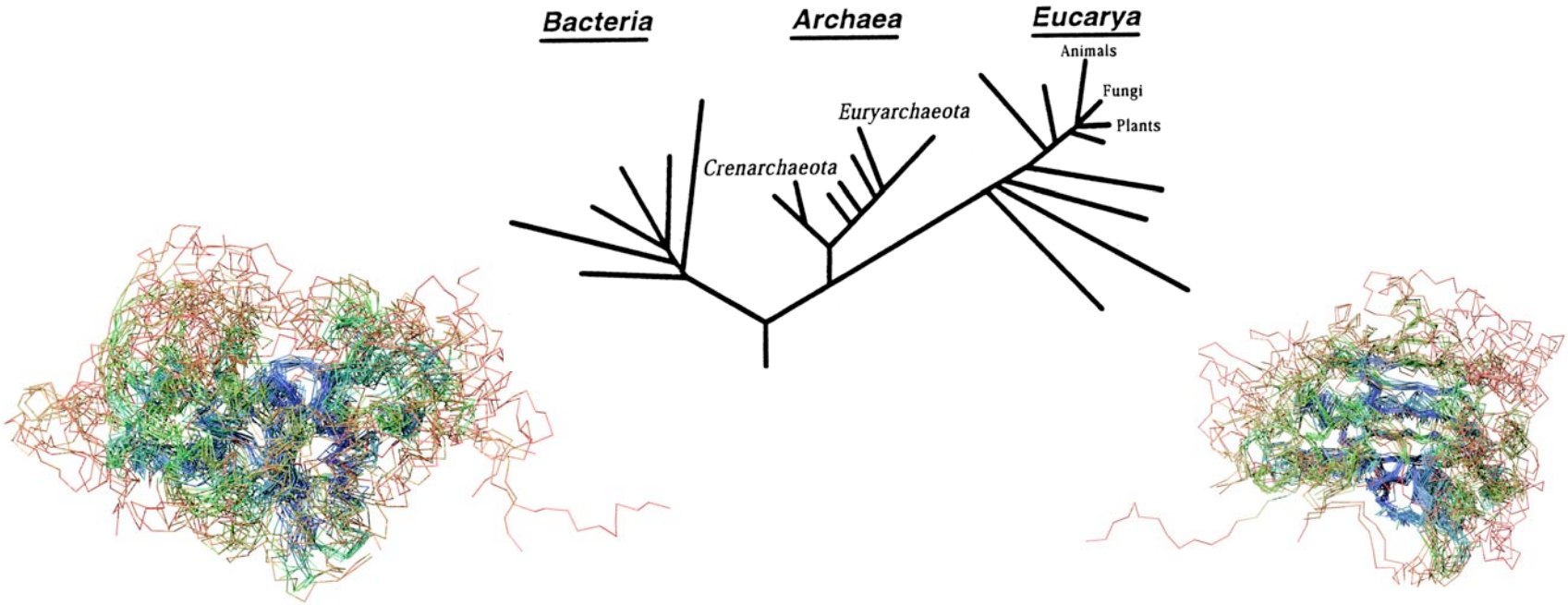
Evolution of Protein Structure

in the Aminoacyl-tRNA Synthetases



class I

class II

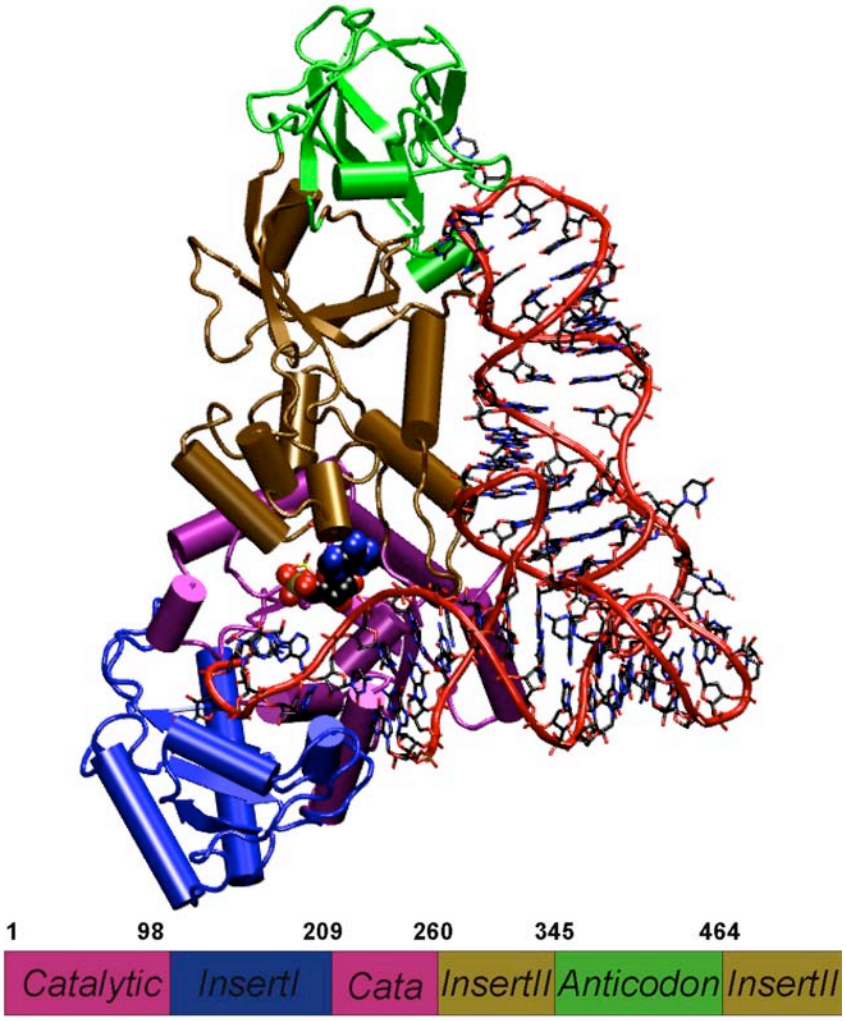
P. O'Donoghue and Z. Luthey-Schulten*

Department of Chemistry, Beckman Institute,
Center for Biophysics and Computational Biology
University of Illinois at Urbana-Champaign

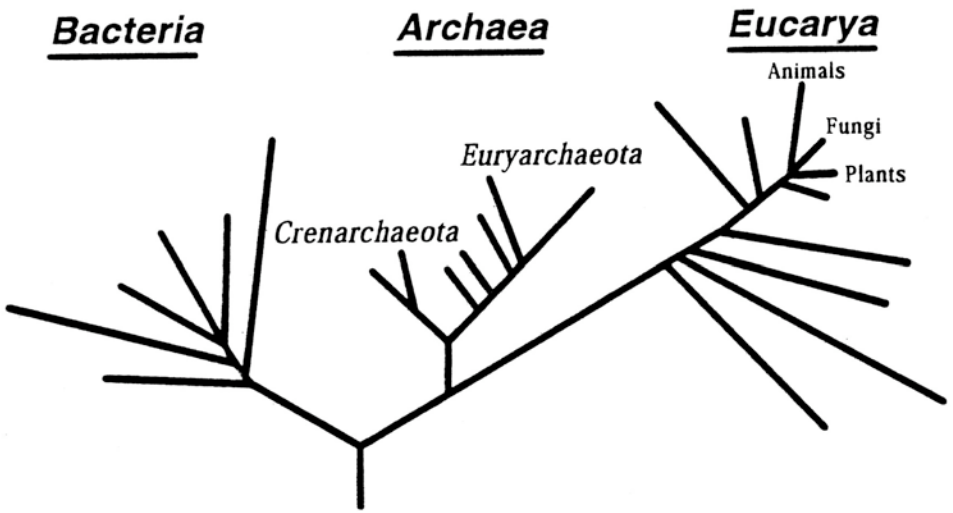
What can be learned from AARS?

- “The aminoacyl-tRNA synthetases, perhaps better than any other molecules in the cell, eptiomize the current situation and help to under standard (the effects) of HGT” Woese (PNAS, 2000; MMBR 2000)

Aminoacyl-tRNA synthetases

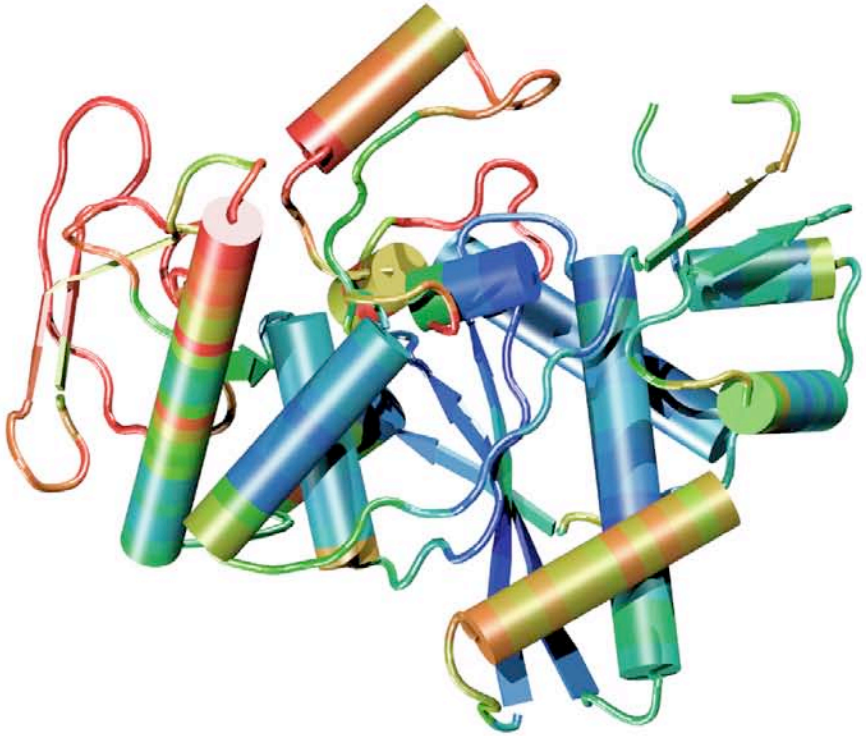


Universal Tree of Life

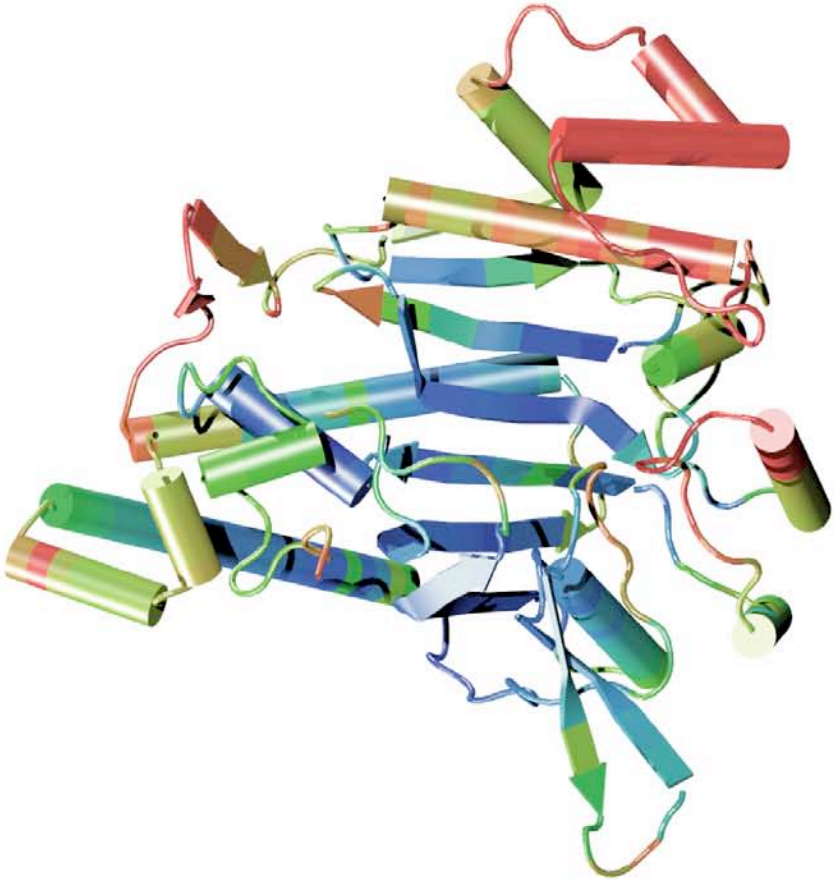


Woese *PNAS* 1990, 2002.

Structural Conservation in the Catalytic Domain of the AARSs



Class I Lysyl-tRNA Synthetase

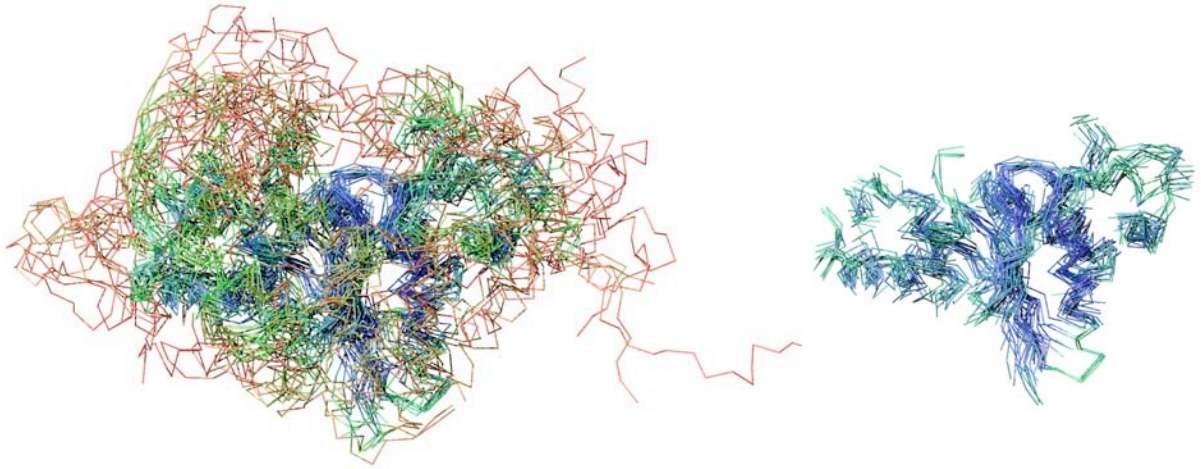


Class II Lysyl-tRNA Synthetase

Why Study the Evolution of Protein Structure?

1. Important for Homology Modeling

Better profiles improve database searches and give better alignments of distant homologs.
Allows mixing of sequence and structure information systematically.



13% sequence id
in the core (blue)

2. Learn how evolutionary dynamics changed protein shape.

Mapping a protein of unknown structure onto a homologous protein of known structure is equivalent to defining the evolutionary pathway connecting the two proteins

3. Impact on protein structure prediction, folding, and function

Evolutionary profiles increase the signal to noise ratio - Evolution is the foundation of bioinformatics.

Outline

1. Summarize evolutionary theory of the universal phylogenetic tree.

Methods

2. Introduce a structure-based metric which accounts for gaps,
   and show that evolutionary information is encoded in protein structure.

3. Introduce multidimensional QR factorization for computing
   non-redundant representative multiple alignments in sequence or structure.

Applications

4. Non-redundant multiple alignments which well represent the evolutionary history
   of a protein group provide better profiles for database searching.

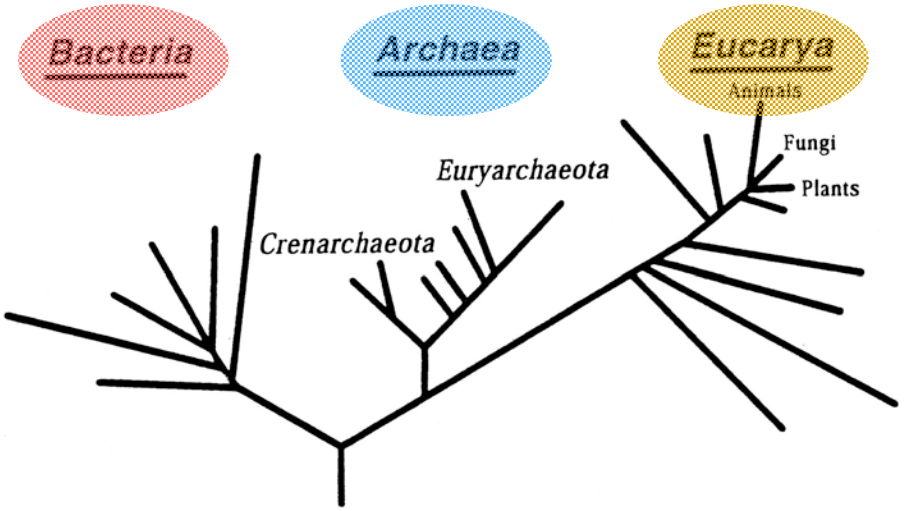
   Eliminate bias inherited from structure or sequence databases.

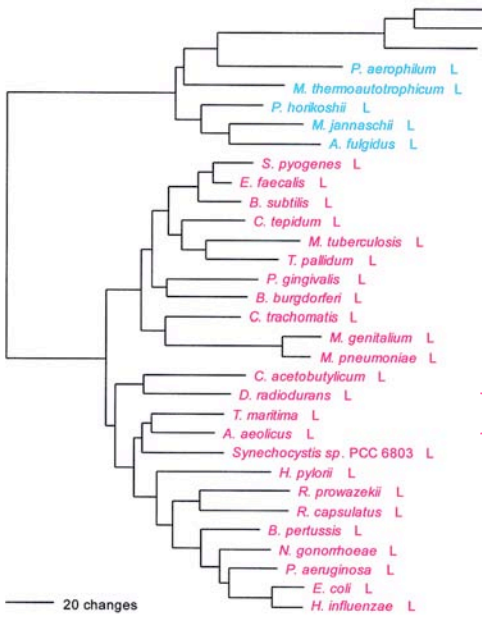
   Important for bioinformatic analysis (substitution matrices, knowledge based potentials structure pred., genome annotation) and evolutionary analysis.

5. Depict the evolution of structure and function in Aspartyl-tRNA synthetase.

Universal Phylogenetic Tree

three domains of life





Eucarya

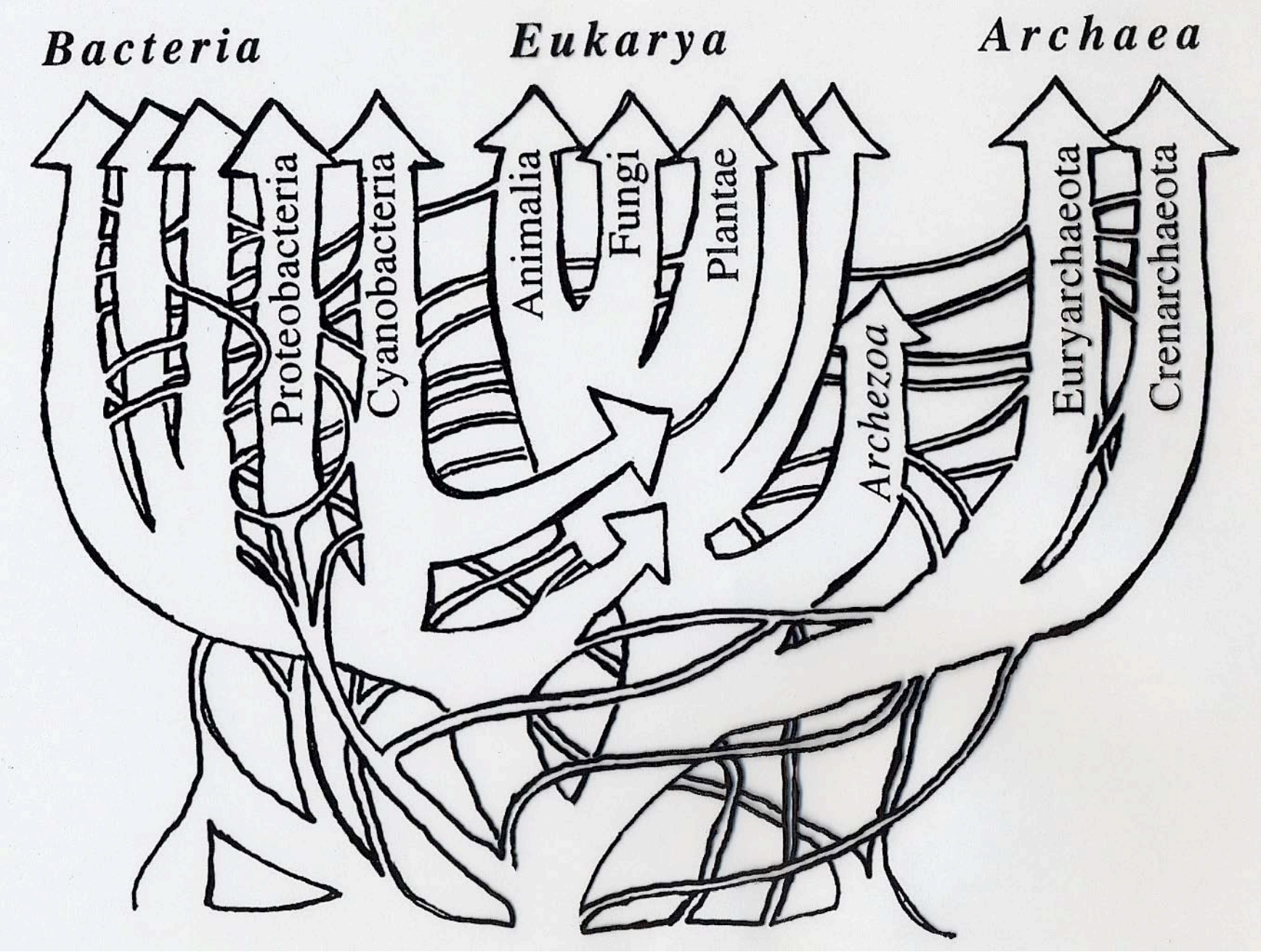
Archaea

Bacteria

Leucyl-tRNA synthetase displays the full canonical phylogenetic distribution.

for review see Woese *PNAS* 2000

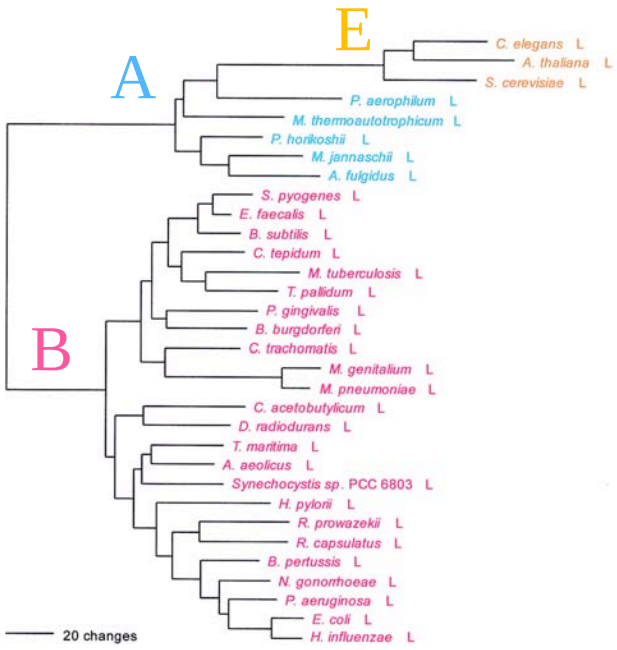
Woese, Olsen, Ibba, Soll *MMBR* 2000



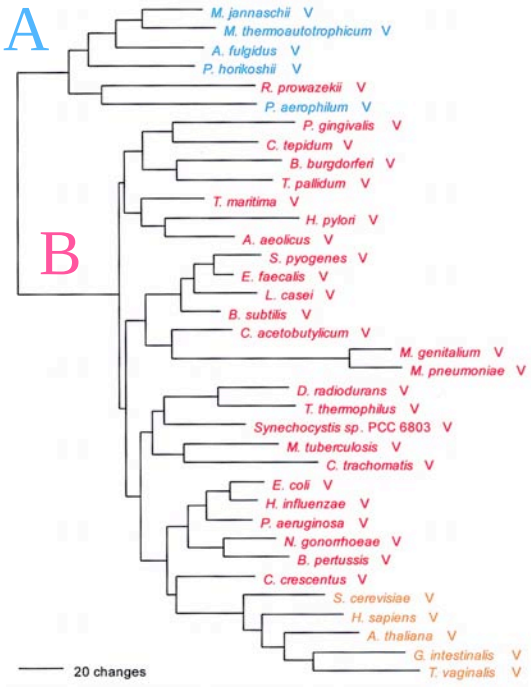
After W. Doolittle, modified by G. Olsen

Phylogenetic Distributions

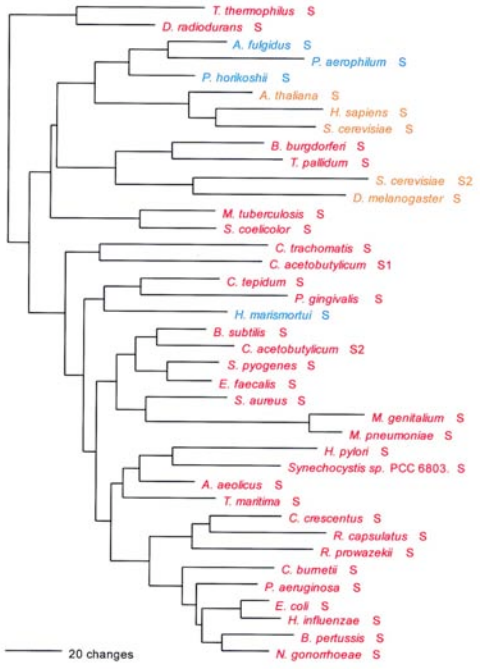
Full Canonical



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

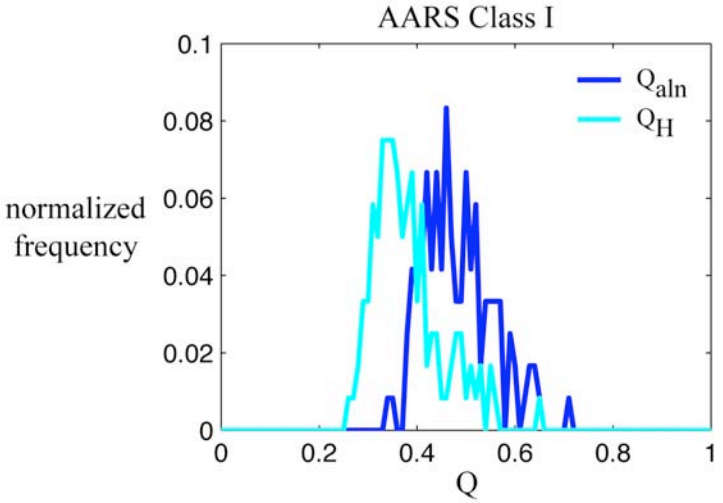
Protein Structure Similarity Measure

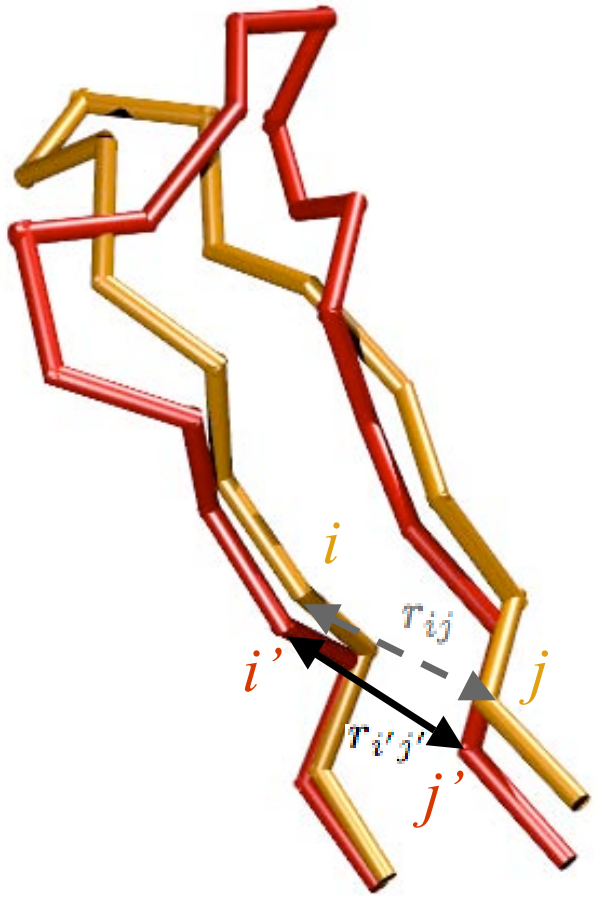
Q_H Structural Homology

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

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

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





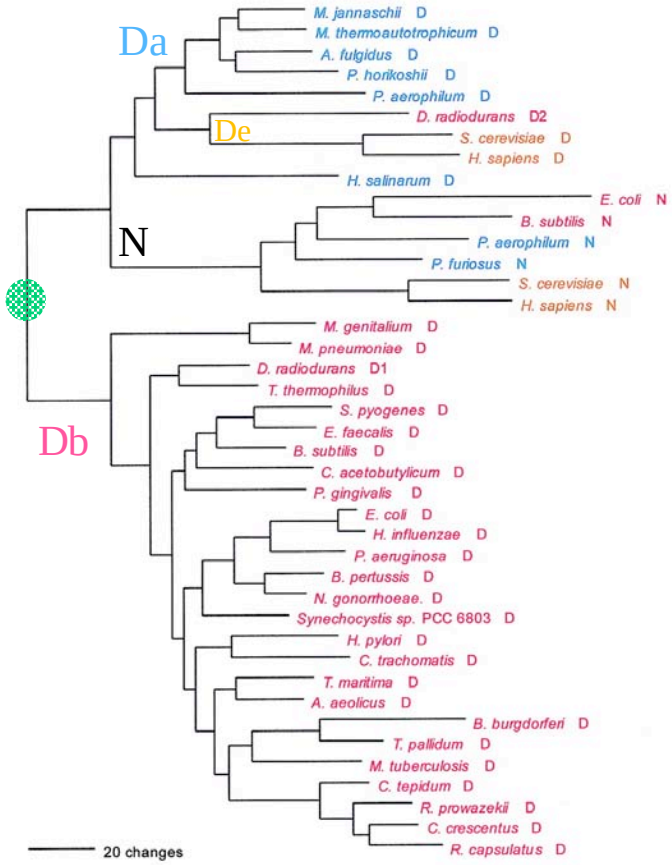
"Gaps should count as a character but not dominate" C. Woese

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

Protein structure encodes evolutionary information

Sequence Phylogeny

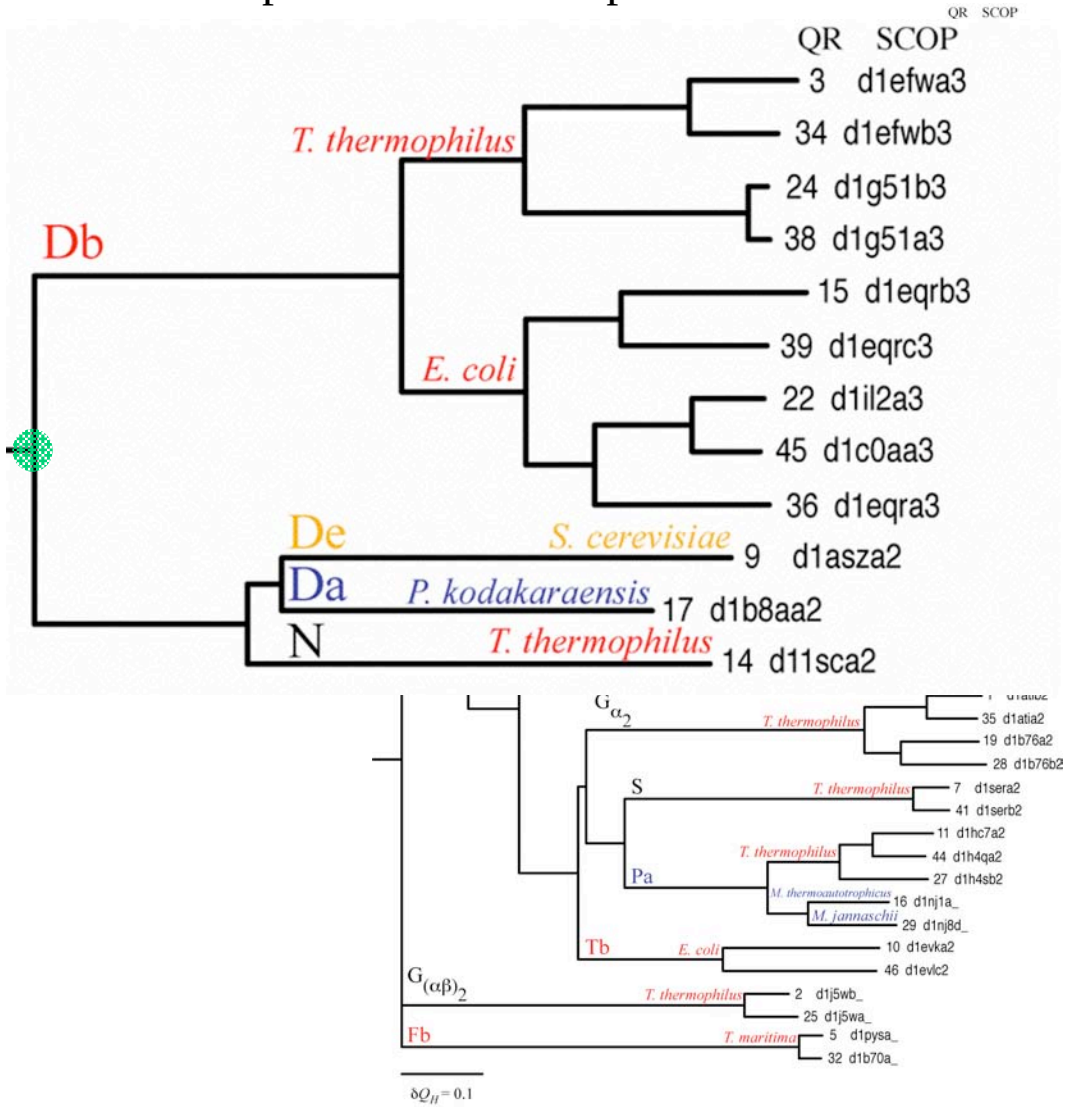
AspRS-AsnRS Group



Woese, Olsen, Ibba, Soll *MMBR* 2000

Structure Phylogeny

AspRS-AsnRS GroupClass II AARSs

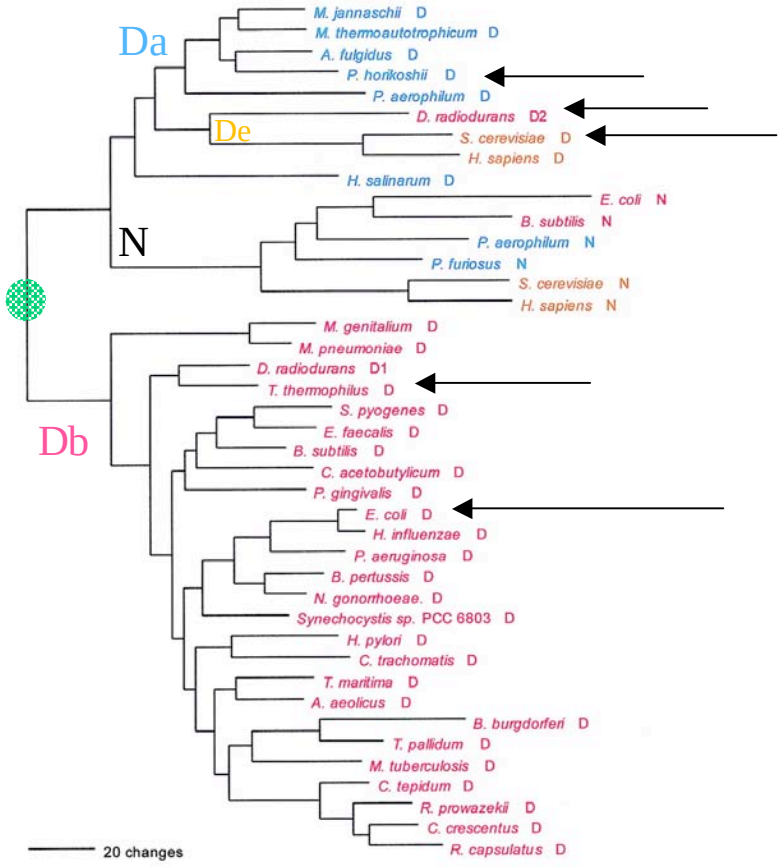


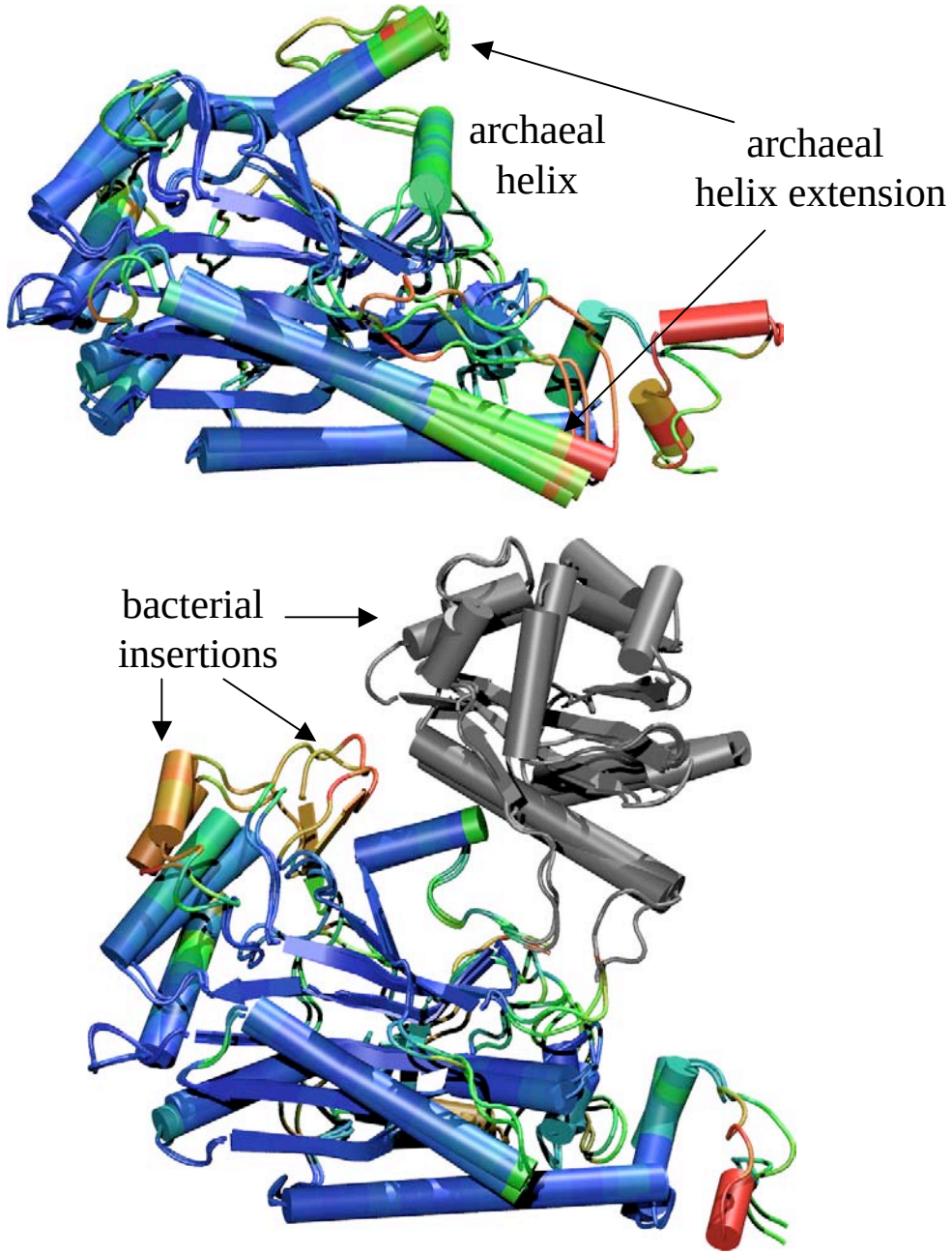
O'Donoghue & Luthey-Schulten *MMBR*.2003.

Horizontal Gene Transfer in Protein Structure

Sequence Phylogeny

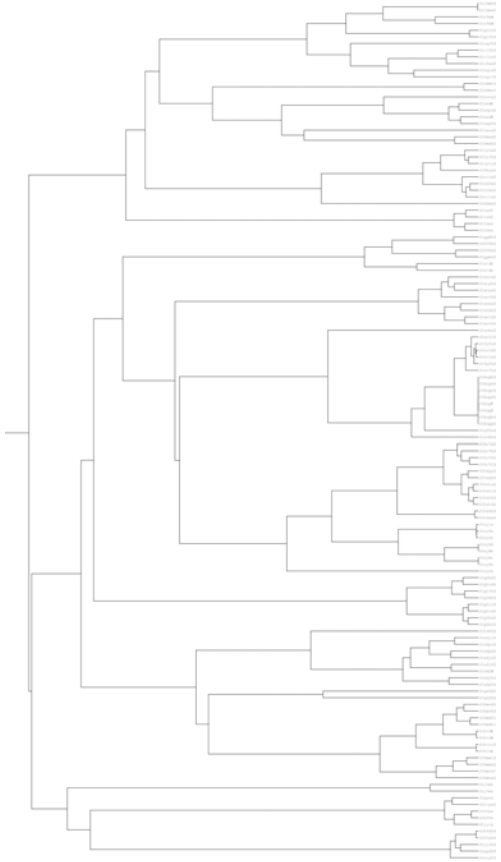
AspRS-AsnRS Group



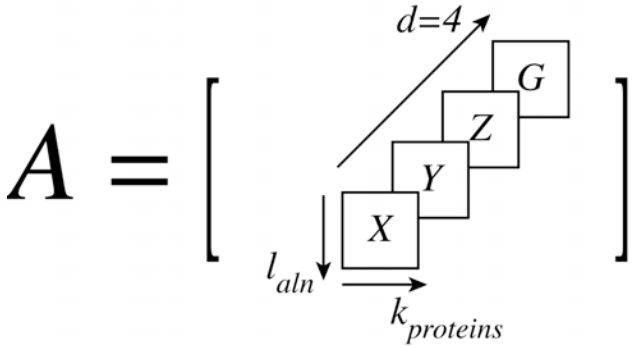


Non-redundant Representative Sets

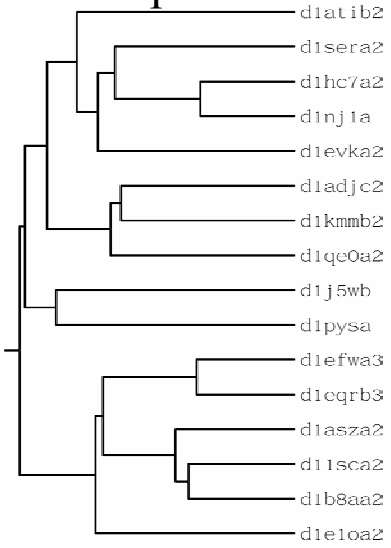
Too much information
129 Structures

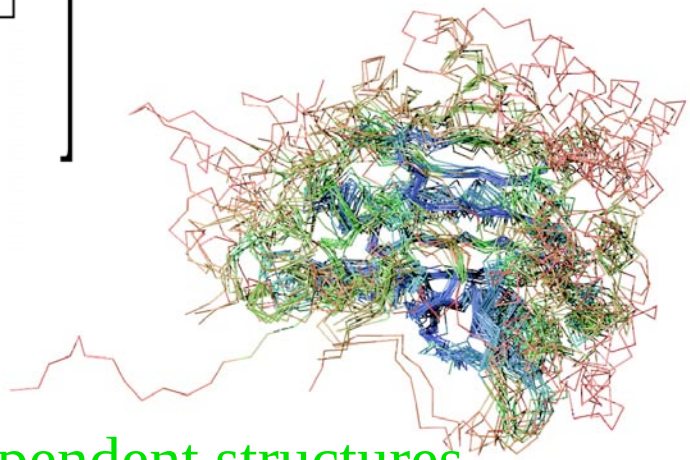


Multidimensional QR factorization of alignment matrix, ***A***.



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, JMB (2004) in press.

Numerical Encoding of Proteins in a Multiple Alignment

Encoding Structure

Rotated Cartesian + Gap = 4-space

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

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

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

adjustable parameter

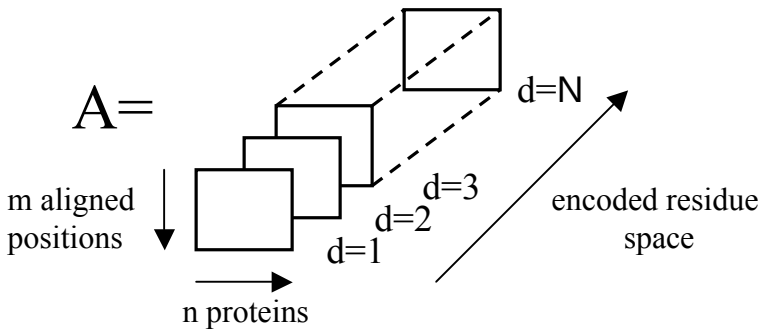
Sequence Space

Orthogonal Encoding = 24-space

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

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

Alignment Matrix



A Multiple Alignment is a Matrix with Linearly Dependent Columns

redundancy is equivalent to linear dependence

QR factorization

Re-orders the columns of A, segregating the linearly independent columns from the dependent ones without scrambling the information in A. SVD not an option.

$$Q^T AP = \tilde{R}$$

$$\tilde{A} = AP$$

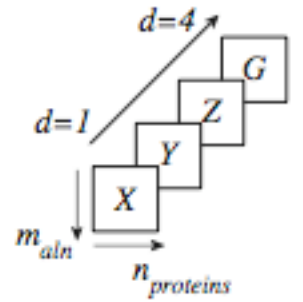
$\boldsymbol{Q^T}$ – orthogonal matrix of product of Householder transformations.

$\boldsymbol{P}$ – permutation matrix encodes column pivoting which exchanges columns of A and puts the redundant or similar proteins to the right hand side.

Multidimensional QR

N **simultaneous** QR factorizations, one for each d-dimension.

$$Q^T_{(d)} A_{(d)} P = Q^T_{(d)} \left[\right] P = \tilde{R}_{(d)}$$



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

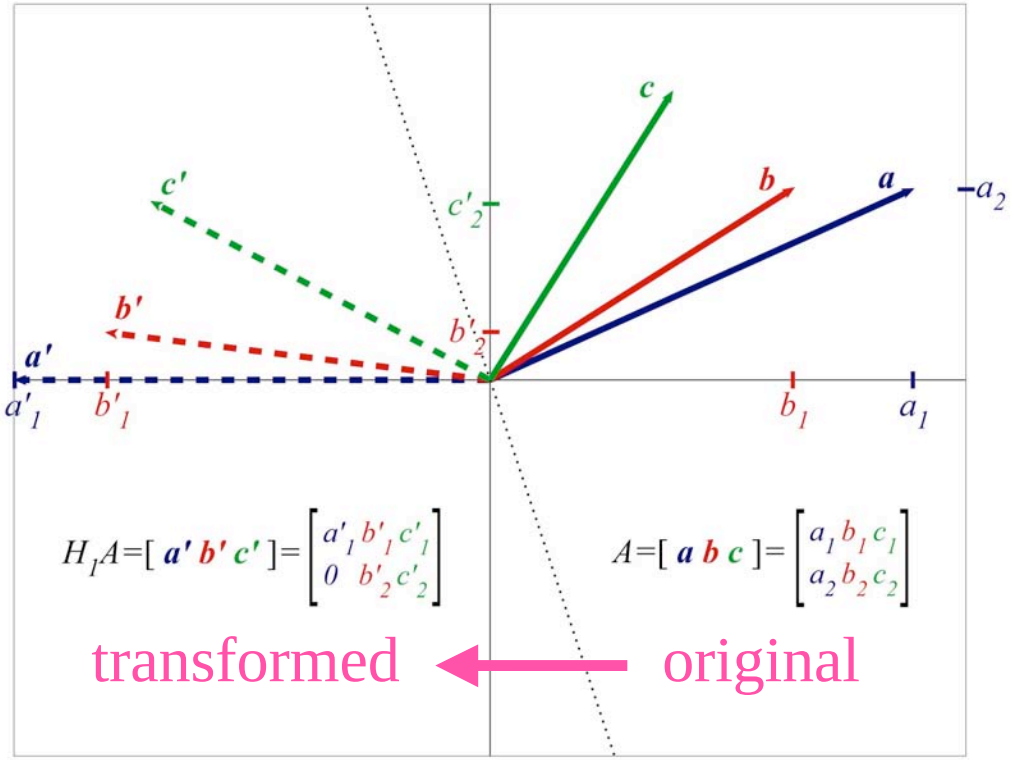
L. Heck, J. Olkin, and K. Nagshineh (1998) *J. Vibration Acoustics* **120**:663.

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

The QR establishes an **order** of linear dependence

by applying Householder transformations and permutations

$$Q^T = H_n \ldots H_1$$



Three 1-D (2 residue) proteins **a** **b** **c.**

a is our measuring stick, reference frame.

The transformation reveals that **b** is more linearly dependent on **a**, so the permutation swaps **b'** with **c'**.

Given **a**, **c** adds more information to the system than **b**.

Multiply aligned proteins exist in a higher dimensional space, so this magnitude is computed with a matrix p-norm:

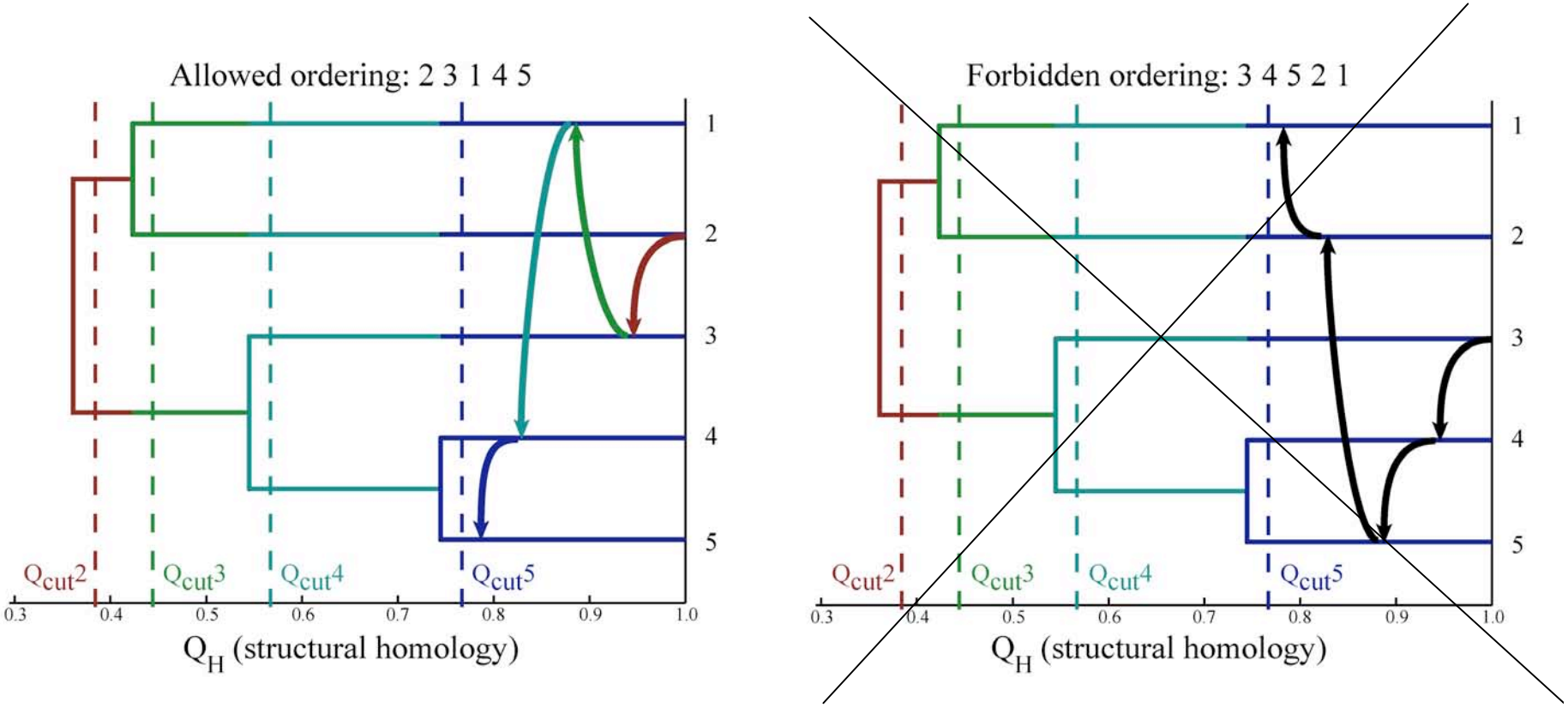
$$\|a_j\|_{F_p} = \left(\sum_{d=1}^{4} \sum_{i=k}^{m_{aln}} |a_{ijd}|^p\right)^{1/p}$$

adjustable parameter

Householder, *J. Assoc. Comput. Mach.*, 1958.

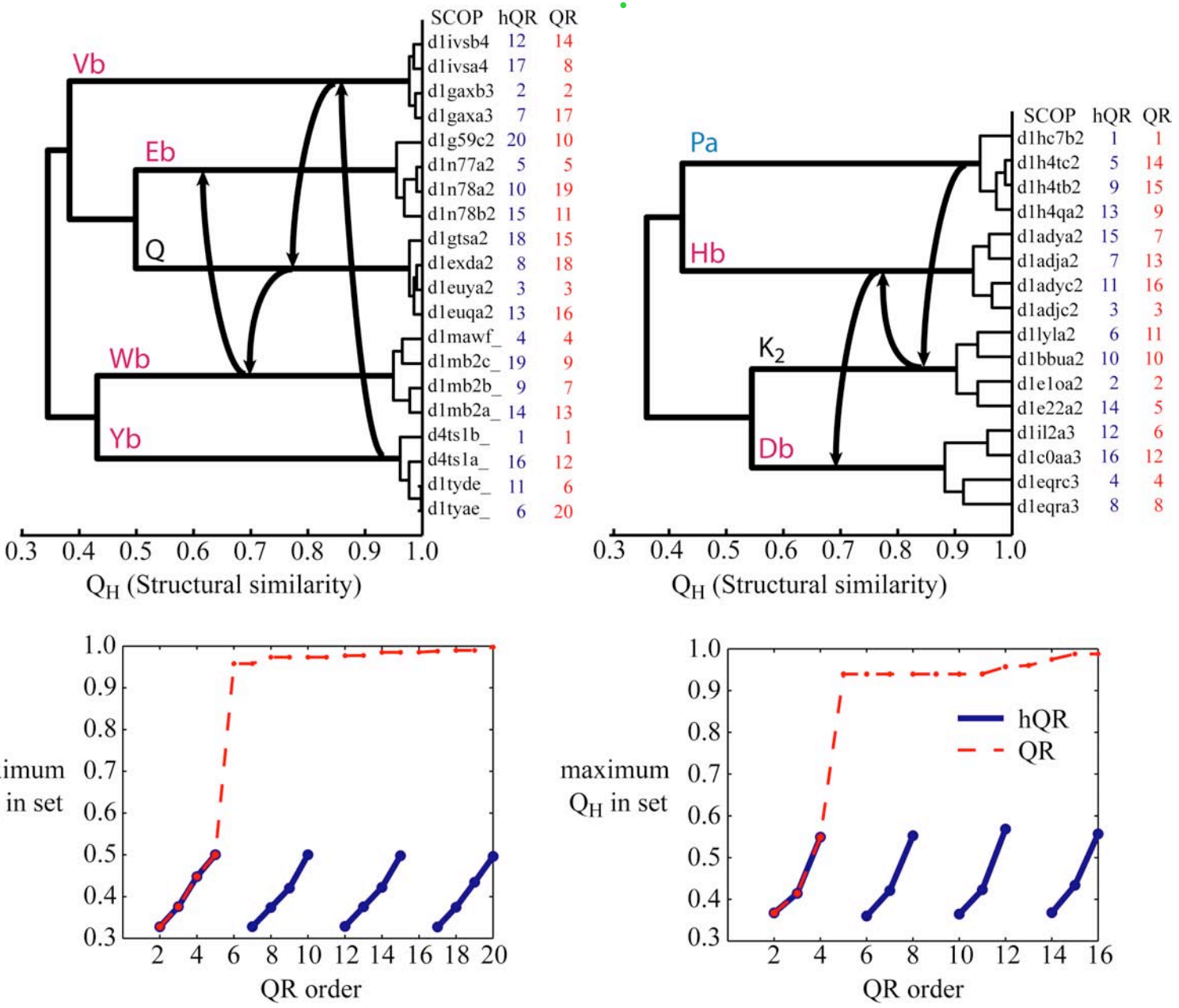
What are the constraints on the parameters?

Must maintain the evolutionary history of the protein group.



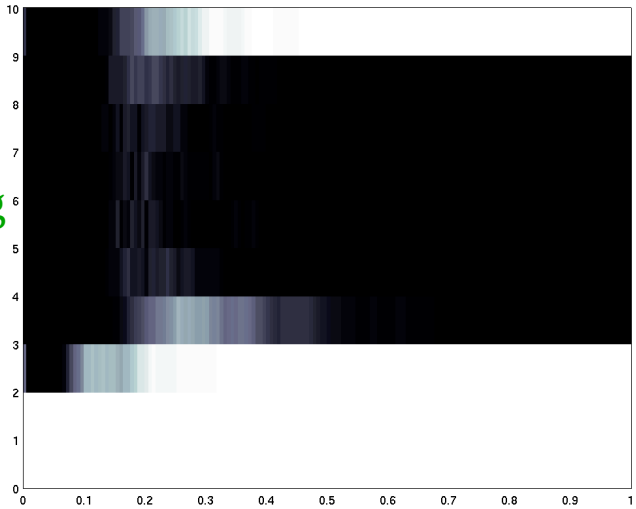
This rule is used to determine the value of
two adjustable parameters in our implementation of the QR.

Hierarchical Multidimensional QR -

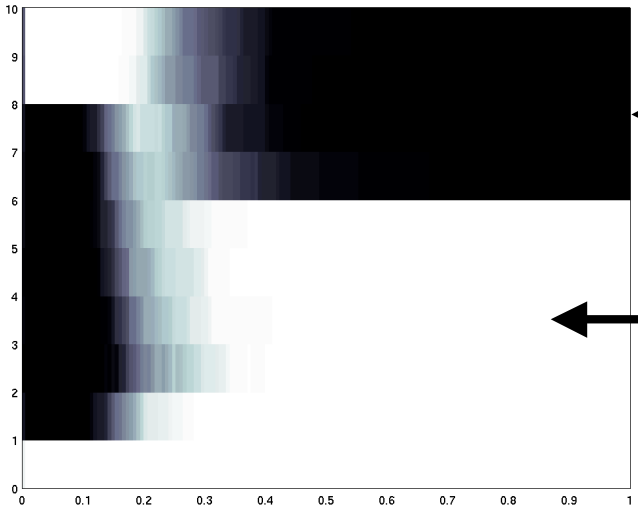


Parameters Define the Measure of Linear Dependence

AARS class I, Rossman fold



AARS class II, Novel fold



forbidden

allowed

ordering
p-norm

γ (normalized)

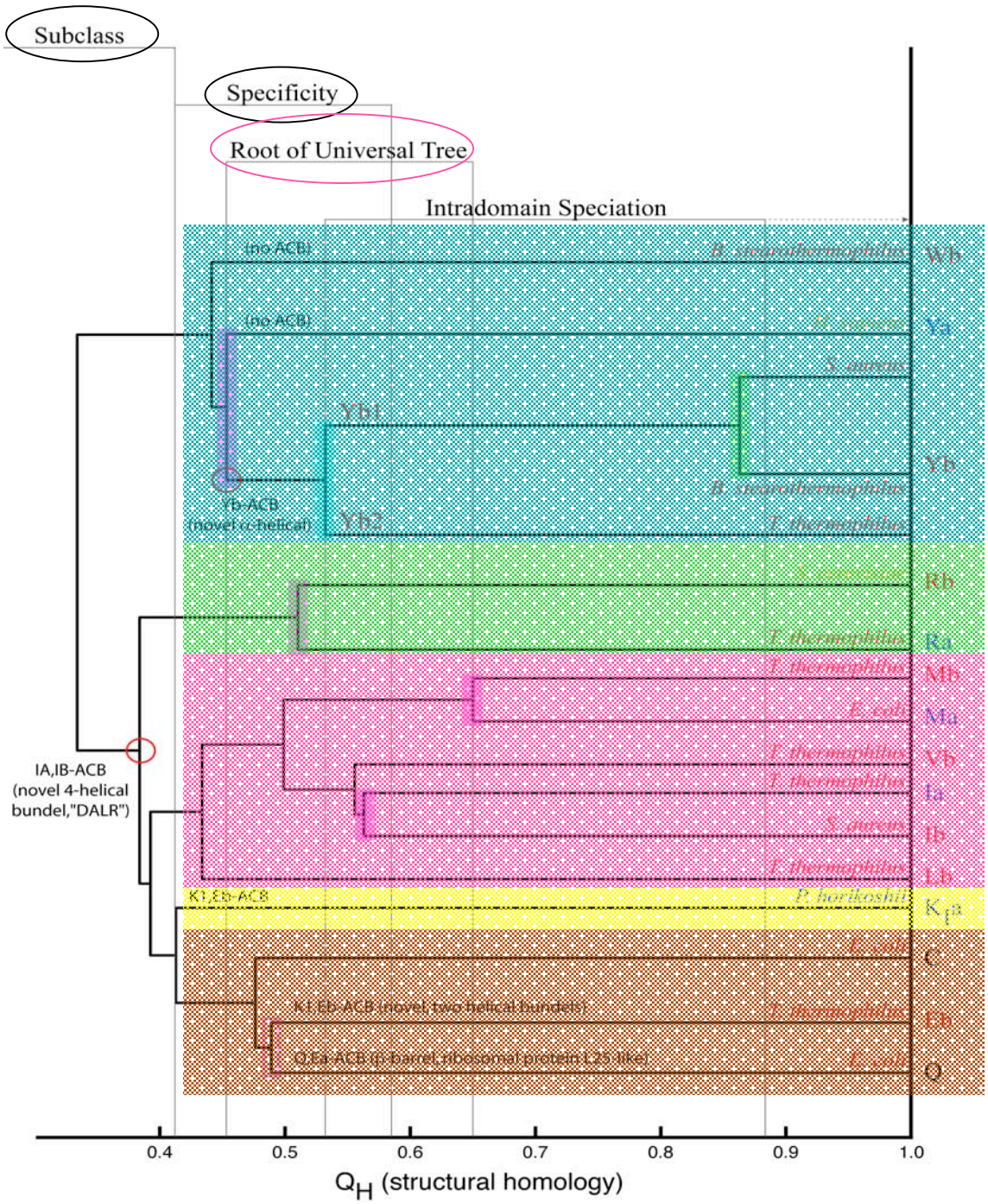
γ (normalized)

ordering norm $\max_{j=k,\ldots,n_{proteins}} (\|a_j\|_{F_p})$

$$\|a_j\|_{F_p} = \left(\sum_{d=1}^{4}\sum_{i=k}^{m_{aln}} |a_{ijd}|^p\right)^{1/p}$$

gap scale

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



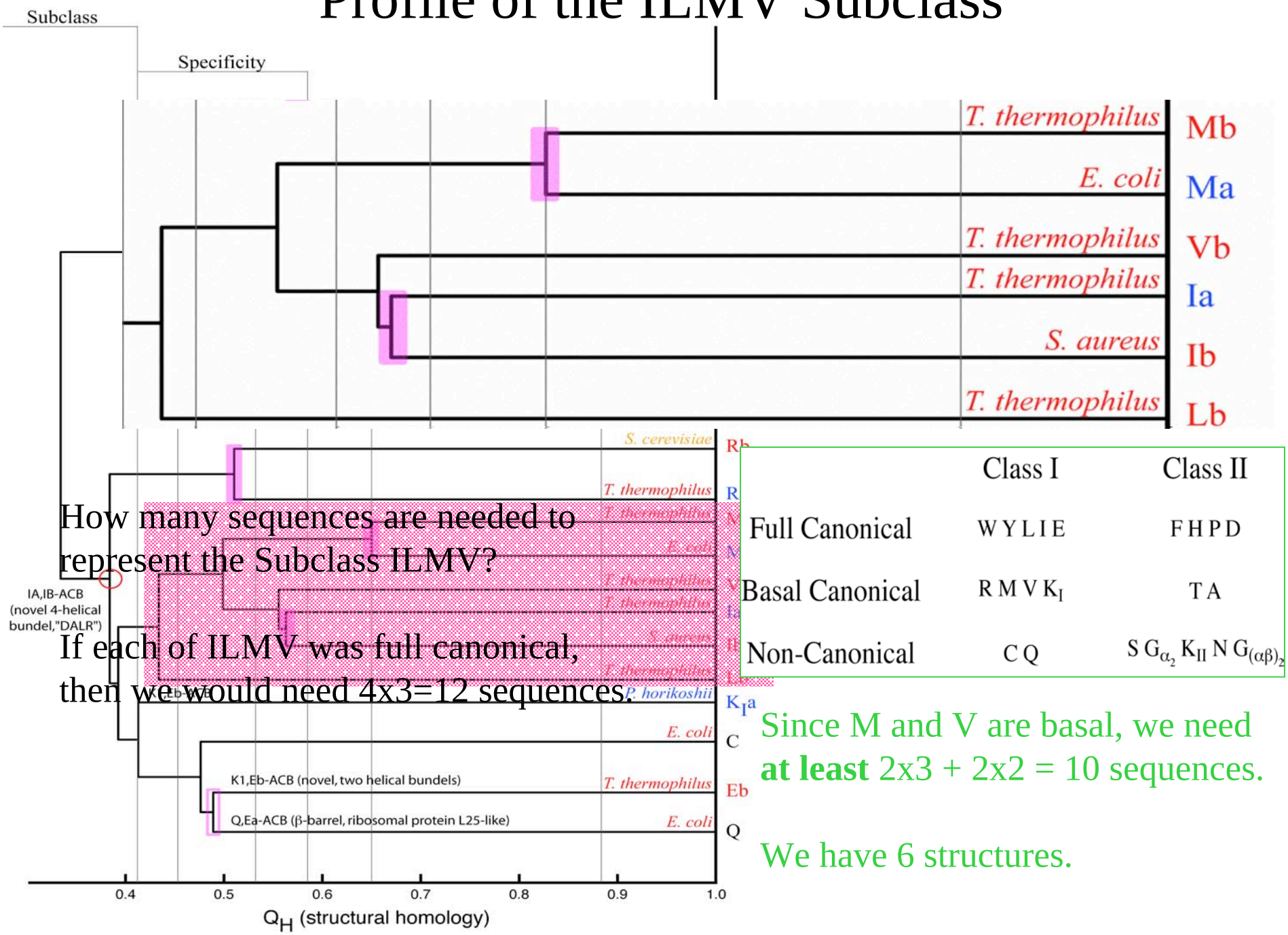
Class I AARSs evolutionary events

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.

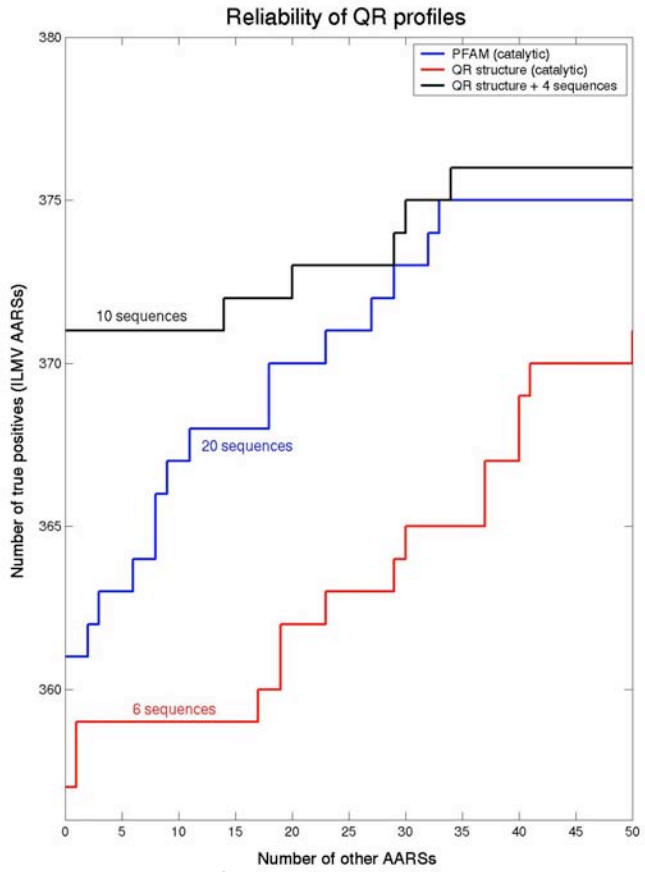
	Class I	Class II
Full Canonical	W Y L I E	F H P D
Basal Canonical	R M V K_I	T A
Non-Canonical	C Q	S G_{α_2} K_{II} N $G_{(\alpha\beta)_2}$

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

We have 6 structures.

Non-Redundant Profiles for Database Searching
AARS Subclass ILMV

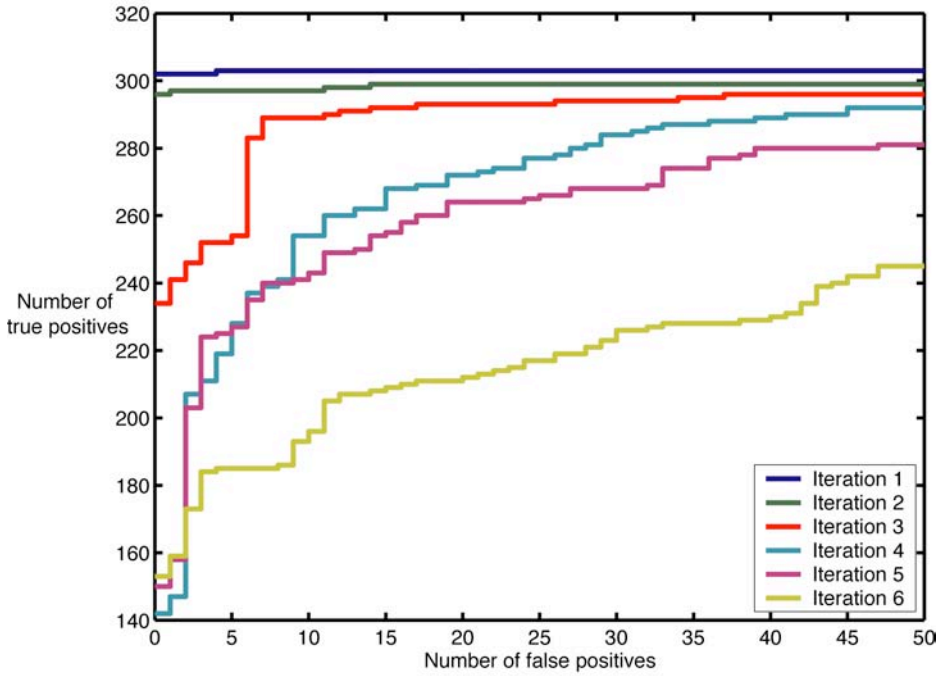
HMMER



false positives

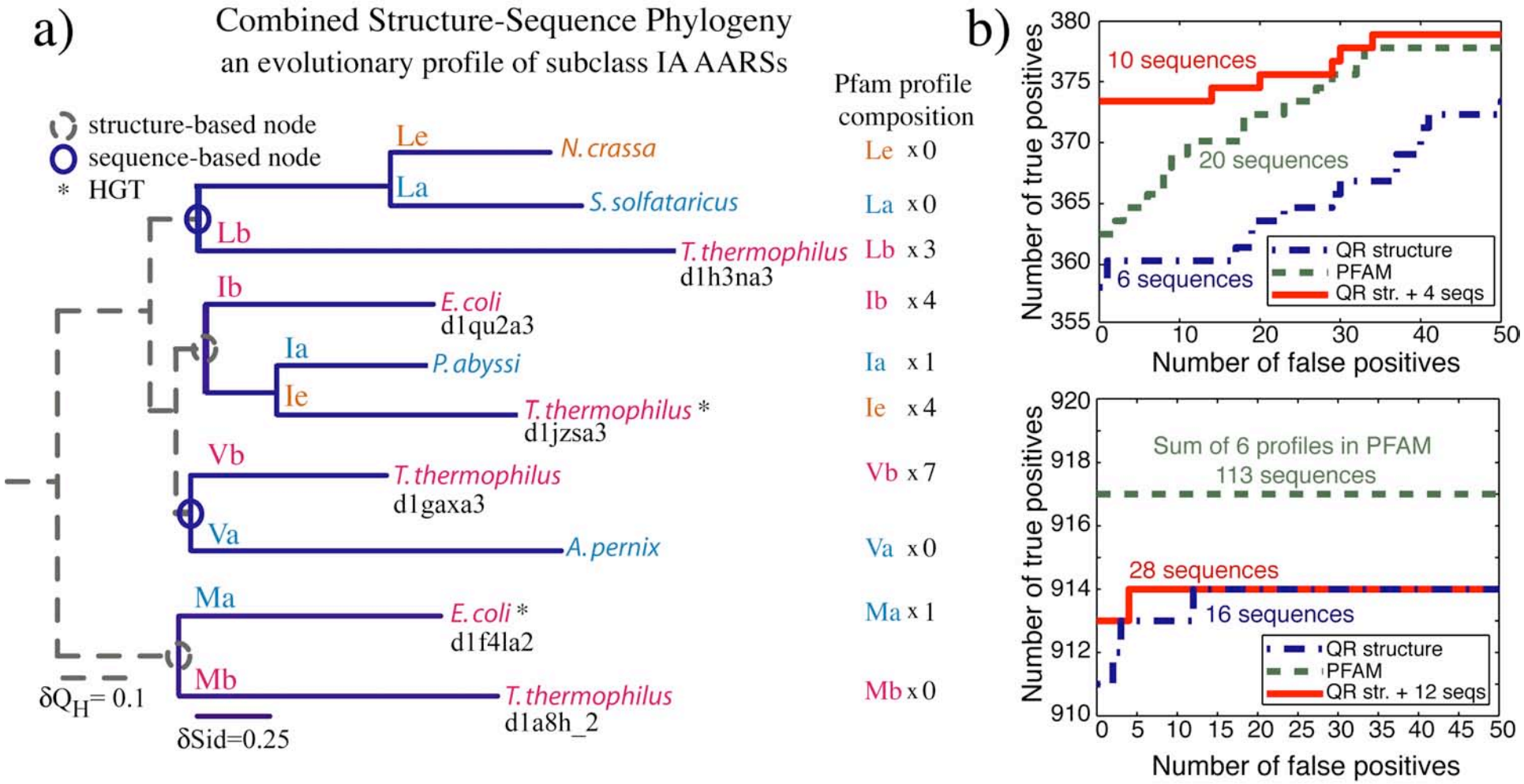
Choosing the right 10 sequence makes all the difference.

Psi-Blast



Starting with a non-redundant profile, **accuracy diminishes with Psi-blast iterations** which add in bias. **Repair with QR filter.**

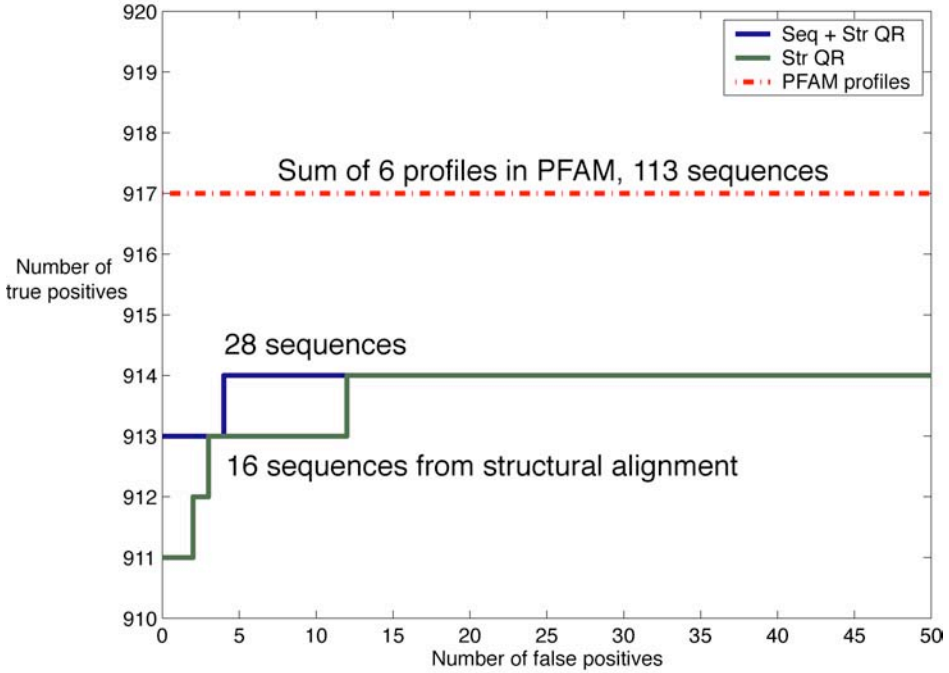
A. Sethi, P. O'Donoghue, Z..Luthey-Schulten


a)
Combined Structure-Sequence Phylogeny
an evolutionary profile of subclass IA AARSs
structure-based node
sequence-based node
* HGT
Le
N. crassa
La
S. solfataricus
Lb
T. thermophilus
d1h3na3
Ib
E. coli
d1qu2a3
Ia
P. abyssi
Ie
T. thermophilus *
d1jzsa3
Vb
T. thermophilus
d1gaxa3
Va
A. pernix
Ma
E. coli *
d1f4la2
Mb
T. thermophilus
d1a8h_2
δQH= 0.1
δSid=0.25
Pfam profile composition
Le x 0
La x 0
Lb x 3
Ib x 4
Ia x 1
Ie x 4
Vb x 7
Va x 0
Ma x 1
Mb x 0
b)
10 sequences
20 sequences
6 sequences
QR structure
PFAM
QR str. + 4 seqs
Number of true positives
Number of false positives
Sum of 6 profiles in PFAM
113 sequences
28 sequences
16 sequences
QR structure
PFAM
QR str. + 12 seqs
Number of true positives
Number of false positives

The Economy of Information

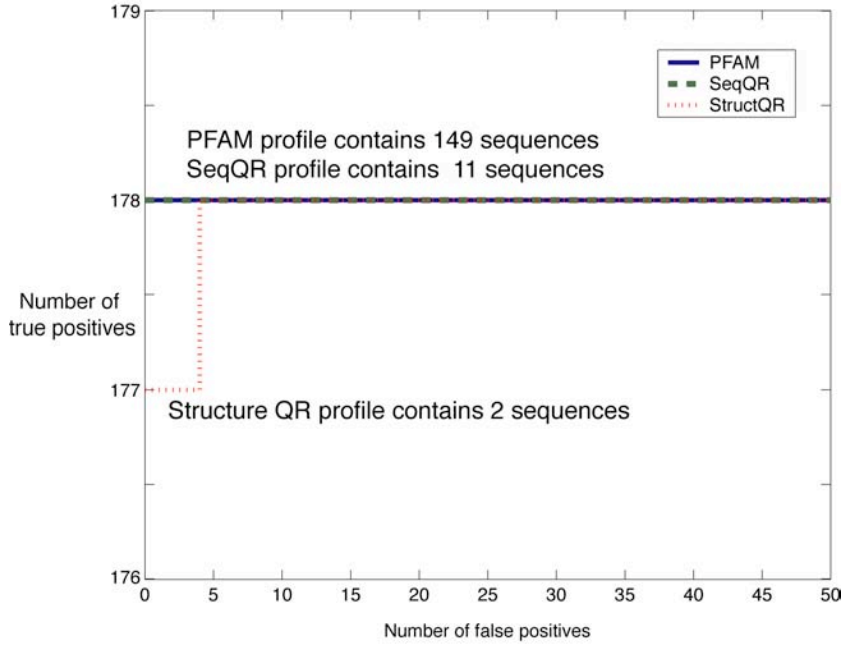
How many sequence are needed for profiles?

A single profile
for class I AARSs



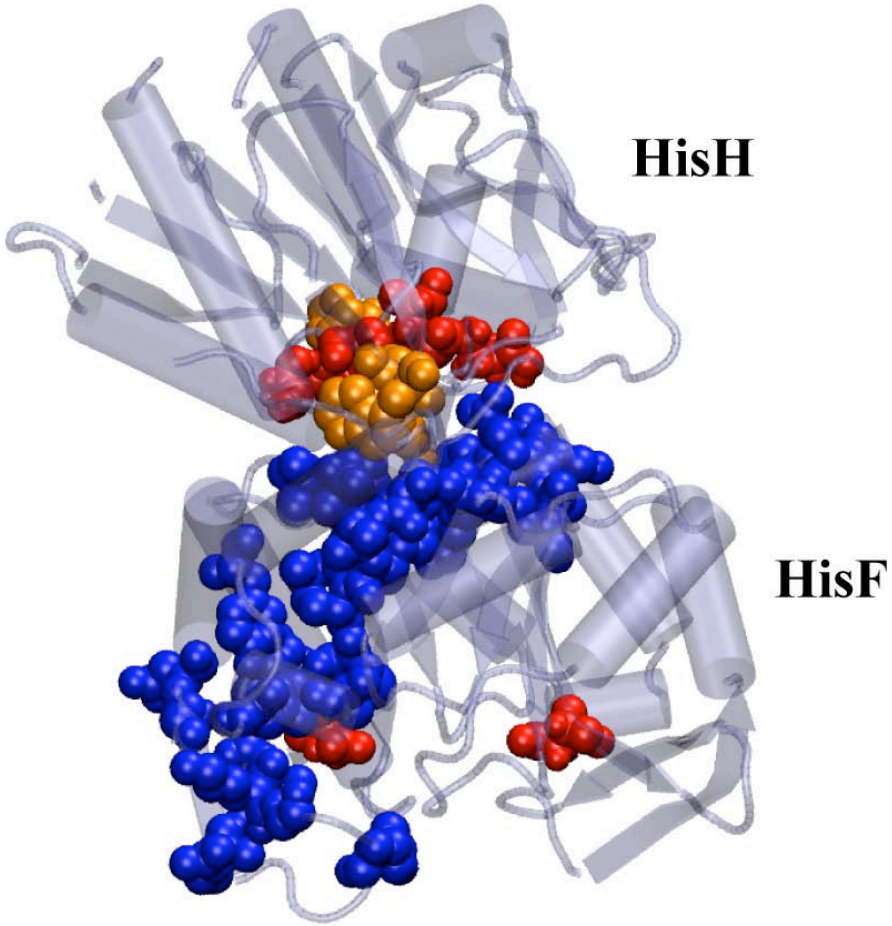
PFAM profile of 113 sequences finds 3 additional sequence fragments compared to the non-redundant profile of 28 sequences.

HisA and HisF Protein Family
TIM Barrel fold



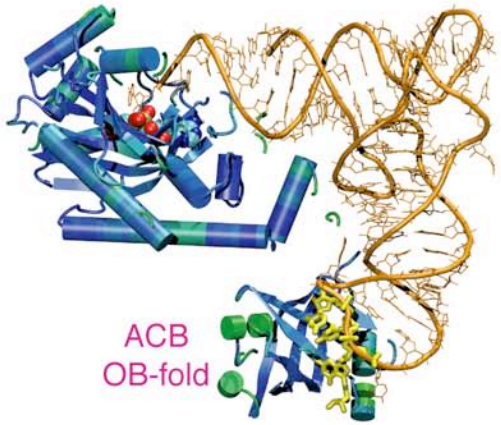
If the sequences well represent the evolutionary history of the protein family, a factor of 10 to 100 less information is required.

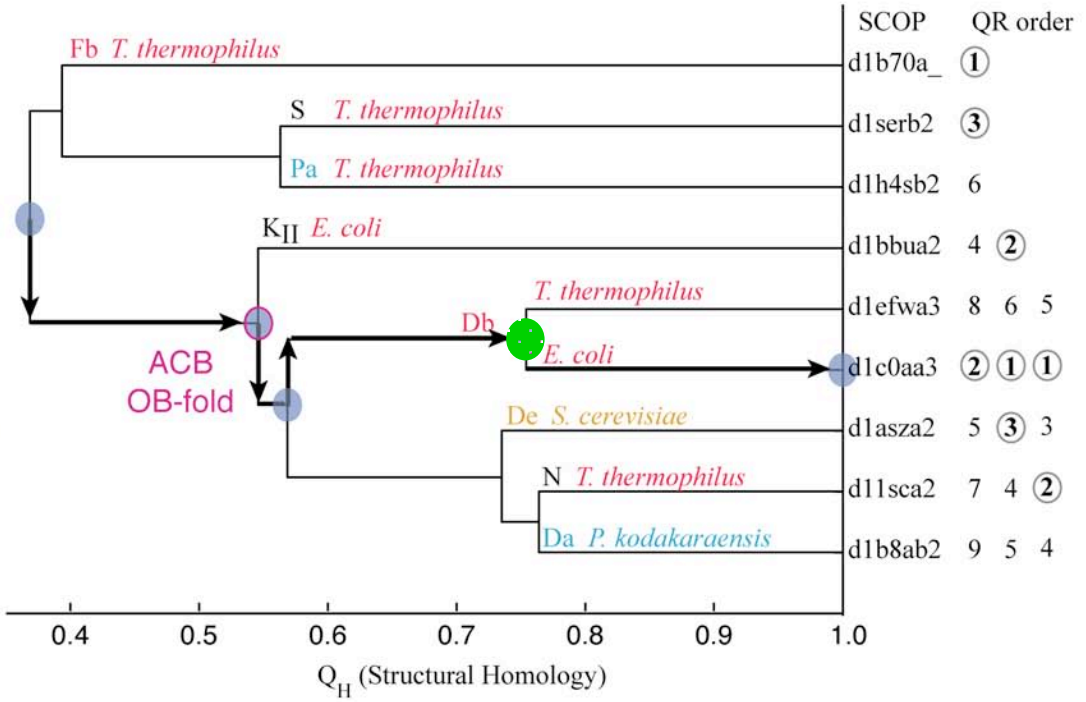
Evolutionary Structure/Sequence Profiles Suggest Reaction Pathway

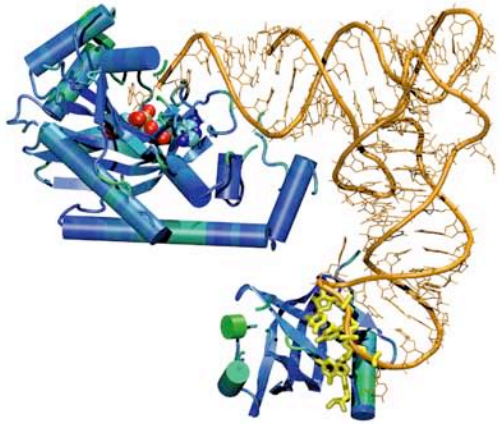


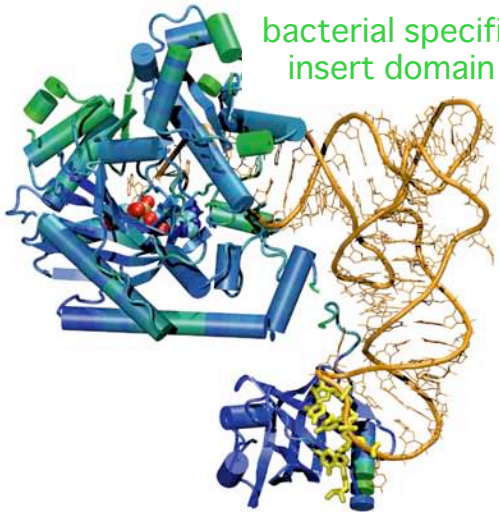
R. Amaro and Z. Schulten, *MD Simulations of Substrate Channeling*, Chemical Physics Special Issue, 2004 (in press). *FE Landscapes of Ammonia Channeling*, PNAS 2003

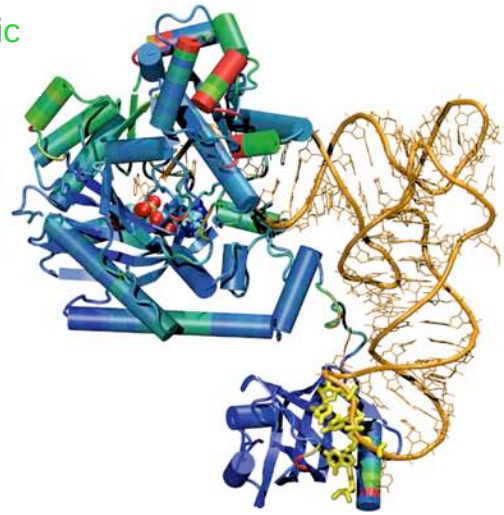
Evolution of Structure and Function in AspRS





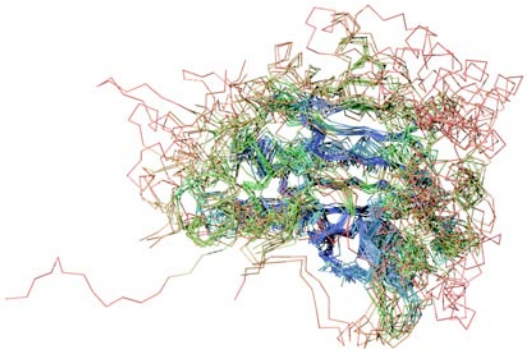


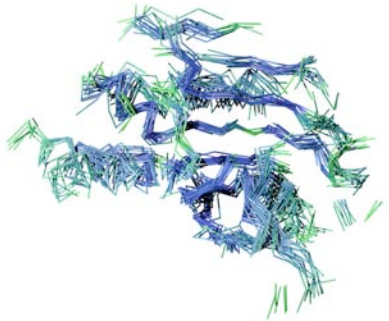


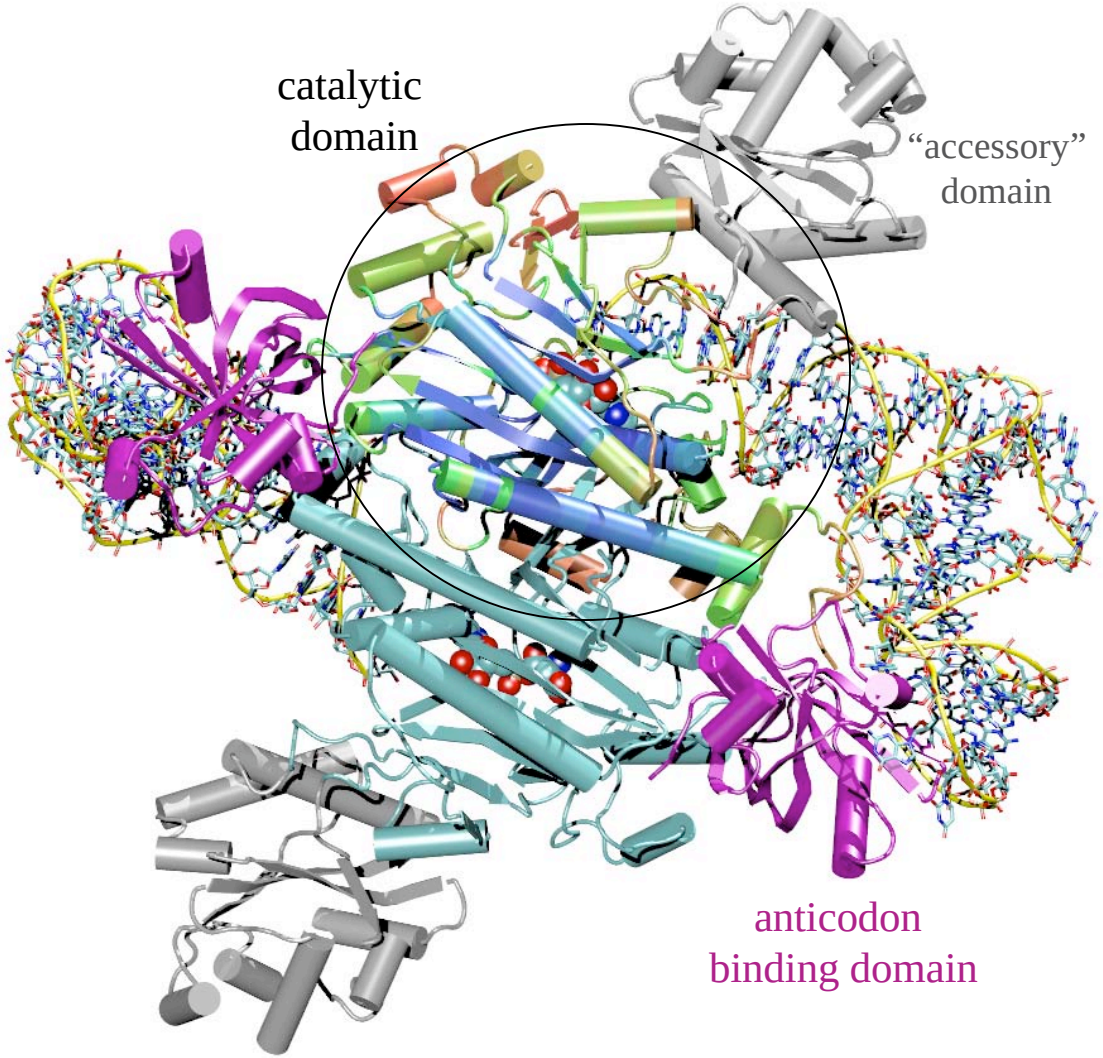


AARS domains have different Evolutionary Histories

catalytic domains
AARSs II







bacterial type aspartyl-tRNA synthetase
E. coli, homodimer

Summary

Evolutionary information is encoded in protein structure.

Protein structure can be used to investigate early evolutionary events.

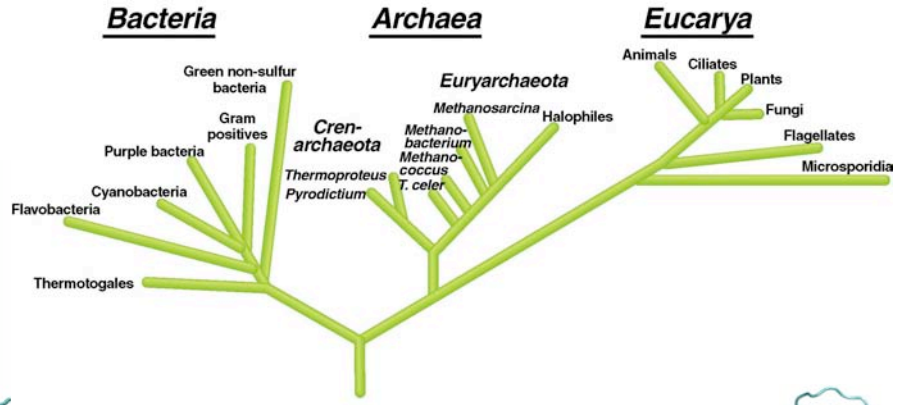
Accounting for gaps is important for comparing homologous structures - structure metric

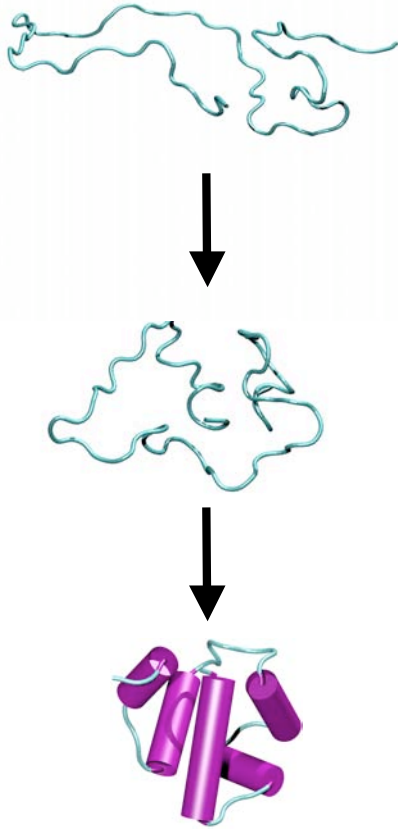
Multidimensional QR factorization computes non-redundant sets from multiple sequence or structure alignments which well represent the evolutionary history of the group as expressed in phylogenetic tree

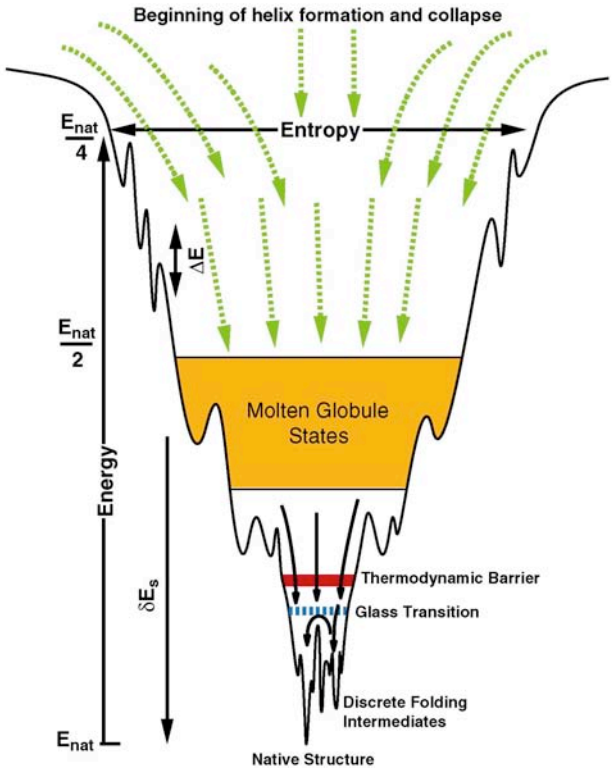
Structure databases are limited, but multiple structural alignments provide accurate alignments, especially in the case of distant homologies

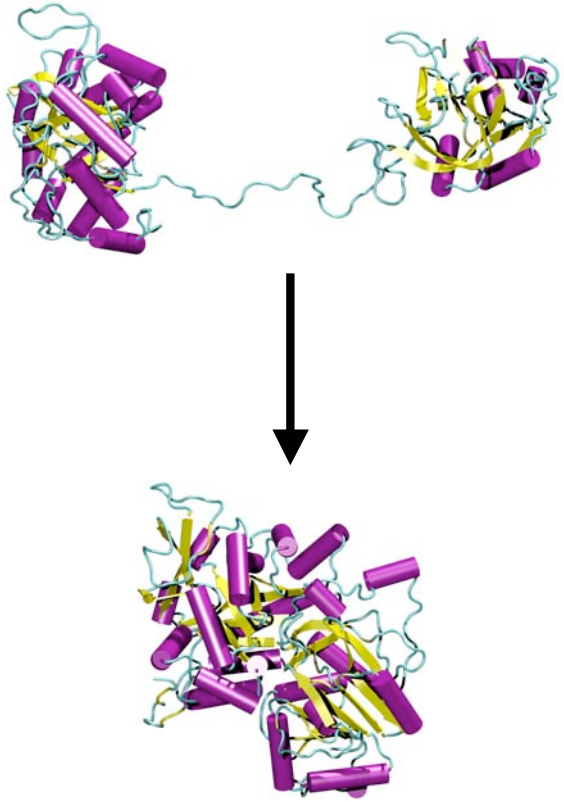
Supplement the structures with an appropriate number and type of sequences (in accord with the phylogenetic topology) to produce minimal representative profiles. Search profiles for foldons!!

Evolution of Protein Structure



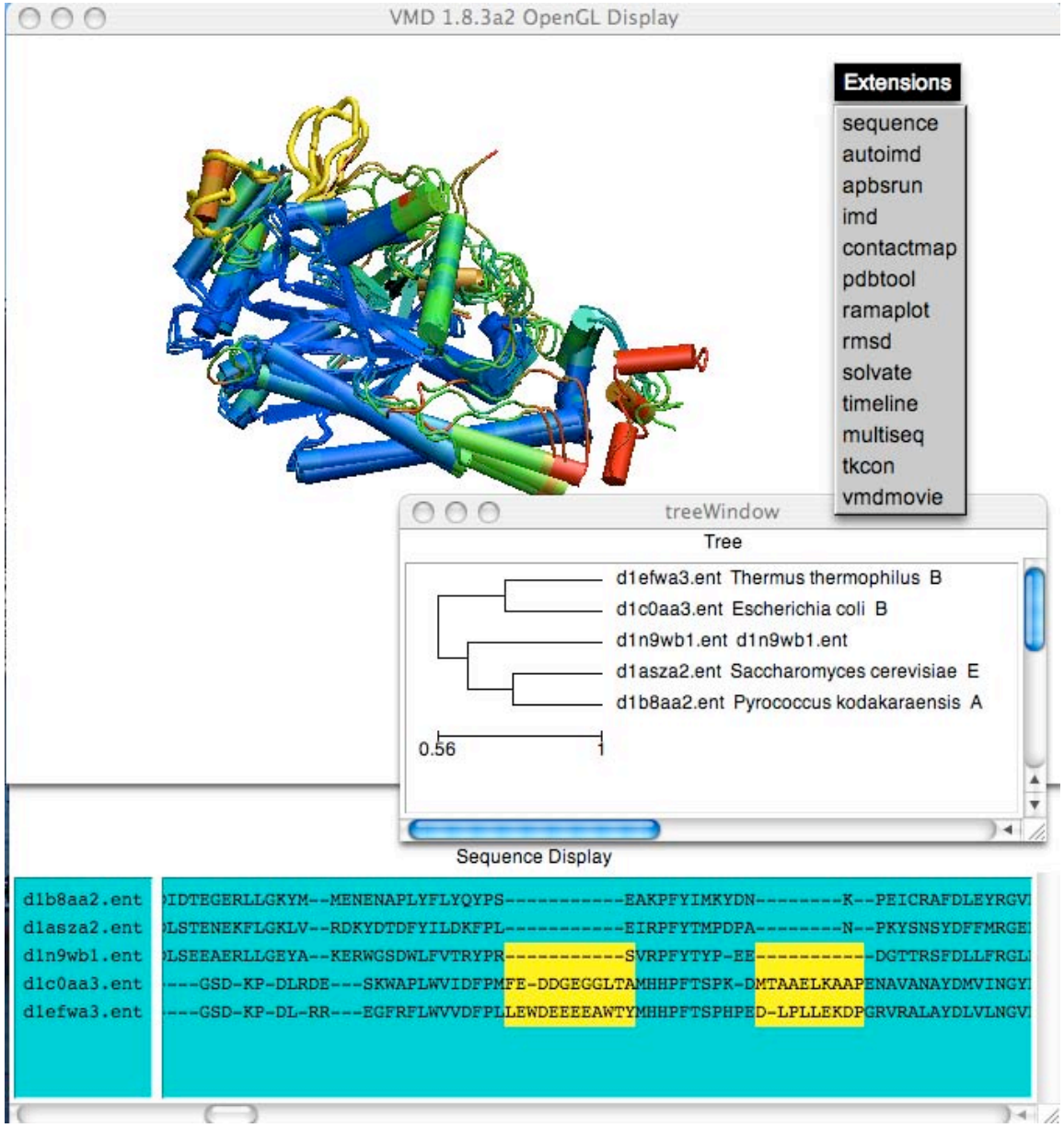






San Diego, 2004

VMD Multiple Sequence Display with Evolution Analysis Algorithms



Acknowledgements

Patrick O'Donoghue

Rommie Amaro
Anurag Sethi
John Eargle
Corey Hardin
Michael Baym
Michael Janusyzk

Felix Autenrieth
Taras Pogorelov

Brijeet Dhaliwal

Graphics Programmers VMD

John Stone, Dan Wright, John Eargle

http://www.ks.uiuc.edu/Research/vmd/alpha/zs04/

Collaborators

Evolutionary Studies

Gary Olsen, Carl Woese (UIUC)

Algorithms

Mike Heath (UIUC)

Rob Russell (EMBL) **STAMP**

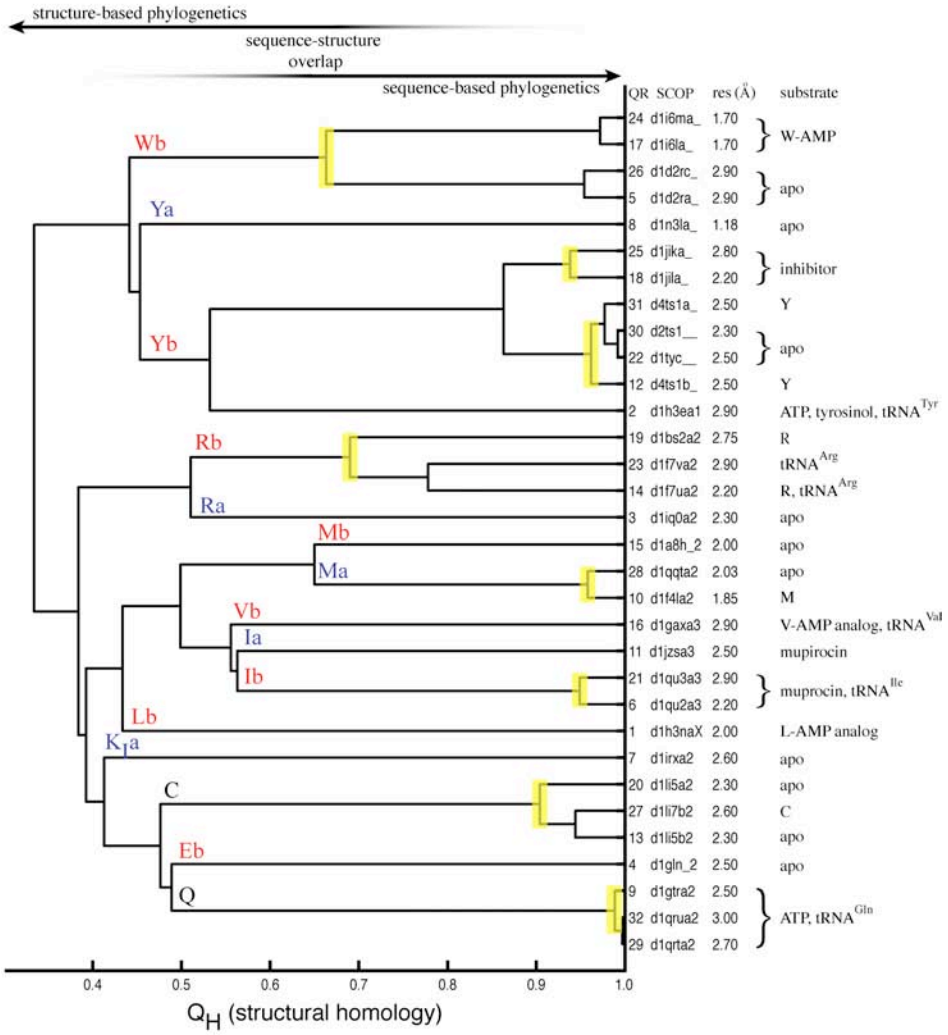
Protein Structure Prediction

Peter Wolynes, Jose Onuchic, Ken Suslick

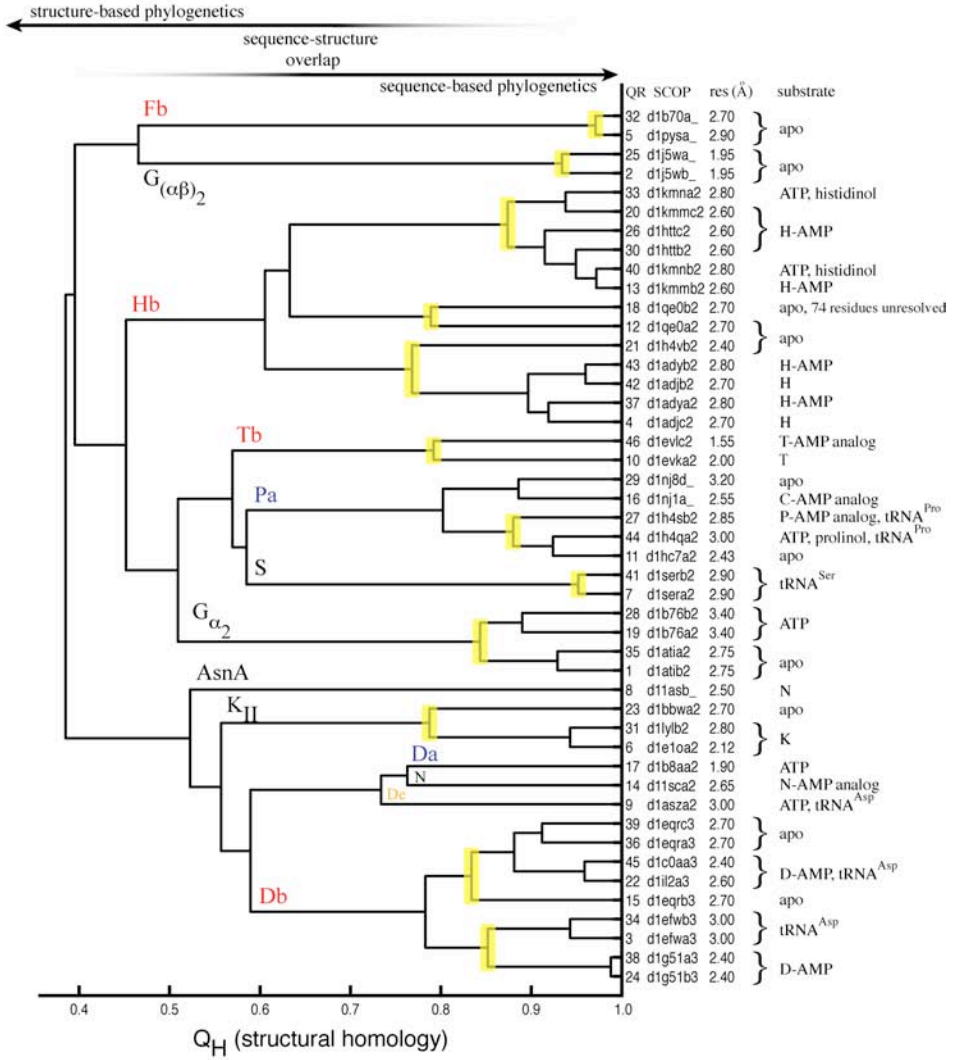
Funding: NSF, NIH, NIH Resource for Macromolecular Modeling and Bioinformatics, NRAC NSF Supercomputer Centers

extra slides

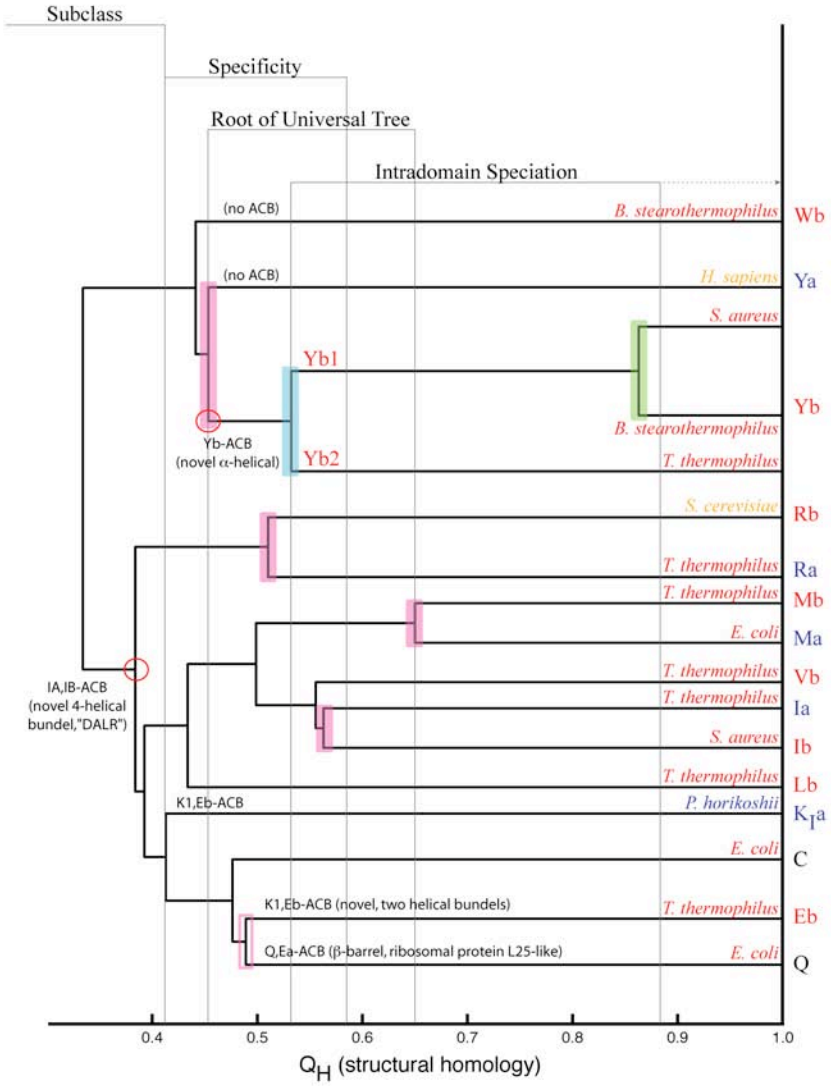
Structure Phylogeny
Class I AARSs



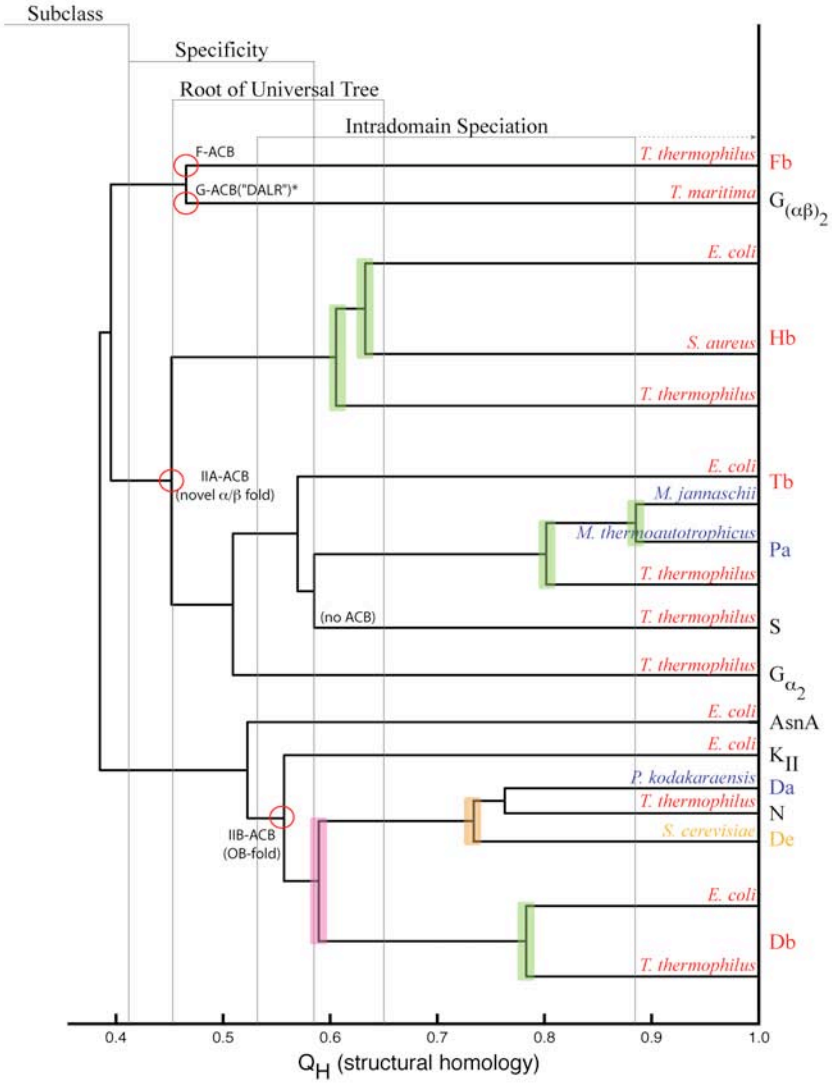
Structure Phylogeny
Class II AARSs



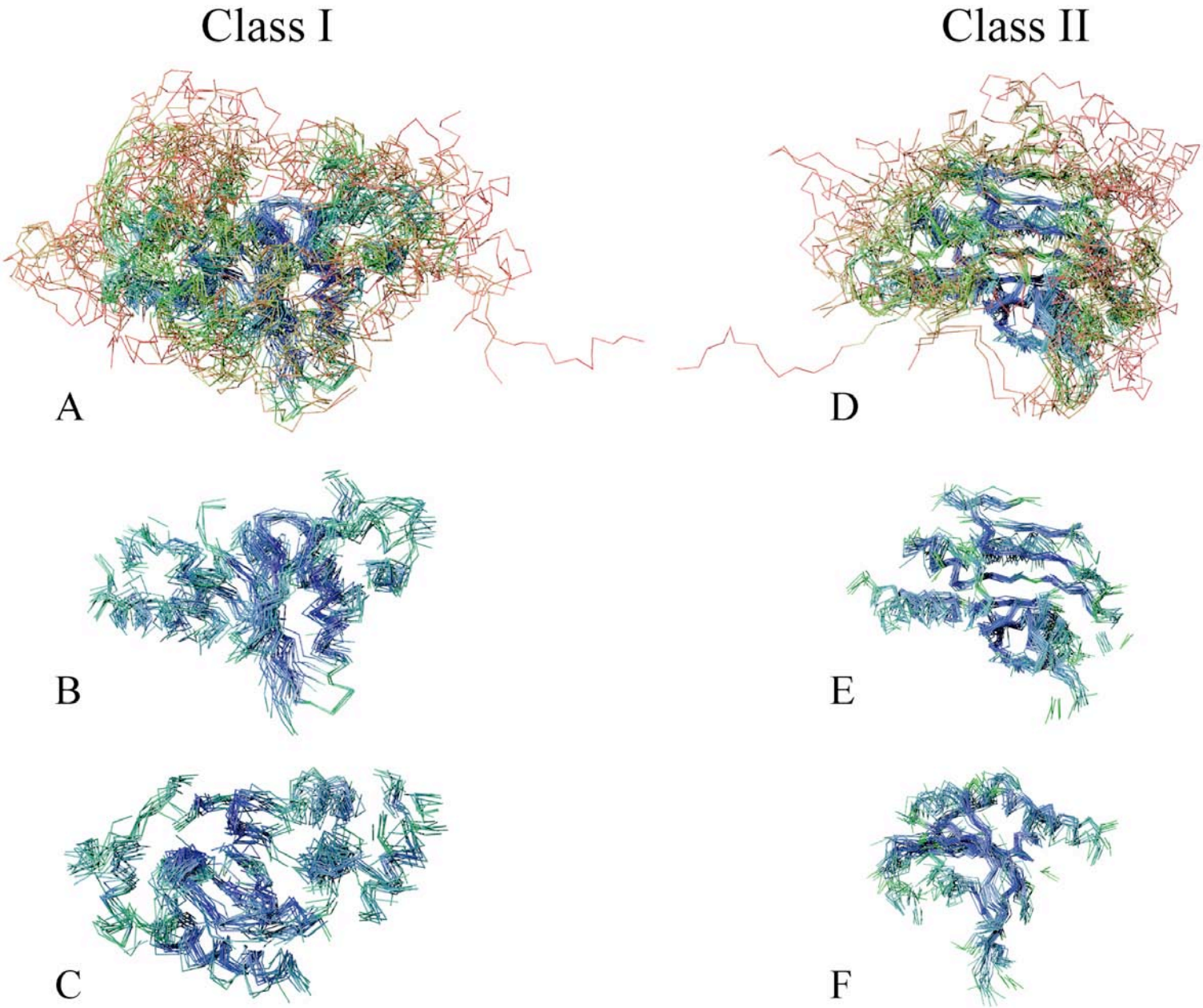
Structure Phylogeny
Class I AARSs



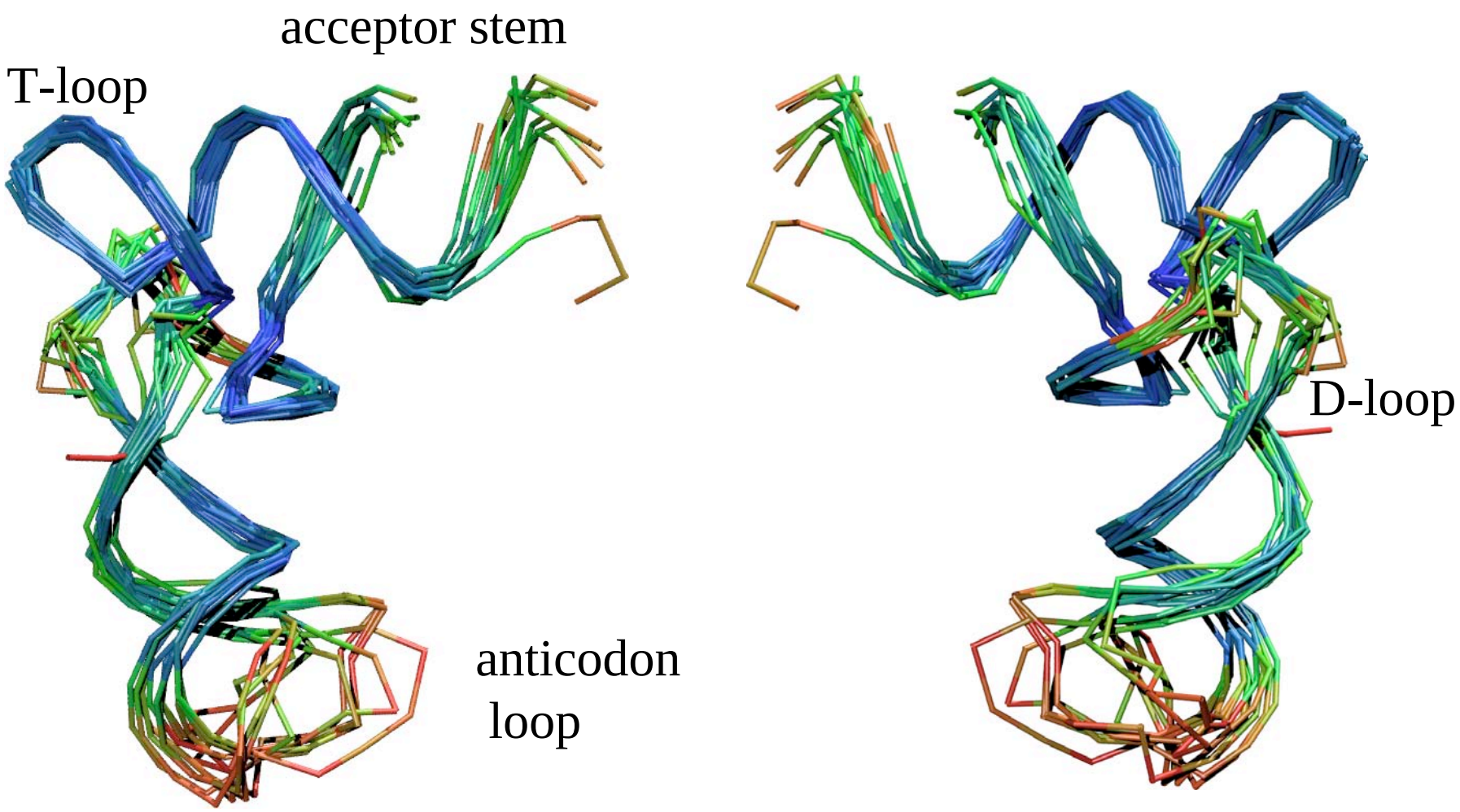
Structure Phylogeny
Class II AARSs



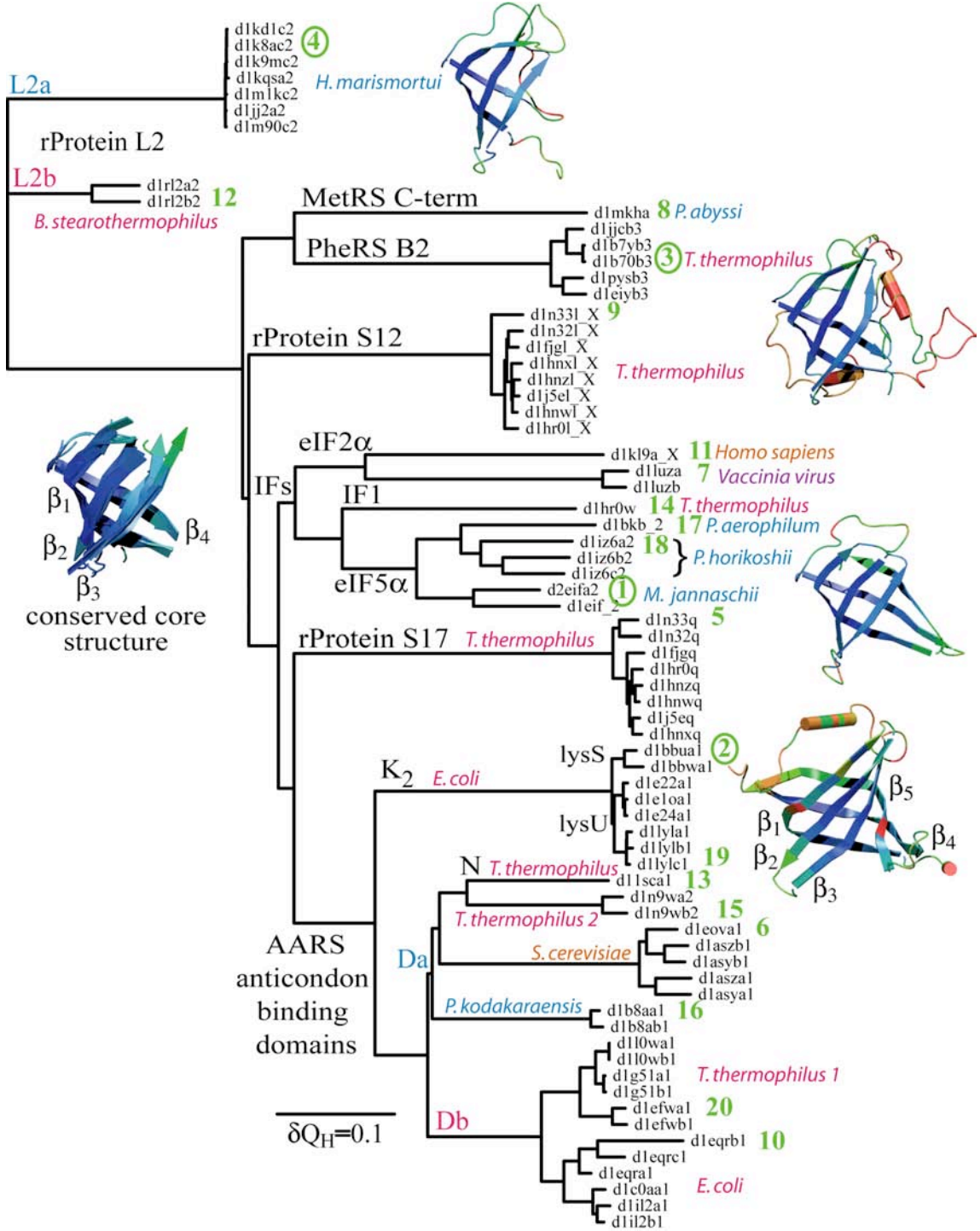
Structural Overlap of the AARSs

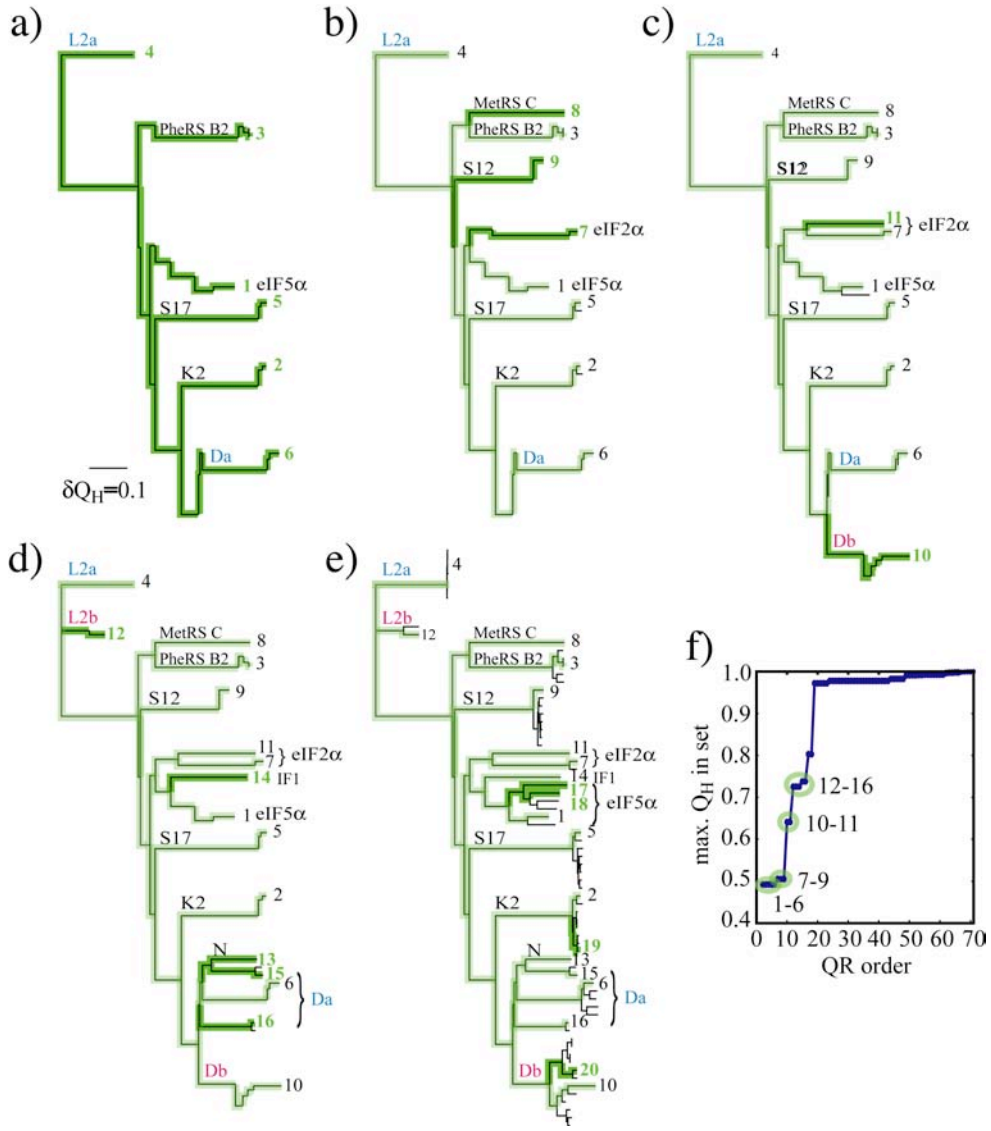


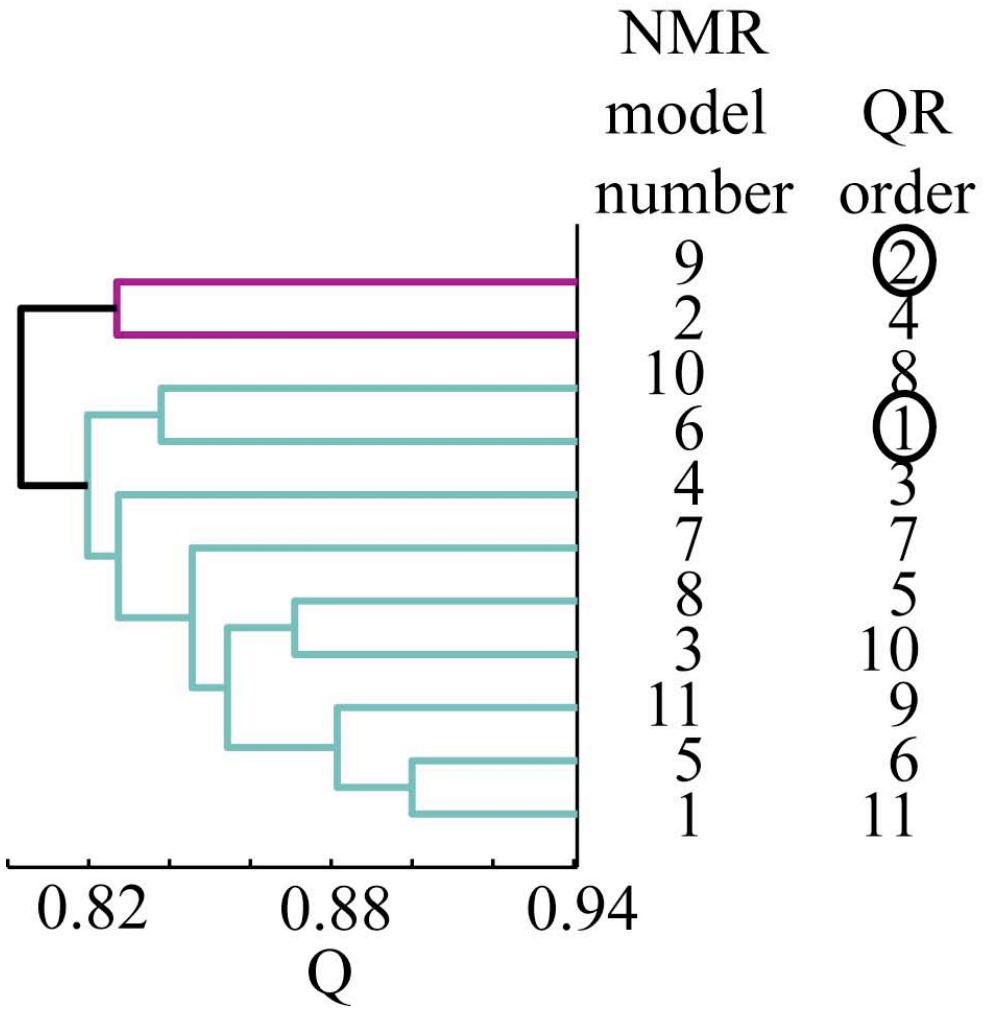
Structural Conservation in tRNA



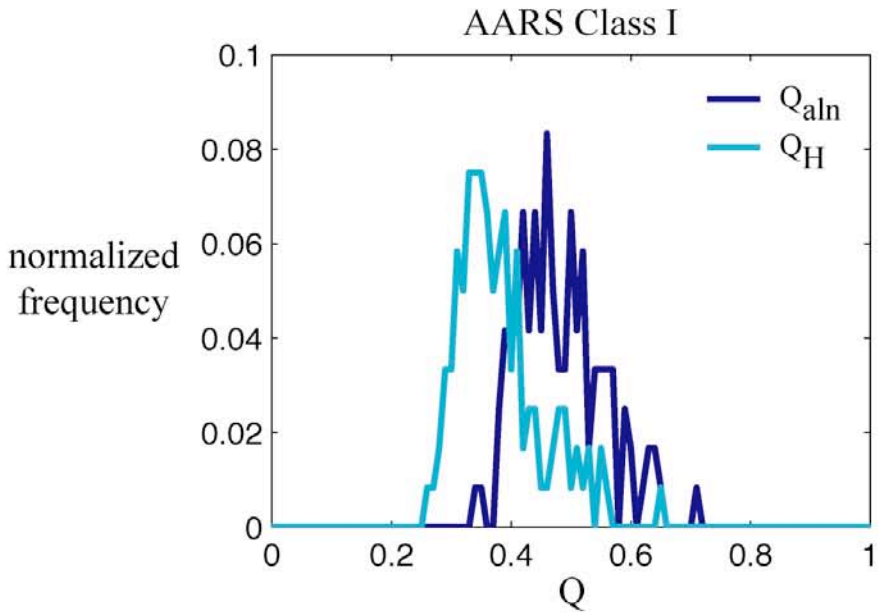
Representative set of OB folds involved in translation ?

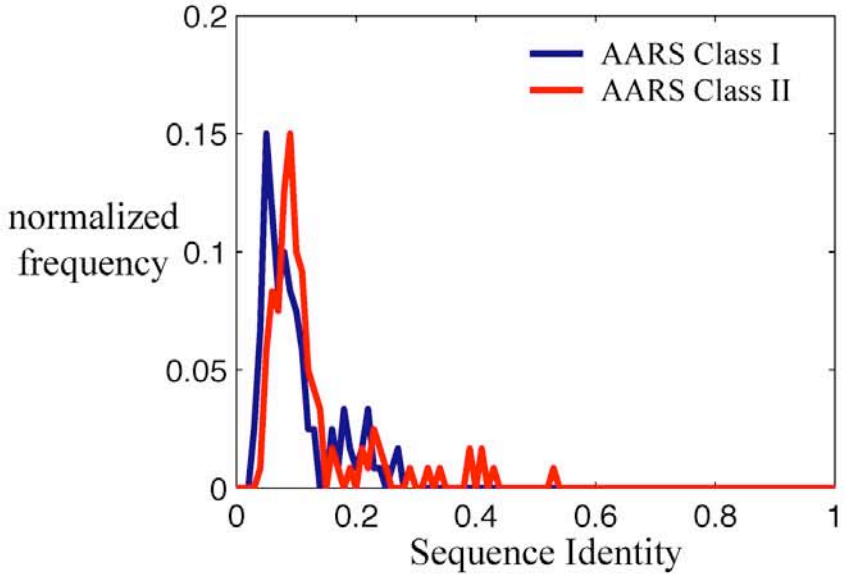


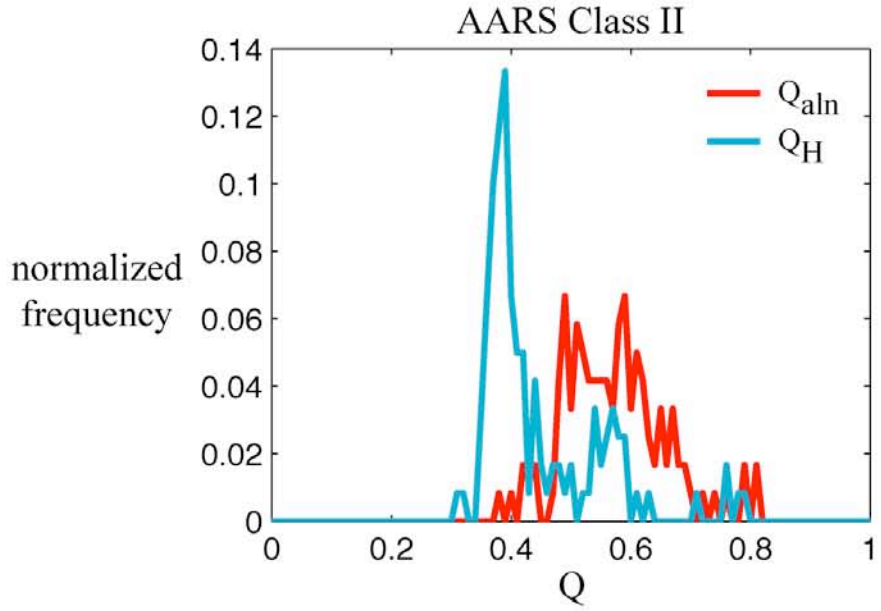

a)
L2a
4
PheRS B2
3
1 eIF5α
S17
5
K2
2
Da
6
δQH=0.1
b)
L2a
4
MetRS C
8
PheRS B2
3
S12
9
7 eIF2α
1 eIF5α
S17
5
K2
2
Da
6
c)
L2a
4
MetRS C
8
PheRS B2
3
S12
9
11
7 } eIF2α
1 eIF5α
S17
5
K2
2
Da
6
Db
10
d)
L2a
4
L2b
12
MetRS C
8
PheRS B2
3
S12
9
11
7 } eIF2α
14 IF1
1 eIF5α
S17
5
K2
2
N
13
15
6 } Da
16
Db
10
e)
L2a
4
L2b
12
MetRS C
8
PheRS B2
3
S12
9
11
7 } eIF2α
14 IF1
17
18 } eIF5α
1
S17
5
K2
2
19
N
13
15
6 } Da
16
Db
20
10
f)
max. QH in set
1.0
0.9
0.8
0.7
0.6
0.5
0.4
12-16
10-11
7-9
1-6
0 10 20 30 40 50 60 70
QR order

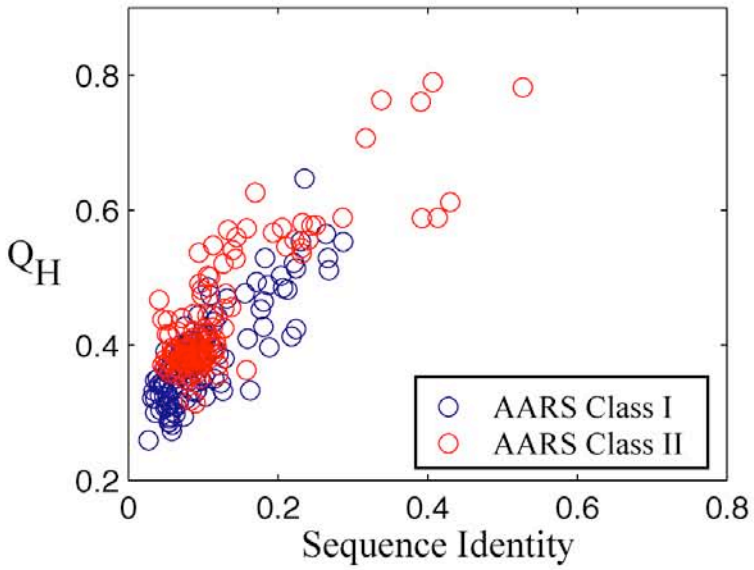

NMR
model
number
QR
order
9 2
2 4
10 8
6 1
4 3
7 7
8 5
3 10
11 9
5 6
1 11
0.82
0.88
0.94
Q

Only structure can reveal distant evolutionary relationships



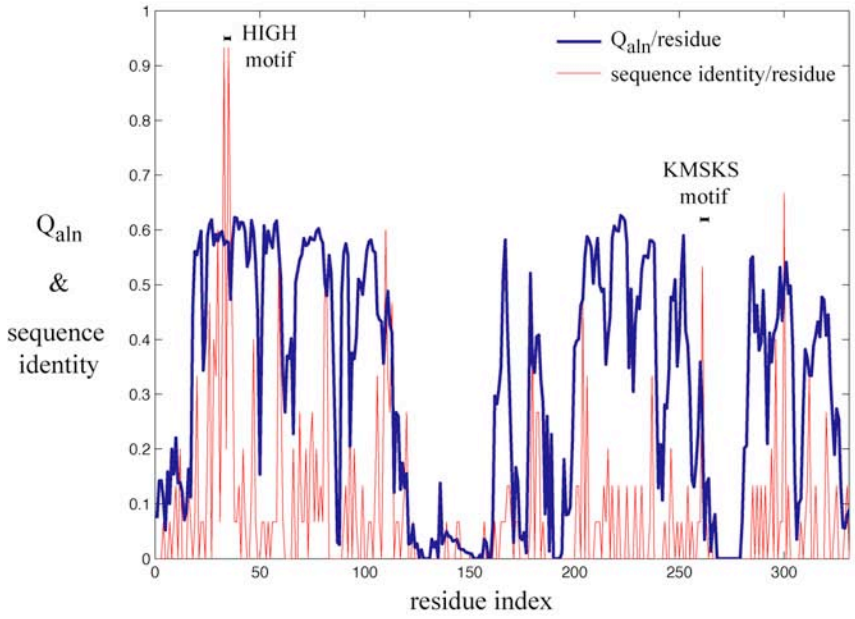




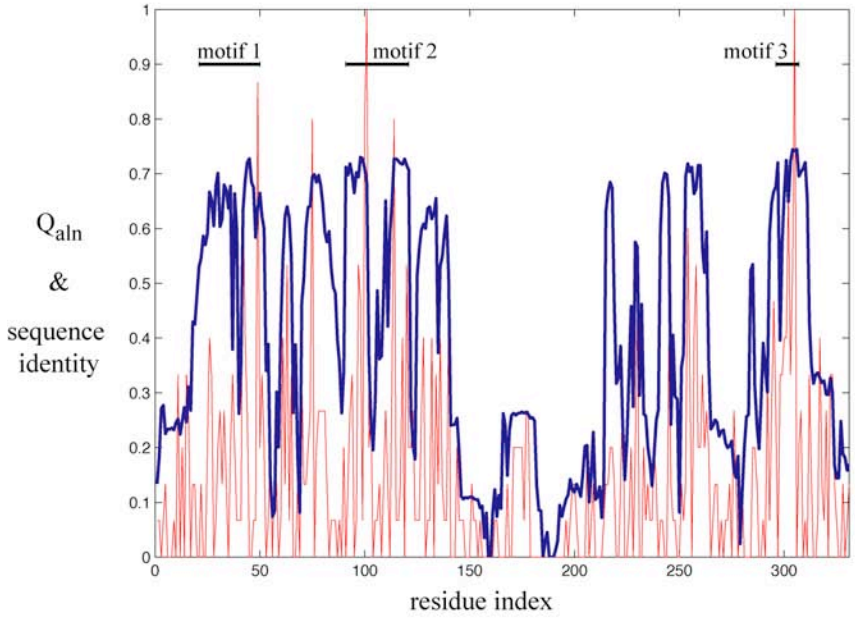


Conservation of Sequence and Structure

GlnRS

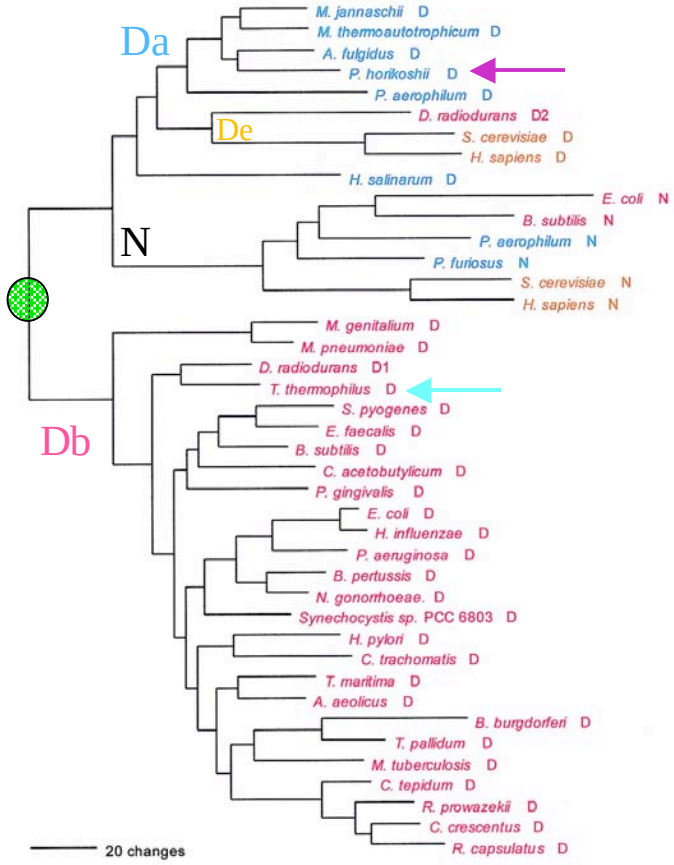


AsnRS

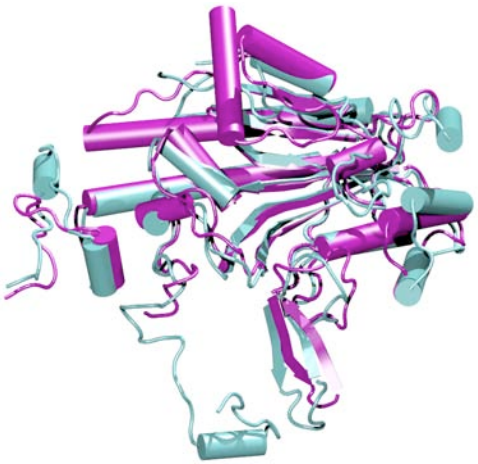


Protein structure encodes evolutionary information

Sequence Phylogeny
Woese et al. 2000

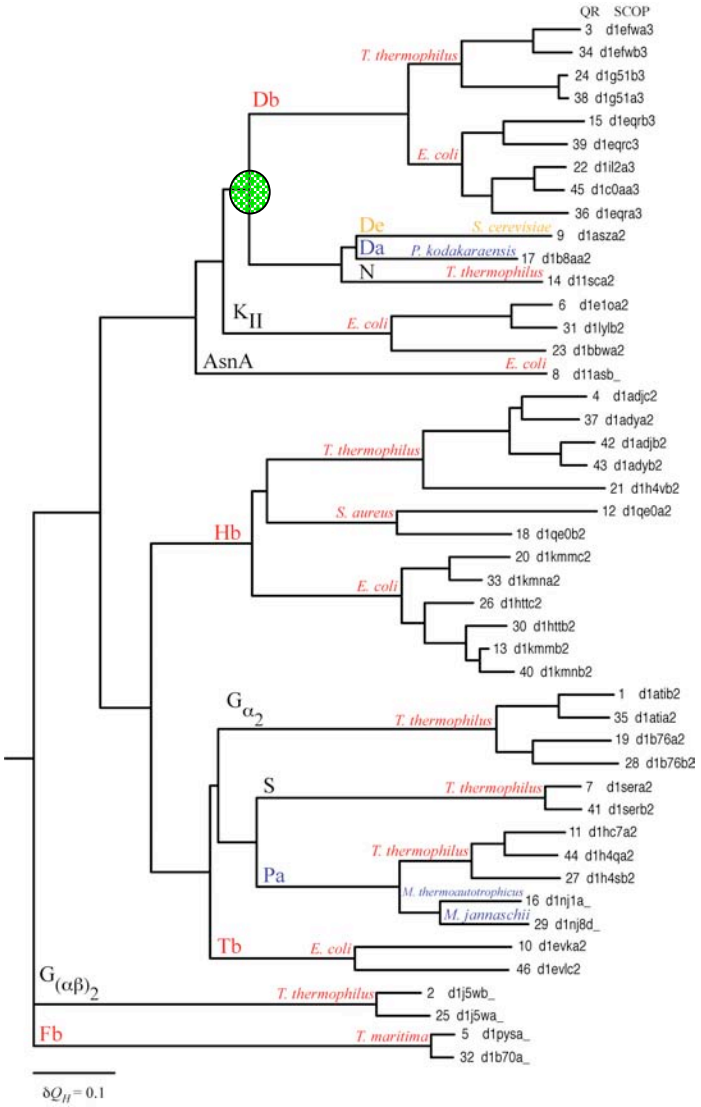


Structural Overlap



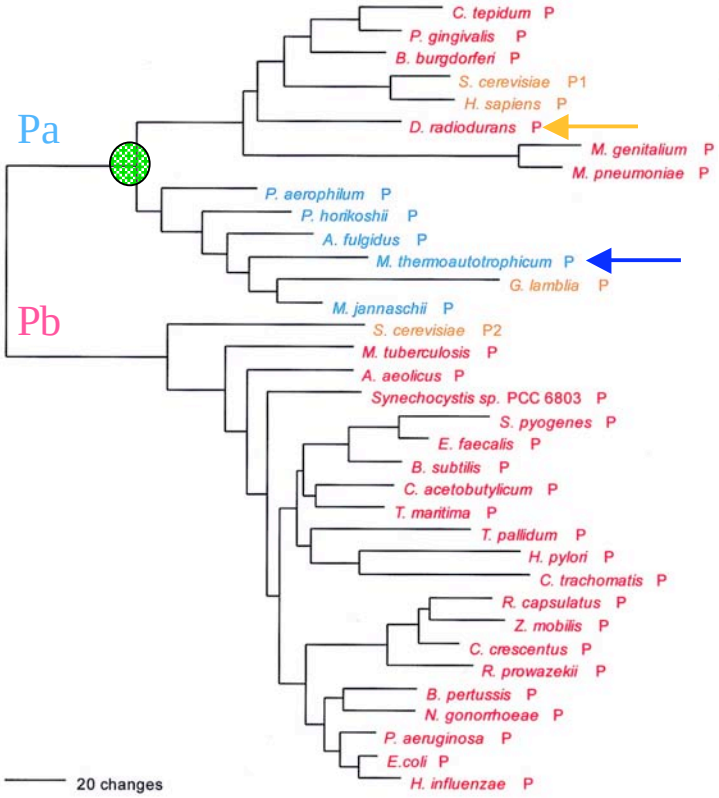
T. thermophilus
P. kodakaraensis

Structure Phylogeny
Class II AARSs

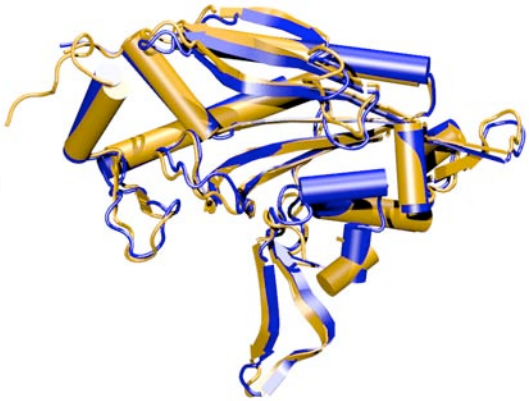


Horizontal Gene Transfer and Protein Structure in ProRS

Sequence Phylogeny
Woese et al. 2000

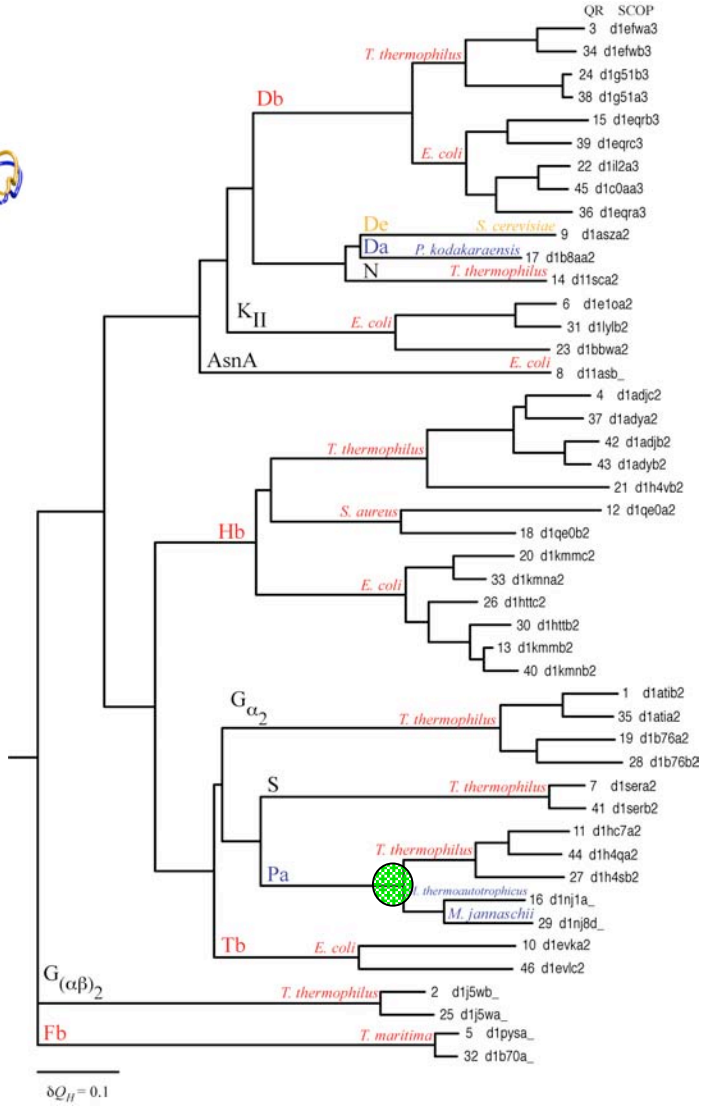


Structural Overlap

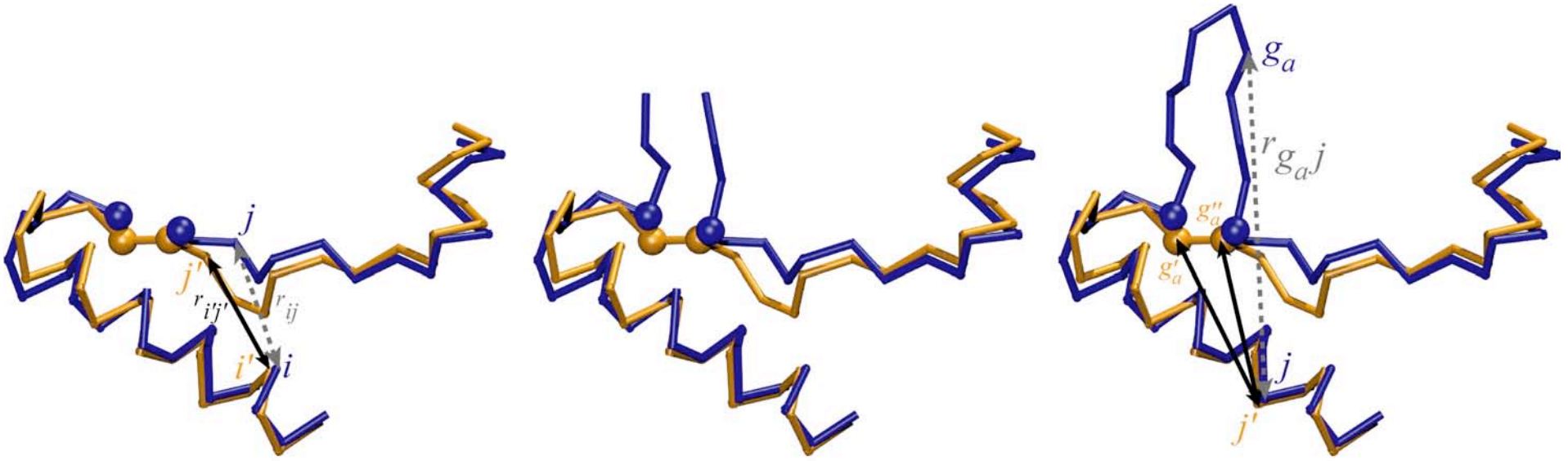


T. thermophilus
M. thermoautotrophicus

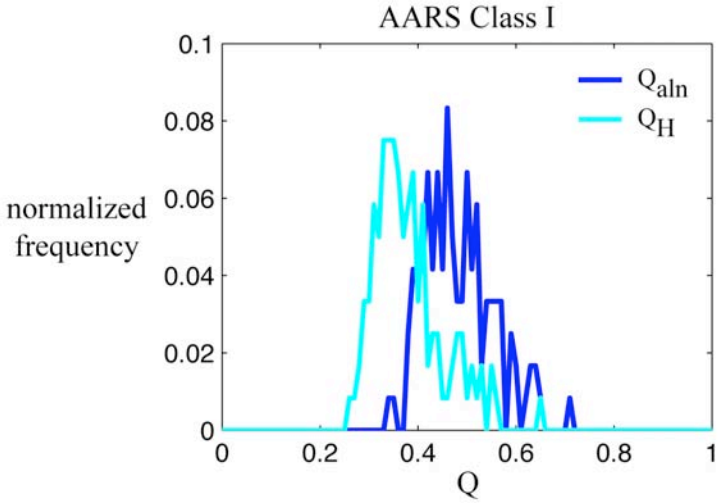
Structure Phylogeny
Class II AARSs



Structural Homology Measure
the effect of insertions



Q_H= 0.82 0.70 0.62



"Gaps influence the analysis
But should not dominate it" CW

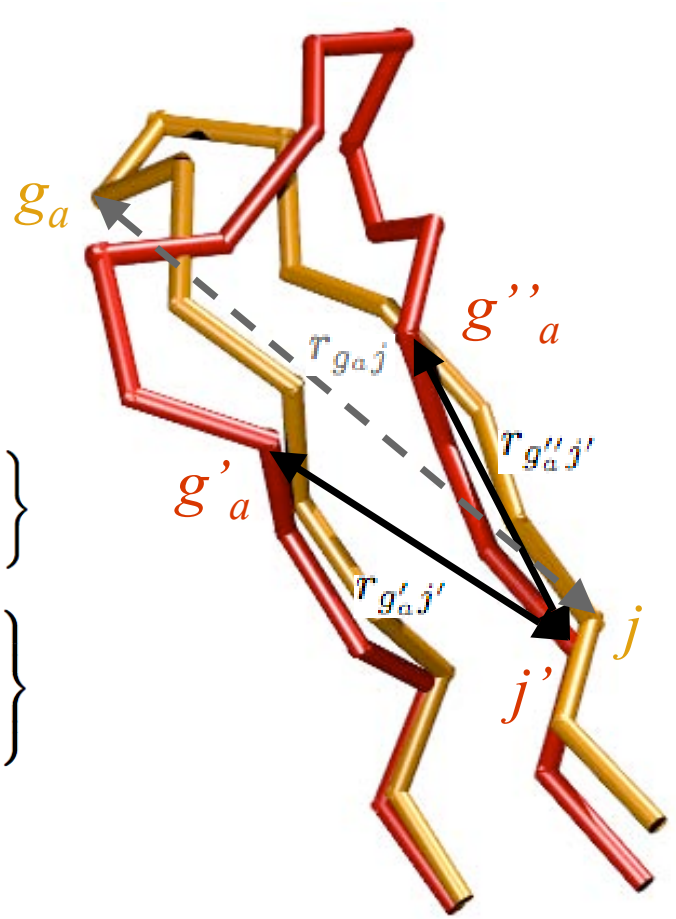
Structural Homology Measure

compare inserted residues to gap edges

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

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

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



QR Factorization

Solve the least squares problem $Ax = b$

by triangularizing A with and orthogonal transformation.

$$Q^T A = \begin{bmatrix} R \\ 0 \end{bmatrix} \qquad Q^T(Ax) = Q^T(b)$$

The system is now solved by back substitution,

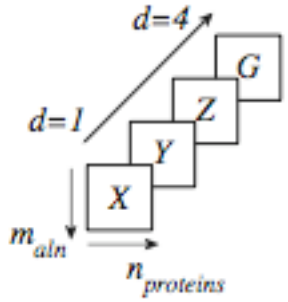
$$\begin{bmatrix} R \\ 0 \end{bmatrix} x = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \qquad Rx = c_1$$

with a minimum residual of

$$\|r\|_2 = \|b - Ax\|_2 = \|c_2\|_2$$

Multi-Dimensional QR

$$Q_{(d)}^T A_{(d)} P = Q_{(d)}^T \left[\right] P = \tilde{R}_{(d)}$$



N-dimensional QR = N one-dimensional QRs.

Permutation matrix is constant for each dimension, ordering norm is Frobenius-like matrix p-norm.

$$\max_{j=k,\ldots,n_{proteins}} (\|a_j\|_{F_p}) \qquad \|a_j\|_{F_p} = (\sum_{d=1}^{4} \sum_{i=k}^{m_{aln}} |a_{ijd}|^p)^{1/p}$$

Encoding Structure

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

Gap "residues": $(0, 0, 0, g)$

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

Encoding Sequence

Orthogonal Encoding = 24-space
23 amino acids symbols (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)

L. Heck, J. Olkin, and K. Nagshineh (1998) *J. Vibration Acoustics* **120**:663.
P. O'Donoghue and Z. Luthey-Schulten (2003) *Micro. Mol. Biol. Rev.* **67**:550-571.

QR Factorization with Column Pivoting

1. Calculate column norm of column *i* and all columns to the right.

$$A^{(k-1)} = H_{k-1}...H_1 A P_1...P_{k-1}$$

$$\max_{j=k,...,n}(s_j^{(k)}) \qquad s_j^{(k)} = (\sum_{i=k}^{m} a_{ij}^2)^{1/2}$$

Ordering Norm

2. Swap column *i* with column to the right of maximum norm and record column permutation.

$$H_{k-1}...H_1 A P_1...P_{k-1} P_k$$

3. Construct and apply H_k

$$A^{(k)} = H_k H_{k-1}...H_1 A P_1...P_{k-1} P_k$$

$$\tilde{A} = AP_1...P_n = AP$$

Original matrix, A, columns ordered by **increasing linear dependence.**

Golub, *Numerische Mathematik*, 1965

Protein Structure Prediction

1-D protein sequence

SISSIRVKSKRIQLG…

Ab Initio protein folding

$$E_{AM} = -\sum_{\mu=1}^{N_\mu}\sum_{i,j}\left\{\gamma_{AM}[P_i, P_j, P_i^\mu, P_j^\mu]\right\}$$

$$X \exp\left[\frac{-(r_{ij} - r_{ij}^{\mu})^2}{2\sigma_{ij}^2}\right]$$

Threading/Profile Alignment

$$E = E_{match} + E_{gap}$$

Target Sequence

SISSRVKSKRIQLGLNQAELAQKV------GTTQ…
QFANEFKVRRIKLGYTQ----TNVGEALAAVHGS…

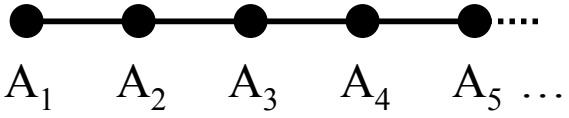
Known structure(s)

3-D protein structure

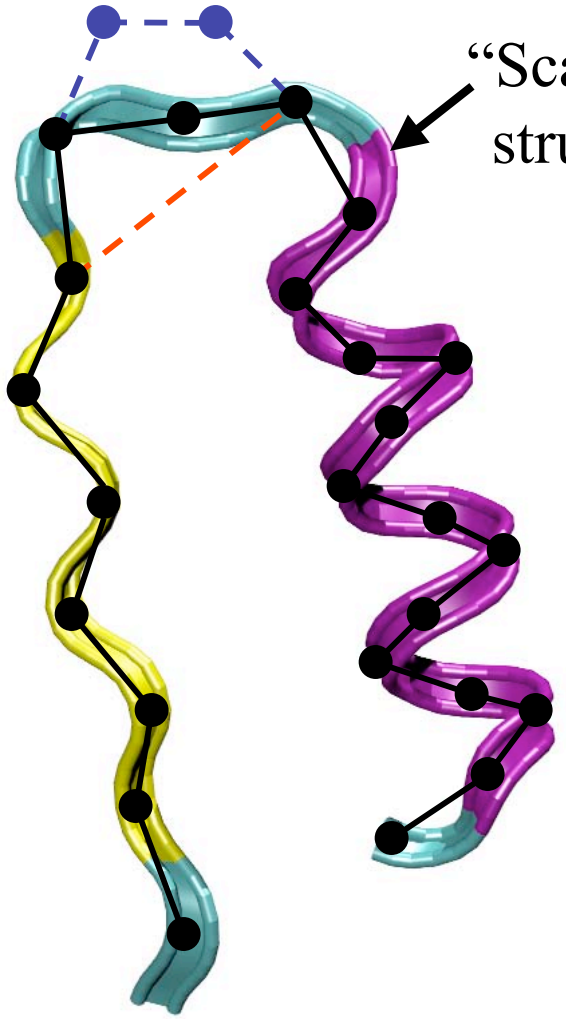
Eastwood,Hardin,Luthey-Schulten,Wolynes (**2001**)
*IBM. J.RES.&DEV.*45:475-497
Papoian, et.al. *PNAS* (2004)

Sequence-Structure Alignment

Target sequence



Alignment between target(s) and scaffold(s)



1. Energy Based Threading*

$$H = E_{contact} + E_{profile} + E_{H-bonds} + E_{gap}$$

$$E_{profile} = \sum_{i}^{n} \gamma^{(p)}\left(A_i, SS_i, SA_i\right)$$

$$E_{contact} = \sum_{i,j} \sum_{k=1}^{2} \gamma_k^{(ct)}\left(A_i, A_j\right) * U\left(r_k - r_{ij}\right)$$

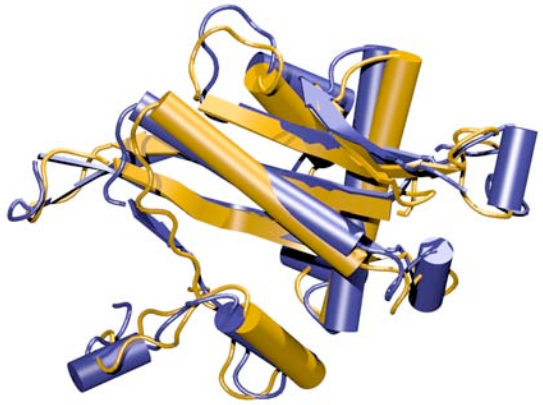
2. Sequence – Structure Profile Alignments

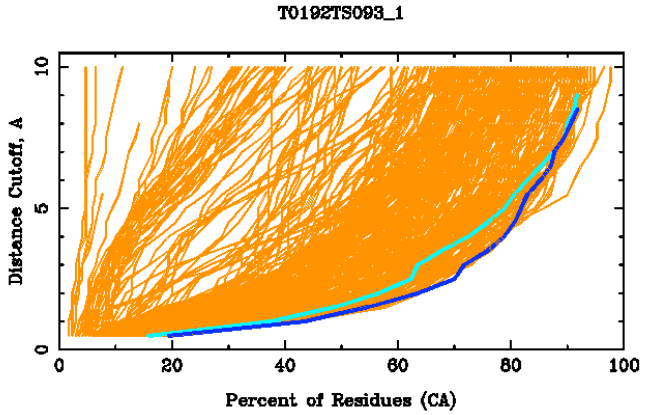
Clustal, Hidden Markov (HMMER, PSSM)
with position dependent gap penalties

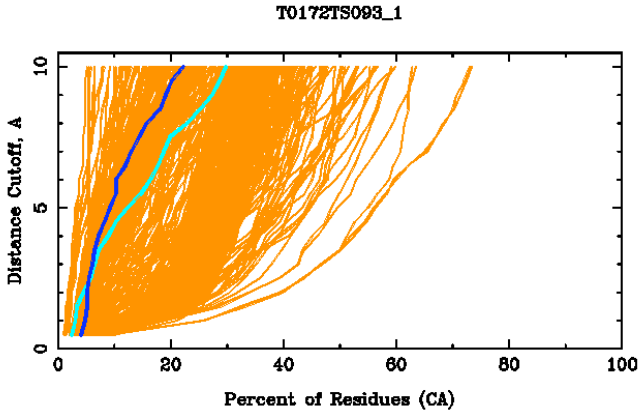
*R. Goldstein, Z. Luthey-Schulten, P. Wolynes (1992, PNAS), K. Koretke et.al. (1996, Proteins)

CASP5
Fold Recognition/Threading

Schulten-Wolynes Group





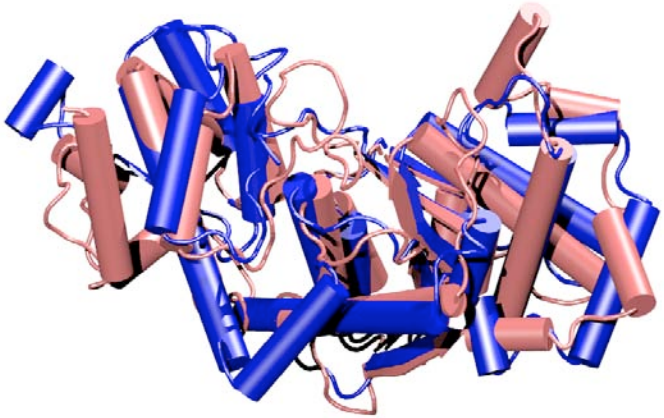


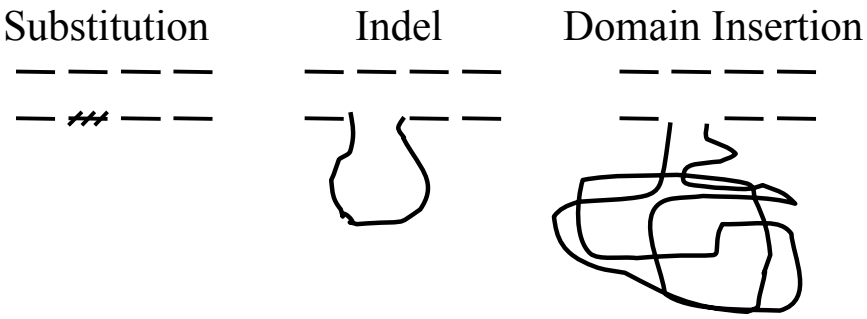
The prediction is never better than the scaffold.

Threading energy function/profiles requires improvement.

Why Study the Evolution of Protein Structure?

In what specific ways has the evolutionary dynamic changed protein shape over time?





implications for protein structure prediction, protein design

What can studying the change in protein shape over time tell us about the evolutionary process?

How did translation evolve?

When, with respect to the root of the universal phylogenetic tree, was translation established in its modern form?

What was the role of the AARSs in the evolution of the translation mechanism, development of the genetic code?