Introduction to Evolutionary Concepts and VMD/MultiSeq - Part I

Zaida (Zan) Luthey-Schulten

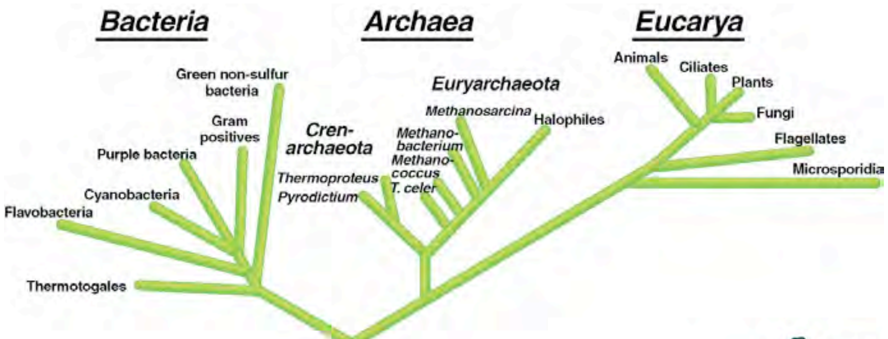
Dept. Chemistry, Beckman Institute, Biophysics, Institute of Genomics Biology, & Physics

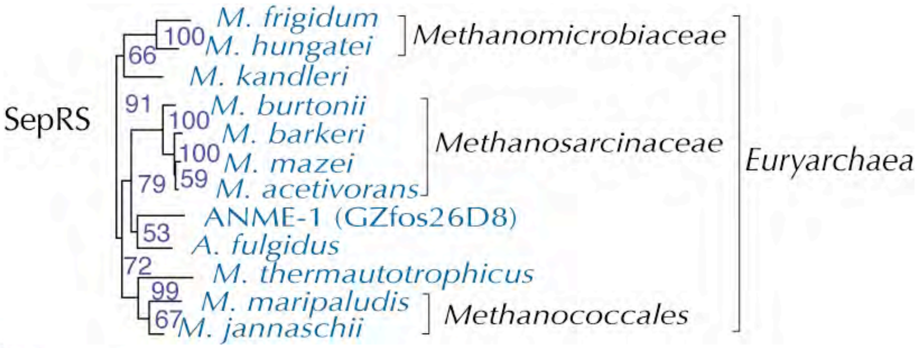
NIH Workshop 2009

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

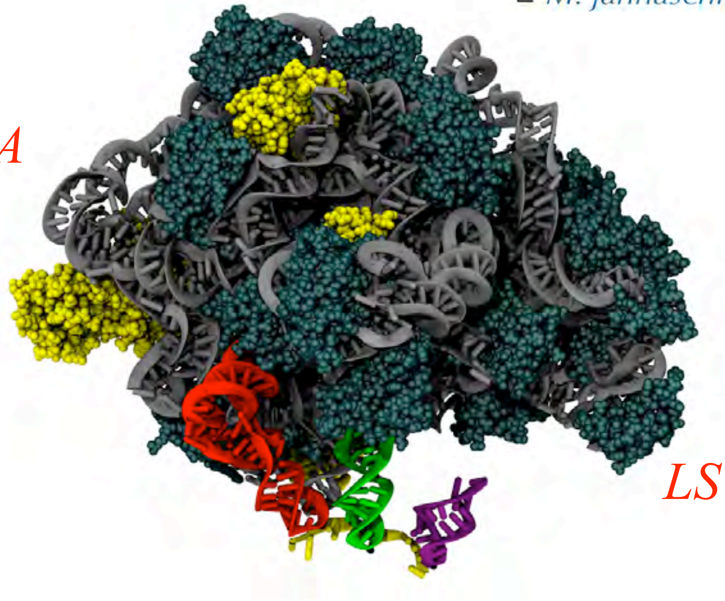
VMD/MultiSeq - “A Tool to Think”

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

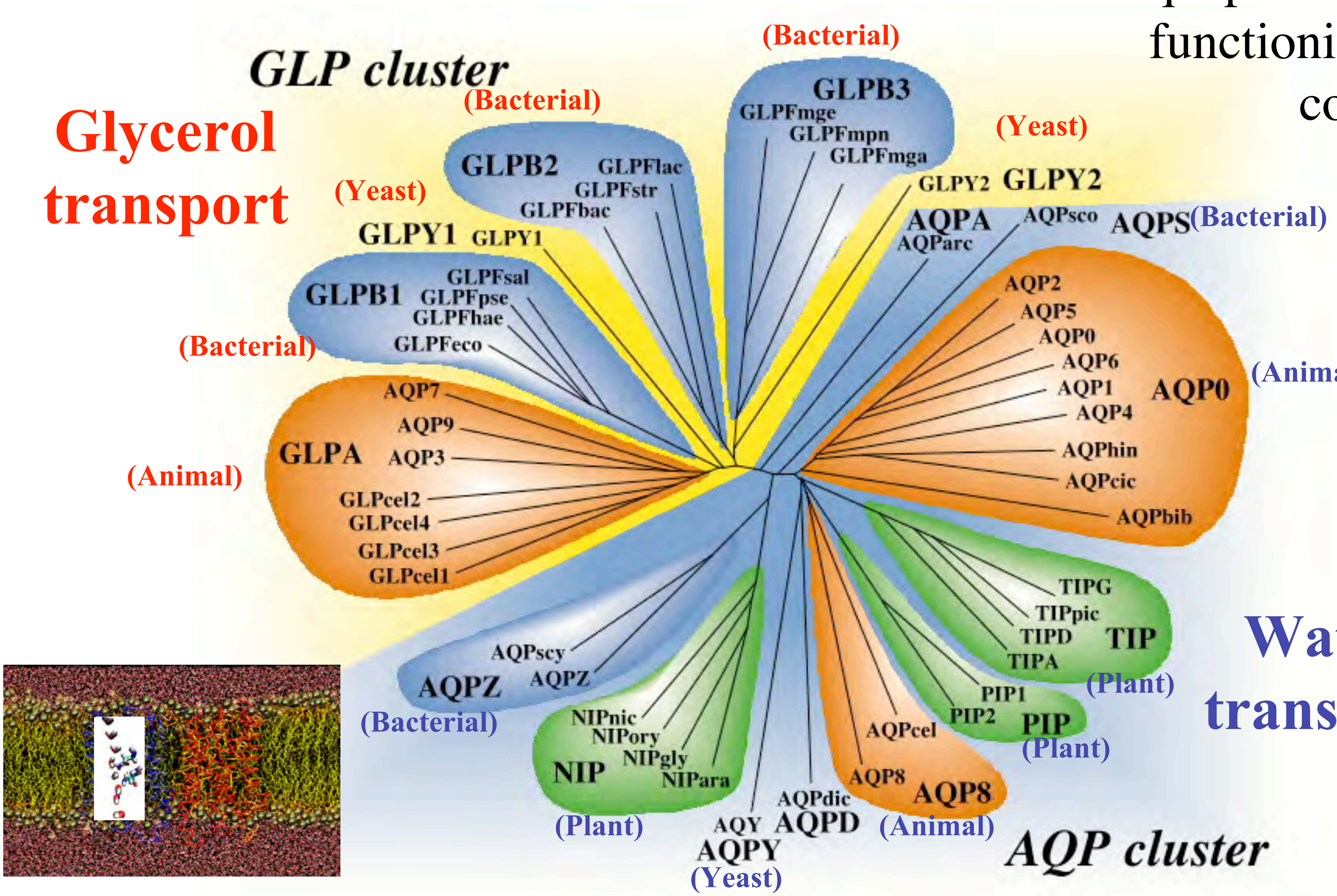


LSU (23S rRNA + rproteins)

The Aquaporin Superfamily

aquaporins are also functioning as gas conductors

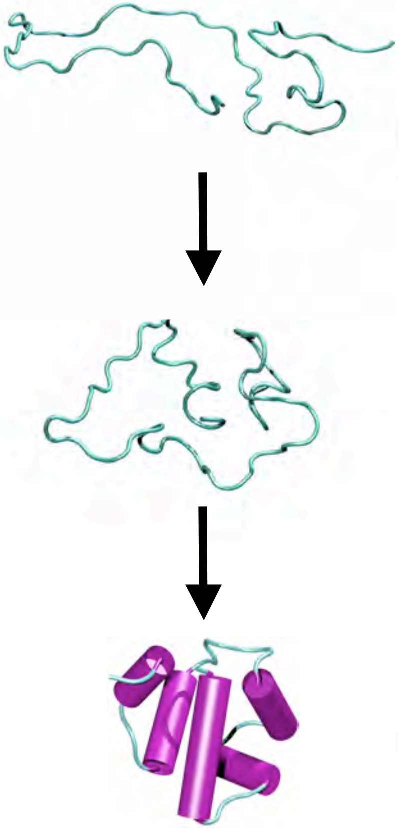
Glycerol transport

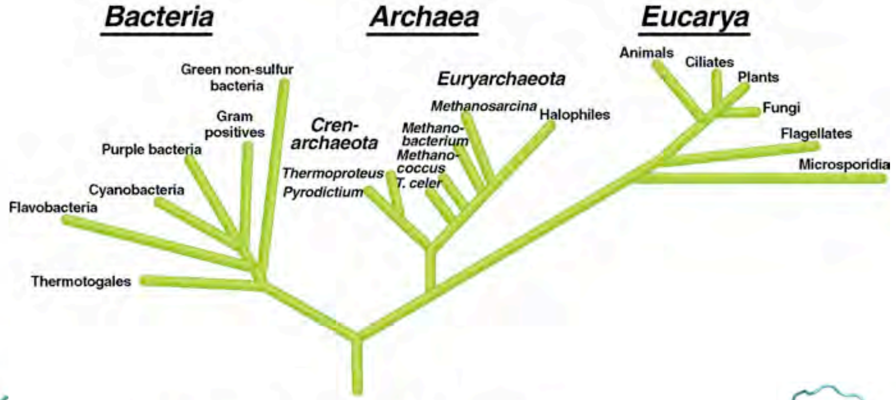


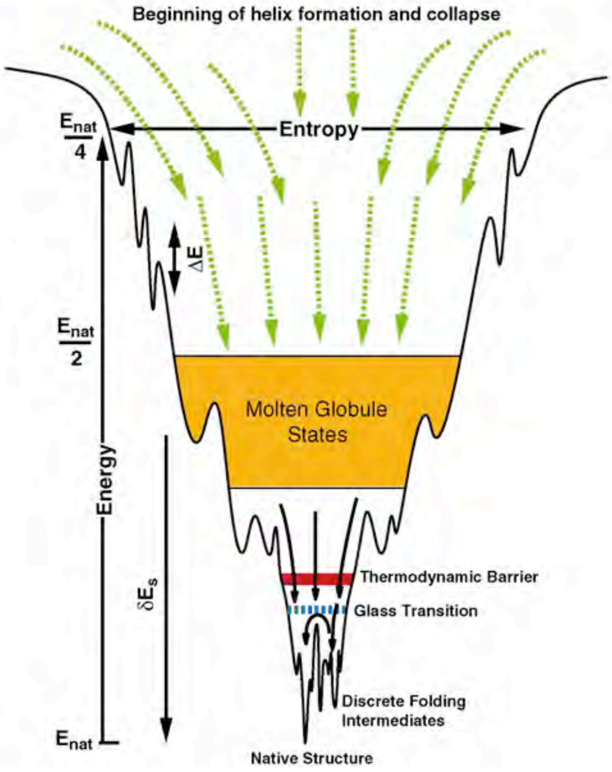
Water transport

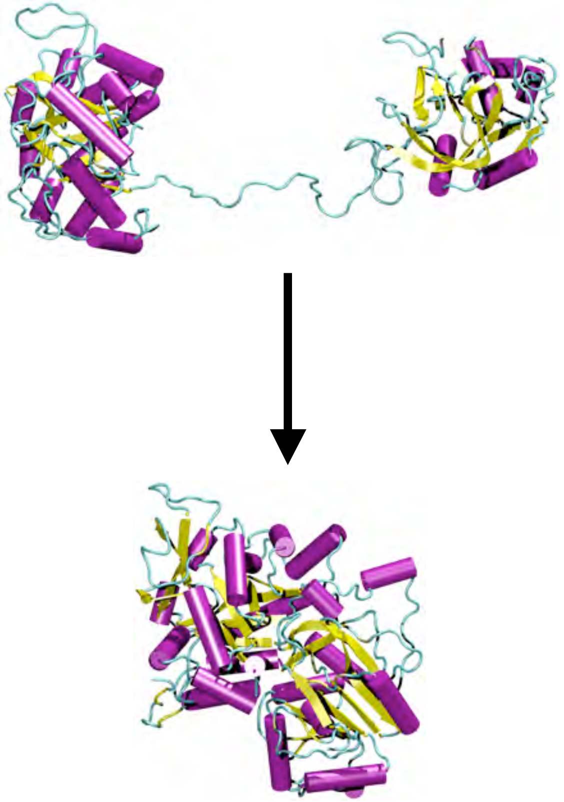
Heymann and Engel *News Physiol. Sci.* **14**, 187 (1999)

Evolution of Protein (RNA) Folding, Structure, & Function

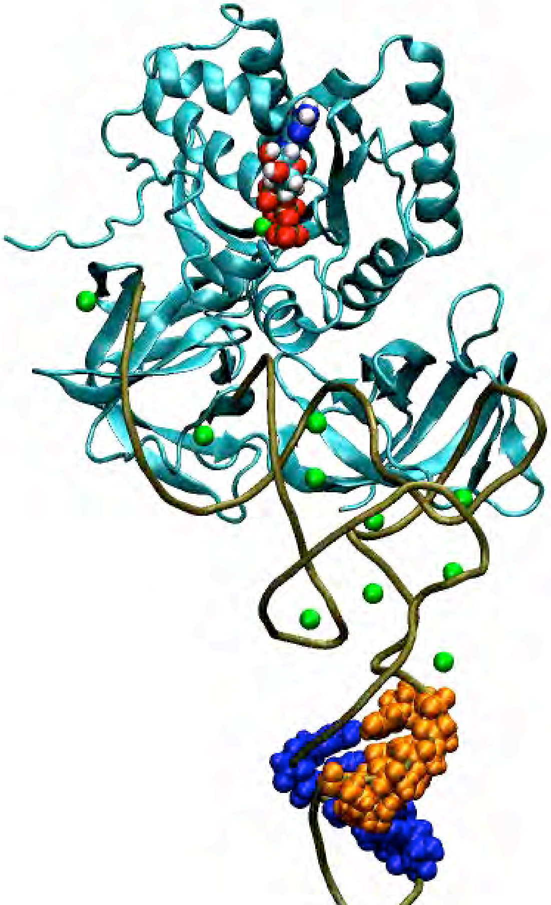




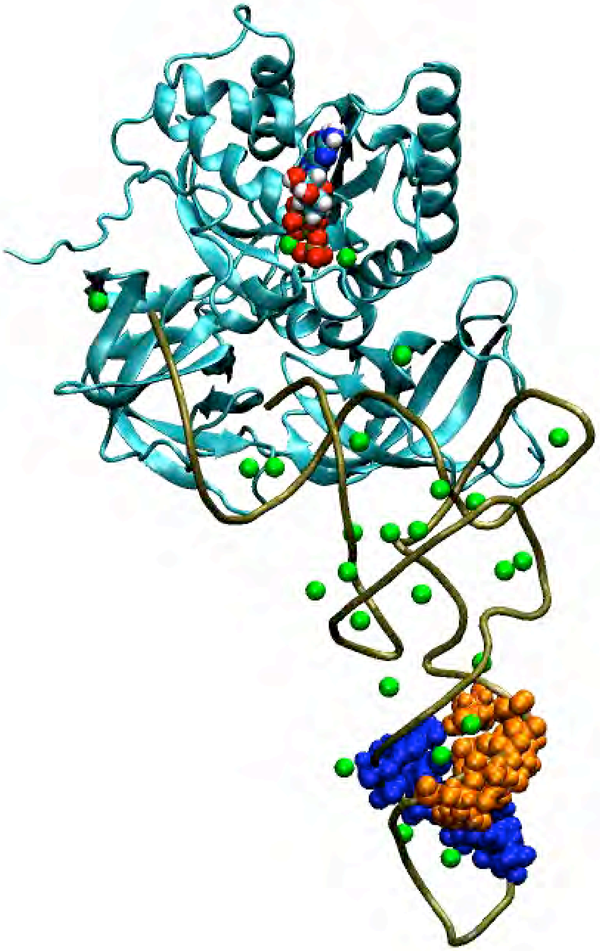




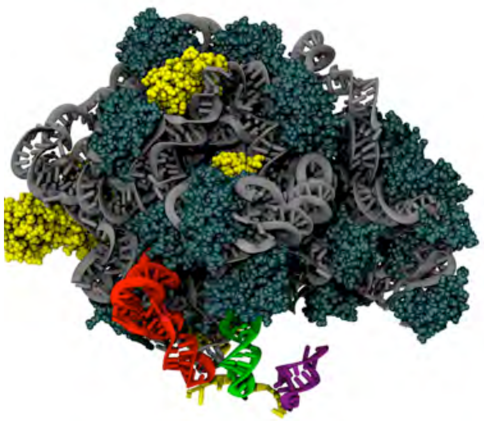
Evolution of Translational Machinery



SepRS/CysRS Evolutionary History
Sethi, O'Donoghue, ZLS, ***PNAS*** 2005
O'Donoghue, Sethi, Woese, ZLS,
PNAS 2005, Sethi, et al. Dynamics of
Allosteric Network, **PNAS** 09



Dynamical Recognition EF-Tu:tRNA-
Novel Amino Acids
Eargle et al., ***JMB*** 2008

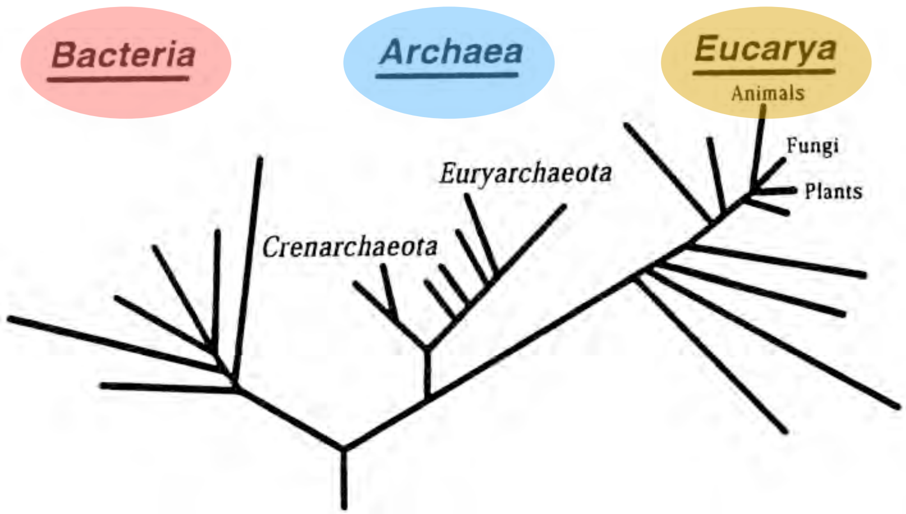


Proteins/RNA
Ribosome

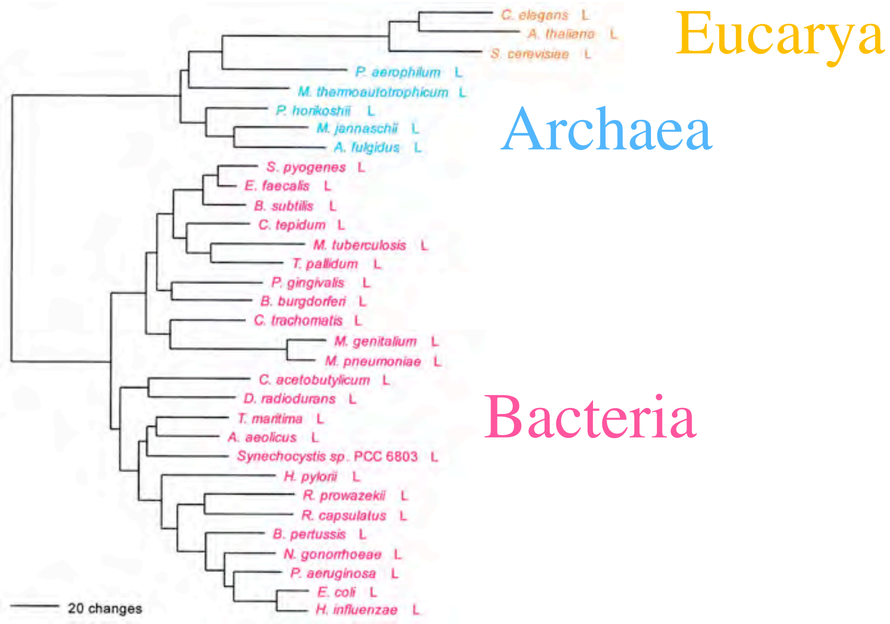
Molecular Signatures of
Ribosomal Evolution***,***
PNAS **2008,** Roberts,
Sethi, Woese, ZLS

Universal Phylogenetic Tree

three domains of life



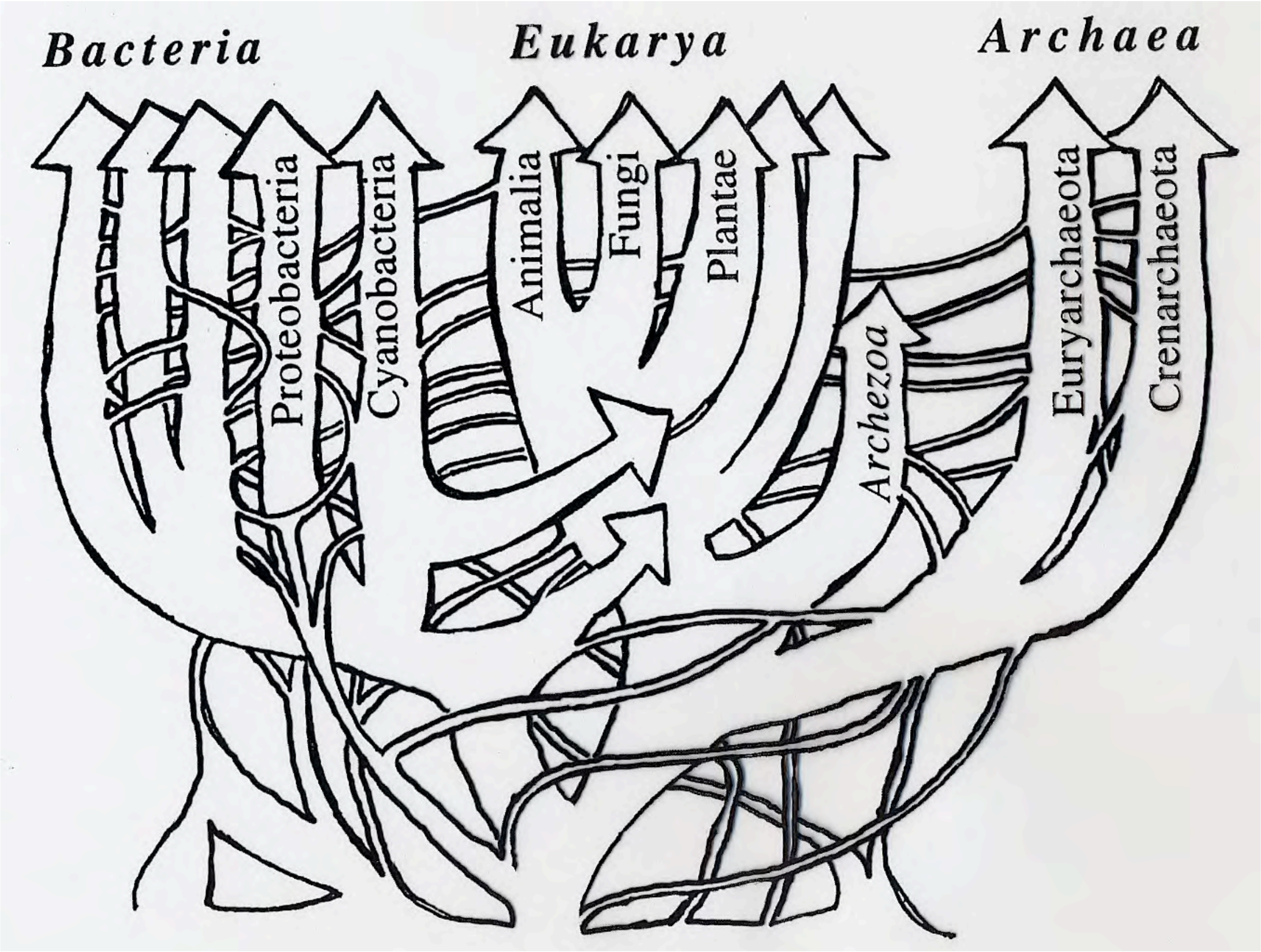
Based on 16S rRNA



Leucyl-tRNA synthetase displays the full canonical phylogenetic distribution.

for review see Woese *PNAS* 2000

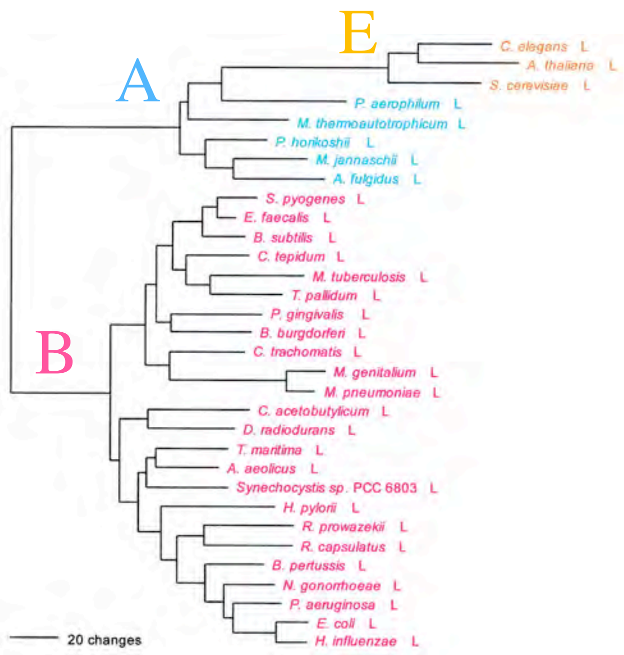
Woese, Olsen, Ibba, Soll *MMBR* 2000



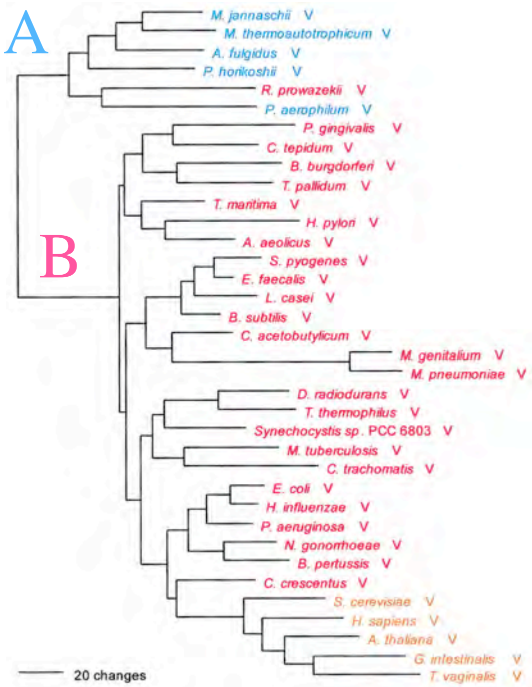
After W. Doolittle, modified by G. Olsen

Phylogenetic Distributions

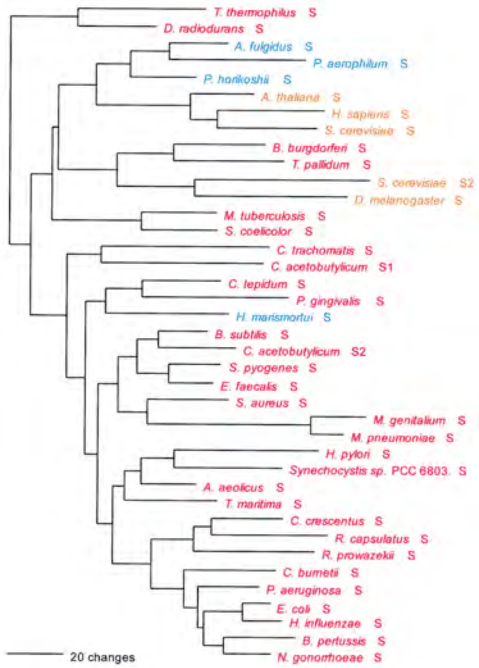
Full Canonical



Basal Canonical



Non-canonical

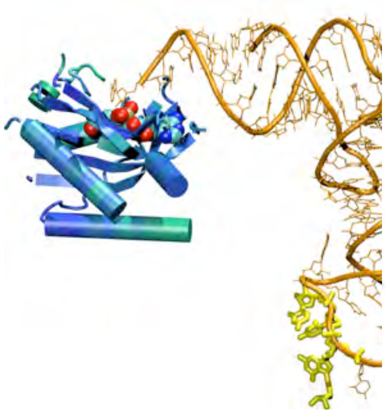


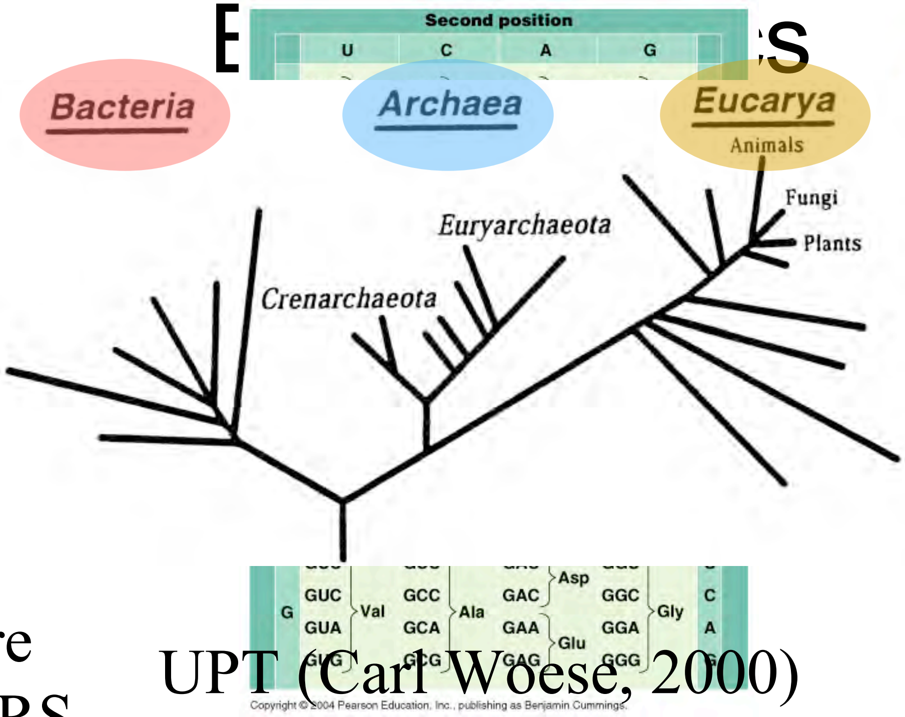
increasing inter-domain of life Horizontal Gene Transfer

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

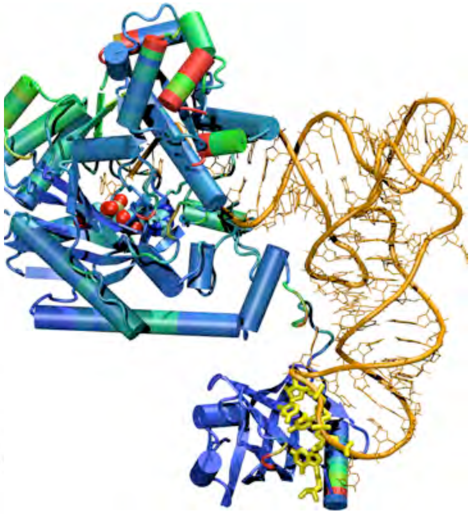
MultiSeq in VMD

Evolutionary Approach to
E s





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



Catalytic Core
of Class II AARS

UPT (Carl Woese, 2000)

Modern Bacterial
AspRS

Evolution of the Translation Machinery

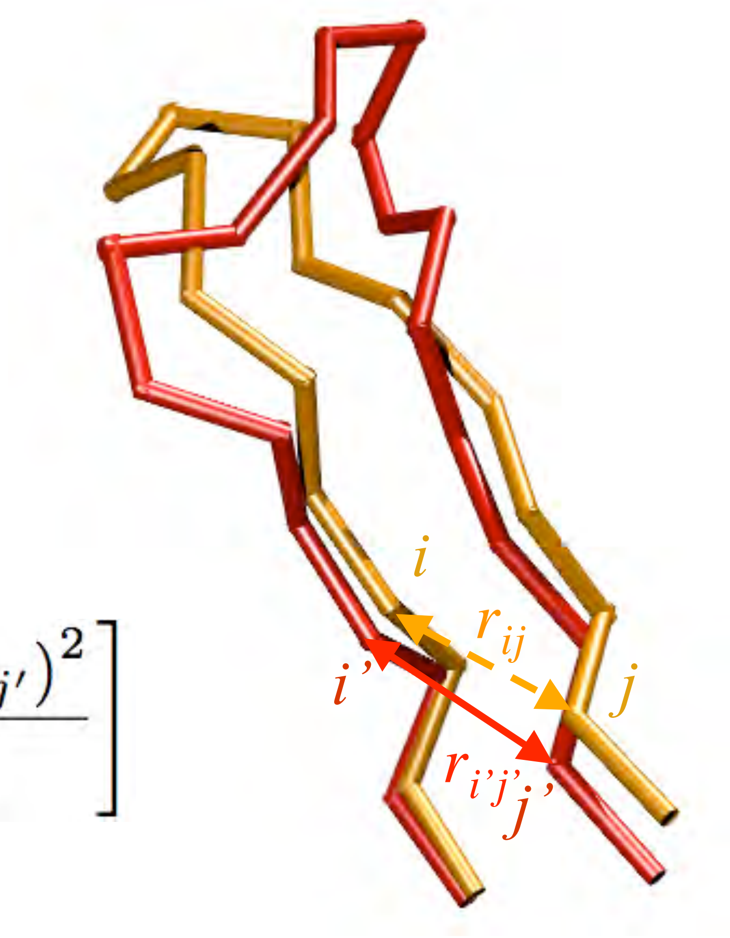
Protein Structure Similarity Measure

Q_H Structural Homology

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

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

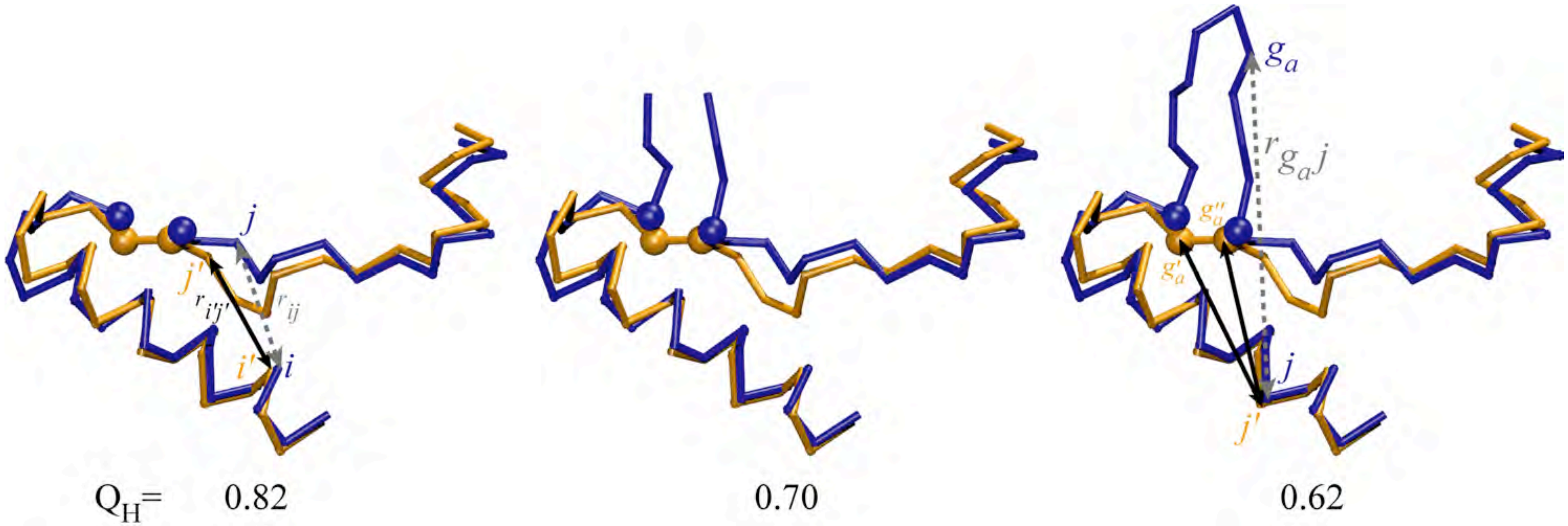
$$q_{aln} = \sum_{i<j-2} \exp\left[-\frac{\left(r_{ij} - r_{i'j'}\right)^2}{2\sigma_{ij}^2}\right]$$

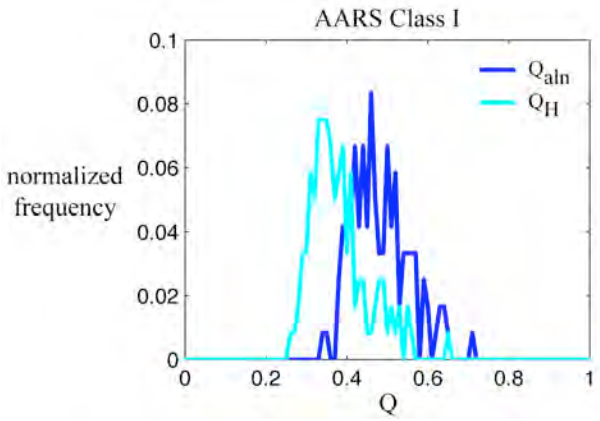


O'Donoghue & Luthey-Schulten *MMBR* 2003.

Structural Similarity Measure
the effect of insertions

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



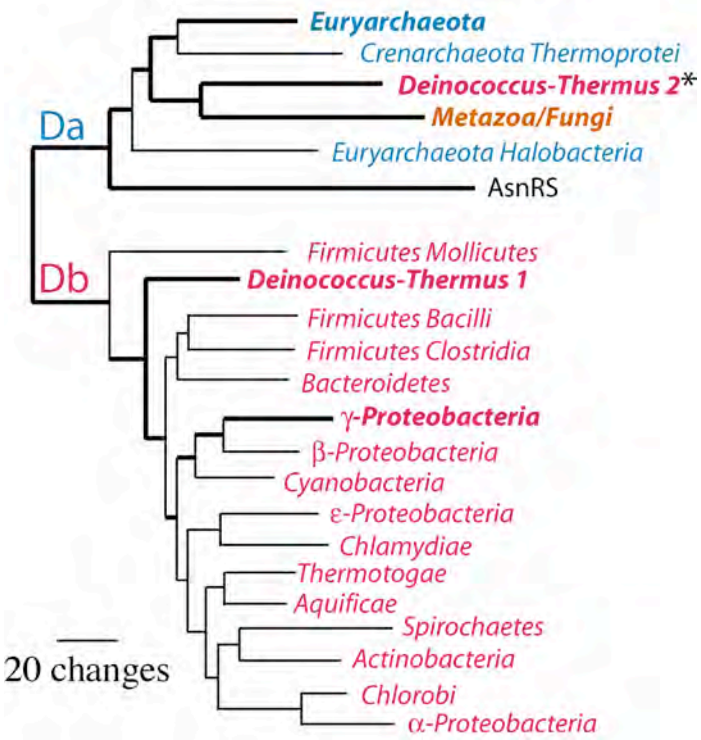


$$q_{gap} = \sum_{g_a} \sum_{j}^{N_{aln}} \max\left\{ \exp\left[-\frac{\left(r_{g_a j} - r_{g'_a j'}\right)^2}{2\sigma^2_{g_a j}}\right], \exp\left[-\frac{\left(r_{g_a j} - r_{g''_a j'}\right)^2}{2\sigma^2_{g_a j}}\right]\right\}$$

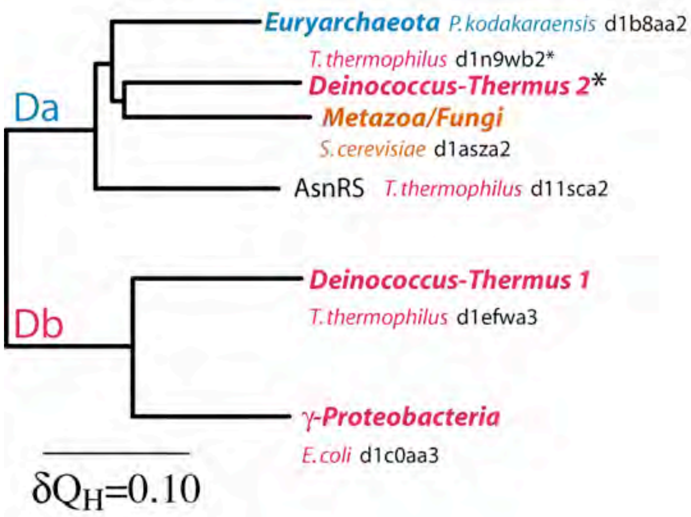
$$+ \sum_{g_b} \sum_{j}^{N_{aln}} \max\left\{ \exp\left[-\frac{\left(r_{g_b j} - r_{g'_b j'}\right)^2}{2\sigma^2_{g_b j}}\right], \exp\left[-\frac{\left(r_{g_b j} - r_{g''_b j'}\right)^2}{2\sigma^2_{g_b j}}\right]\right\}$$

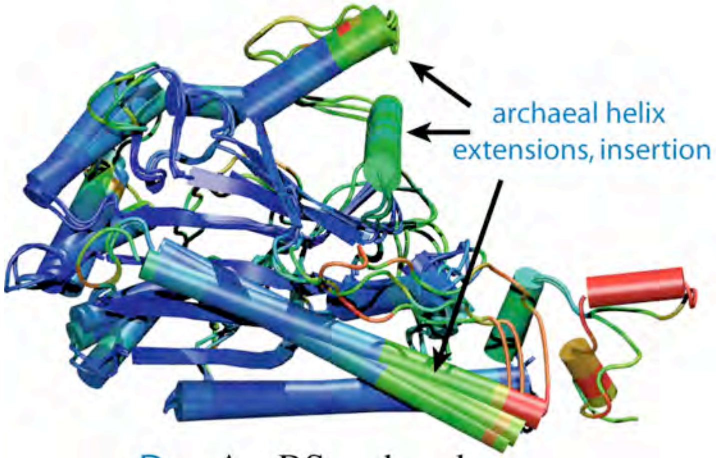
Protein structure encodes evolutionary information

sequence-based phylogeny

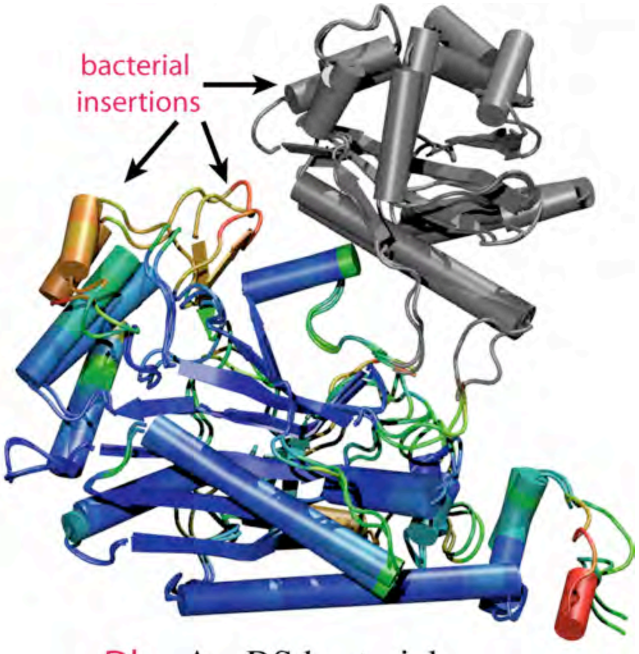


structure-based phylogeny





Da - AspRS archaeal genre

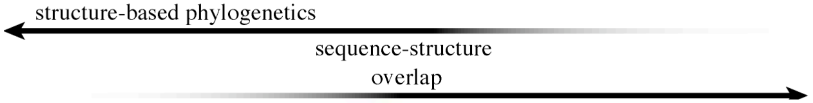


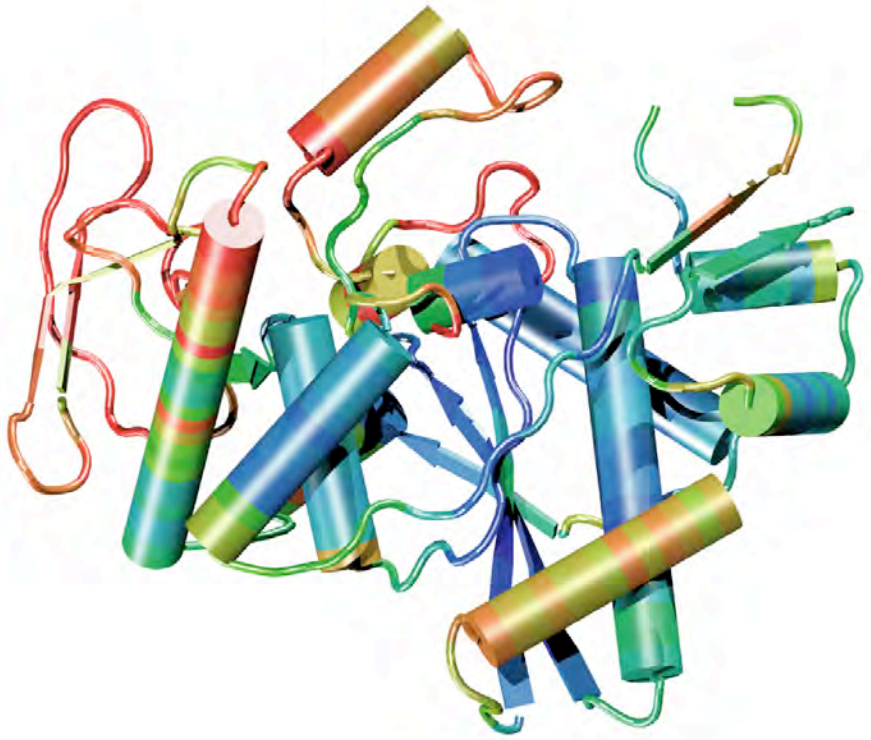
Db - AspRS bacterial genre

JMB 2005
MMBR 2003

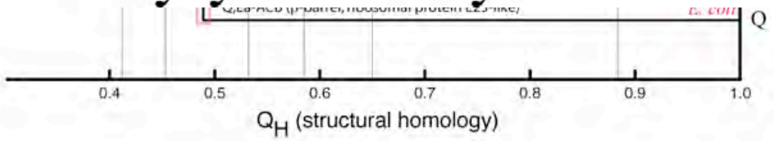
Protein structure reveals distant evolutionary events

Class I AARSs

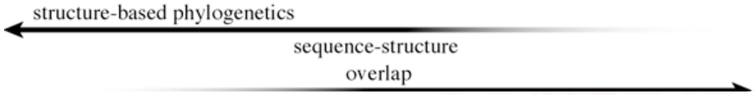


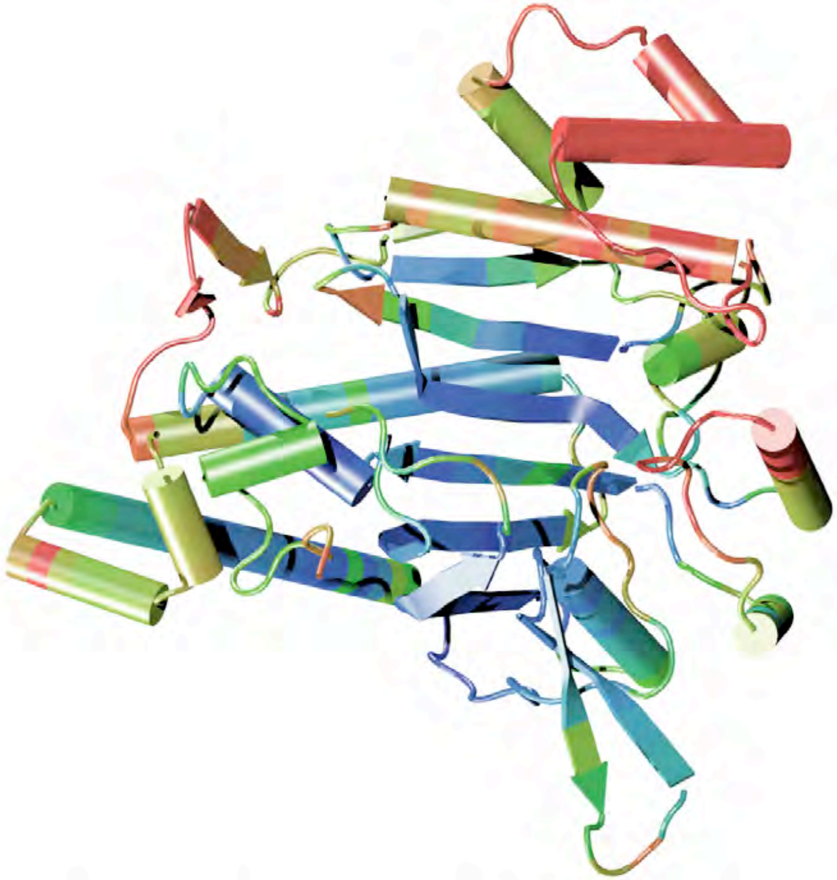


Class I Lysyl-tRNA Synthetase

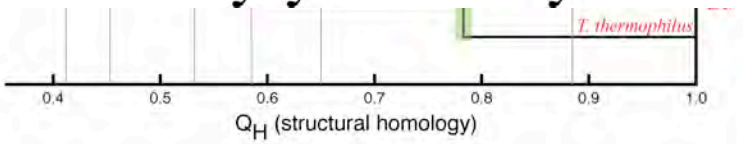


Class II AARSs

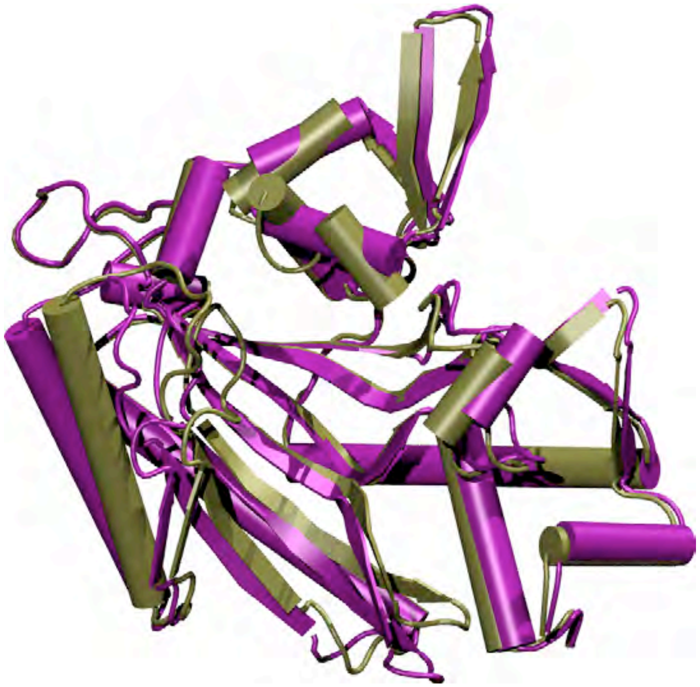




Class II Lysyl-tRNA Synthetase



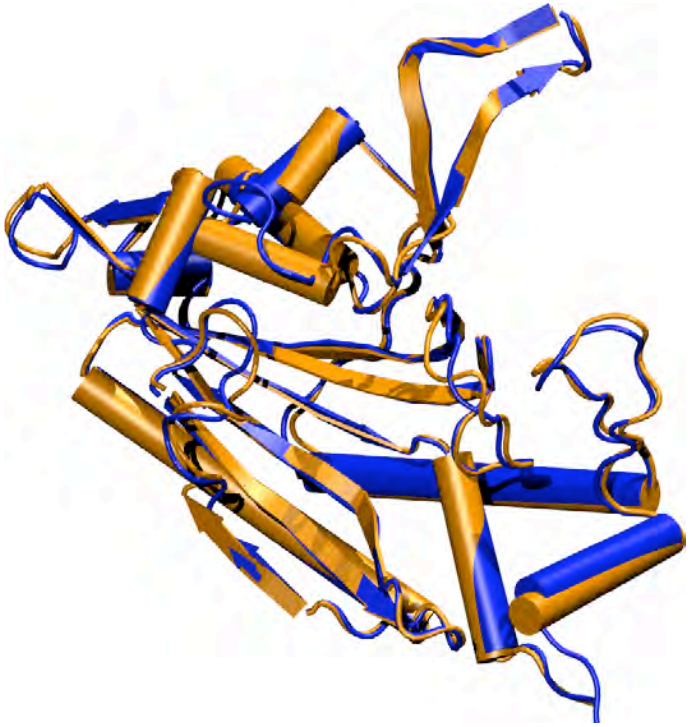
Sequences define more recent evolutionary events



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

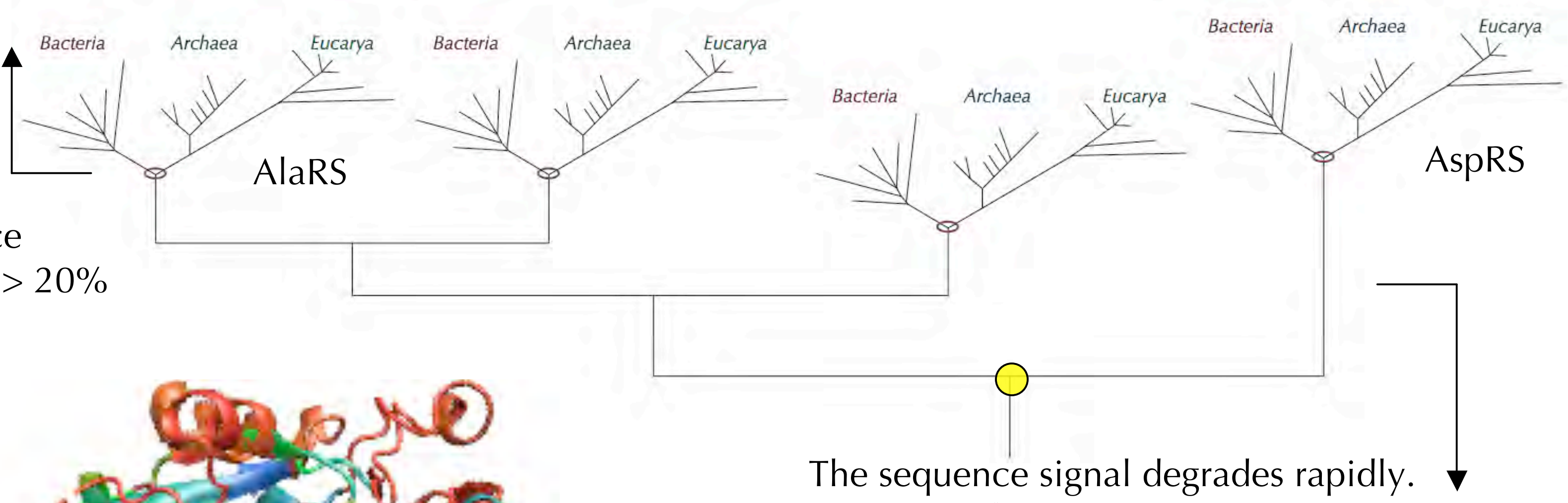


Structures for two different species.

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

$Q_H = 0.89$
Sequence identity = 0.69

The Relationship Between Sequence & Structure



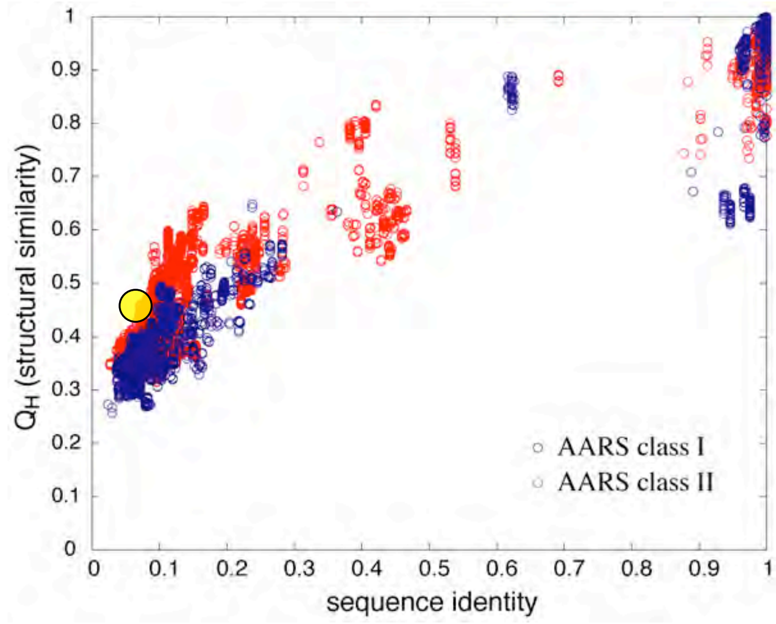
sequence
identity > 20%

The sequence signal degrades rapidly.
sequence identity < 10%

Structural superposition of AlaRS & AspRS.

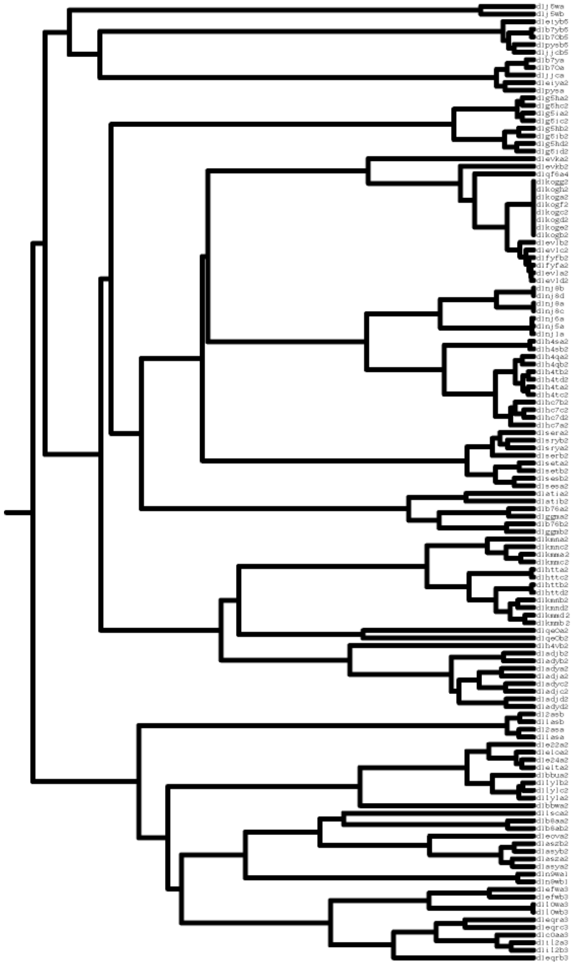
● Sequence id = 0.055, Q_H= 0.48

O’Donoghue & Luthey-Schulten (2003) *MMBR* 67: 550–73.
Structural alignment & visualization software MultiSeq/VMD



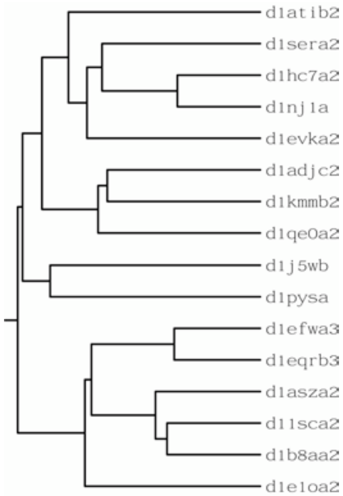
Non-redundant Representative Sets

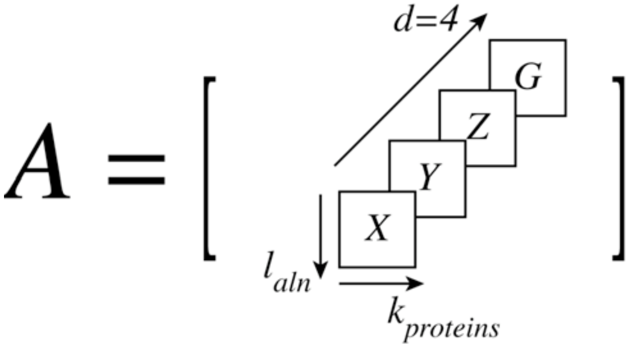
Too much information
129 Structures

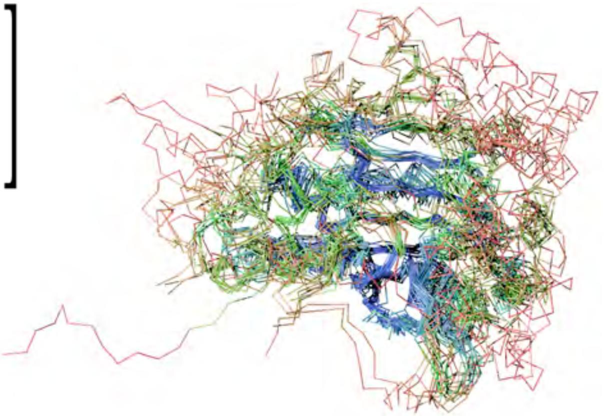


Economy of information
16 representatives

Multidimensional QR factorization of alignment matrix, *A*.







QR computes a set of maximal linearly independent structures.

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

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

Numerical Encoding of Proteins in a Multiple Alignment

Encoding Structure

Rotated Cartesian + Gap = 4-space

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

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

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

adjustable parameter

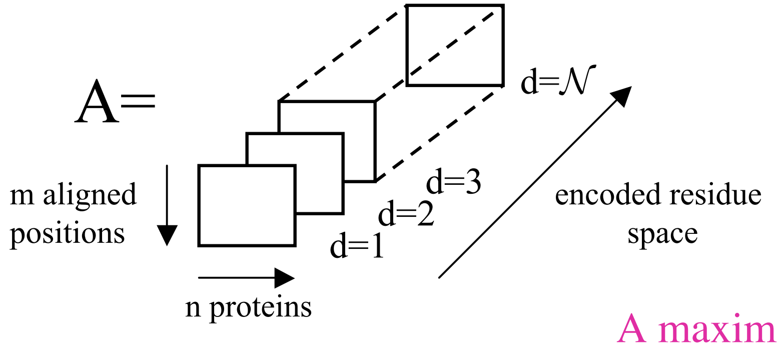
Sequence Space

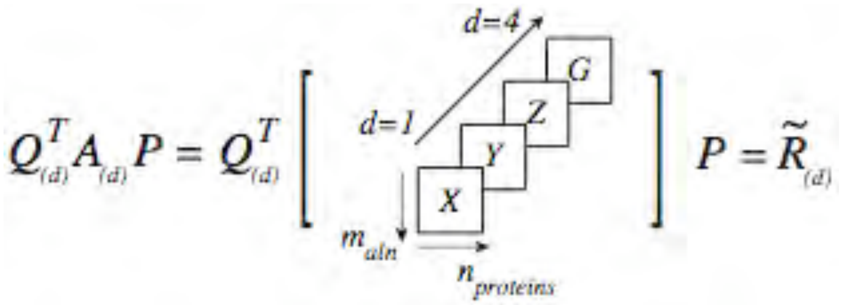
Orthogonal Encoding = 24-space

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

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

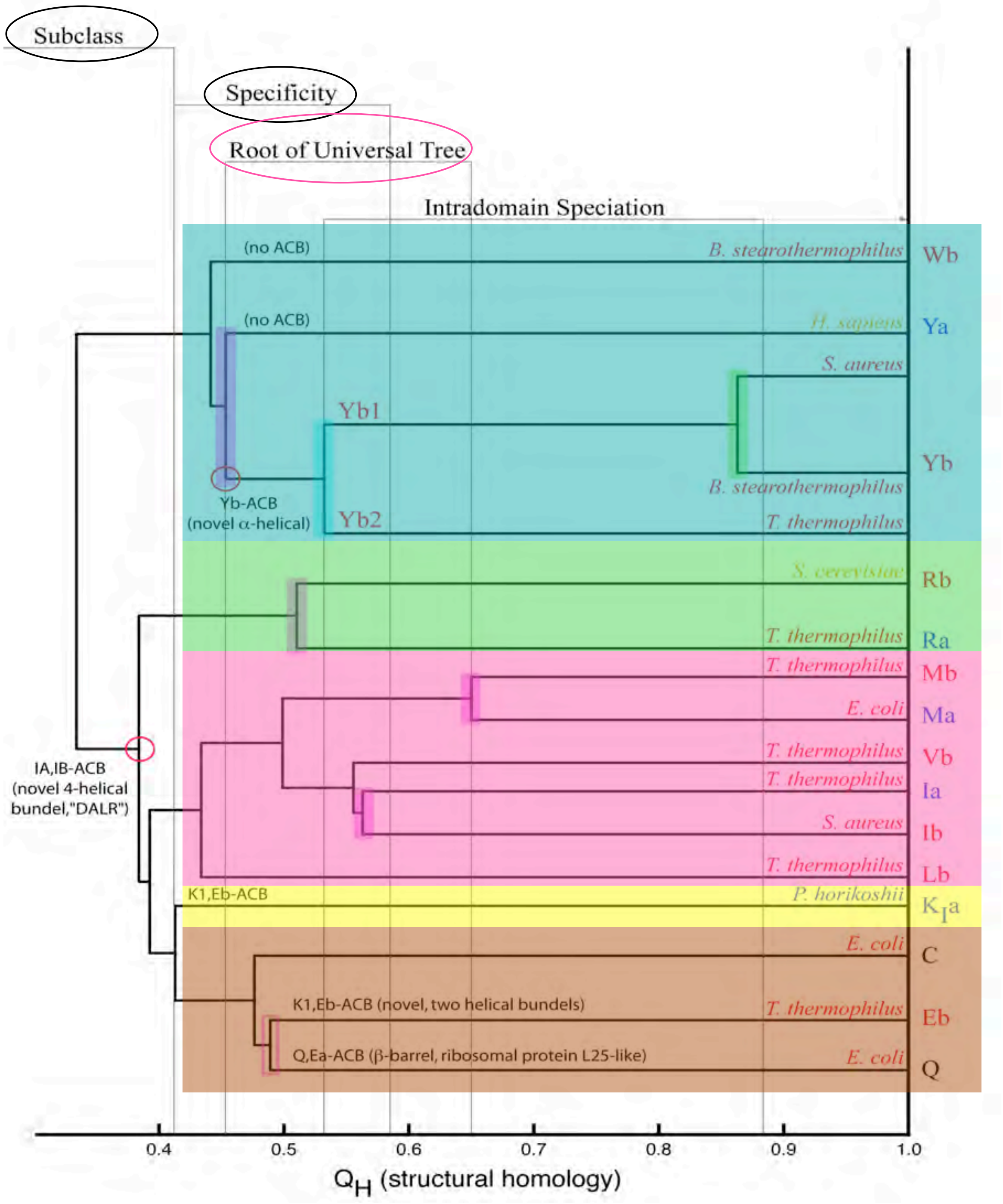
Alignment is a Matrix with Linearly Dependent Columns





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

Class I AARSs evolutionary events

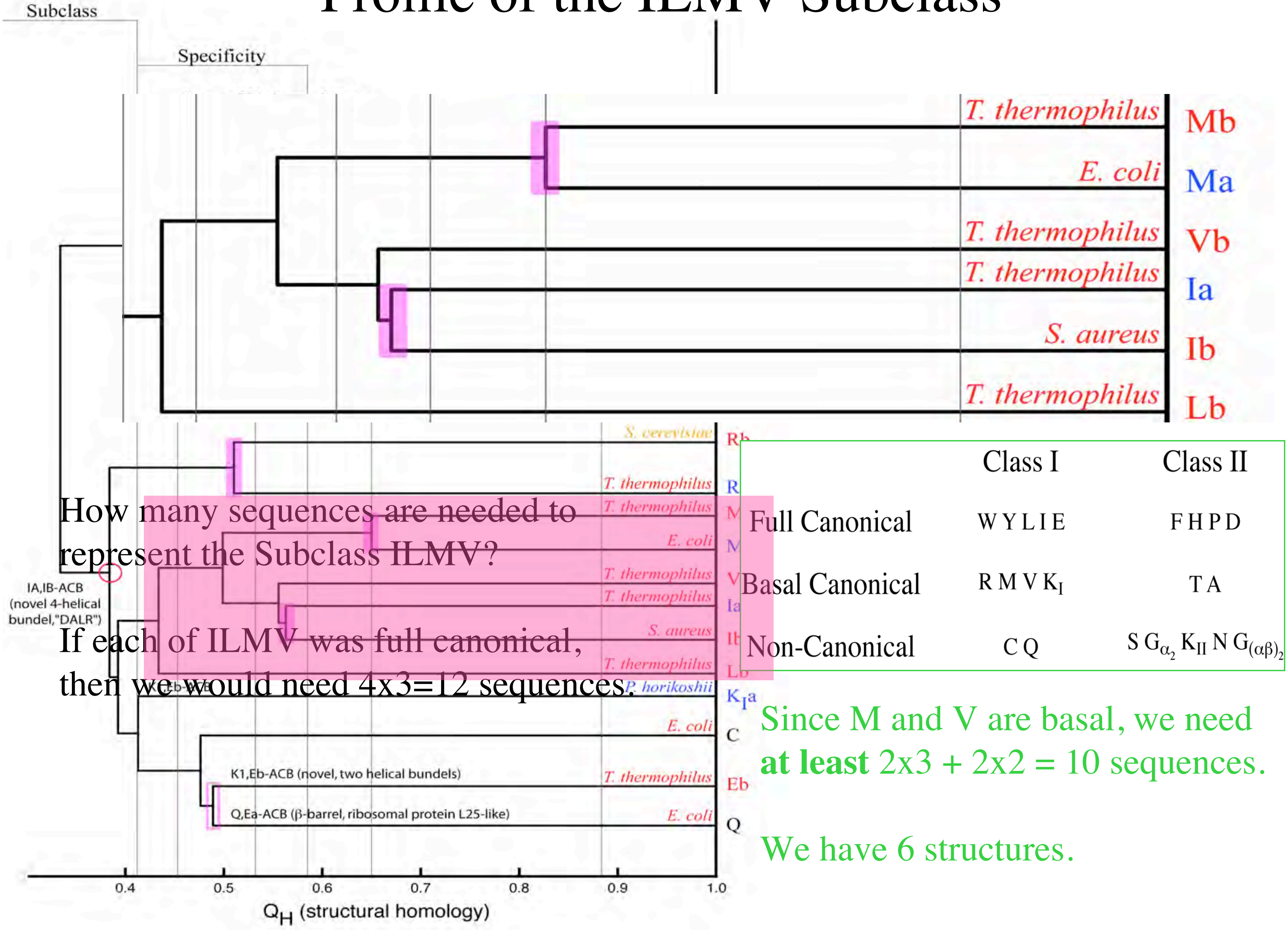


5 Subclasses

Specificity – 11 Amino acids

Domain of life A,B,E

Profile of the ILMV Subclass



How many sequences are needed to represent the Subclass ILMV?

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

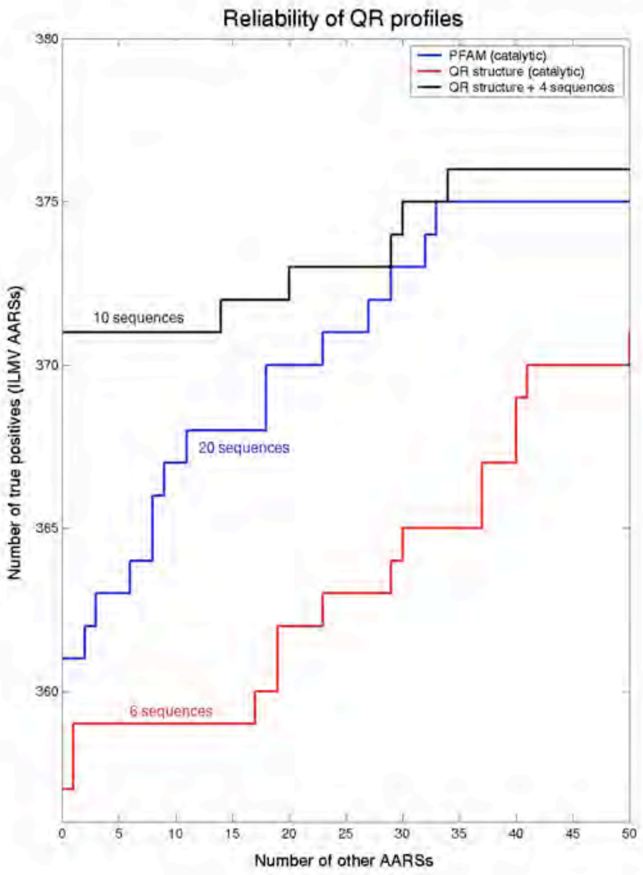
	Class I	Class II
Full Canonical	W Y L I E	F H P D
Basal Canonical	R M V K_I	T A
Non-Canonical	C Q	S G_{α_2} K_{II} N $G_{(\alpha\beta)_2}$

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

We have 6 structures.

Evolutionary Profiles for Homology Recognition

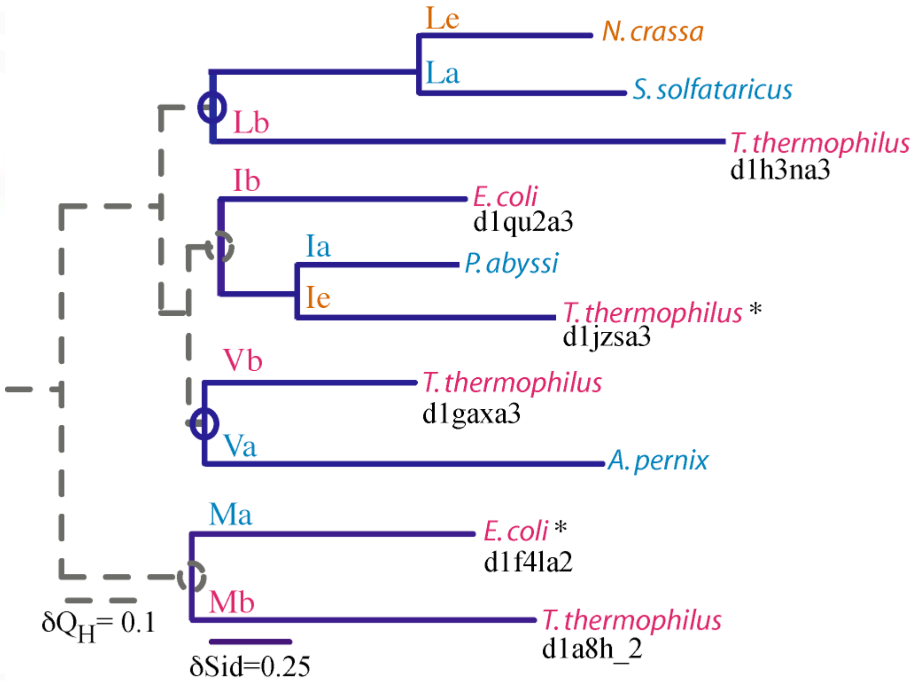
AARS Subclass ILMV



false positives

Combined Structure-Sequence Phylogeny

an evolutionary profile of the AARS subclass IA



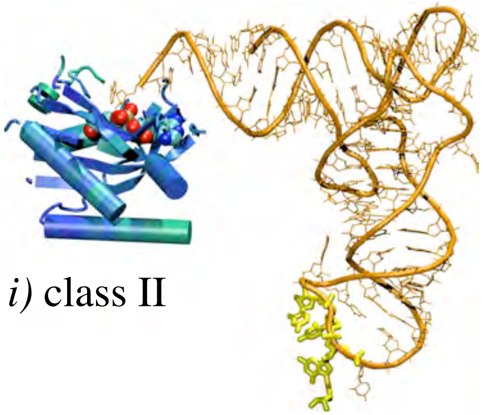
Pfam profile composition

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

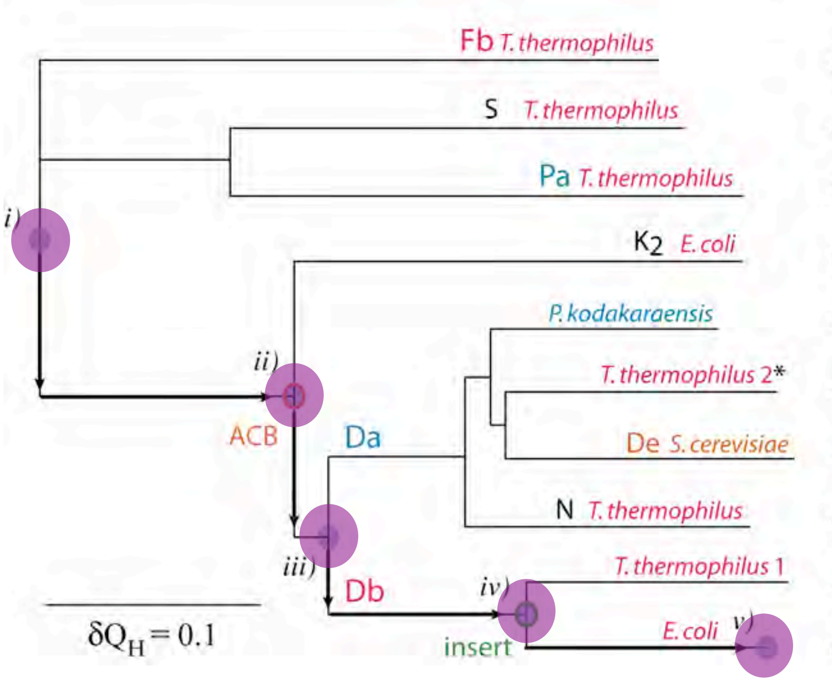
The composition of the profile matters.
Choosing the right 10 sequence makes all the difference.

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

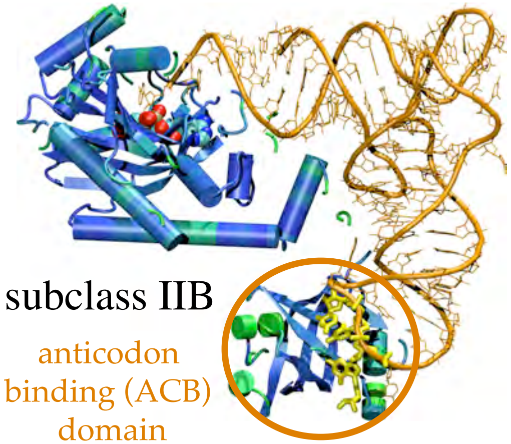
Evolution of Structure and Function in AspRS



i) class II

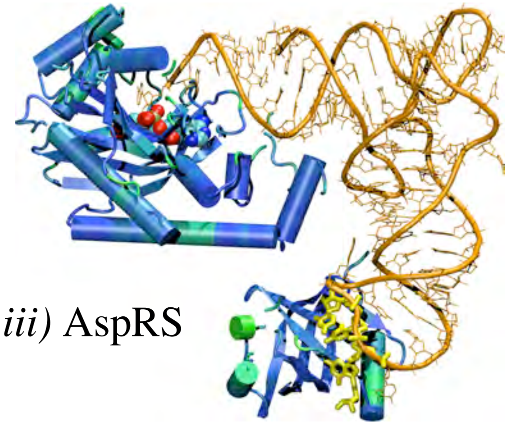


SCOP	QR order		
d1b70a_	①		
d1serb2	③		
d1h4sb2	6		
d1bbua2	4	②	
d1b8ab2	9	5	4
d1n9wb2	10	7	6
d1asza2	5	③	3
d11sca2	7	4	②
d1efwa3	8	6	5
d1c0aa3	②	①	①

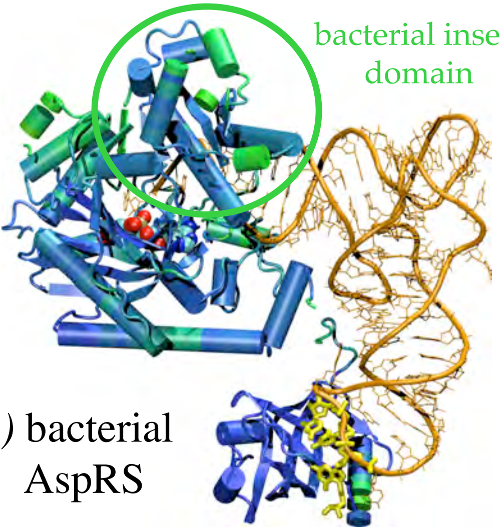


ii) subclass IIB

anticodon binding (ACB) domain

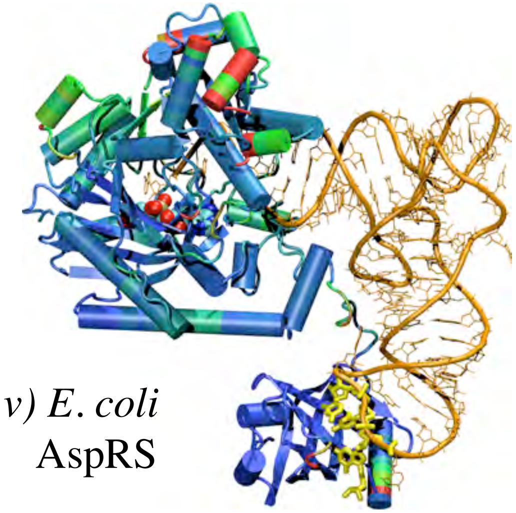


iii) AspRS



bacterial insert domain

iv) bacterial AspRS



v) E. coli AspRS

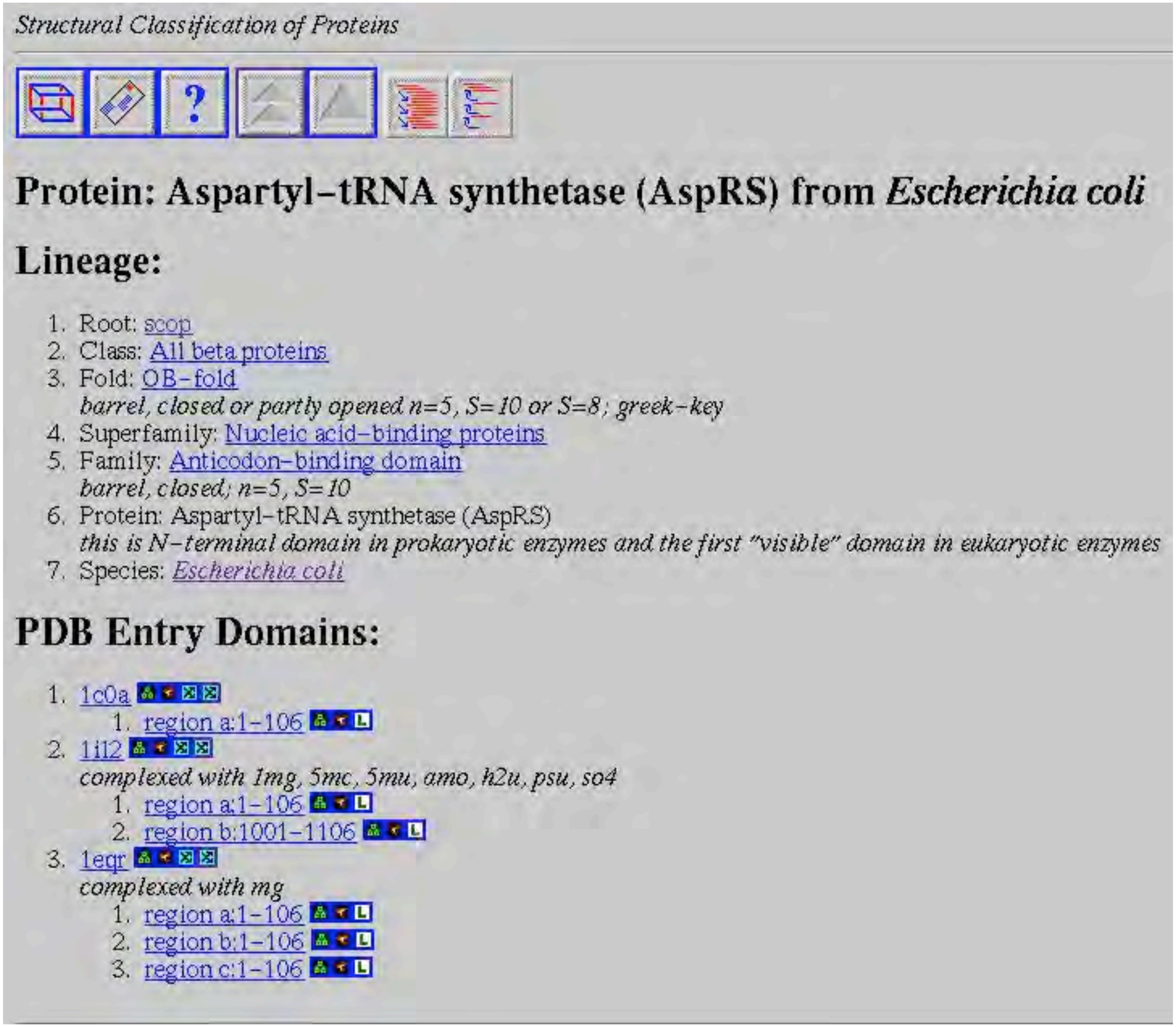
Structural Profiles

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

   Add more structural information

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

Structural Domains



Structural Classification of Proteins

Protein: Aspartyl-tRNA synthetase (AspRS) from *Escherichia coli*

Lineage:

1. Root: scop
2. Class: All beta proteins
3. Fold: OB-fold
   barrel, closed or partly opened n=5, S=10 or S=8; greek-key
4. Superfamily: Nucleic acid-binding proteins
5. Family: Anticodon-binding domain
   barrel, closed; n=5, S=10
6. Protein: Aspartyl-tRNA synthetase (AspRS)
   this is N-terminal domain in prokaryotic enzymes and the first "visible" domain in eukaryotic enzymes
7. Species: *Escherichia coli*

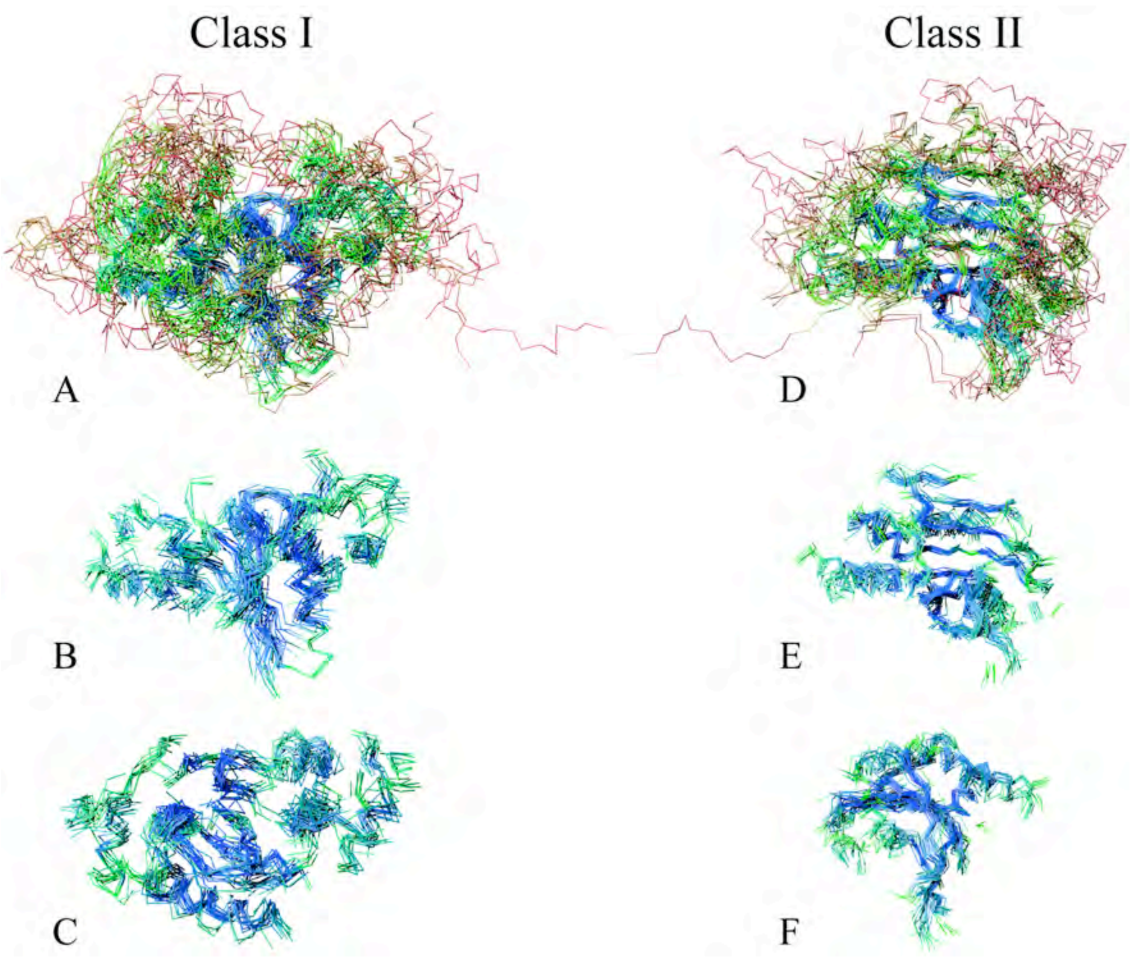
PDB Entry Domains:

1. 1c0a
   1. region a:1-106
2. 1il2
   complexed with 1mg, 5mc, 5mu, amo, h2u, psu, so4
   1. region a:1-106
   2. region b:1001-1106
3. 1eqr
   complexed with mg
   1. region a:1-106
   2. region b:1-106
   3. region c:1-106

Profile - Multiple Structural Alignments

Representative Profile of AARS Family

Catalytic Domain

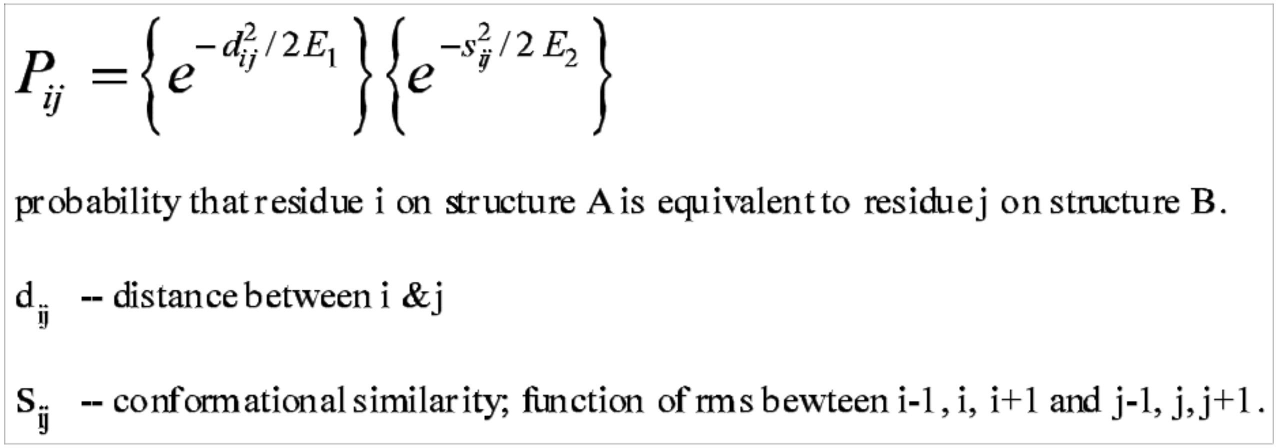


STAMP - Multiple Structural Alignments

1. Initial Alignment Inputs

- Multiple Sequence alignment
- Ridged Body “Scan”

2. Refine Initial Alignment & Produce Multiple Structural Alignment



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

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 bewteen 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.

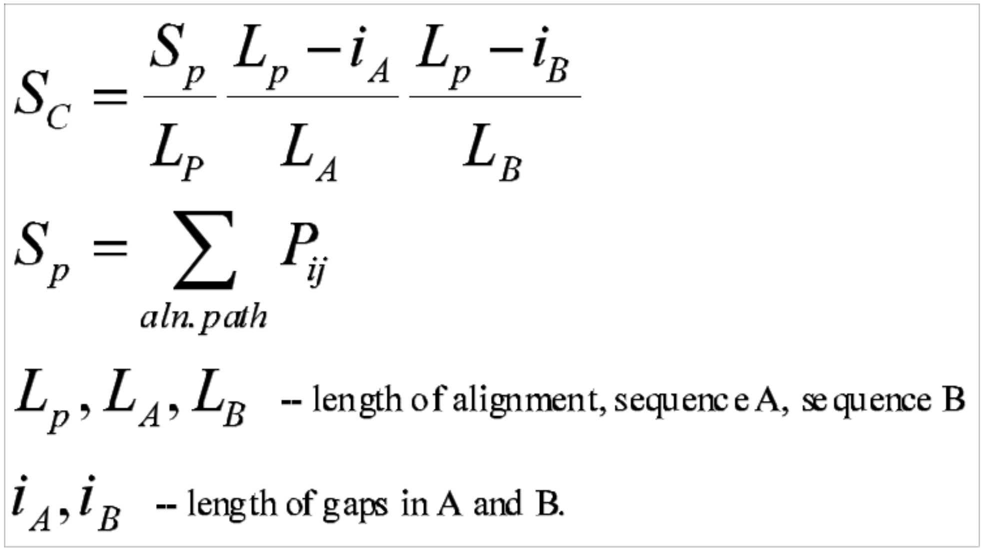
R. Russell, G. Barton (1992) *Proteins* **14:** 309.

Multiple Structural Alignments

STAMP – cont'd

2. Refine Initial Alignment & Produce Multiple Structural Alignment

Alignment score:



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

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

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.

Planned Tools in MultiSeq

Protein / RNA Sequence Data

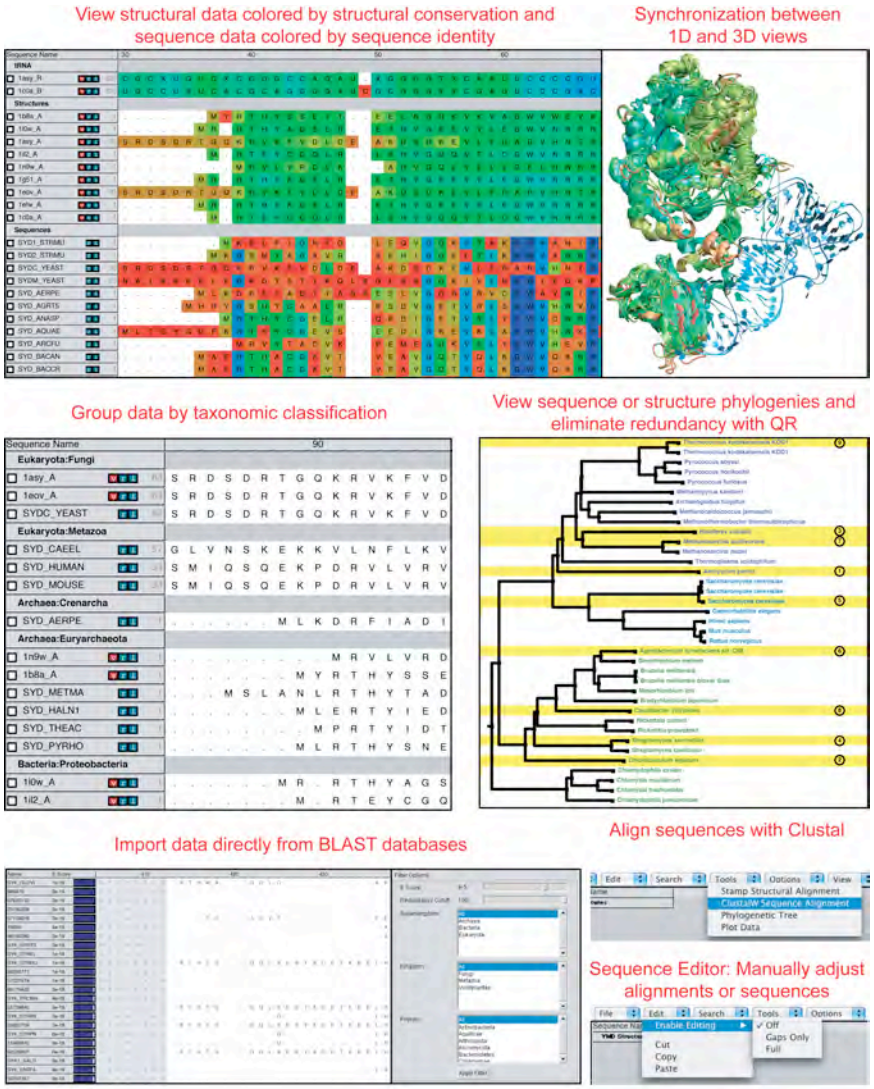
Entire SwissProt DB, 100,000+ RNA seqs

Metadata Information, Clustal & Phylogenetic Trees

Incorporate genomic content

Blast & PsiBlast

Sequence Editor


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

Synchronization between 1D and 3D views

Group data by taxonomic classification

View sequence or structure phylogenies and eliminate redundancy with QR

Import data directly from BLAST databases

Align sequences with Clustal

Sequence Editor: Manually adjust alignments or sequences

Sequence / Structure Alignment

RNA Secondary Structure

QR non-redundant seq / str sets

Cluster analysis / Bioinformatics scripting

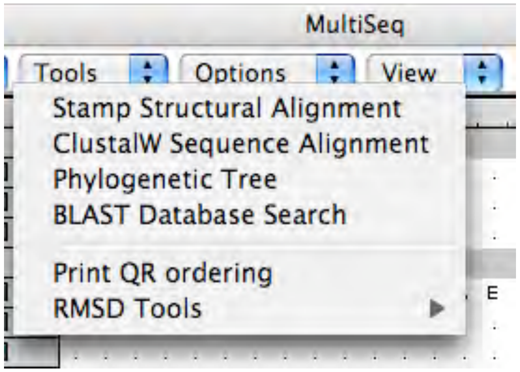
Tutorials MultiSeq/AARS

EF-Tu/Ribosome

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)

What is MultiSeq?

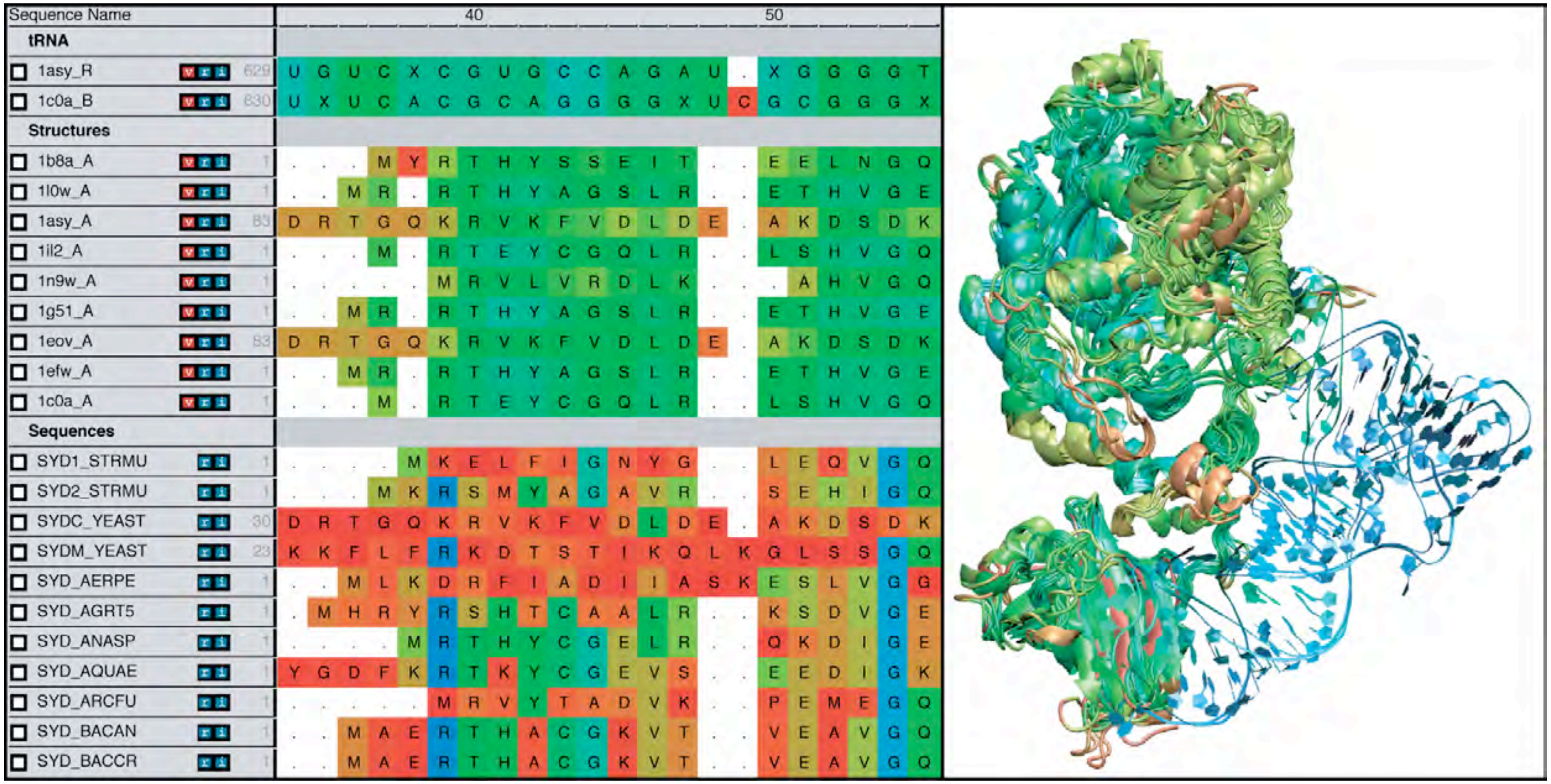
- MultiSeq is an extension to VMD that provides an environment to combine sequence and structure data
- A platform for performing bioinformatics analyses within the framework of evolution
- Provides software for improving the signal-to-noise ratio in an evolutionary analysis by eliminating redundancy (StructQR, SeqQR, Evolutionary Profiles “EP”)
- Visualizes computationally derived metrics (Q_{res}, Q_{H},..) or imported experimental properties



- Integrates popular bioinformatics tools making them easier to use and reducing the barrier to performing bioinformatics analysis (ClustalW, STAMP, BLAST)

MultiSeq Combines Sequence and Structure

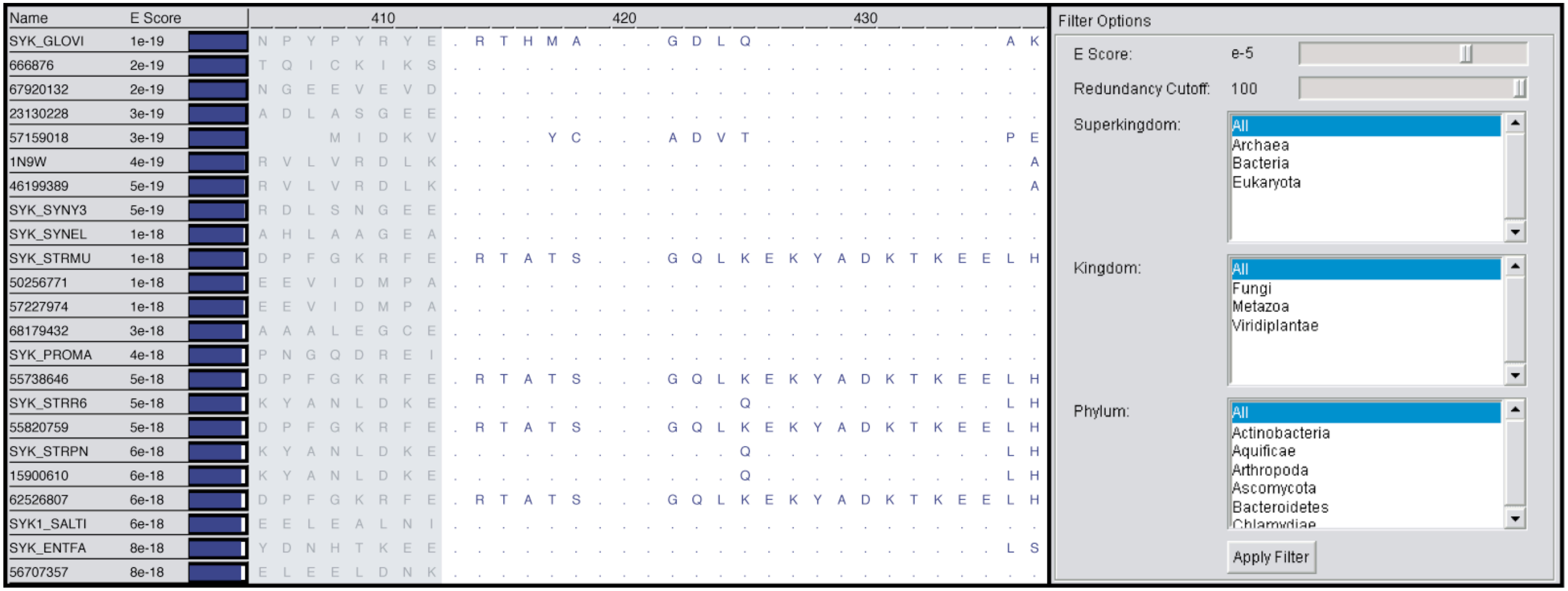
- Align sequences or structures; manually edit alignments
- View data colored by numerous metrics including structural conservation and sequence similarity
- Synchronized coloring between 1D and 3D views



Sequence Name									40										50					
tRNA																								
□ 1asy_R	v r i	629	U	G	U	C	X	C	G	U	G	C	C	A	G	A	U	.	X	G	G	G	G	T
□ 1c0a_B	v r i	630	U	X	U	C	A	C	G	C	A	G	G	G	G	X	U	C	G	C	G	G	G	X
Structures																								
□ 1b8a_A	v r i	1	.	.	.	M	Y	R	T	H	Y	S	S	E	I	T	.	.	E	E	L	N	G	Q
□ 1l0w_A	v r i	1	.	.	M	R	.	R	T	H	Y	A	G	S	L	R	.	.	E	T	H	V	G	E
□ 1asy_A	v r i	83	D	R	T	G	Q	K	R	V	K	F	V	D	L	D	E	.	A	K	D	S	D	K
□ 1il2_A	v r i	1	.	.	.	M	.	R	T	E	Y	C	G	Q	L	R	.	.	L	S	H	V	G	Q
□ 1n9w_A	v r i	1	.	.	.	.	.	M	R	V	L	V	R	D	L	K	.	.	.	A	H	V	G	Q
□ 1g51_A	v r i	1	.	.	M	R	.	R	T	H	Y	A	G	S	L	R	.	.	E	T	H	V	G	E
□ 1eov_A	v r i	83	D	R	T	G	Q	K	R	V	K	F	V	D	L	D	E	.	A	K	D	S	D	K
□ 1efw_A	v r i	1	.	.	M	R	.	R	T	H	Y	A	G	S	L	R	.	.	E	T	H	V	G	E
□ 1c0a_A	v r i	1	.	.	.	M	.	R	T	E	Y	C	G	Q	L	R	.	.	L	S	H	V	G	Q
Sequences																								
□ SYD1_STRMU	r i	1	.	.	.	.	M	K	E	L	F	I	G	N	Y	G	.	.	L	E	Q	V	G	Q
□ SYD2_STRMU	r i	1	.	.	.	M	K	R	S	M	Y	A	G	A	V	R	.	.	S	E	H	I	G	Q
□ SYDC_YEAST	r i	30	D	R	T	G	Q	K	R	V	K	F	V	D	L	D	E	.	A	K	D	S	D	K
□ SYDM_YEAST	r i	23	K	K	F	L	F	R	K	D	T	S	T	I	K	Q	L	K	G	L	S	S	G	Q
□ SYD_AERPE	r i	1	.	.	M	L	K	D	R	F	I	A	D	I	I	A	S	K	E	S	L	V	G	G
□ SYD_AGRT5	r i	1	.	M	H	R	Y	R	S	H	T	C	A	A	L	R	.	.	K	S	D	V	G	E
□ SYD_ANASP	r i	1	.	.	.	.	M	R	T	H	Y	C	G	E	L	R	.	.	Q	K	D	I	G	E
□ SYD_AQUAE	r i	1	Y	G	D	F	K	R	T	K	Y	C	G	E	V	S	.	.	E	E	D	I	G	K
□ SYD_ARCFU	r i	1	.	.	.	.	.	M	R	V	Y	T	A	D	V	K	.	.	P	E	M	E	G	Q
□ SYD_BACAN	r i	1	.	.	M	A	E	R	T	H	A	C	G	K	V	T	.	.	V	E	A	V	G	Q
□ SYD_BACCR	r i	1	.	.	M	A	E	R	T	H	A	C	G	K	V	T	.	.	V	E	A	V	G	Q

BLAST DB Searching

- Import sequence data directly from BLAST databases
- Search using a single sequence or an EP profile
- Filter results based on taxonomy or redundancy (QR)



```
Name        E Score            410                 420                 430
SYK_GLOVI   1e-19     N P Y P Y R Y E . R T H M A . . . G D L Q . . . . . . . . . . A K
666876      2e-19     T Q I C K I K S . . . . . . . . . . . . . . . . . . . . . . . . .
67920132    2e-19     N G E E V E V D . . . . . . . . . . . . . . . . . . . . . . . . .
23130228    3e-19     A D L A S G E E . . . . . . . . . . . . . . . . . . . . . . . . .
57159018    3e-19           M I D K V . . . . Y C . . . A D V T . . . . . . . . . . P E
1N9W        4e-19     R V L V R D L K . . . . . . . . . . . . . . . . . . . . . . . . A
46199389    5e-19     R V L V R D L K . . . . . . . . . . . . . . . . . . . . . . . . A
SYK_SYNY3   5e-19     R D L S N G E E . . . . . . . . . . . . . . . . . . . . . . . . .
SYK_SYNEL   1e-18     A H L A A G E A . . . . . . . . . . . . . . . . . . . . . . . . .
SYK_STRMU   1e-18     D P F G K R F E . R T A T S . . . G Q L K E K Y A D K T K E E L H
50256771    1e-18     E E V I D M P A . . . . . . . . . . . . . . . . . . . . . . . . .
57227974    1e-18     E E V I D M P A . . . . . . . . . . . . . . . . . . . . . . . . .
68179432    3e-18     A A A L E G C E . . . . . . . . . . . . . . . . . . . . . . . . .
SYK_PROMA   4e-18     P N G Q D R E I . . . . . . . . . . . . . . . . . . . . . . . . .
55738646    5e-18     D P F G K R F E . R T A T S . . . G Q L K E K Y A D K T K E E L H
SYK_STRR6   5e-18     K Y A N L D K E . . . . . . . . . . . . Q . . . . . . . . . . L H
55820759    5e-18     D P F G K R F E . R T A T S . . . G Q L K E K Y A D K T K E E L H
SYK_STRPN   6e-18     K Y A N L D K E . . . . . . . . . . . . Q . . . . . . . . . . L H
15900610    6e-18     K Y A N L D K E . . . . . . . . . . . . Q . . . . . . . . . . L H
62526807    6e-18     D P F G K R F E . R T A T S . . . G Q L K E K Y A D K T K E E L H
SYK1_SALTI  6e-18     E E L E A L N I . . . . . . . . . . . . . . . . . . . . . . . . .
SYK_ENTFA   8e-18     Y D N H T K E E . . . . . . . . . . . . . . . . . . . . . . . L S
56707357    8e-18     E L E E L D N K . . . . . . . . . . . . . . . . . . . . . . . . .
```

Filter Options

E Score: e-5

Redundancy Cutoff: 100

Superkingdom: All
Archaea
Bacteria
Eukaryota

Kingdom: All
Fungi
Metazoa
Viridiplantae

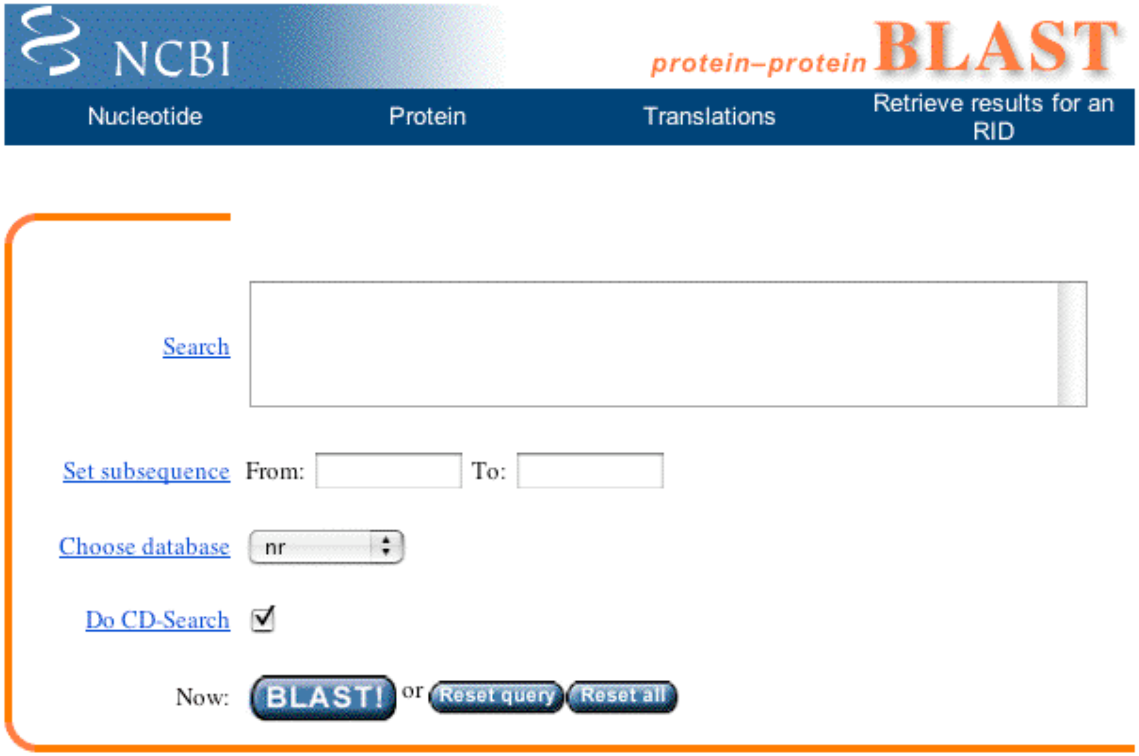
Phylum: All
Actinobacteria
Aquificae
Arthropoda
Ascomycota
Bacteroidetes
Chlamydiae

Protein sequence alignment
How do I align two similar, but different sequences

Sequence 1: $a_1\ a_2\ a_3 - \ - \ a_4\ a_5 \ldots a_n$

Sequence 2: $c_1\ - \ c_2\ c_3\ c_4\ c_5 - \ldots c_m$

There exist web accessible tools, e.g., BLAST search: **http://www.ncbi.nlm.nih.gov/**



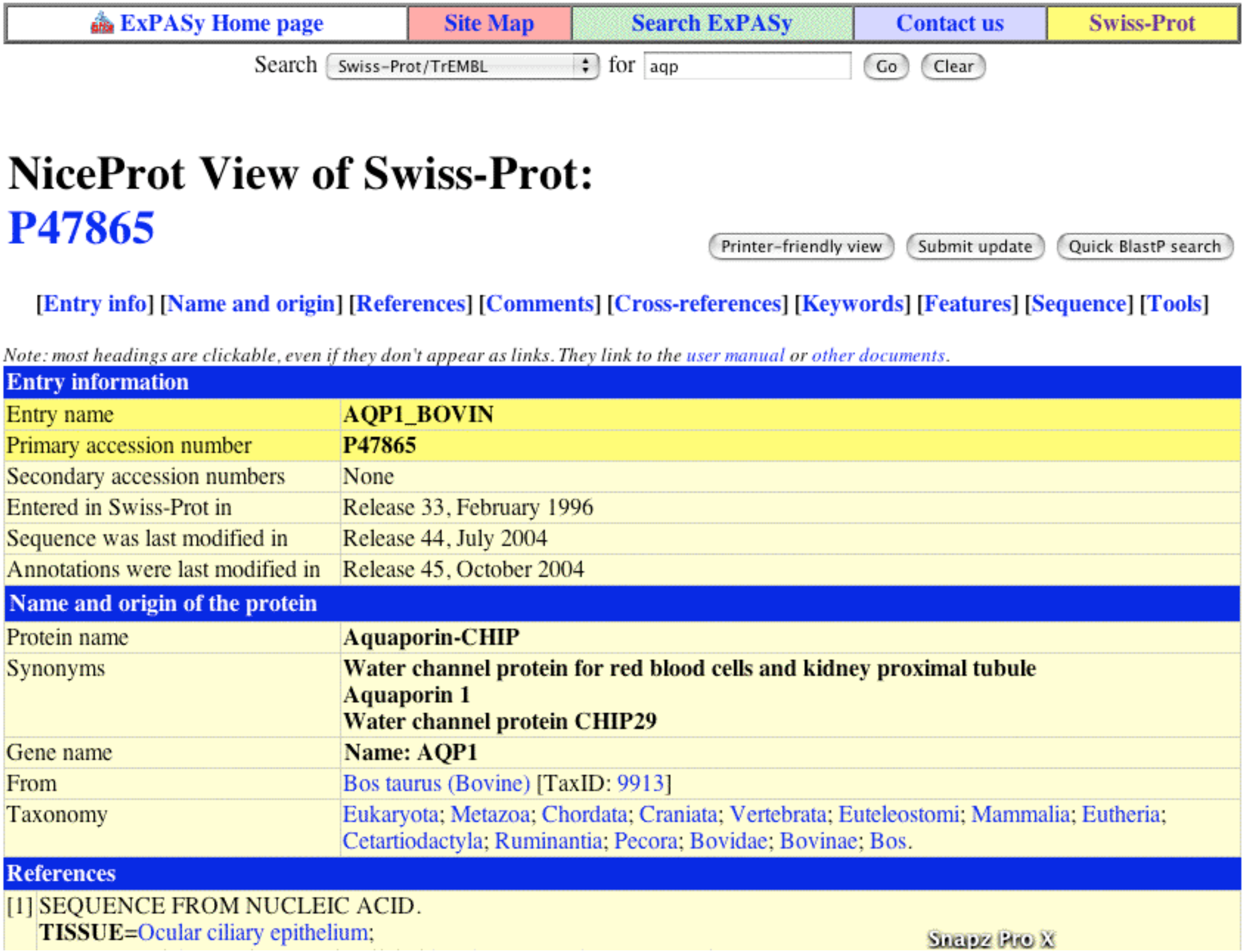
Search

Set subsequence From: To:

Choose database nr

Do CD-Search ☑

Now:



Search Swiss-Prot/TrEMBL for aqp

NiceProt View of Swiss-Prot: P47865

[Entry info] [Name and origin] [References] [Comments] [Cross-references] [Keywords] [Features] [Sequence] [Tools]

Note: most headings are clickable, even if they don't appear as links. They link to the user manual or other documents.

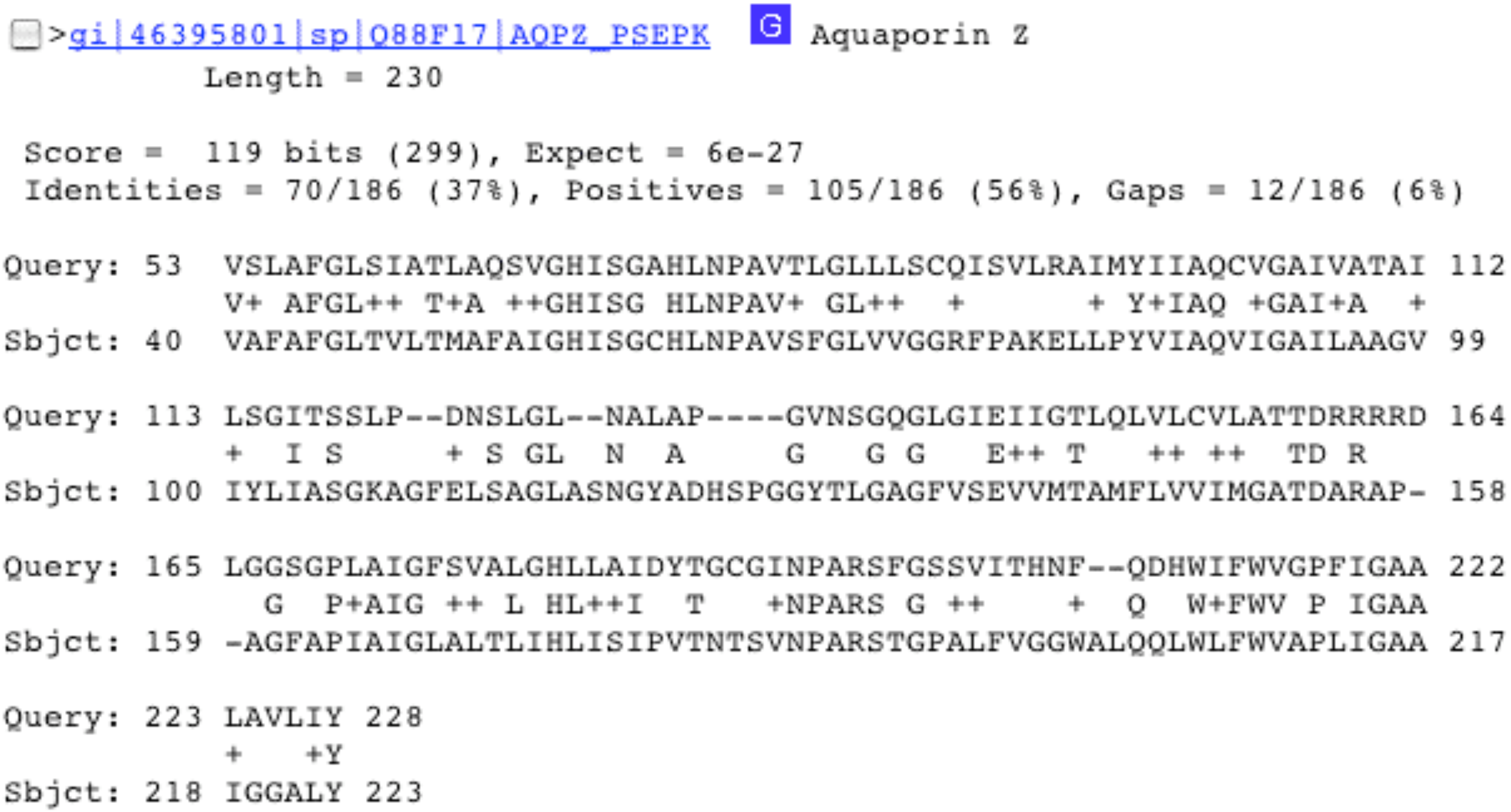
Entry information	
Entry name	**AQP1_BOVIN**
Primary accession number	**P47865**
Secondary accession numbers	None
Entered in Swiss-Prot in	Release 33, February 1996
Sequence was last modified in	Release 44, July 2004
Annotations were last modified in	Release 45, October 2004
Name and origin of the protein	
Protein name	**Aquaporin-CHIP**
Synonyms	**Water channel protein for red blood cells and kidney proximal tubule** **Aquaporin 1** **Water channel protein CHIP29**
Gene name	**Name: AQP1**
From	Bos taurus (Bovine) [TaxID: 9913]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Cetartiodactyla; Ruminantia; Pecora; Bovidae; Bovinae; Bos.

References

[1] SEQUENCE FROM NUCLEIC ACID.
TISSUE=Ocular ciliary epithelium;

Scroll down to the sequence:

Final Result: Sequence Alignment

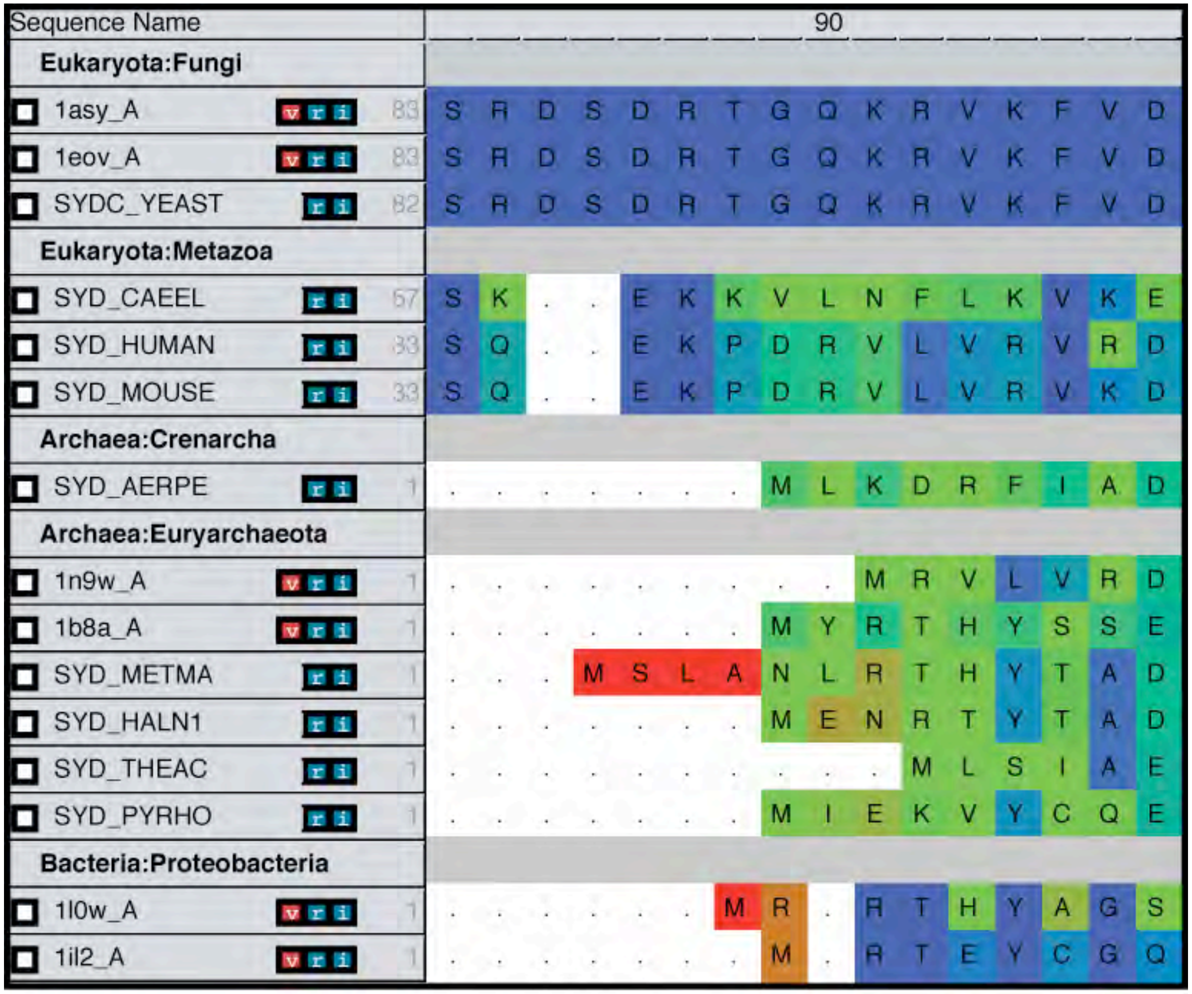


```
□>gi|46395801|sp|Q88F17|AQPZ_PSEPK  G  Aquaporin Z
          Length = 230

 Score =  119 bits (299), Expect = 6e-27
 Identities = 70/186 (37%), Positives = 105/186 (56%), Gaps = 12/186 (6%)

Query: 53  VSLAFGLSIATLAQSVGHISGAHLNPAVTLGLLLSCQISVLRAIMYIIAQCVGAIVATAI 112
            V+ AFGL++ T+A ++GHISG HLNPAV+ GL++  +      + Y+IAQ +GAI+A  +
Sbjct: 40  VAFAFGLTVLTMAFAIGHISGCHLNPAVSFGLVVGGRFPAKELLPYVIAQVIGAILAAGV 99

Query: 113 LSGITSSLP--DNSLGL--NALAP----GVNSGQGLGIEIIGTLQLVLCVLATTDRRRRD 164
            +  I S     + S GL  N  A     G   G G   E++ T   ++ ++  TD R
Sbjct: 100 IYLIASGKAGFELSAGLASNGYADHSPGGYTLGAGFVSEVVMTAMFLVVIMGATDARAP- 158

Query: 165 LGGSGPLAIGFSVALGHLLAIDYTGCGINPARSFGSSVITHNF--QDHWIFWVGPFIGAA 222
              G  P+AIG ++ L HL++I  T   +NPARS G ++    +  Q  W+FWV P IGAA
Sbjct: 159 -AGFAPIAIGLALTLIHLISIPVTNTSVNPARSTGPALFVGGWALQQLWLFWVAPLIGAA 217

Query: 223 LAVLIY 228
            +   +Y
Sbjct: 218 IGGALY 223
```

Search method returns approximate alignments - needing refinement

Flexible Grouping of Data

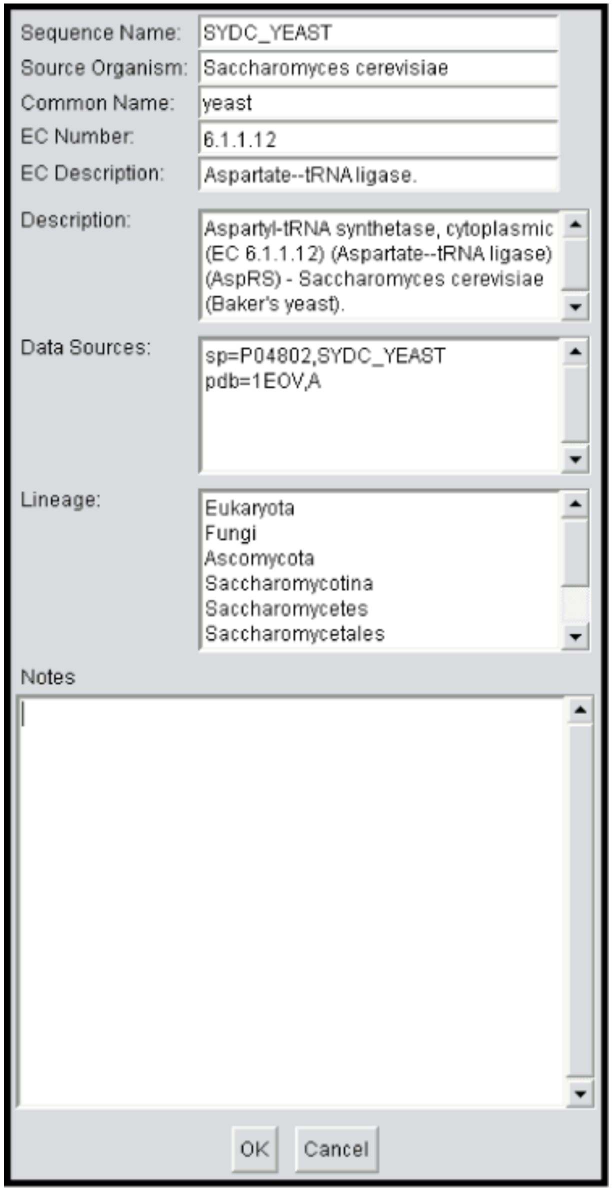
- Automatically group data by taxonomic classification to assist in evolutionary analysis (HGT) or create custom groups
- Apply metrics to groups independently, e.g bacterial signal



Sequence Name											90							
Eukaryota:Fungi																		
□ 1asy_A	v r i	83	S	R	D	S	D	R	T	G	Q	K	R	V	K	F	V	D
□ 1eov_A	v r i	83	S	R	D	S	D	R	T	G	Q	K	R	V	K	F	V	D
□ SYDC_YEAST	r i	82	S	R	D	S	D	R	T	G	Q	K	R	V	K	F	V	D
Eukaryota:Metazoa																		
□ SYD_CAEEL	r i	57	S	K	.	.	E	K	K	V	L	N	F	L	K	V	K	E
□ SYD_HUMAN	r i	33	S	Q	.	.	E	K	P	D	R	V	L	V	R	V	R	D
□ SYD_MOUSE	r i	33	S	Q	.	.	E	K	P	D	R	V	L	V	R	V	K	D
Archaea:Crenarcha																		
□ SYD_AERPE	r i	1	.	.	.	.	.	.	.	M	L	K	D	R	F	I	A	D
Archaea:Euryarchaeota																		
□ 1n9w_A	v r i	1	.	.	.	.	.	.	.	.	.	M	R	V	L	V	R	D
□ 1b8a_A	v r i	1	.	.	.	.	.	.	.	M	Y	R	T	H	Y	S	S	E
□ SYD_METMA	r i	1	.	.	.	M	S	L	A	N	L	R	T	H	Y	T	A	D
□ SYD_HALN1	r i	1	.	.	.	.	.	.	.	M	E	N	R	T	Y	T	A	D
□ SYD_THEAC	r i	1	.	.	.	.	.	.	.	.	.	.	M	L	S	I	A	E
□ SYD_PYRHO	r i	1	.	.	.	.	.	.	.	M	I	E	K	V	Y	C	Q	E
Bacteria:Proteobacteria																		
□ 1l0w_A	v r i	1	.	.	.	.	.	.	M	R	.	R	T	H	Y	A	G	S
□ 1il2_A	v r i	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 are preserved for future sessions
- Electronic Notebook: Notes and annotations about a specific sequence or structure can be added



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

MultiSeq Tutorials