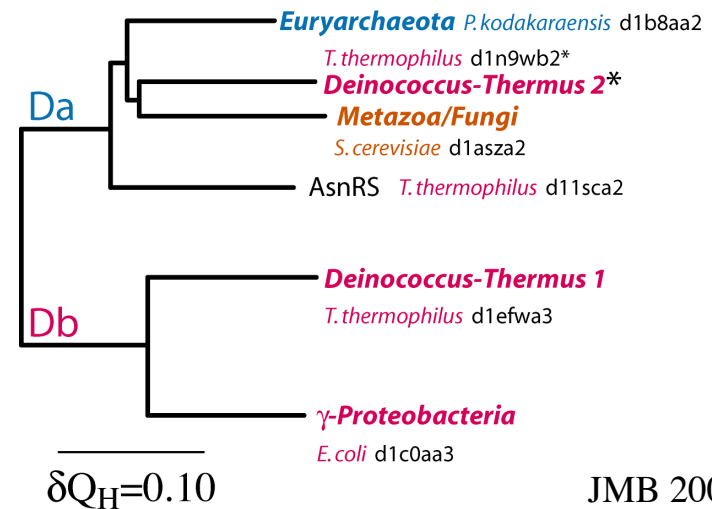
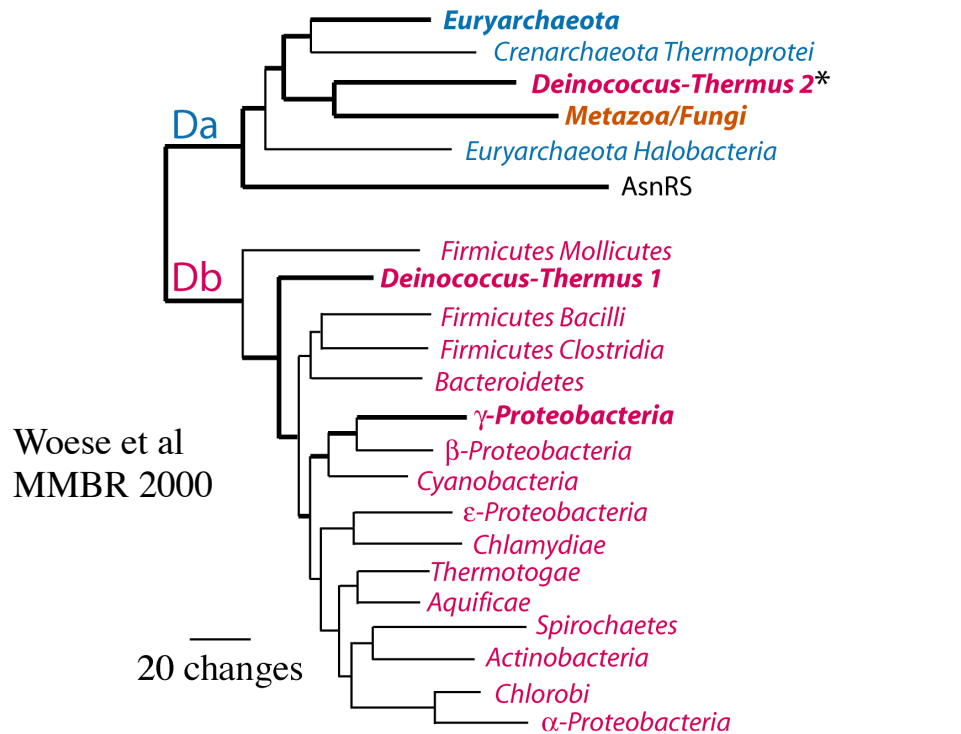


$$\begin{aligned}
 q_{gap} = & \sum_{g_a} \sum_j^{N_{aln}} \max \left\{ \exp \left[-\frac{(r_{g_a j} - r_{g'_a j'})^2}{2\sigma_{g_a j}^2} \right], \exp \left[-\frac{(r_{g_a j} - r_{g''_a j'})^2}{2\sigma_{g_a j}^2} \right] \right\} \\
 & + \sum_{g_b} \sum_j^{N_{aln}} \max \left\{ \exp \left[-\frac{(r_{g_b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

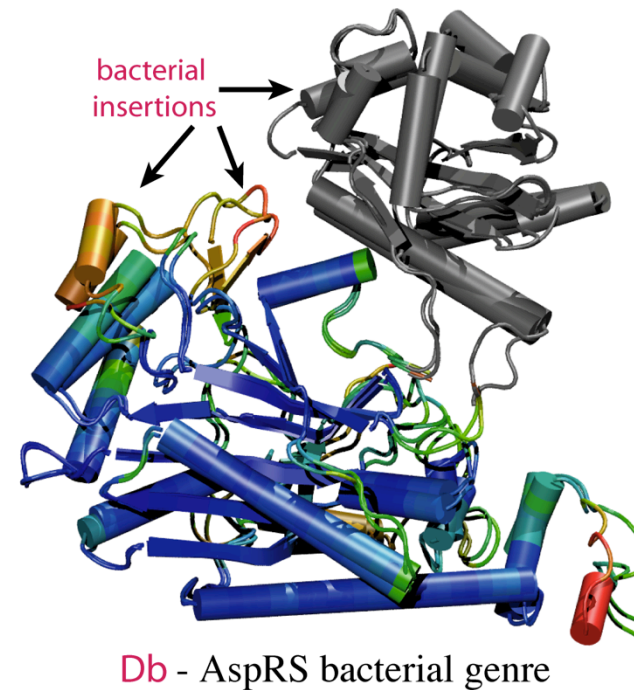
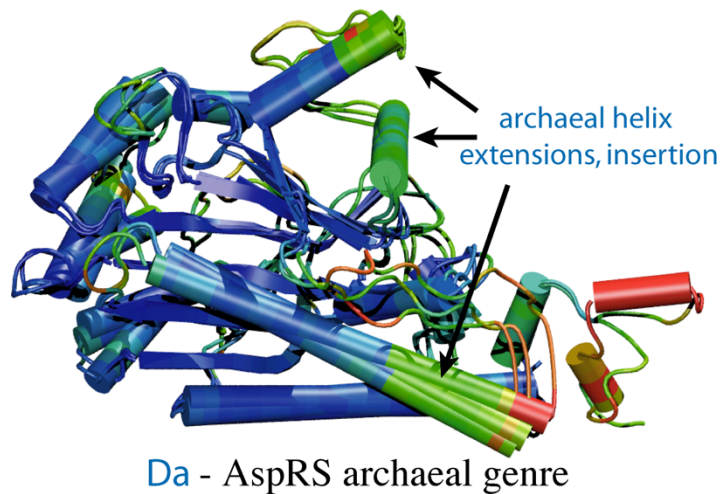
Structure encodes evolutionary information!

sequence-based phylogeny

structure-based phylogeny



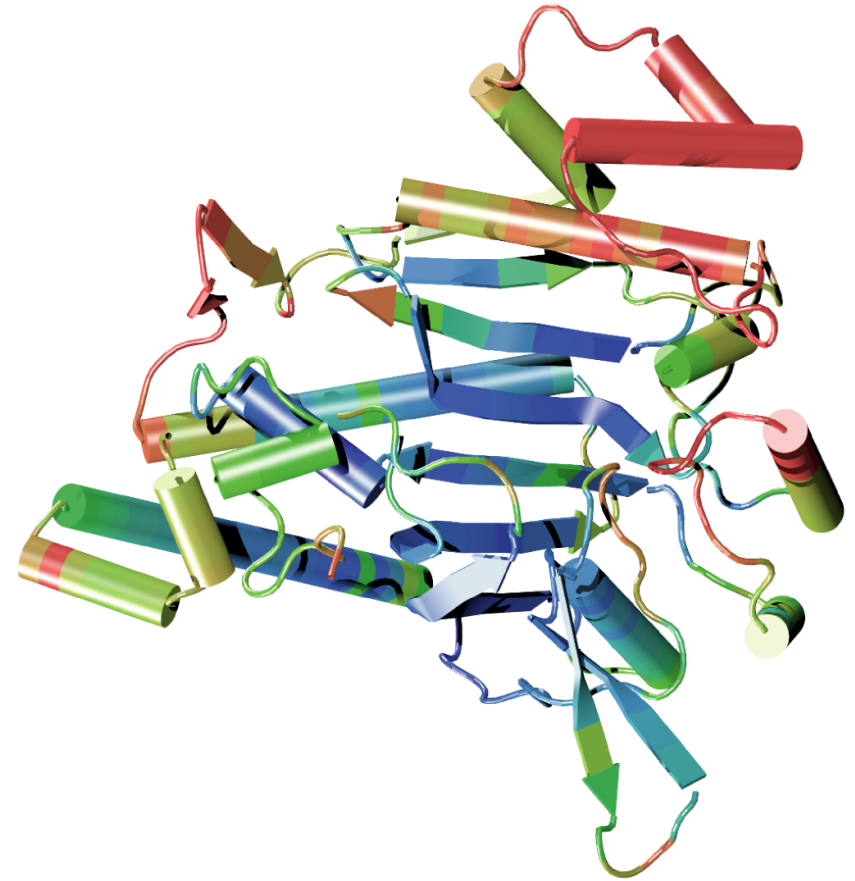
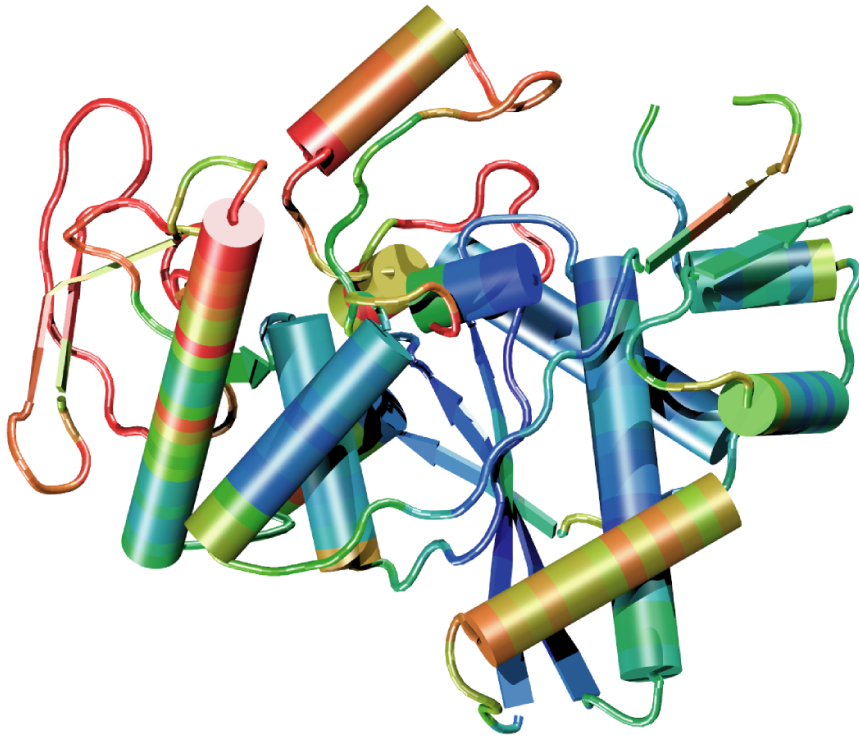
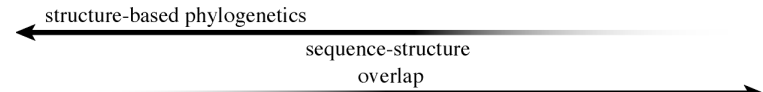
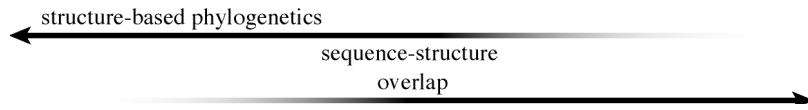
JMB 2005
MMBR 2003



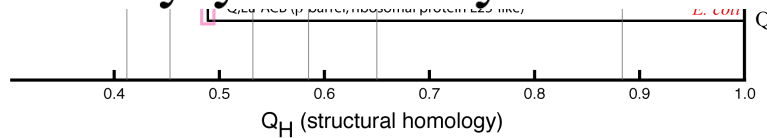
Structure reveals distant evolutionary events

Class I AARSs

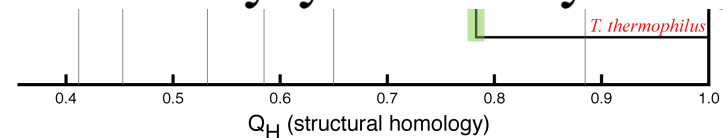
Class II AARSs



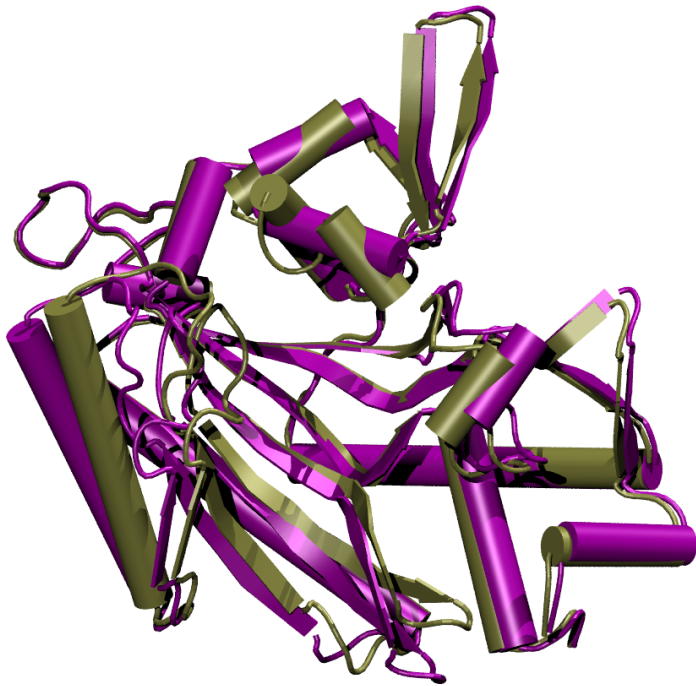
Class I Lysyl-tRNA Synthetase



Class II Lysyl-tRNA Synthetase



Sequences define more recent evolutionary events:



Conformational changes
in the same protein.

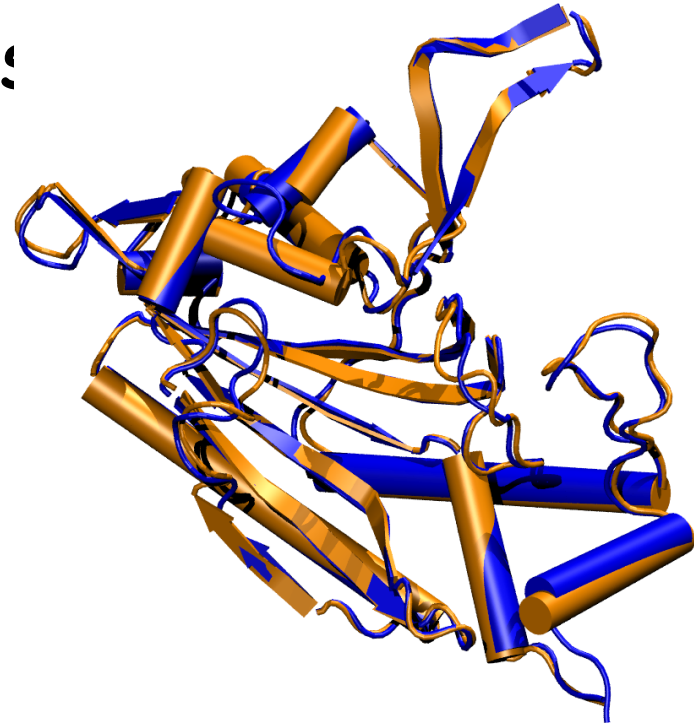
ThrRS

T-AMP analog, 1.55 Å.

T, 2.00 Å.

$Q_H = 0.80$

Sequence identity = 1.00



Structures for two
different species.

ProRS

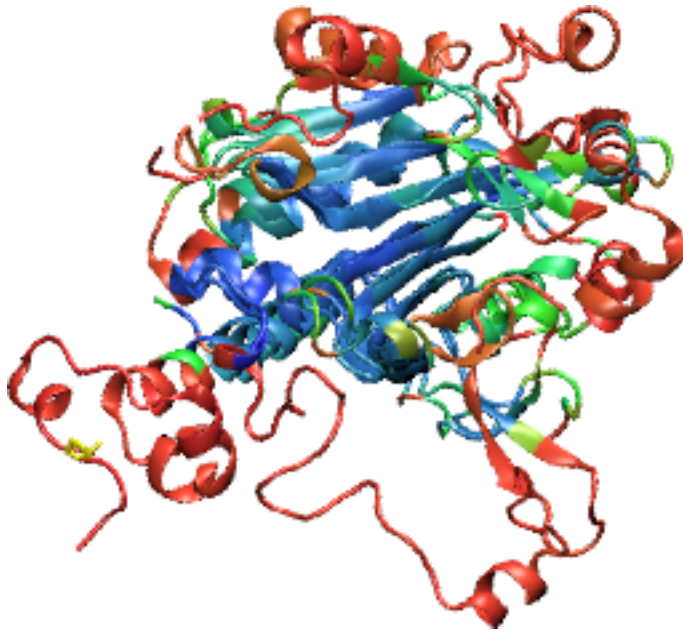
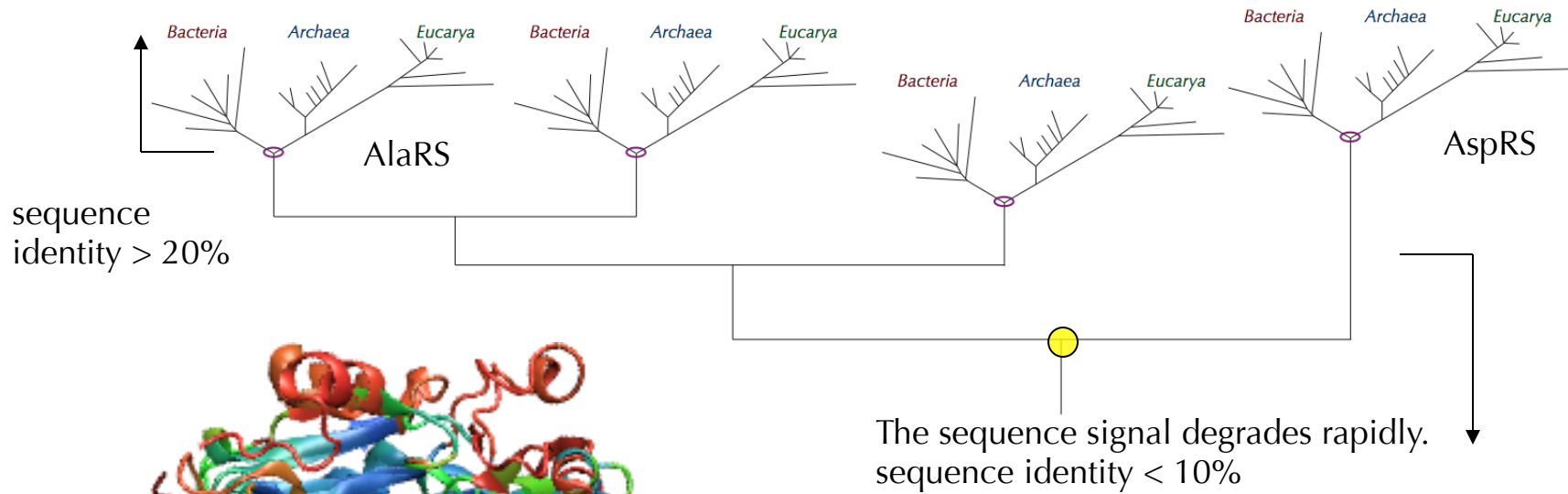
M. jannaschii, 2.55 Å.

M. thermoautotrophicus, 3.20 Å.

$Q_H = 0.89$

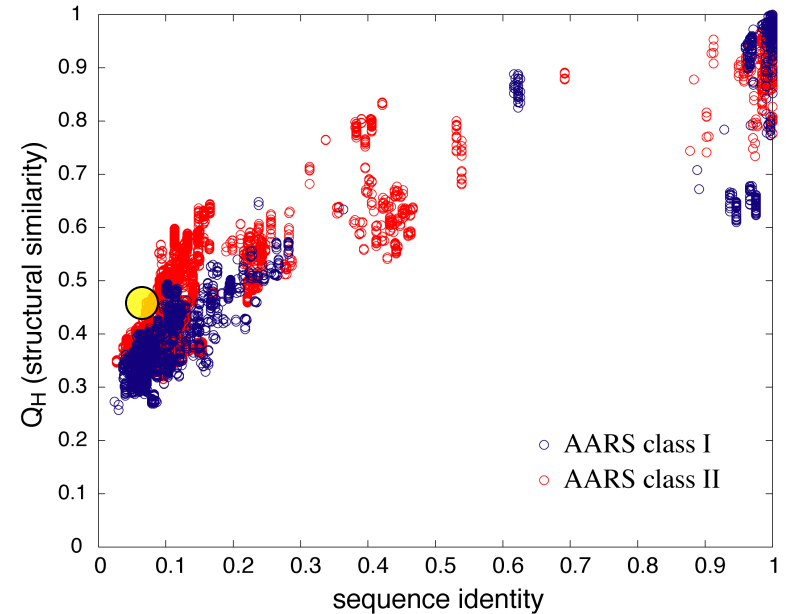
Sequence identity = 0.69

Relationship Between Sequence & Structure



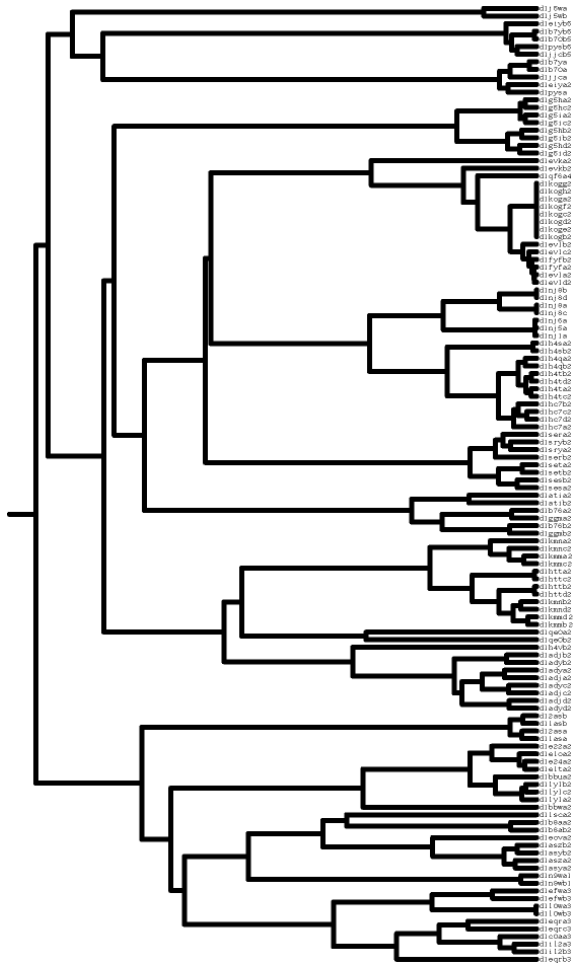
Structural superposition of AlaRS & AspRS.

● Sequence id = 0.055, $Q_H = 0.48$



Non-redundant Representative Profiles

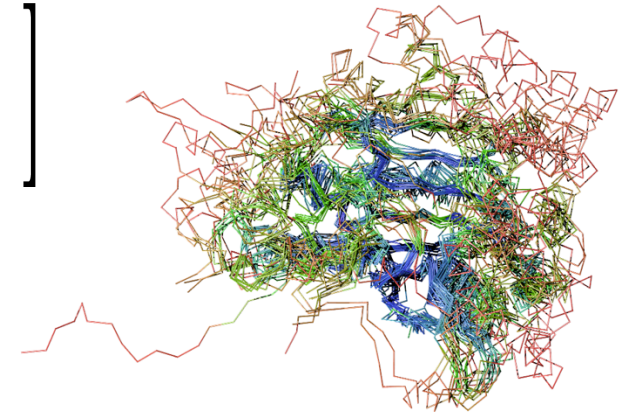
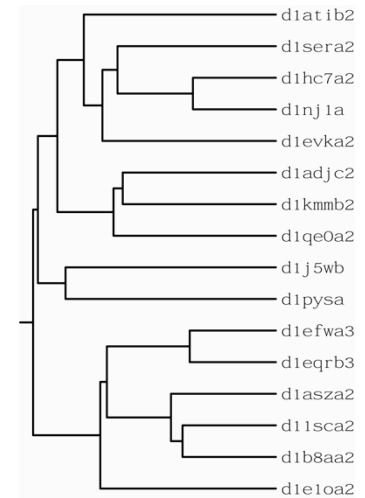
Too much information
129 Structures



Multidimensional QR
factorization
of alignment matrix, A .

$$A = \left[\begin{array}{c} \nearrow d=4 \\ \downarrow l_{aln} \\ \xrightarrow{k_{proteins}} \end{array} \begin{array}{c} G \\ Z \\ Y \\ X \end{array} \right]$$

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 $g = \gamma \frac{\|X\|_{F_4} + \|Y\|_{F_4} + \|Z\|_{F_4}}{\|G\|_{F_4}}$

adjustable
parameter

Sequence Space

Orthogonal Encoding = 24-space

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

A = (1,0)

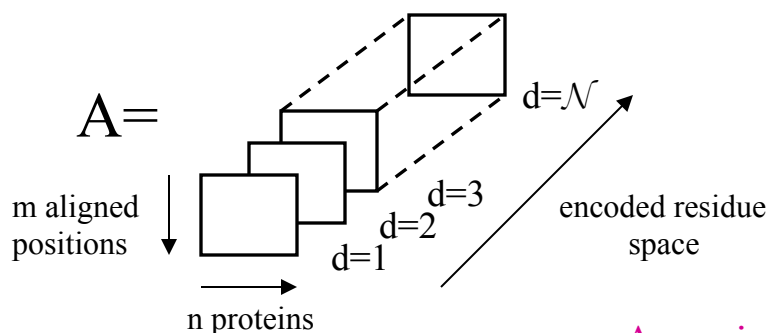
B = (0,1,0)

C = (0,0,1,0)

...

GAP = (0,1)

Alignment is a Matrix with Linearly Dependent Columns

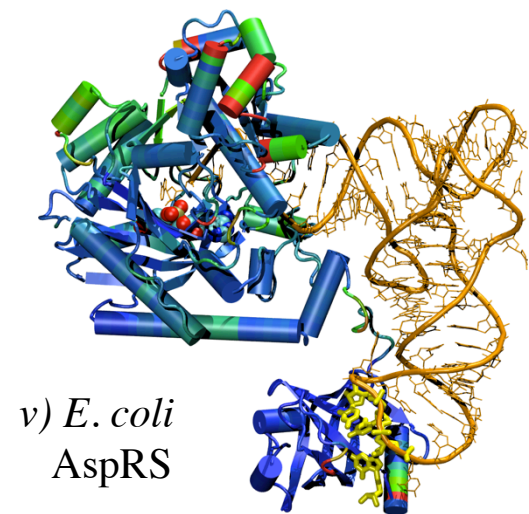
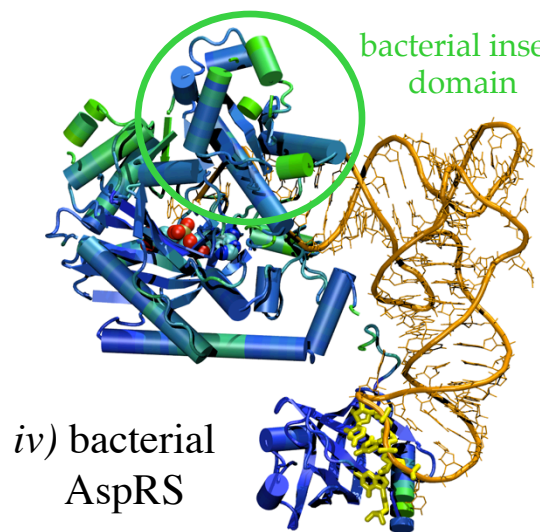
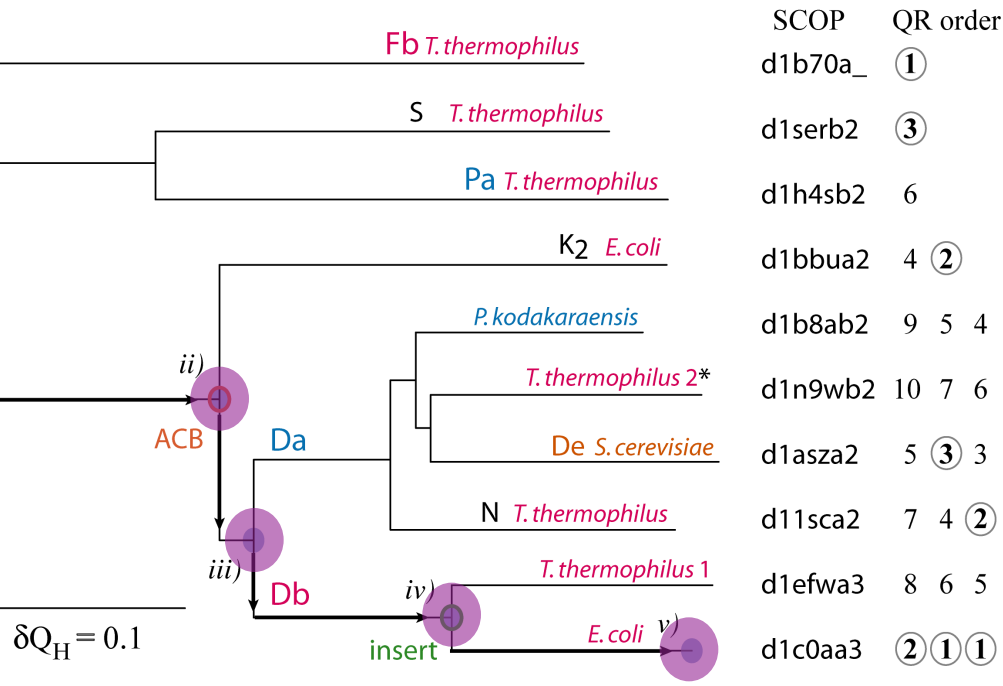
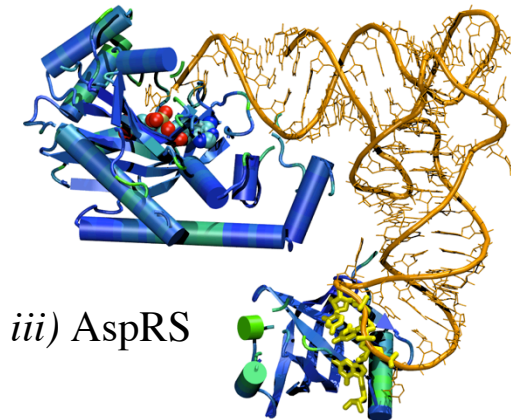
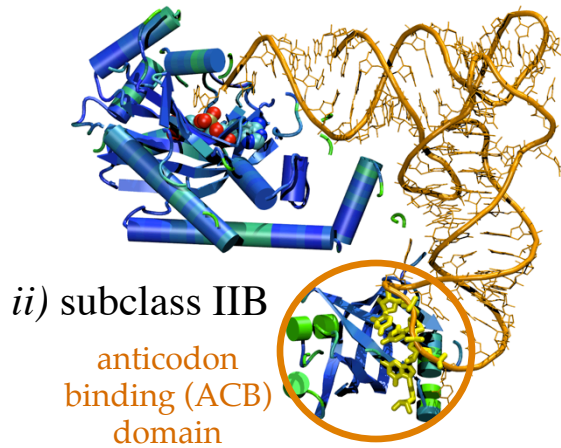
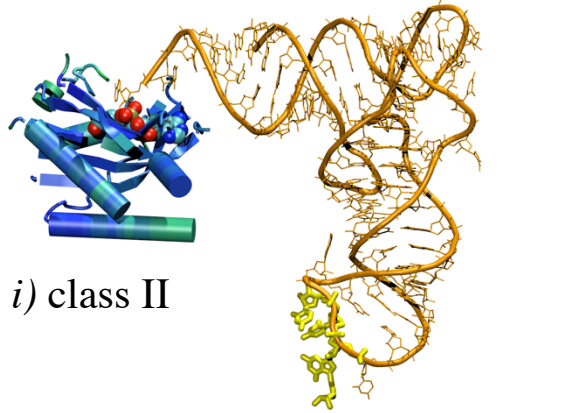


$$Q_{(d)}^T A_{(d)} P = Q_{(d)}^T \begin{bmatrix} & & & & \\ & & & & G \\ & & & & Z \\ & & & & Y \\ & & & & X \\ & & & & \\ & & & & \\ & & & & \\ & & & & \end{bmatrix} P = \tilde{R}_{(d)}$$

m_{aln} $n_{proteins}$

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

Evolution of Structure and Function in AspRS



Summary Structural Evolutionary Profiles

1. Structures often more conserved than sequences!! Similar structures at the Family and Superfamily levels.

Add more structural information to identify core and variable regions

2. Which structures and sequences to include? Use evolution and eliminate redundancy with QR factorization

New Tools in VMD/MultiSeq

Protein / RNA
Sequence Data

SwissProt DB (400K),
Greengenes RNA (100K)
Signatures, Zoom

Metadata Information,
Clustal &
Phylogenetic Trees

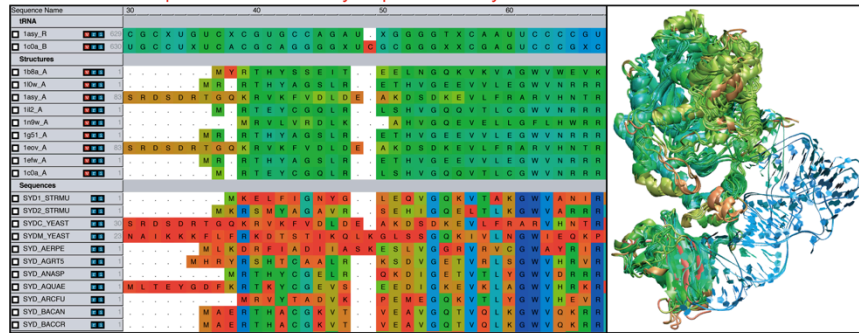
RAXml Trees,
Genomic Content,
Temperature DB

Blast & PsiBlast

Sequence Editor

View structural data colored by structural conservation and
sequence data colored by sequence identity

Synchronization between
1D and 3D views



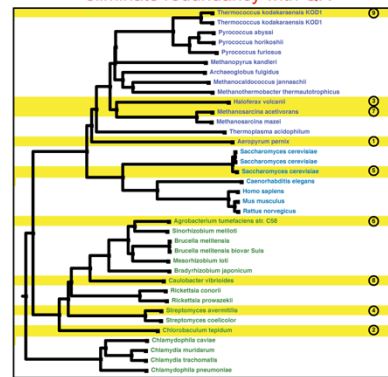
Sequence /Structure
Alignment

Protein & RNA
secondary structure

Group data by taxonomic classification

View sequence / structure phylogenies and
eliminate redundancy with QR

Sequence Name		90
Eukaryota:Fungi		
1asy_A	83	S R D S D R T G Q K R V K F V D
1teov_A	83	S R D S D R T G Q K R V K F V D
SYDC_YEAST	82	S R D S D R T G Q K R V K F V D
Eukaryota:Metazoa		
SYD_CAEL	57	G L V N S K E K K V L N F L K V
SYD_HUMAN	55	S M I Q S Q E K P D R V L V R V
SYD_MOUSE	55	S M I Q S Q E K P D R V L V R V
Archaea:Crenarcha		
SYD_AERPE	1 M L K D R F I A D I
Archaea:Euryarchaeota		
1n9w_A	1 M R V L V R D
1b8a_A	1 M Y R T H Y S S E
SYD_METMA	1	. . . M S L A N L R T H Y T A D
SYD_HALN1	1 M L E R T Y I E D
SYD_THEAC	1 M P R T Y I D T
SYD_PVRHO	1 M L R T H Y S N E
Bacteria:Proteobacteria		
110w_A	1 M R . R T H Y A G S
112_A	1 M . R T E Y C G Q

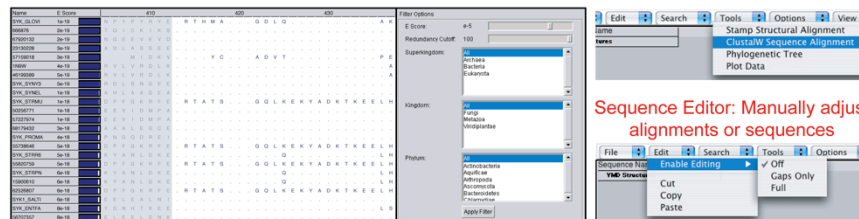


QR non-redundant
seq / str sets

Cluster
analysis /
Bioinformatics
scripting
Tutorials MultiSeq/
AARS
EF-Tu/Ribosome

Import data directly from BLAST databases

Align sequences with Clustal



J. Eargle, D. Wright, Z. Luthey-Schulten, *Bioinformatics*, 22:504 (2006)

E. Roberts, J. Eargle, D. Wright, Z. Luthey-Schulten, *BMC Bioinformatics*, 7:382 (2006)

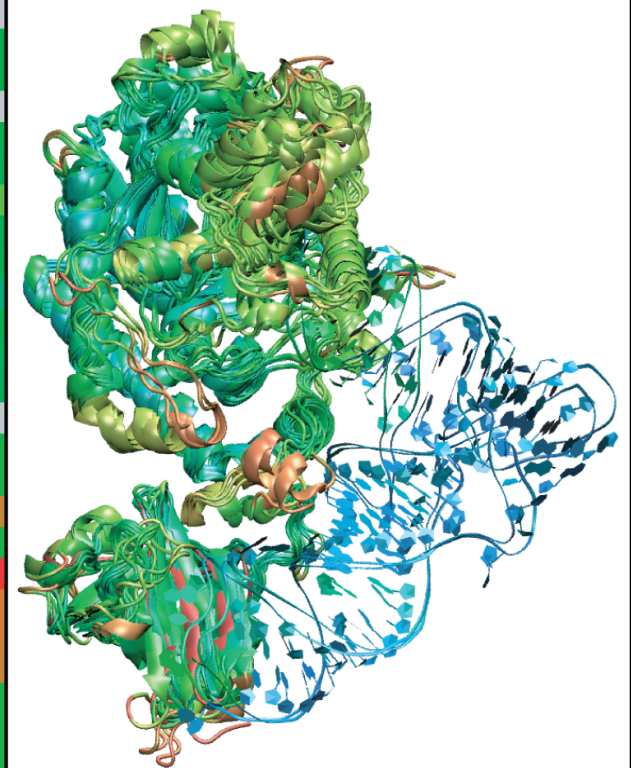
MultiSeq Combines Sequence and Structure

- Align sequences or structures; manually edit alignments
- View data colored by numerous metrics including structural conservation and sequence similarity
- Synchronized coloring between 1D and 3D views

Variation
in structures

Variation
in sequences

Sequence Name		40	50
tRNA			
<input type="checkbox"/> 1asy_R	629	U G U C X C G U G C C A G A U .	X G G G G T
<input type="checkbox"/> 1c0a_B	630	U X U C A C G C A G G G G X U C	G C G G G X
Structures			
<input type="checkbox"/> 1b8a_A	1	. . . M Y R T H Y S S E I T . . .	E E L N G Q
<input type="checkbox"/> 110w_A	1	. . . M R . R T H Y A G S L R . . .	E T H V G E
<input type="checkbox"/> 1asy_A	83	D R T G Q K R V K F V D L D E . . .	A K D S D K
<input type="checkbox"/> 1il2_A	1	. . . M . R T E Y C G Q L R . . .	L S H V G Q
<input type="checkbox"/> 1n9w_A	1	. . . M R V L V R D L K . . .	A H V G Q
<input type="checkbox"/> 1g51_A	1	. . . M R . R T H Y A G S L R . . .	E T H V G E
<input type="checkbox"/> 1eov_A	83	D R T G Q K R V K F V D L D E . . .	A K D S D K
<input type="checkbox"/> 1efw_A	1	. . . M R . R T H Y A G S L R . . .	E T H V G E
<input type="checkbox"/> 1c0a_A	1	. . . M . R T E Y C G Q L R . . .	L S H V G Q
Sequences			
<input type="checkbox"/> SYD1_STRMU	1	. . . M K E L F I G N Y G . . .	L E Q V G Q
<input type="checkbox"/> SYD2_STRMU	1	. . . M K R S M Y A G A V R . . .	S E H I G Q
<input type="checkbox"/> SYDC_YEAST	30	D R T G Q K R V K F V D L D E . . .	A K D S D K
<input type="checkbox"/> SYDM_YEAST	23	K K F L F R K D T S T I K Q L K G L S S G Q	
<input type="checkbox"/> SYD_AERPE	1	. . . M L K D R F I A D I I A S K . . .	E S L V G G
<input type="checkbox"/> SYD_AGR5	1	. . . M H R Y R S H T C A A L R . . .	K S D V G E
<input type="checkbox"/> SYD_ANASP	1	. . . M R T H Y C G E L R . . .	Q K D I G E
<input type="checkbox"/> SYD_AQUAE	1	Y G D F K R T K Y C G E V S . . .	E E D I G K
<input type="checkbox"/> SYD_ARCFU	1	. . . M R V Y T A D V K . . .	P E M E G Q
<input type="checkbox"/> SYD_BACAN	1	. . . M A E R T H A C G K V T . . .	V E A V G Q
<input type="checkbox"/> SYD_BACCR	1	. . . M A E R T H A C G K V T . . .	V E A V G Q



Load large sequence sets*

Swiss-Prot (Proteins)

Curated sequences

392,667 sequences

Unaligned

177 MB on disk

2 minutes to load

2.4 GB memory used

Greengenes (RNA)*

Environmental 16S rRNA

90,654 entries

Aligned (7682 positions)

670 MB on disk

2.5 minutes to load *

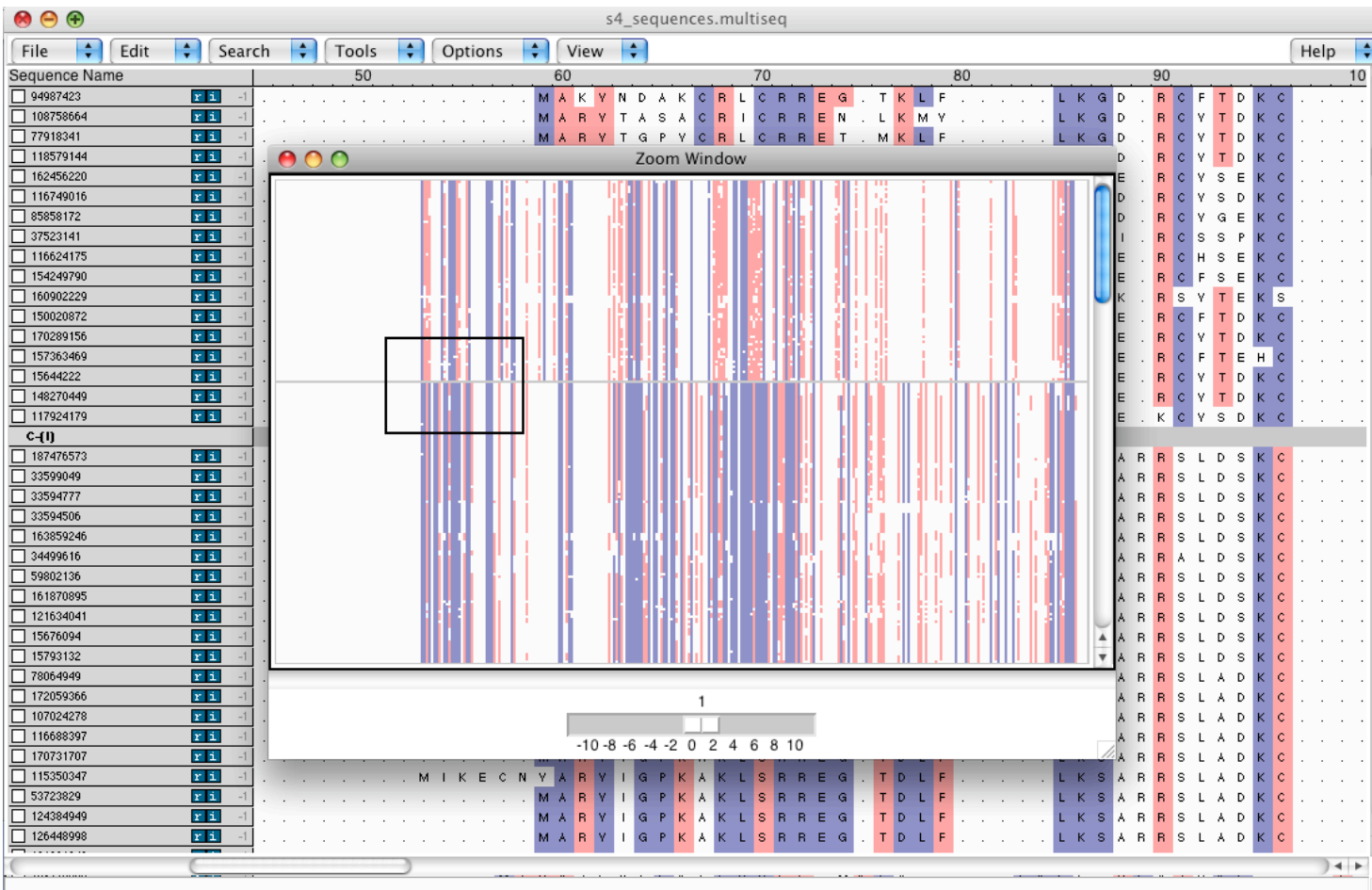
4.0 GB memory used*

*"Signatures of ribosomal evolution" with Carl Woese, PNAS (2008)

*Release May 2013 contains 1.2 million sequences – Memory??

Sequence editor

- New sequence API allows editing of large alignments. Align closely related sequences by group, combine groups, and then manually correct.
- Zoom window gives an overview of the alignment, quickly move the editing window to any part of the alignment.

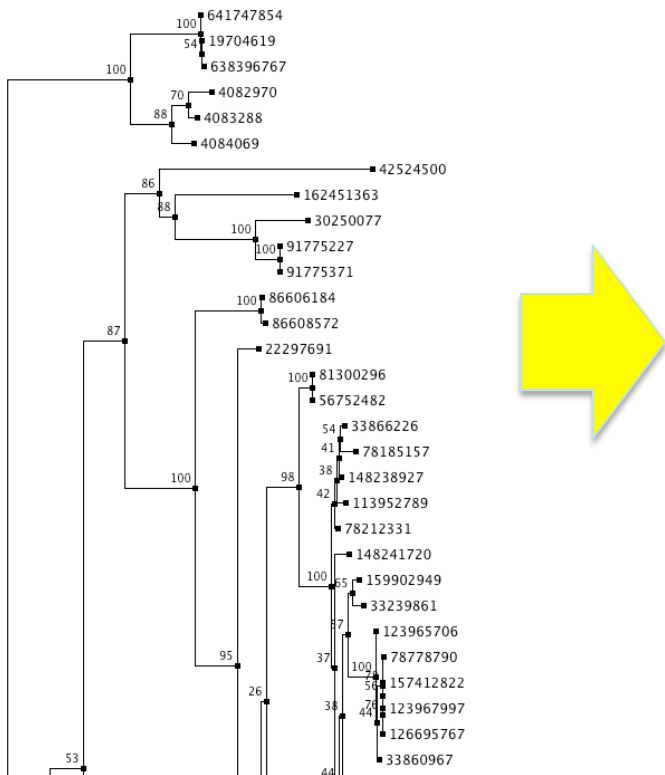


660 sequences of ribosomal protein S4 from all complete bacterial genomes*.

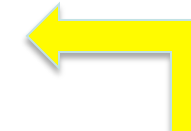
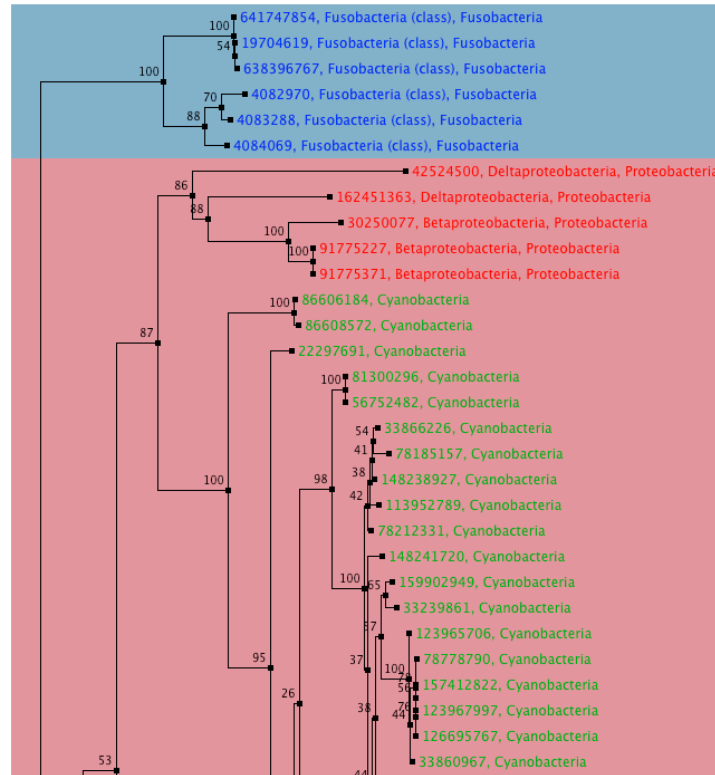
* K. Chen, E. Roberts, Z Luthey-Schulten (2009) BMC Bioinformatics

Phylogenetic tree editor

- Automatically add annotations and colors to phylogenetic trees based on taxonomy, enzyme, temperature class, and/or MultiSeq groupings.



Maximum likelihood tree of 660 S4 sequences reconstructed using RAXML.



A cluster of five proteobacterial sequences branch near the cyanobacterial sequences. These are cases of horizontal gene transfer.

Leaf Colors

■ Fusobacteria	■ Proteobacteria	■ Cyanobacteria
■ Chlamydiae	■ Firmicutes	■ Planctomycetes
■ Spirochaetes	■ Verrucomicrobia	■ Tenericutes
■ Chlorobi	■ Acidobacteria	■ Chloroflexi
■ Thermotogae		

Background Colors

■ C-(IV)_in	■ C-(IV)_out	■ C-(V)_out	■ C-(III)
■ C-(II)	■ C+	■ C-(I)	

Elijah Roberts 2009

Scripting MultiSeq

- All MultiSeq functions can be scripted.
- Scripting an analysis provides benefits:
 - It can be checked for correctness.
 - It can be quickly repeated by anyone.
 - It can be modified later with new functionality.
 - It can be run on a cluster in VMD text mode.
(if it can be easily broken into independent chunks)
- Many functions are too user specific and/or too complex to be turned into a GUI.
- Some examples of MultiSeq scripts...

Genome content

- When using sequence from fully sequenced genomes, additional information is available in the genome content.
- Conservation of gene ordering, neighbors, or intergenic regions can provide additional evolutionary information not contained in the sequence.
- Gene names and ordering can be obtained from the genome PTT files, want to organize the information in an evolutionarily meaningful manner.

Location	Strand	Length	PID	Gene	Synonym	Code	COG	Product
3437638..3438021	-	127	16131173	rplQ	b3294	-	COG0203J	50S ribosomal subunit protein L17
3438062..3439051	-	329	16131174	rpoA	b3295	-	COG0202K	RNA polymerase, alpha subunit
3439077..3439697	-	206	16131175	rpsD	b3296	-	COG0522J	30S ribosomal subunit protein S4
3439731..3440120	-	129	16131176	rpsK	b3297	-	COG0100J	30S ribosomal subunit protein S11
3440137..3440493	-	118	16131177	rpsM	b3298	-	COG0099J	30S ribosomal subunit protein S13
3440640..3440756	-	38	16131178	rpmJ	b3299	-	COG0257J	50S ribosomal subunit protein L36
3440788..3442119	-	443	16131179	secY	b3300	-	COG0201U	preprotein translocase membrane subunit
3442127..3442561	-	144	16131180	rplO	b3301	-	COG0200J	50S ribosomal subunit protein L15
3442565..3442744	-	59	16131181	rpmD	b3302	-	COG1841J	50S ribosomal subunit protein L30
3442748..3443251	-	167	16131182	rpsE	b3303	-	COG0098J	30S ribosomal subunit protein S5

Combined genomic context/phylogenetic tree

- Use a script to walk through a phylogenetic tree, find the genome content near the source gene, create a graphical representation of the combined data.

```
proc draw_genome_context_of_phylogeny {args} {  
  
    # Load the sequences.  
    set alignment [::SeqData::Fasta::loadSequences $alignmentFilename]  
  
    # Load the tree  
    set tree [::PhyloTree::Newick::loadTreeFile $treeFilename]  
  
    # Reorder the alignment by the tree.  
    set treeAlignment {}  
    set leafNodes [::PhyloTree::Data::getLeafNodes $tree]  
    foreach node $leafNodes {  
        set foundNode 0  
        set nodeName [::PhyloTree::Data::getNodeName $tree $node]  
        foreach sequence $alignment {  
            if {$nodeName == [::SeqData::getName $sequence]} {  
                lappend treeAlignment $sequence  
                set foundNode 1  
                break  
            }  
        }  
    }  
  
    # Draw the genomic context.  
    drawGenomicContextOfAlignment $outputFilename $treeAlignment $contextDistance $scaling $genomeDirectory  
}
```


Combined genomic context/phylogenetic tree

```

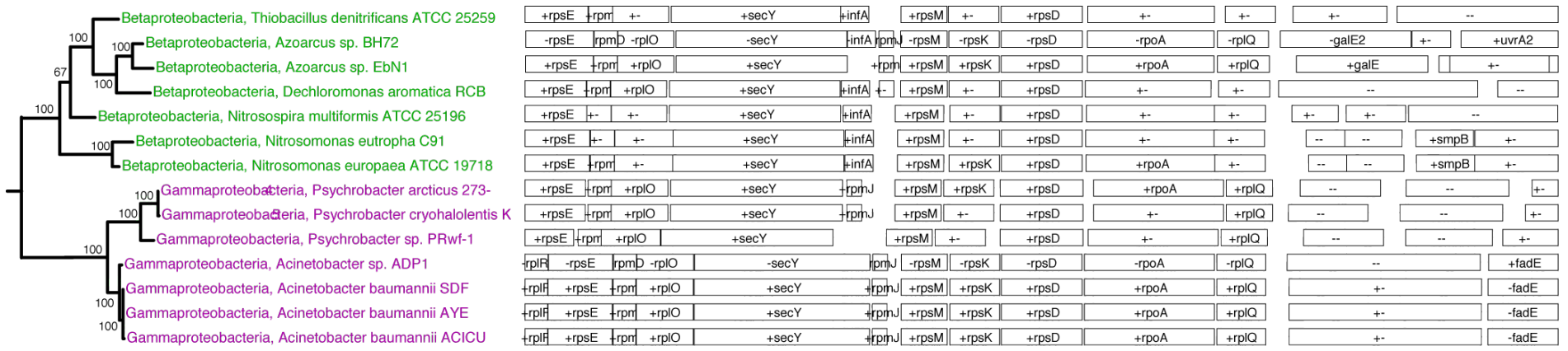
proc drawGenomicContextOfAlignment {outputFilename alignment contextDistance scaling genomeDirectory} {
  foreach sequence $alignment {
    # Make sure we have the GI number for this sequence.
    set giNumber [::SeqData::getSourceData $sequence "gi"]

    # Make sure we can tell which genome this sequence is from.
    set taxonomy [join [::SeqData::getLineage $sequence 1 0 1] ","]
    if ![info exists genomeTaxonomyMap($taxonomy)] {
      error "ERROR) Unknown genome for sequence [::SeqData::getName $sequence]: $taxonomy"
    }

    # Go through each of the genome context files for the genome.
    set foundGene 0
    foreach genomeName $genomeTaxonomyMap($taxonomy) {
      ...
    }
  }

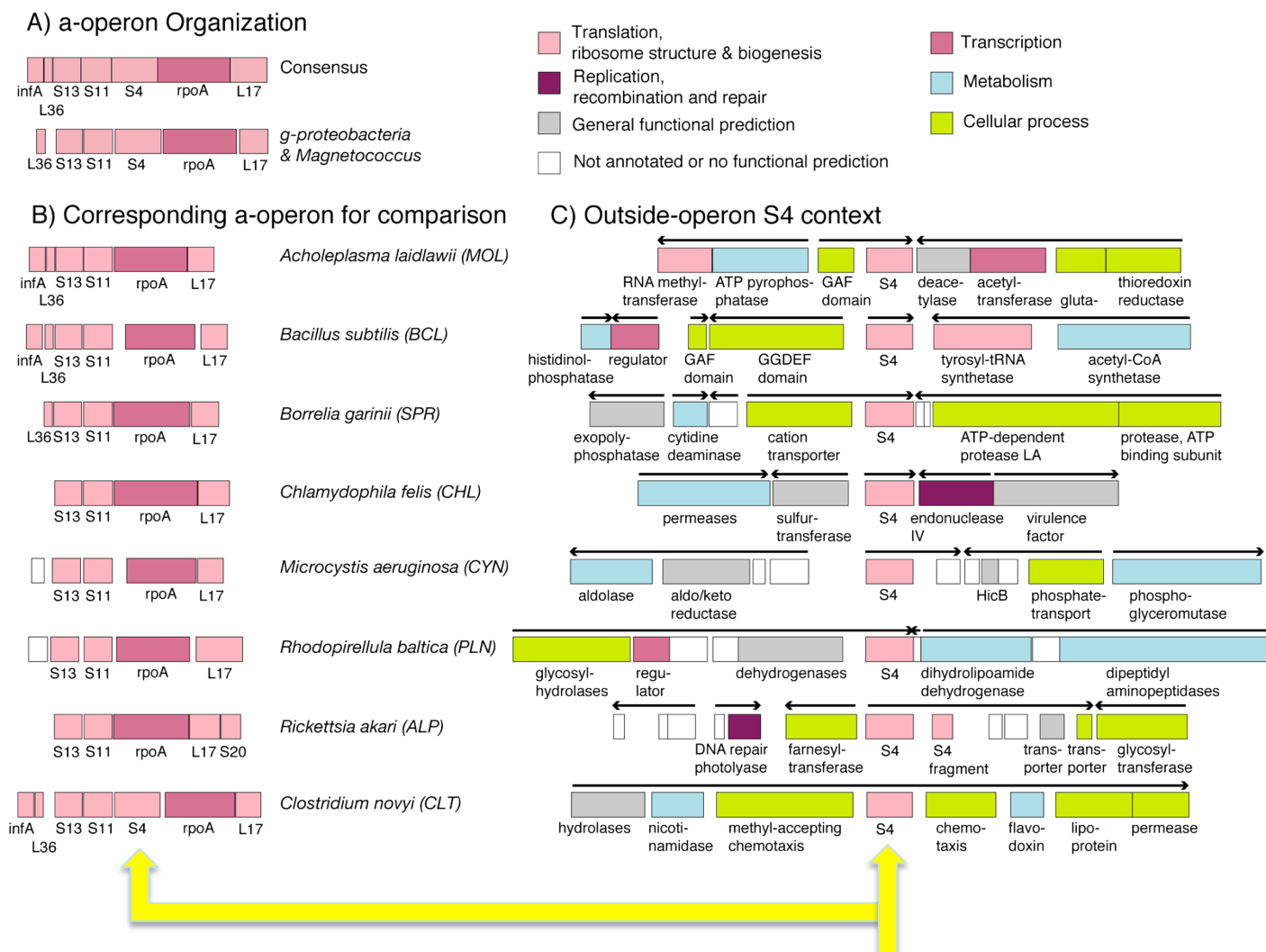
  # Draw the genomic context.
  drawMultipleGenomicContext $outputFilename $alignment $geneFiles $genePositions $geneStrands $contextDistance
}

```



Genome content future directions

- Genome content still a work in progress.
- Good candidate for a GUI: combined phylogenetic tree/ genome content viewer.
- Can also use COG codes to color by gene function.
- Still need API for manipulating PTT files.



Roberts, Chen, ZLS,
BMC Evol. Bio. 2009

See also ITEP for microbial genomes, Benedict et al.
BMC Genomics 2014

Genome content of ribosomal protein S4 by occurrence of the gene in the alpha operon.

Fifteen Clostridia genomes contain two copies of S4: one zinc-binding and one zinc-free.

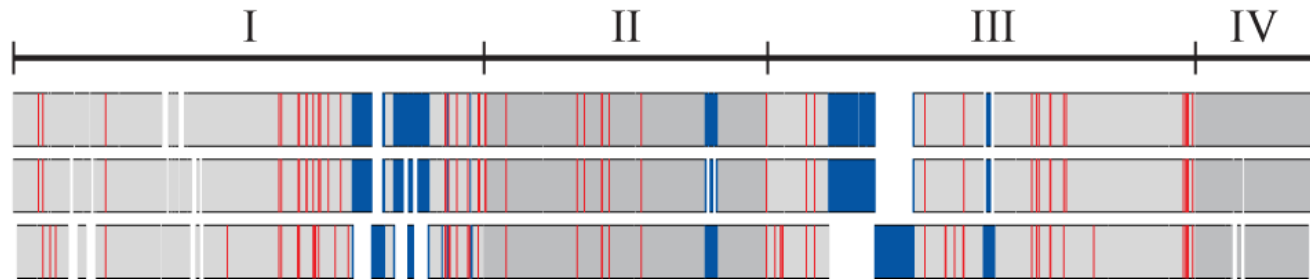
Molecular Signatures of Translation- Drug Targets

16S rRNA

E. coli

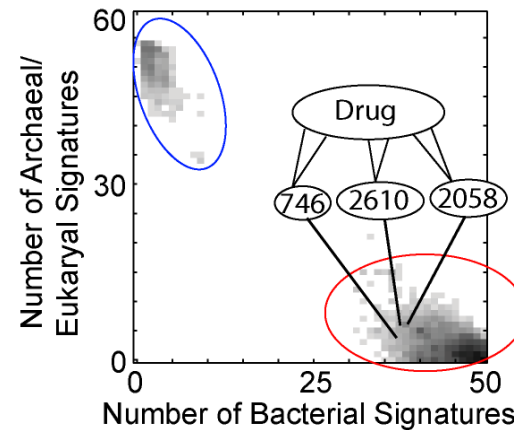
T. thermophilus

H. marismortui



Ribosomal Signatures: Idiosyncrasies in rRNA and/or r-proteins characteristic of the domains of life

69 (119) & 6 (14) in 16S (23S)



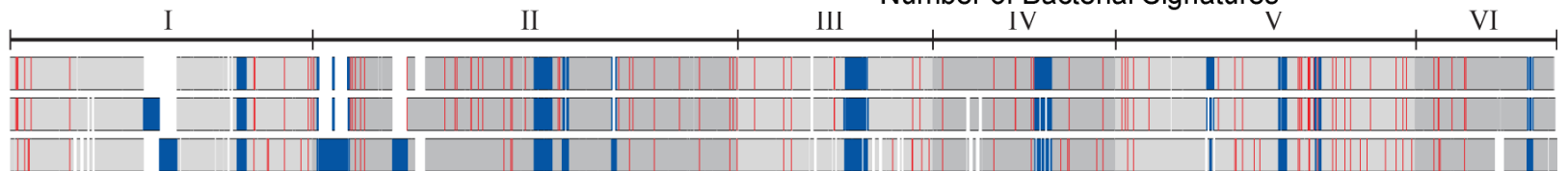
MultiSeq
Zoom

23S rRNA

E. coli

T. thermophilus

H. marismortui



E. Roberts, A. Sethi, J. Montoya, C. R. Woese & Z. Luthey-Schulten. *PNAS*
 "Molecular Signatures of Ribosomal Evolution" (2008)

Kim,... Luthey-Schulten, Ha, and Woodson, *Nature* "Protein-guided RNA dynamics during early ribosome assembly (2014)

Flexible Grouping of Data

- Automatically group data by taxonomic classification to assist in evolutionary analysis (HGT) or create custom groups
- Apply metrics to groups independently, e.g bacterial signal

Sequence Name		90
Eukaryota:Fungi		
<input type="checkbox"/> 1asy_A		83 S R D S D R T G Q K R V K F V D
<input type="checkbox"/> 1eov_A		83 S R D S D R T G Q K R V K F V D
<input type="checkbox"/> SYDC_YEAST		82 S R D S D R T G Q K R V K F V D
Eukaryota:Metazoa		
<input type="checkbox"/> SYD_CAEEL		57 S K . . E K K V L N F L K V K E
<input type="checkbox"/> SYD_HUMAN		33 S Q . . E K P D R V L V R V R D
<input type="checkbox"/> SYD_MOUSE		33 S Q . . E K P D R V L V R V K D
Archaea:Crenarcha		
<input type="checkbox"/> SYD_AERPE		1 M L K D R F I A D
Archaea:Euryarchaeota		
<input type="checkbox"/> 1n9w_A		1 M R V L V R D
<input type="checkbox"/> 1b8a_A		1 M Y R T H Y S S E
<input type="checkbox"/> SYD_METMA		1 . . . M S L A N L R T H Y T A D
<input type="checkbox"/> SYD_HALN1		1 M E N R T Y T A D
<input type="checkbox"/> SYD_THEAC		1 M L S I A E
<input type="checkbox"/> SYD_PYRHO		1 M I E K V Y C Q E
Bacteria:Proteobacteria		
<input type="checkbox"/> 110w_A		1 M R . R T H Y A G S
<input type="checkbox"/> 1il2_A		1 M . R T E Y C G Q

MultiSeq: Display and Edit Metadata

- External databases are **cross-referenced** to display **metadata** such as taxonomic information and enzymatic function
- Changes to metadata should periodically be updated!!!
- **Electronic Notebook**: Notes and annotations about a specific sequence or structure can be added – and saved

The screenshot shows a metadata editor window for the sequence SYDC_YEAST. The fields are as follows:

Sequence Name:	SYDC_YEAST
Source Organism:	Saccharomyces cerevisiae
Common Name:	yeast
EC Number:	6.1.1.12
EC Description:	Aspartate--tRNA ligase.
Description:	Aspartyl-tRNA synthetase, cytoplasmic (EC 6.1.1.12) (Aspartate--tRNA ligase) (AspRS) - Saccharomyces cerevisiae (Baker's yeast).
Data Sources:	sp=P04802,SYDC_YEAST pdb=1EOV,A
Lineage:	Eukaryota Fungi Ascomycota Saccharomycotina Saccharomycetes Saccharomycetales
Notes	There were missing residues

At the bottom of the window are buttons for 'OK' and 'Cancel'.

VMD/MultiSeq - Summary

1. Visualization, analysis tools, modeling large and long timescale biomolecular simulations, coarse-grained particles - *VMD*
2. Evolutionary and genetic information integration with structural information *Proteins/RNA – MultiSeq (Msalign, Metadata)*
3. Integrate simulation data and databases with graphical interface - *MultiSeq (Translation Tutorials, Metadata) & VMD*
4. Support high performance interactive and batch mode analysis - *MultiSeq & VMD (e.g analyze all rproteins, genomic content)*
5. Improve graphics quality and performance using emerging technologies (*GPU acceleration, programmable shading*) *VMD*