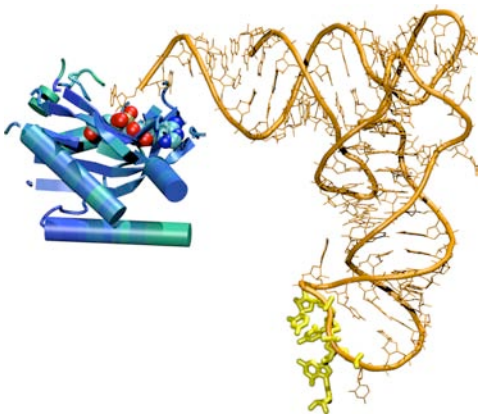
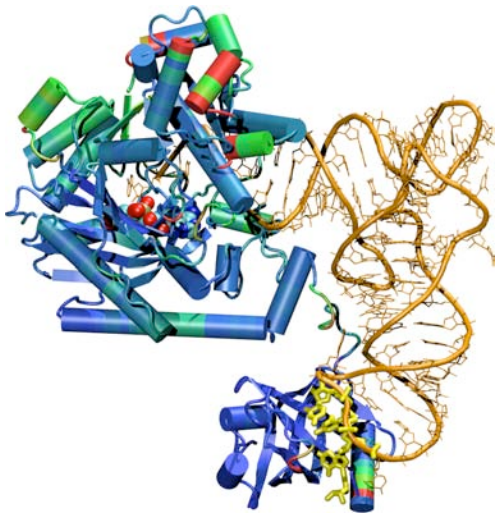
# Evolutionary concepts in bioinformatics

## evolution of protein structure in the aminoacyl-tRNA synthetases



| First position | Second position: U | Second position: C | Second position: A | Second position: G | Third position |
|---|---|---|---|---|---|
| U | UUU Phe | UCU Ser | UAU Tyr | UGU Cys | U |
| | UUC Phe | UCC Ser | UAC Tyr | UGC Cys | C |
| | UUA Leu | UCA Ser | UAA Stop | UGA Stop | A |
| | UUG Leu | UCG Ser | UAG Stop | UGG Trp | G |
| C | CUU Leu | CCU Pro | CAU His | CGU Arg | U |
| | CUC Leu | CCC Pro | CAC His | CGC Arg | C |
| | CUA Leu | CCA Pro | CAA Gln | CGA Arg | A |
| | CUG Leu | CCG Pro | CAG Gln | CGG Arg | G |
| A | AUU Ile | ACU Thr | AAU Asn | AGU Ser | U |
| | AUC Ile | ACC Thr | AAC Asn | AGC Ser | C |
| | AUA Ile | ACA Thr | AAA Lys | AGA Arg | A |
| | AUG Met/start | ACG Thr | AAG Lys | AGG Arg | G |
| G | GUU Val | GCU Ala | GAU Asp | GGU Gly | U |
| | GUC Val | GCC Ala | GAC Asp | GGC Gly | C |
| | GUA Val | GCA Ala | GAA Glu | GGA Gly | A |
| | GUG Val | GCG Ala | GAG Glu | GGG Gly | G |

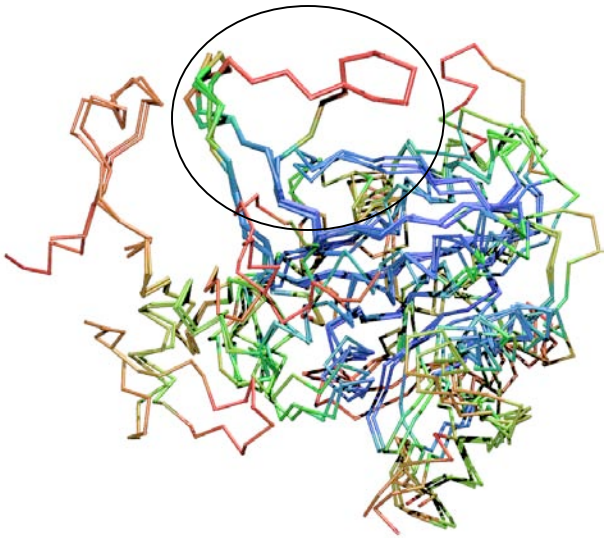




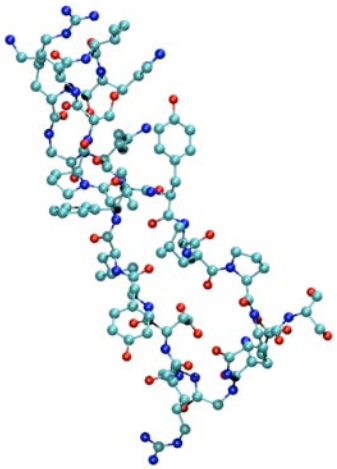
Patrick O'Donoghue* and Zan Luthey-Schulten

Department of Chemistry
University of Illinois at Urbana-Champaign

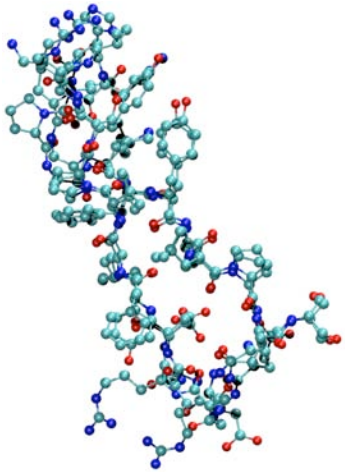
# Protein Homology in Structure and Sequence



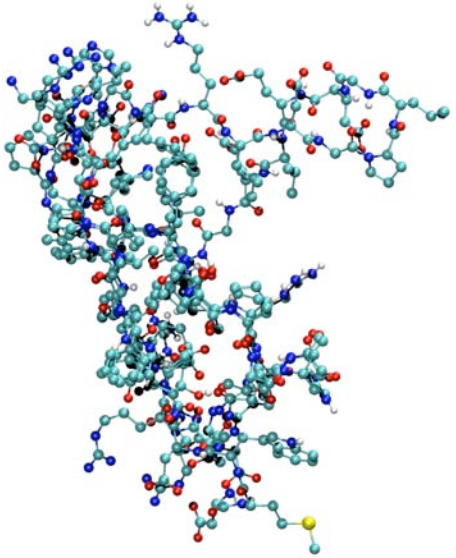
3 homologous structures;
2 closely related (Db1, Db2),
1 more distant (Fb).



Db1



Db1, Db2



Db1, Db2, Fb

Db1, Db2, Fb
backbone only

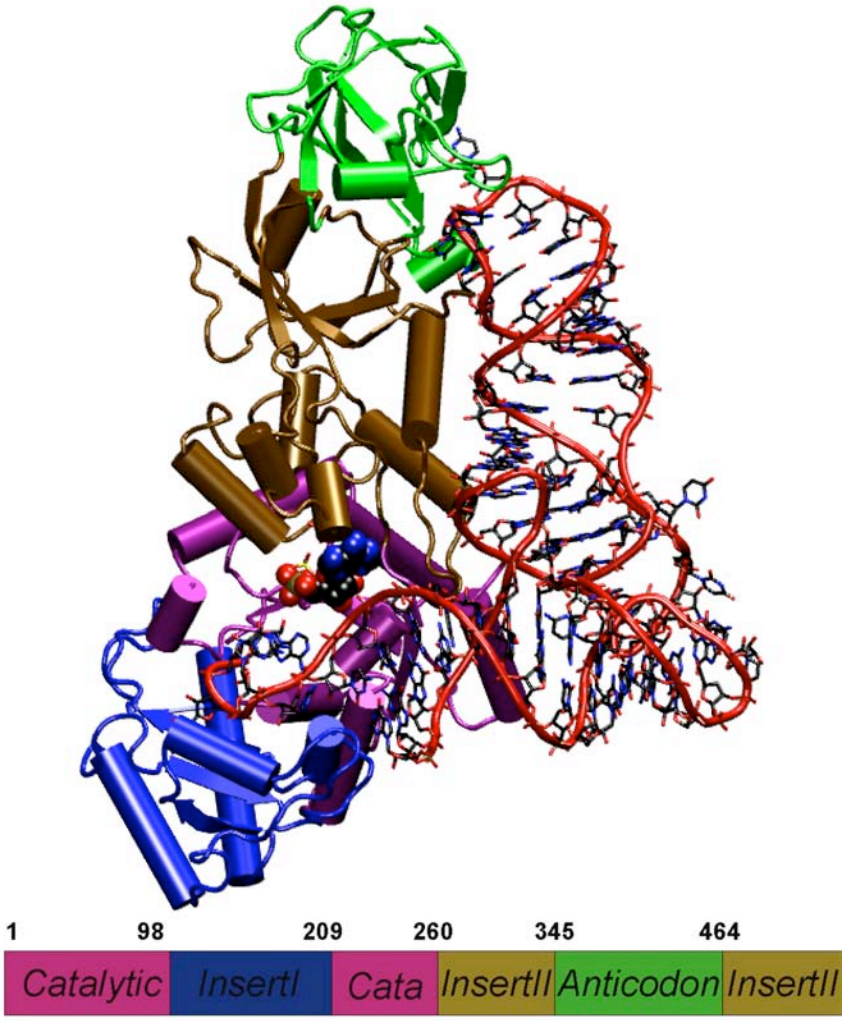
```
Db1      E---GARDFLV-PYRHE----------PGLFYALPQS
Db2      -E--GARDYLV-PSRVH----------KGKFYALPQS
Fb       ---DMWDTFWLT-GE--GFRLEGPLGEEVEGRLLLRTH
```

# What can be learned from AARSs?

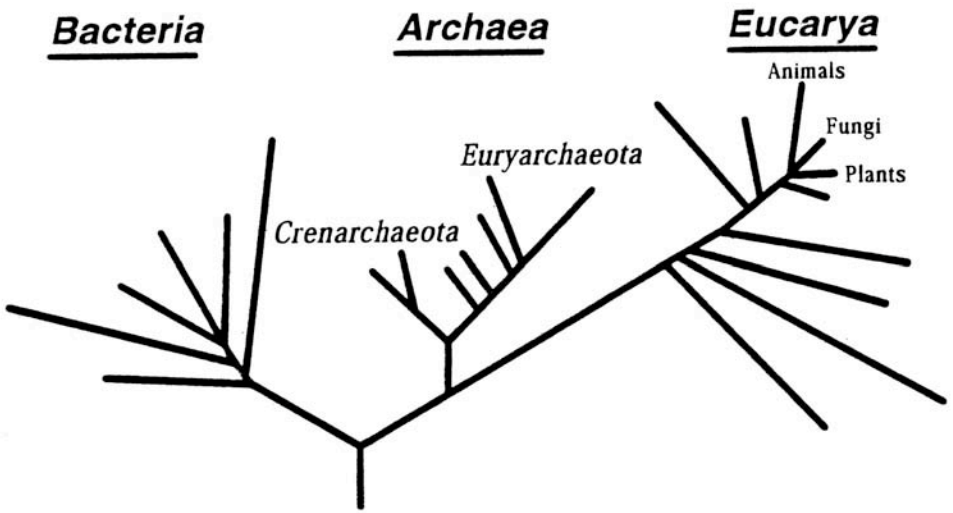
"The aminoacyl-tRNA synthetases, perhaps better than any other molecules in the cell, eptiomize the current situation and help to understandard (the effects) of Horizontal Gene Transfer (HGT)."

Carl Woese (PNAS, 2000; MMBR 2000)

# Aminoacyl-tRNA synthetases




## Universal Tree of Life

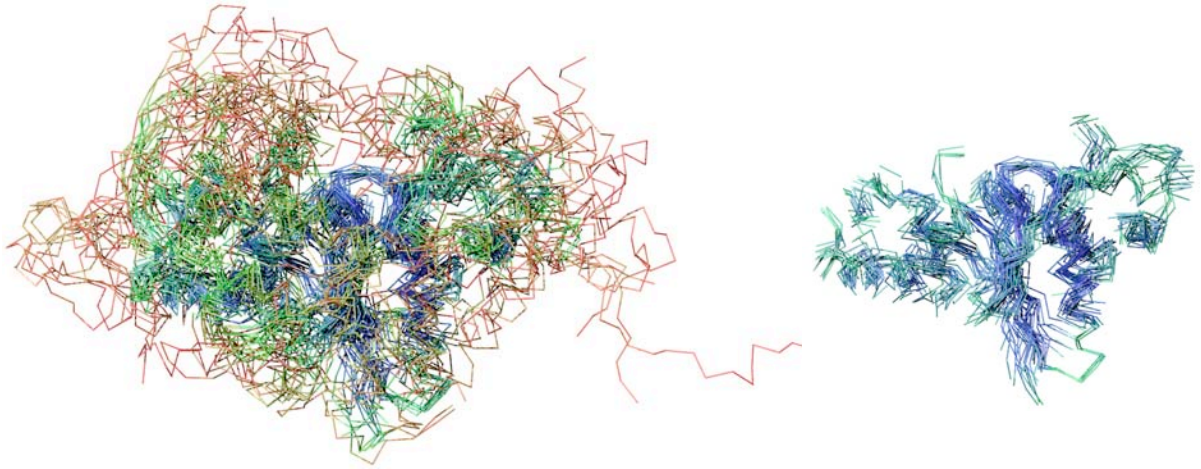



Woese *PNAS* 1990, 2002.

# 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.

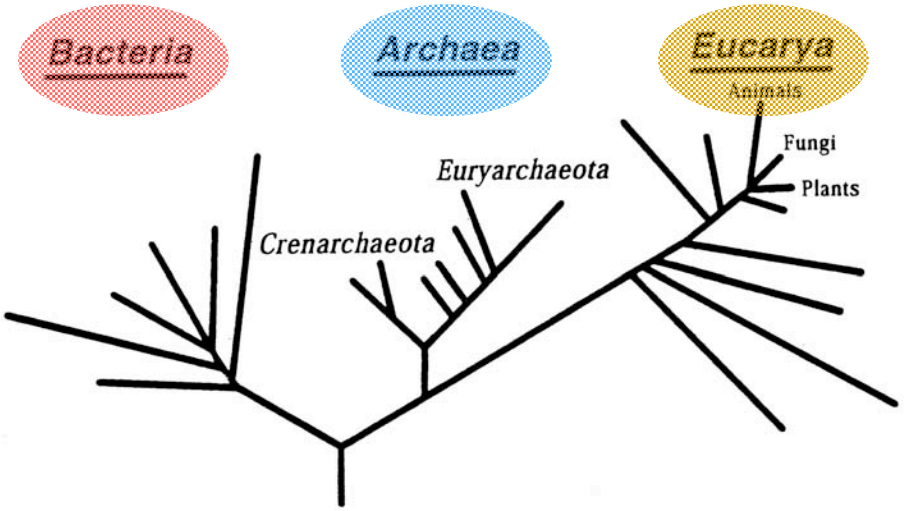
   Eliminate bias inherited from structure or sequence databases.

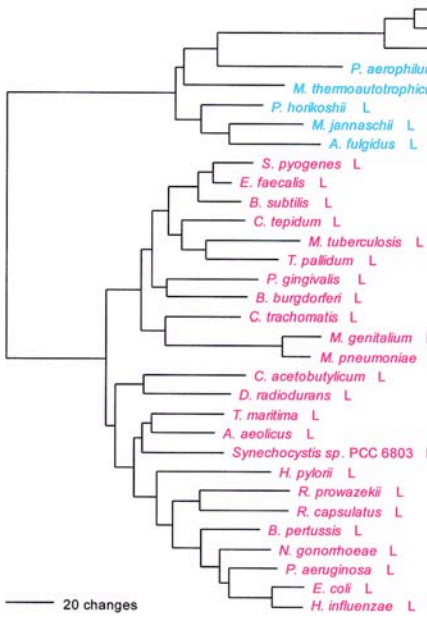
   Important for bioinformatic analysis (substitution matrices, knowledge based potentials structure pred., genome annotation) and evolutionary analysis.

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

# Universal Phylogenetic Tree

## three domains of life




Eucarya

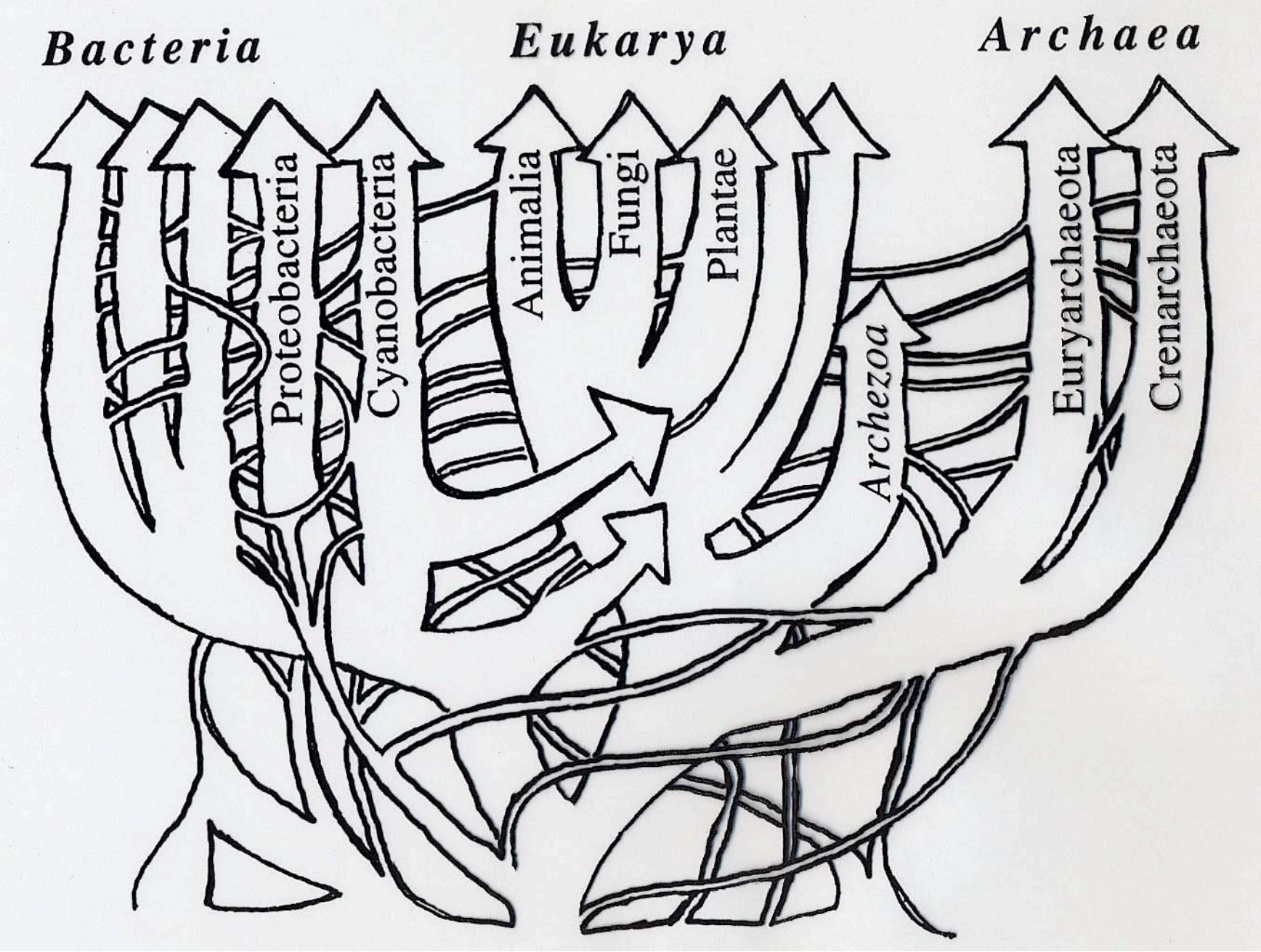
Archaea

Bacteria

Leucyl-tRNA synthetase displays the full canonical phylogenetic distribution.

for review see Woese *PNAS* 2000

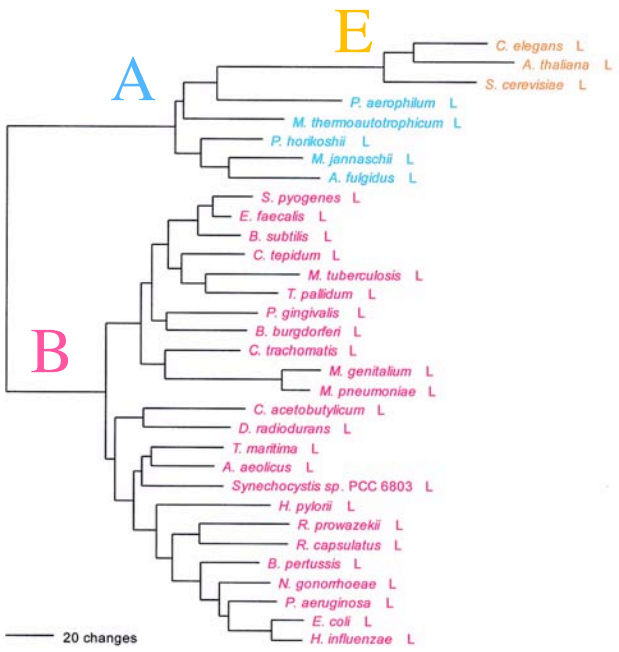
Woese, Olsen, Ibba, Soll *MMBR* 2000



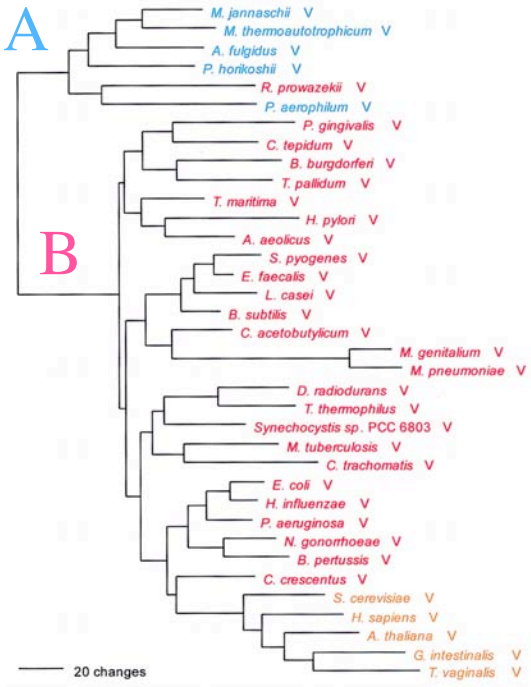

After W. Doolittle, modified by G. Olsen

# Phylogenetic Distributions

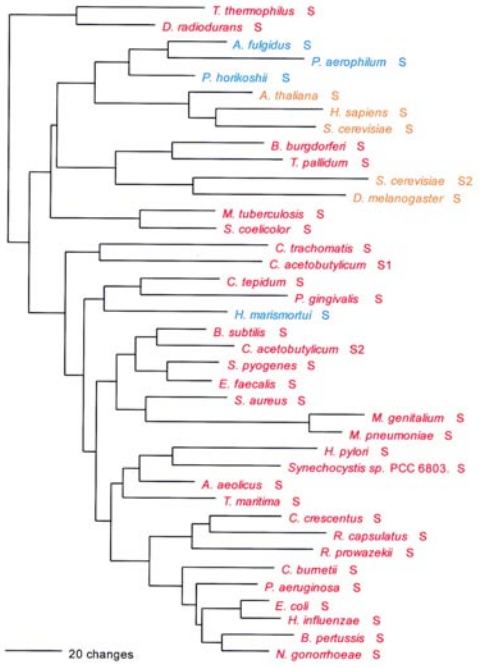
Full Canonical




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

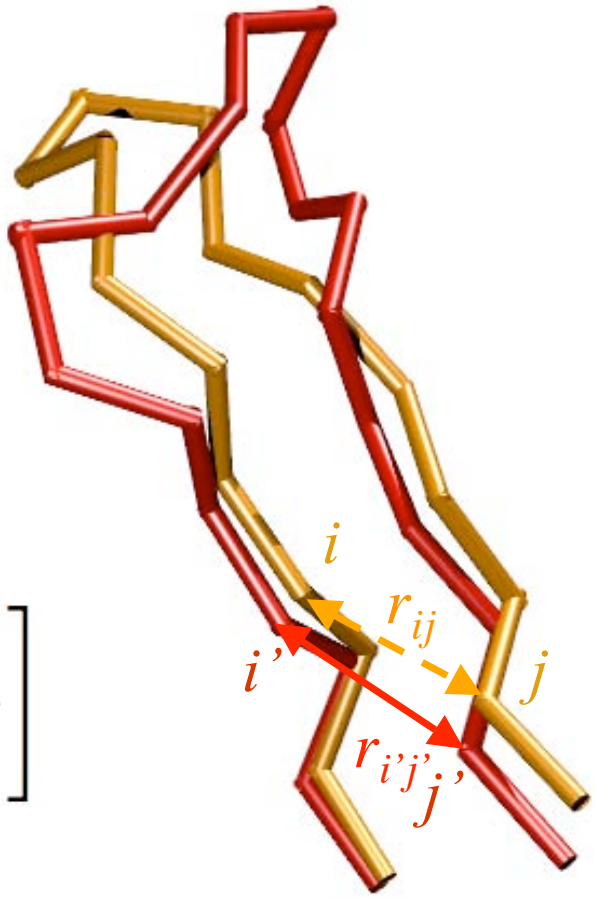
# Protein Structure Similarity Measure

## $Q_H$ Structural Homology

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

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

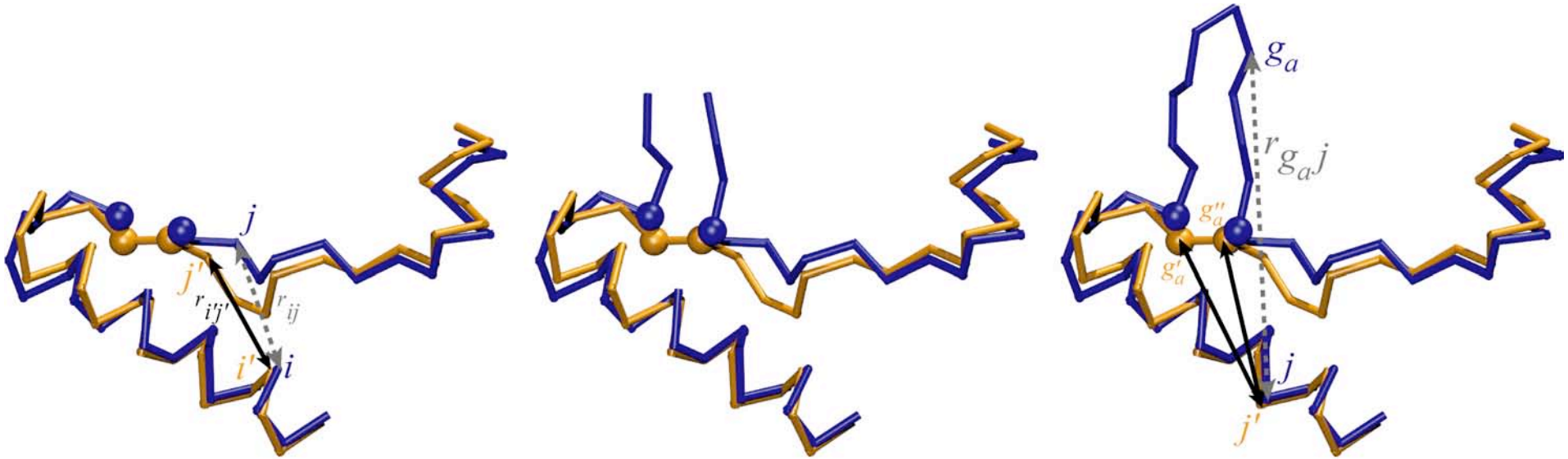
$$q_{aln} = \sum_{i<j-2} \exp\left[-\frac{(r_{ij} - r_{i'j'})^2}{2\sigma_{ij}^2}\right]$$



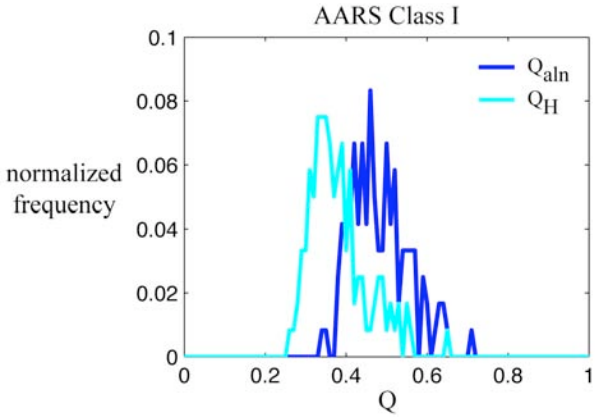

O'Donoghue & Luthey-Schulten *MMBR* 2003.

# Structural Similarity Measure the effect of insertions

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



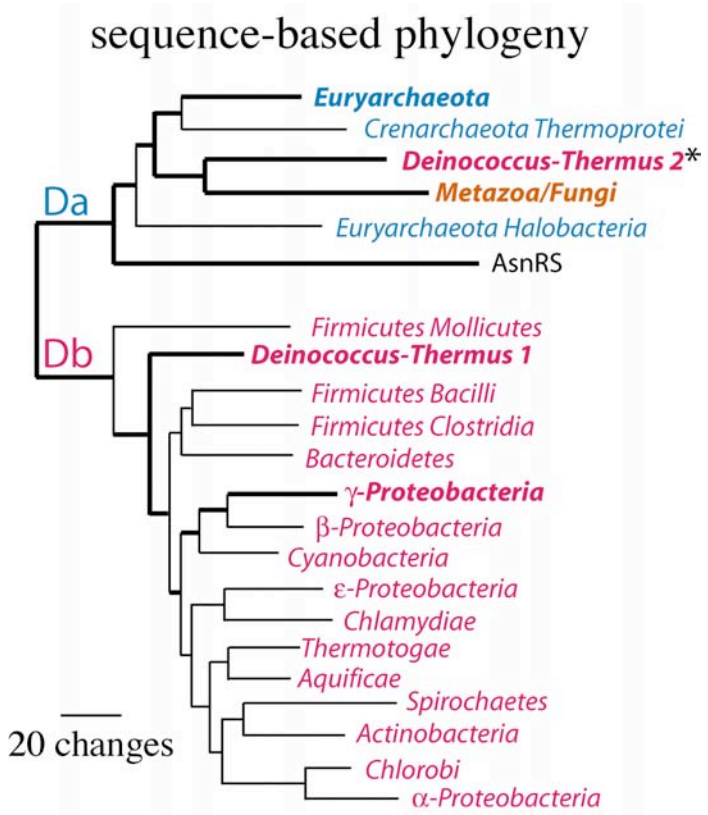

$Q_H$= 0.82 0.70 0.62

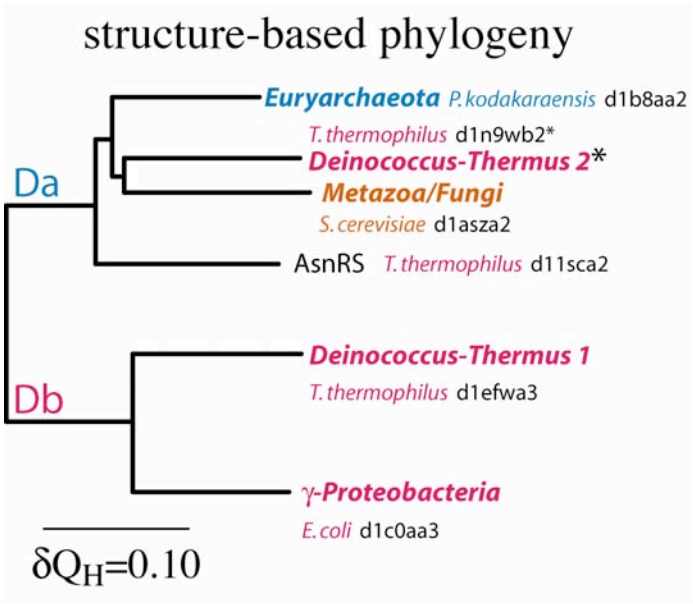


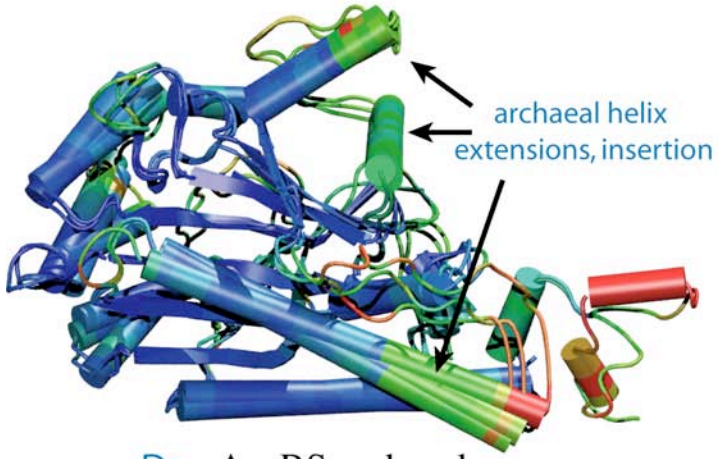

$$q_{gap} = \sum_{g_a}\sum_{j}^{N_{aln}} \max\left\{\exp\left[-\frac{\left(r_{g_aj} - r_{g'_aj'}\right)^2}{2\sigma^2_{g_aj}}\right], \exp\left[-\frac{\left(r_{g_aj} - r_{g''_aj'}\right)^2}{2\sigma^2_{g_aj}}\right]\right\}$$

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

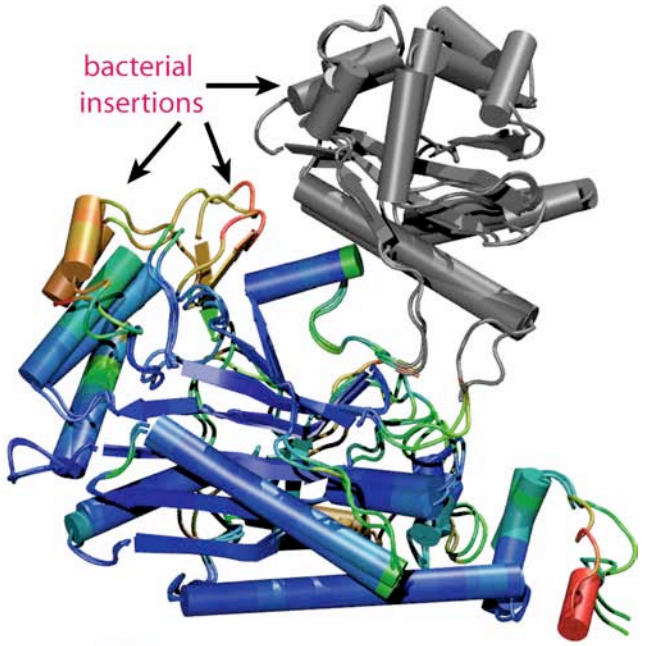
# Protein structure encodes evolutionary information



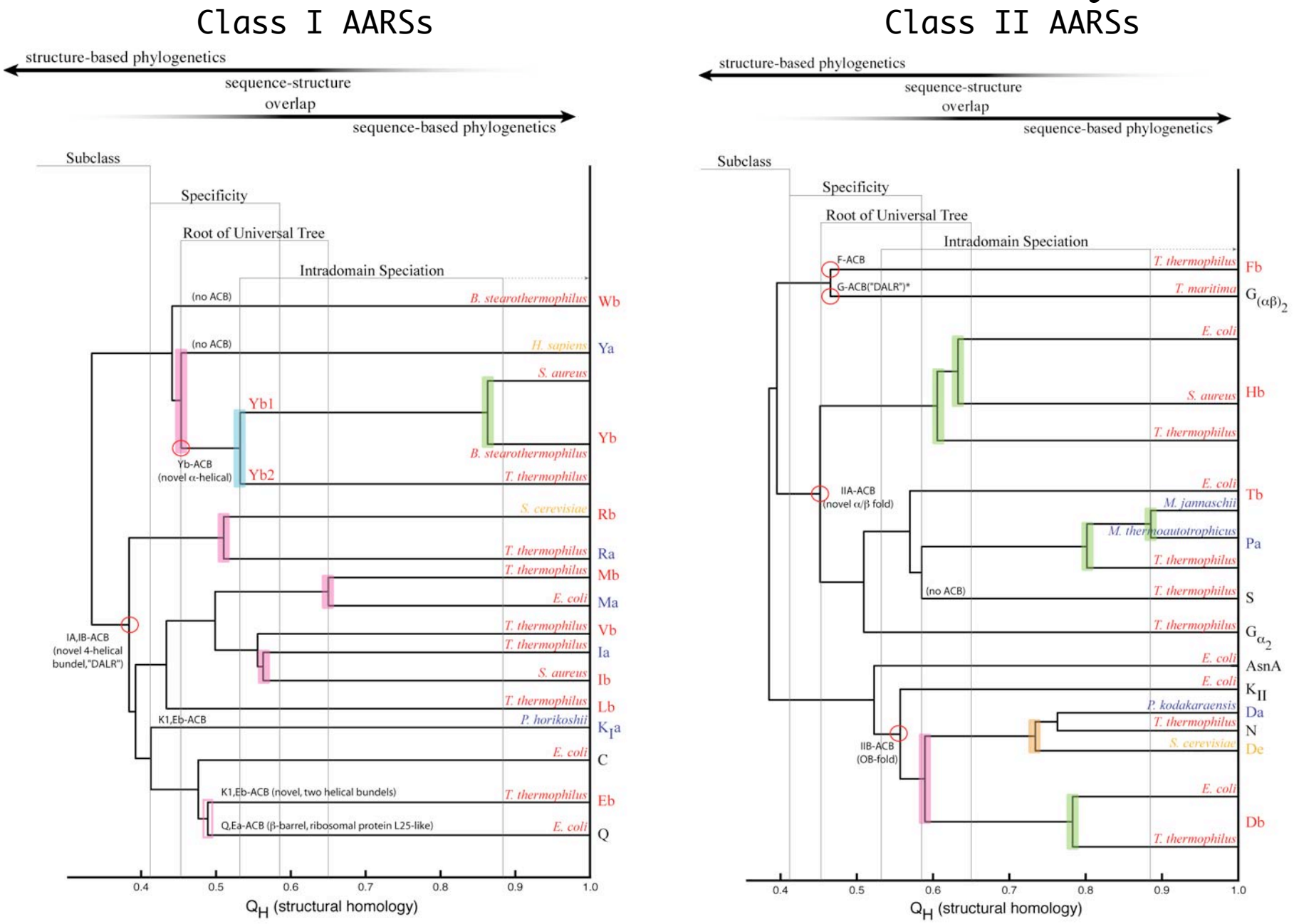




Da - AspRS archaeal genre

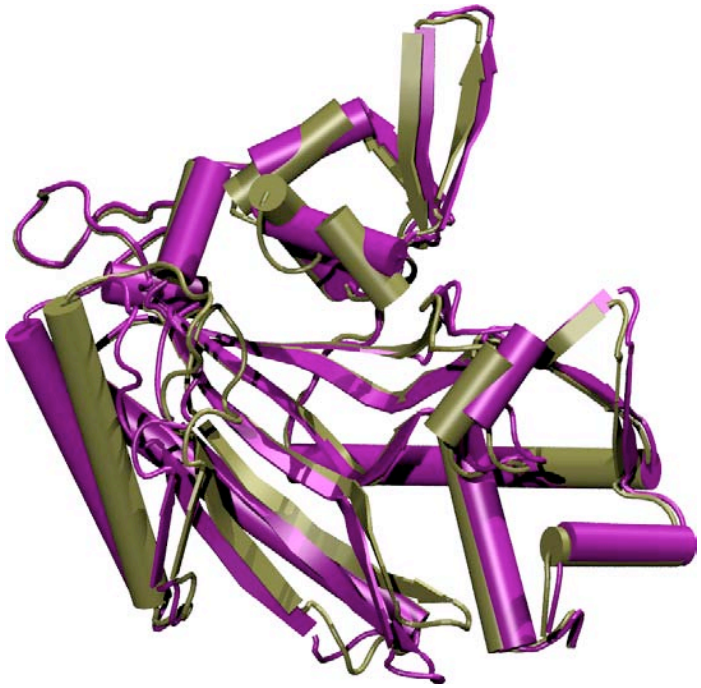



Db - AspRS bacterial genre

# Protein structure reveals distant evolutionary events



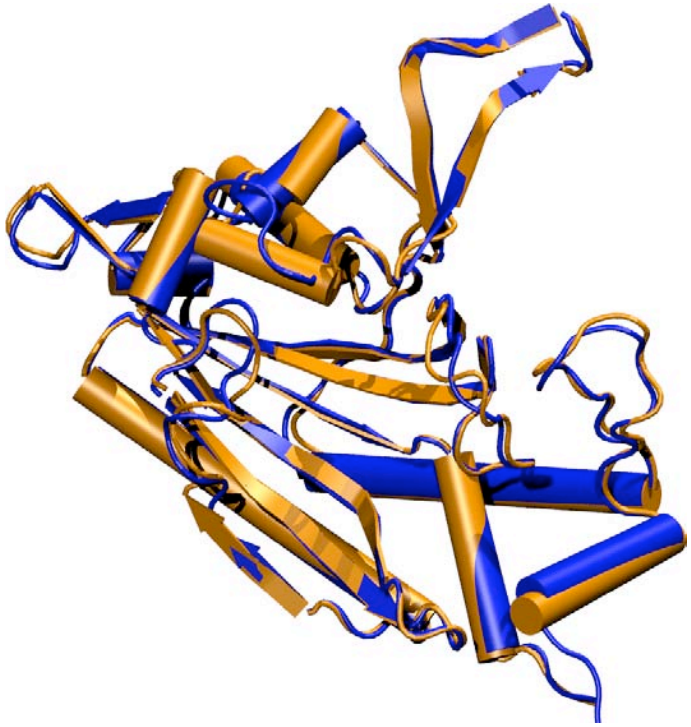
# Sequences define more recent evolutionary events



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



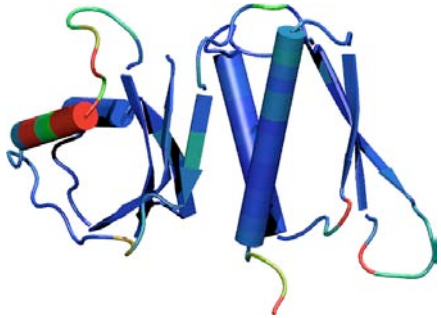
Structures for two
different species.

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

$Q_H = 0.89$
Sequence identity = 0.69

# Using structure to correct sequence alignments

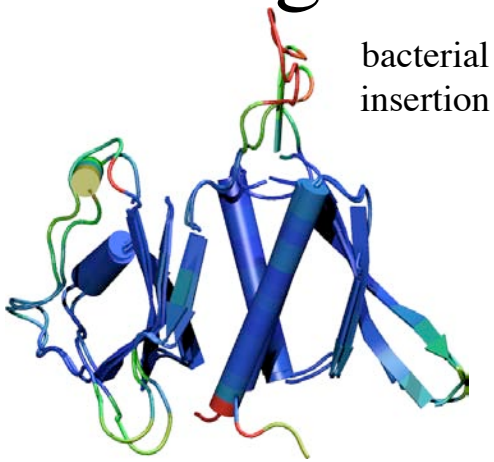
archaeal helix



archaeal S8 (S8a)

overlap of S8a and S8b

bacterial insertion

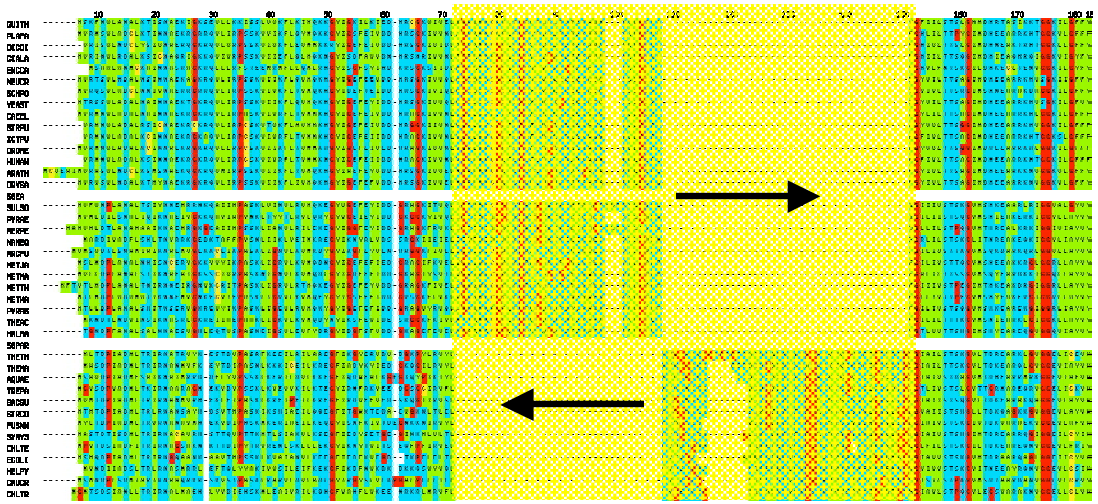


bacterial S8 (S8b)

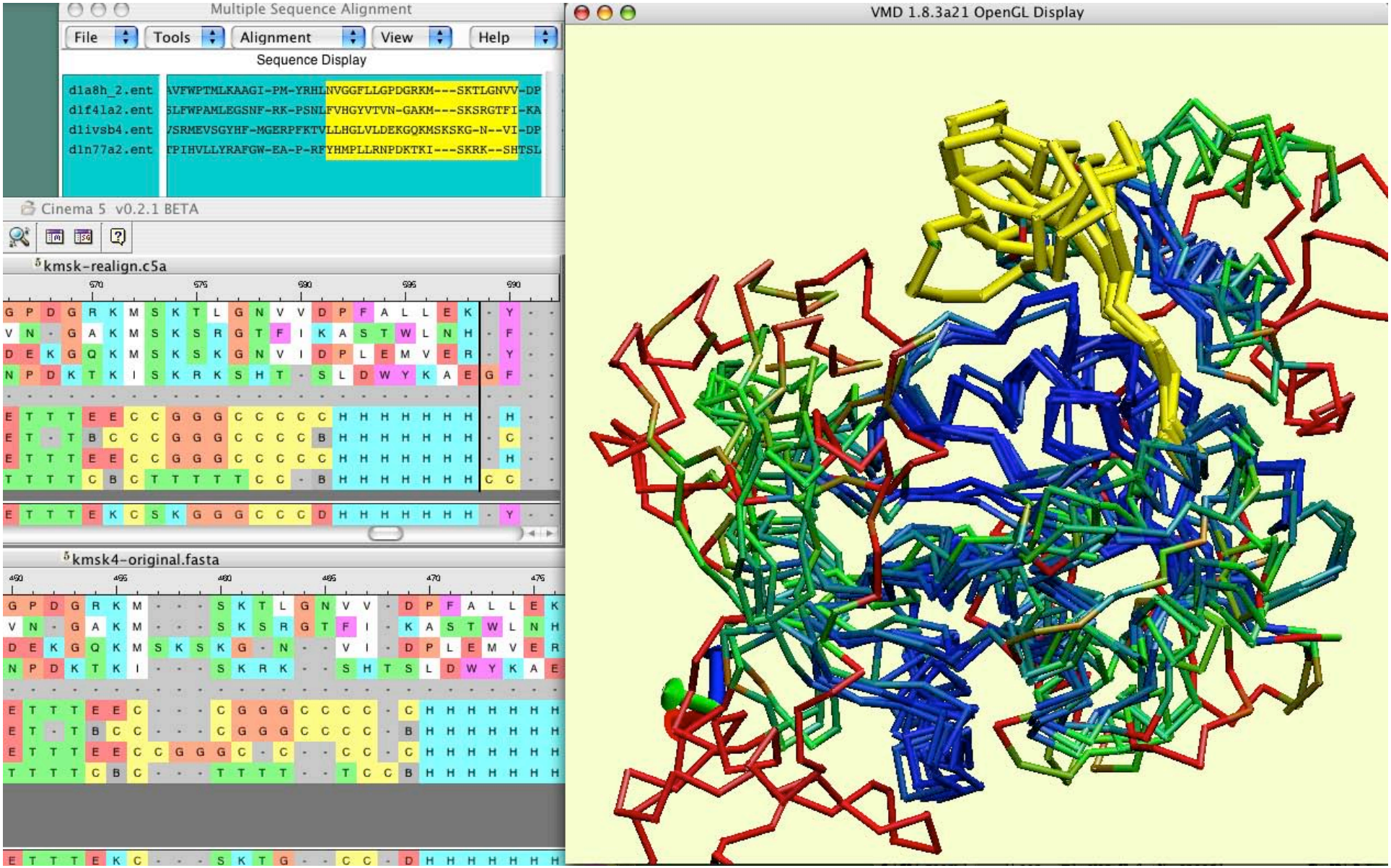
*Eucarya*

*Archaea*

*Bacteria*

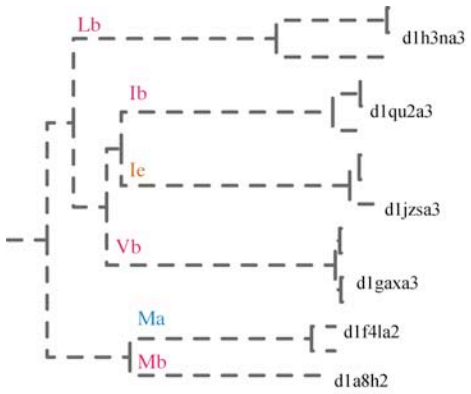


# Conformational versus evolutionary change

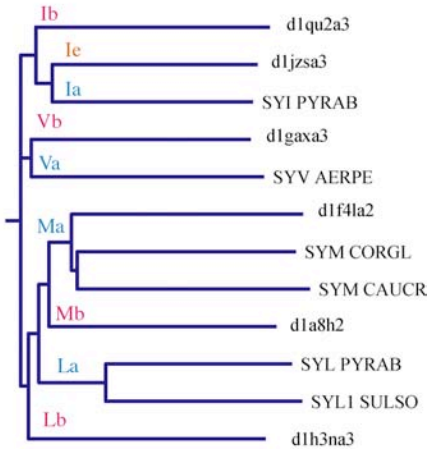


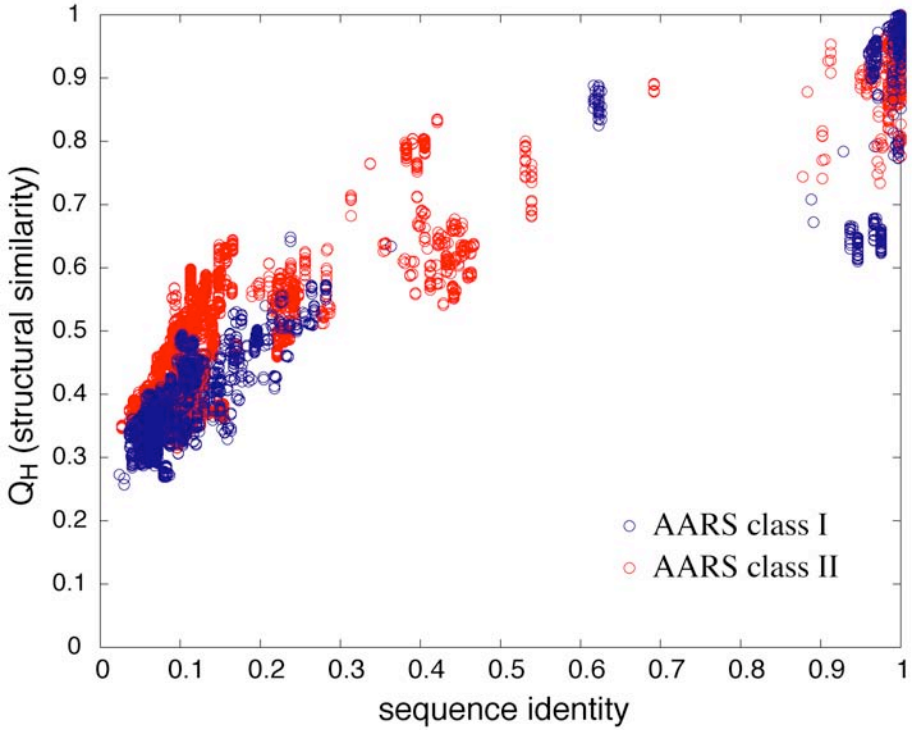
# Towards a unified phylogenetic framework in sequence and structure

Structure-based tree



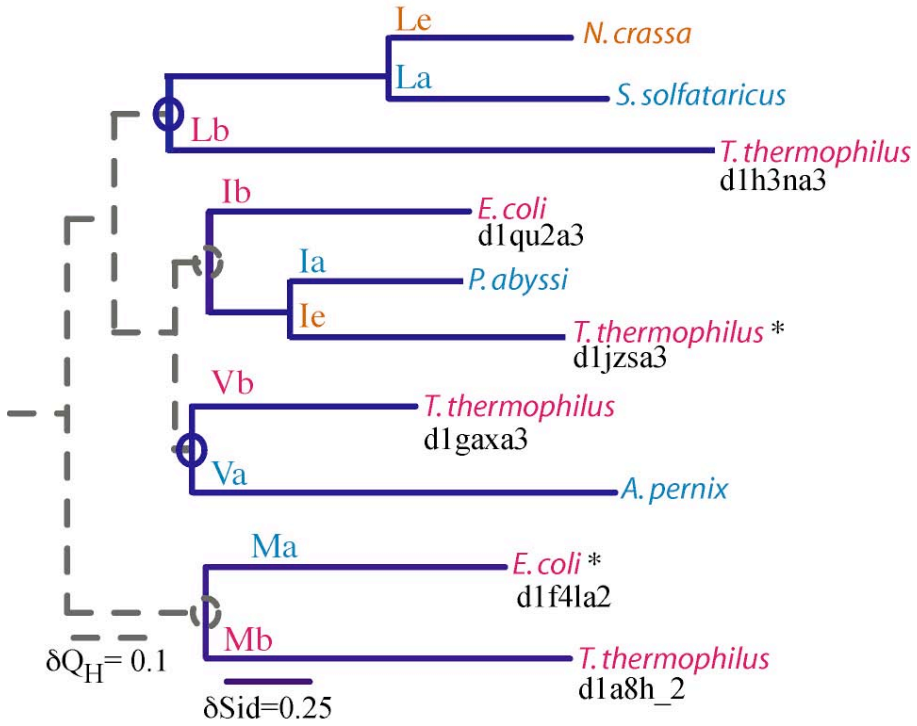

Sequence-based tree



# Towards a unified phylogenetic framework in sequence and structure

Combined sequence-structure tree

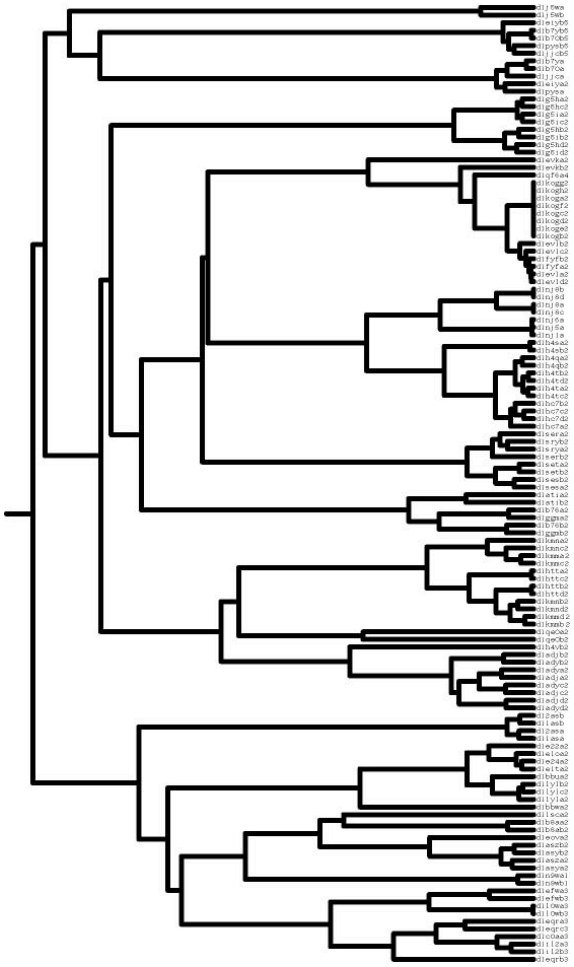



Structure is used to infer distant evolutionary events, i.e., the development of basic structures and functions.

Sequences supplement the missing structure data, and define more recent evolutionary events, i.e., speciation.

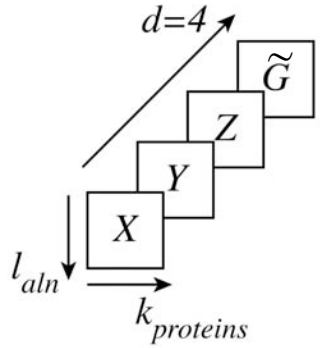
# Non-redundant representative sets

Too much information
129 Structures

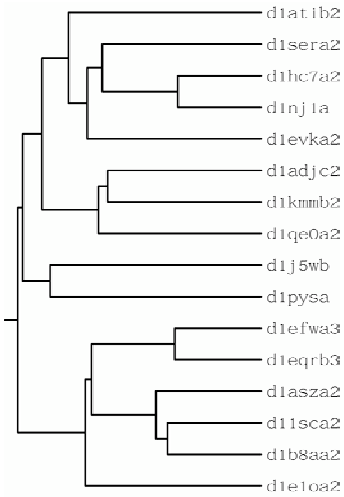


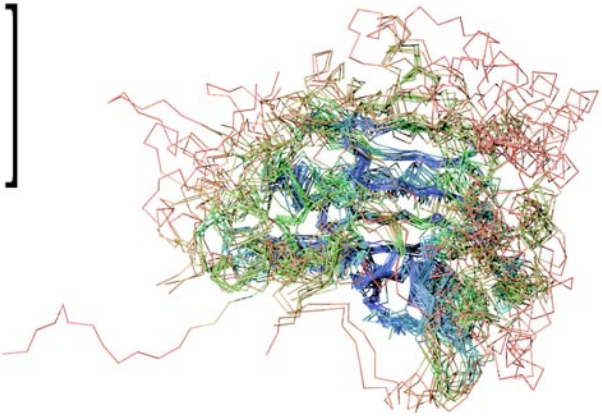
Multidimensional QR factorization of alignment matrix, $A$.

$$A = \left[ \quad \right]$$




Economy of information
16 representatives



QR computes a set of minimal linearly dependent structures.


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

P. O'Donoghue and Z. Luthey-Schulten (2004) *J. Mol. Biol.*, 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 $\tilde{G} = gG \quad (0, 0, 0, g)$

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

adjustable parameter

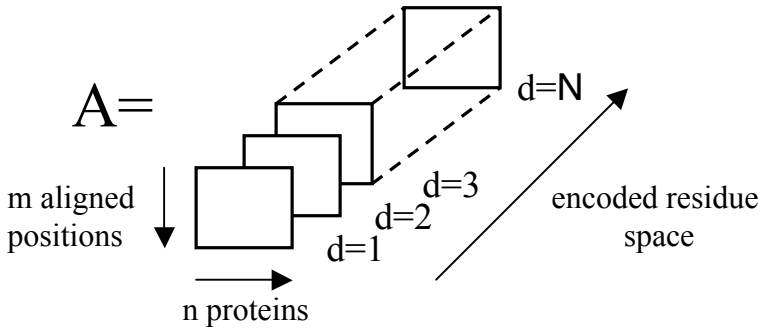
## Sequence Space

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.

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

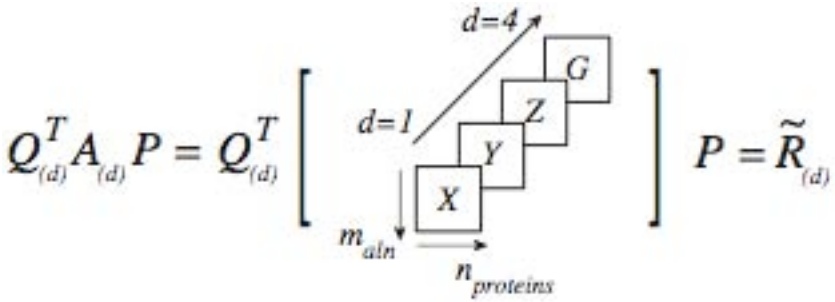
$$\tilde{A} = AP$$

$\boldsymbol{Q^T}$ – 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.




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

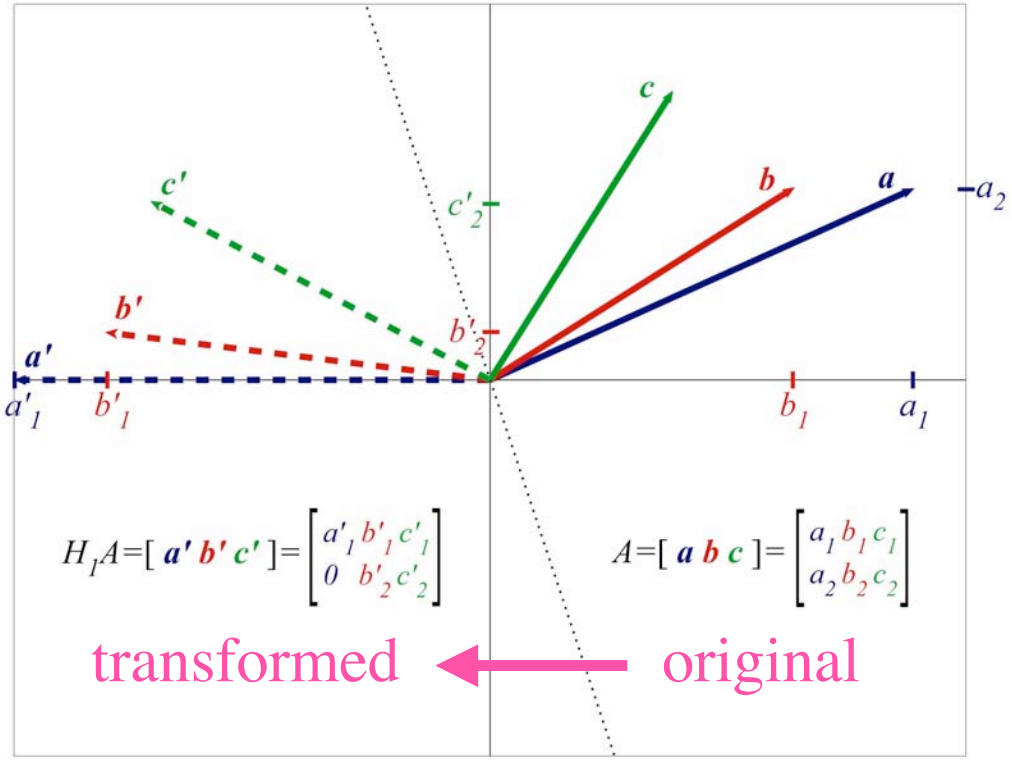

L. Heck, J. Olkin, and K. Nagshineh (1998) *J. Vibration Acoustics* **120**:663.

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:

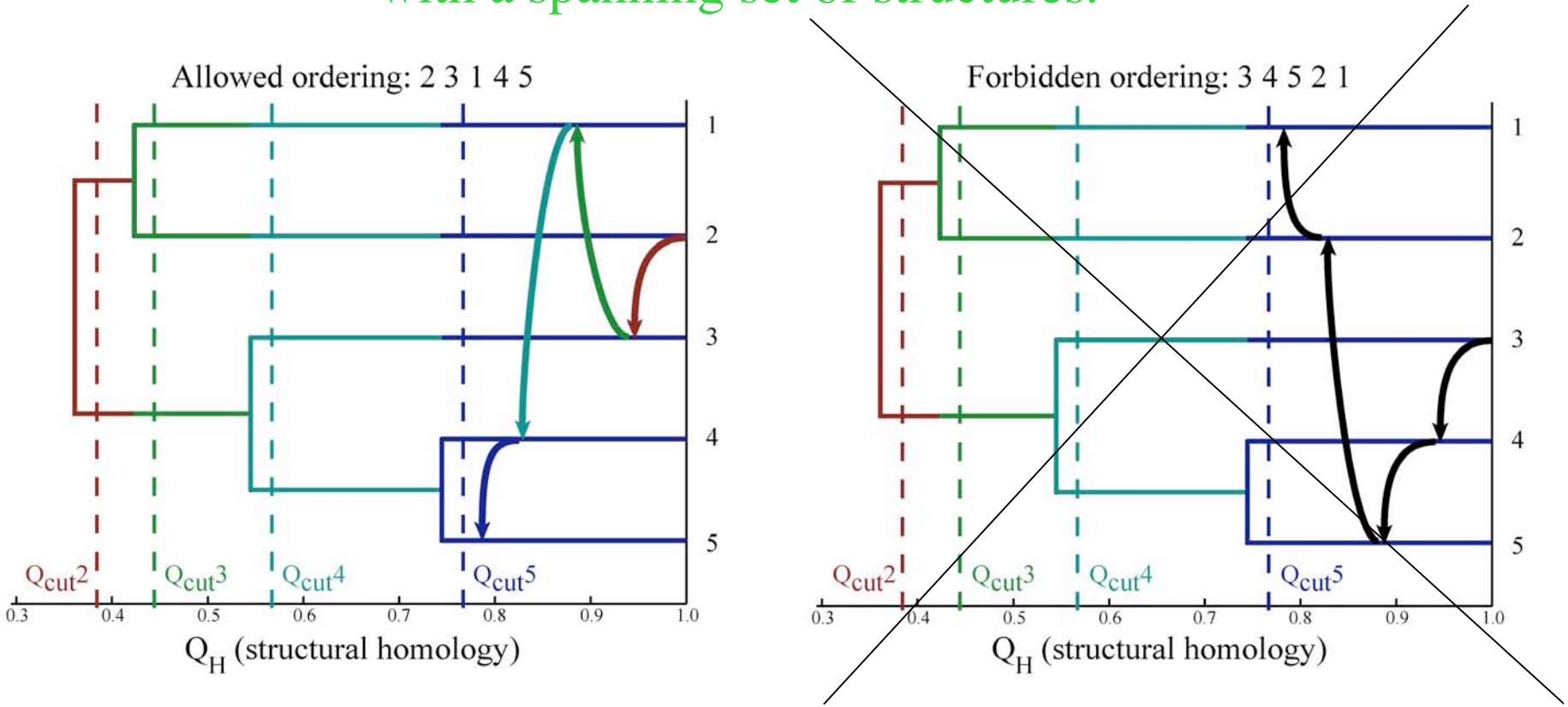
$$\|a_j\|_{F_p} = \left(\sum_{d=1}^{4} \sum_{i=k}^{m_{aln}} |a_{ijd}|^p\right)^{1/p}$$

adjustable parameter

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

# What are the constraints on the parameters?

Represent the evolutionary history of the protein group
with a spanning set of structures.

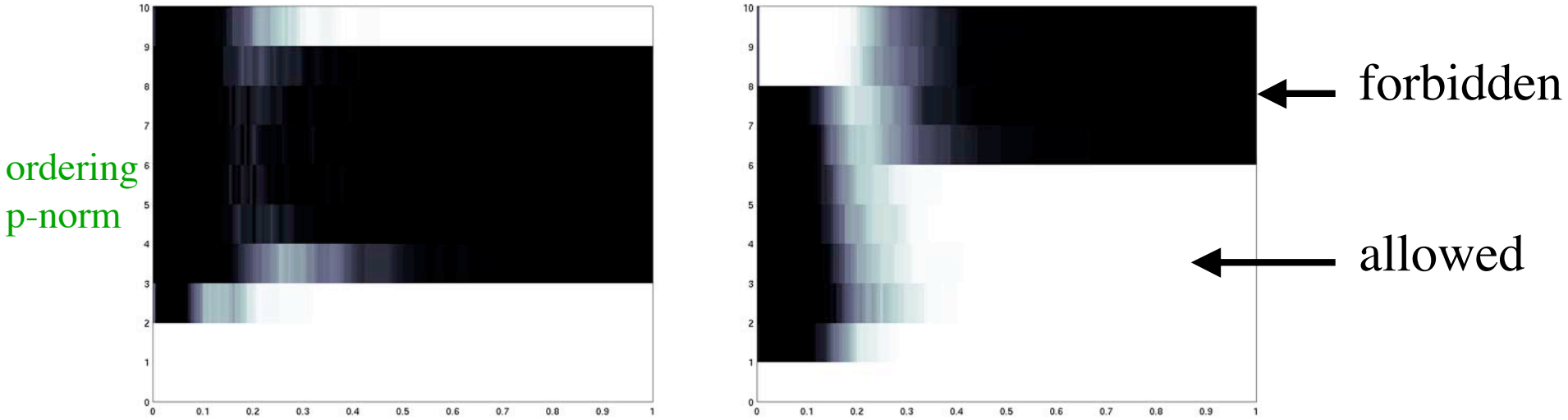



This rule is used to determine the value of
two adjustable parameters in our implementation of the QR.

# Parameters Define the Measure of Linear Dependence

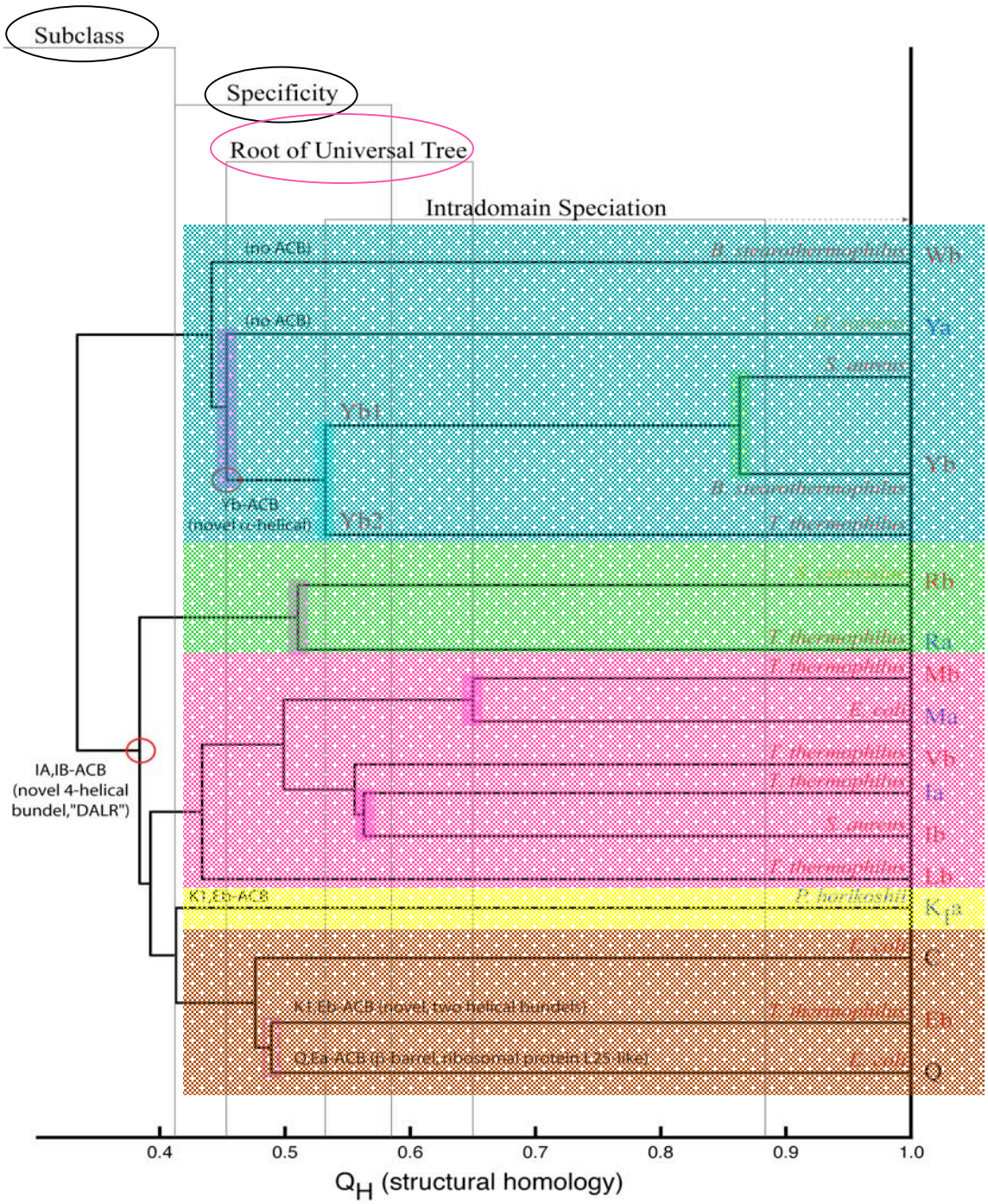
AARS class I, Rossman fold

AARS class II, Novel fold




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

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



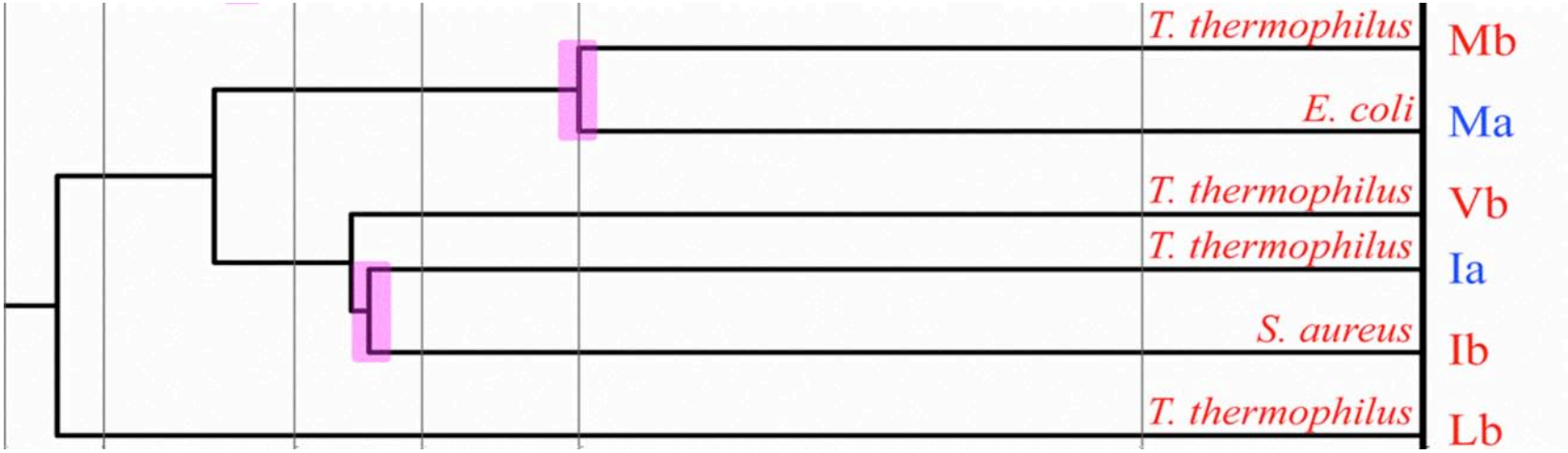

# Class I AARSs evolutionary events

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?

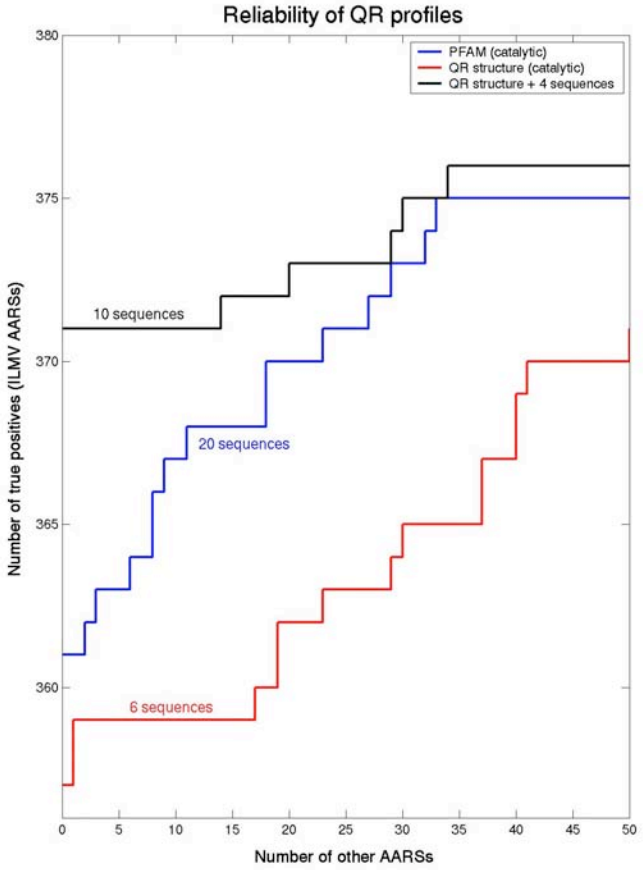
If each of ILMV was full canonical, then we would need 4x3=12 sequences.

| | 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 $G_{\alpha_2}$ $K_{II}$ N $G_{(\alpha\beta)_2}$ |

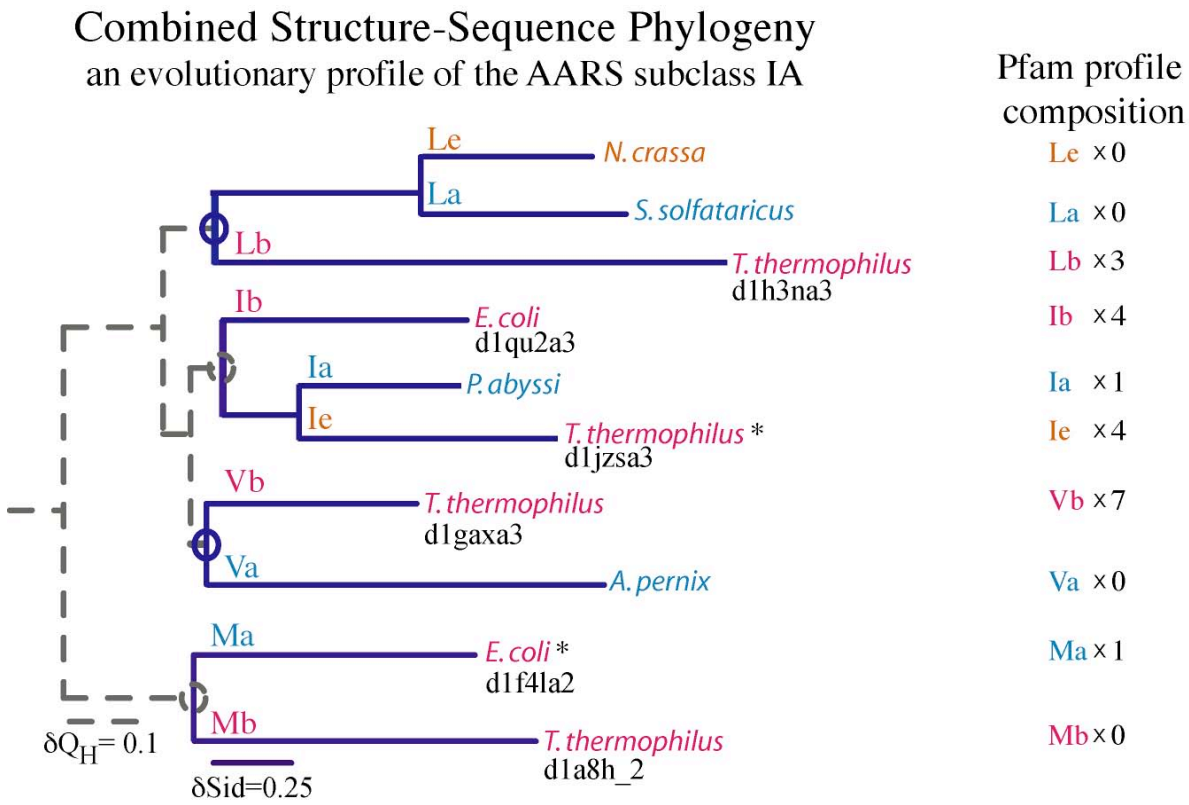
Since M and V are basal, we need **at least** 2x3 + 2x2 = 10 sequences.

We have 6 structures.

# Evolutionary Profiles for Homology Recognition
## AARS Subclass ILMV




false positives



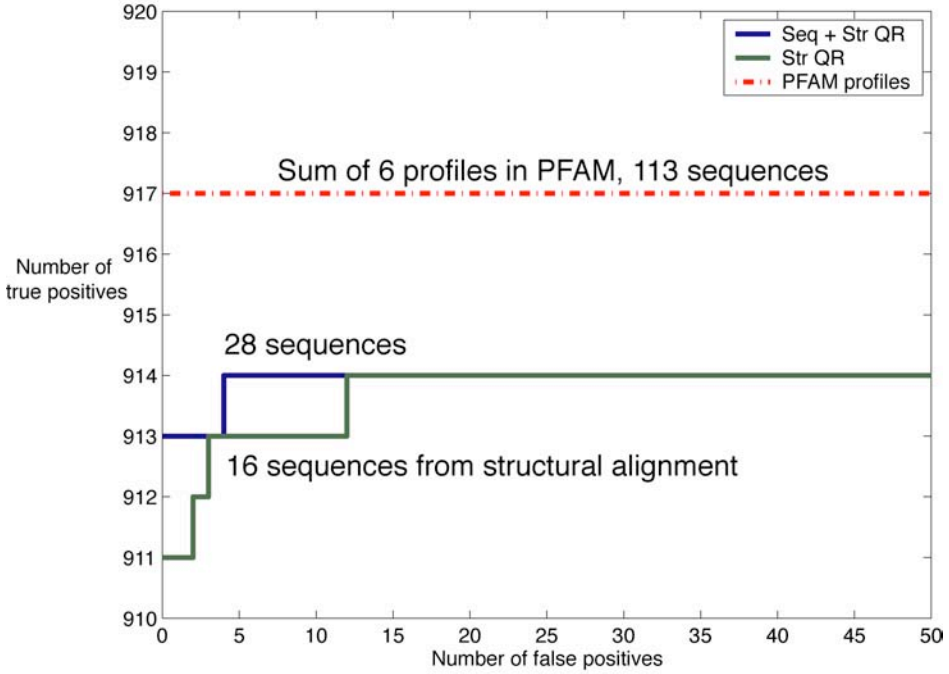

The composition of the profile matters.
Choosing the right 10 sequence makes all the difference.

**A. Sethi**, P. O'Donoghue, Z. Luthey-Schulten (2004) *Nucl. Acids Res*, submitted.

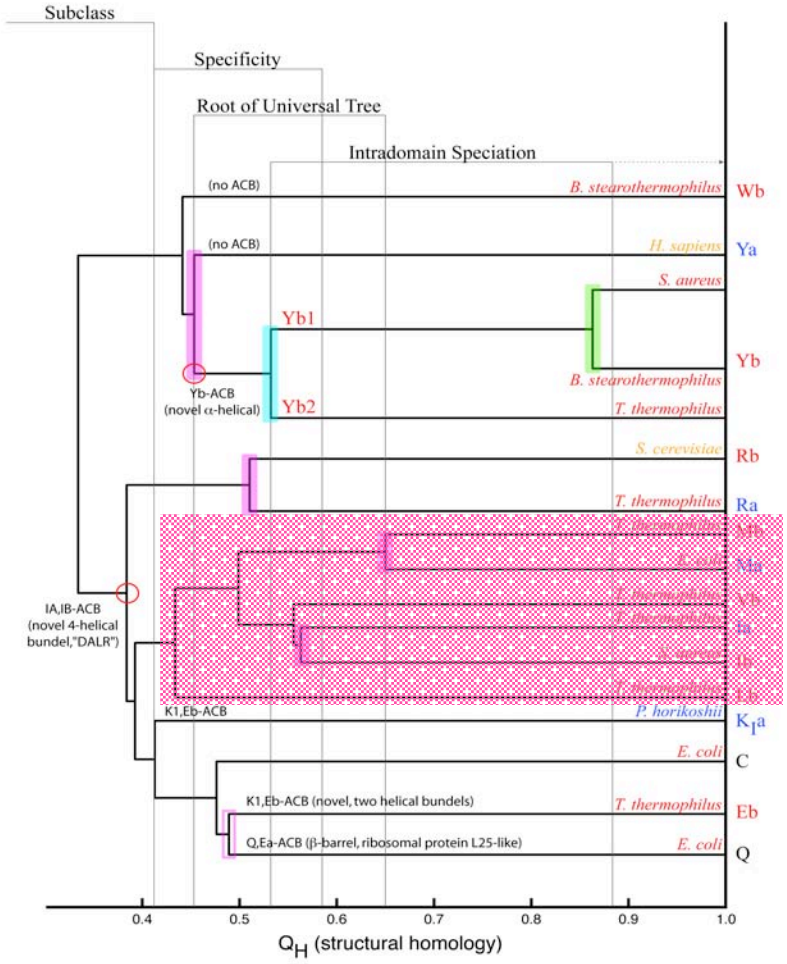
# The Economy of Information

## How many sequence are needed for profiles?

A single profile
for class I AARSs



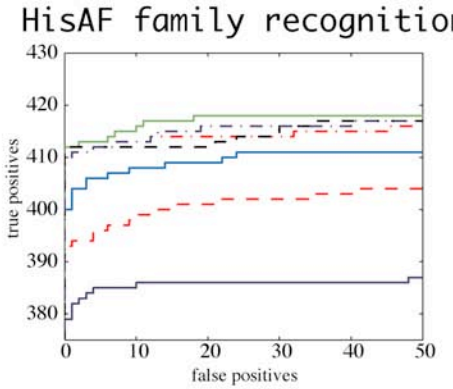

PFAM profile of 113 sequences finds 3 additional sequence fragments compared to the non-redundant profile of 28 sequences.

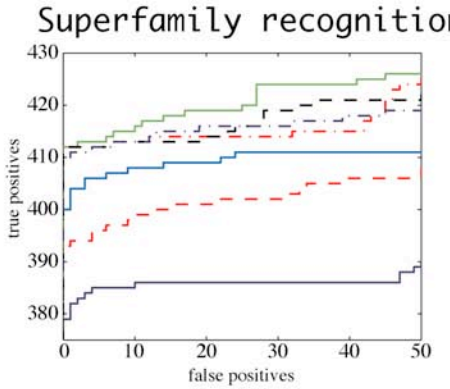


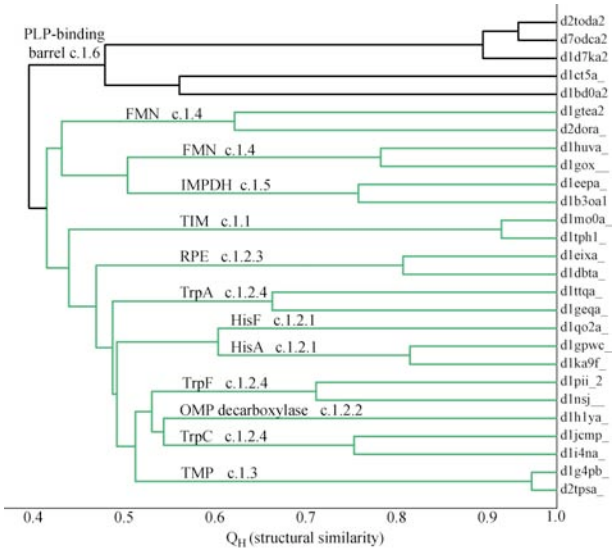
# Economy of Information

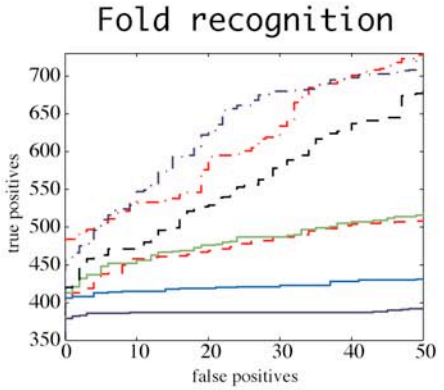
## How many sequences are needed for profiles?

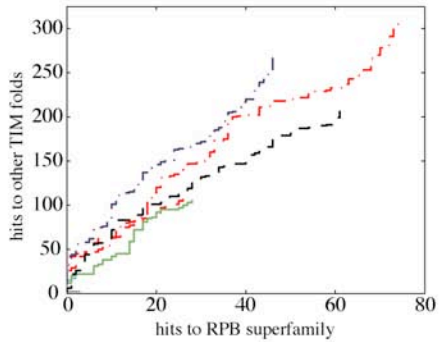
| Profile | $N_{seq}$ |
|---|---|
| Pfam full | 250 |
| Pfam seed | 147 |
| Pfam/QR 40% | 21 |
| QR 15% | 2 |
| QR 30% | 4 |
| QR 40% | 13 |
| QR 100% | 238 |



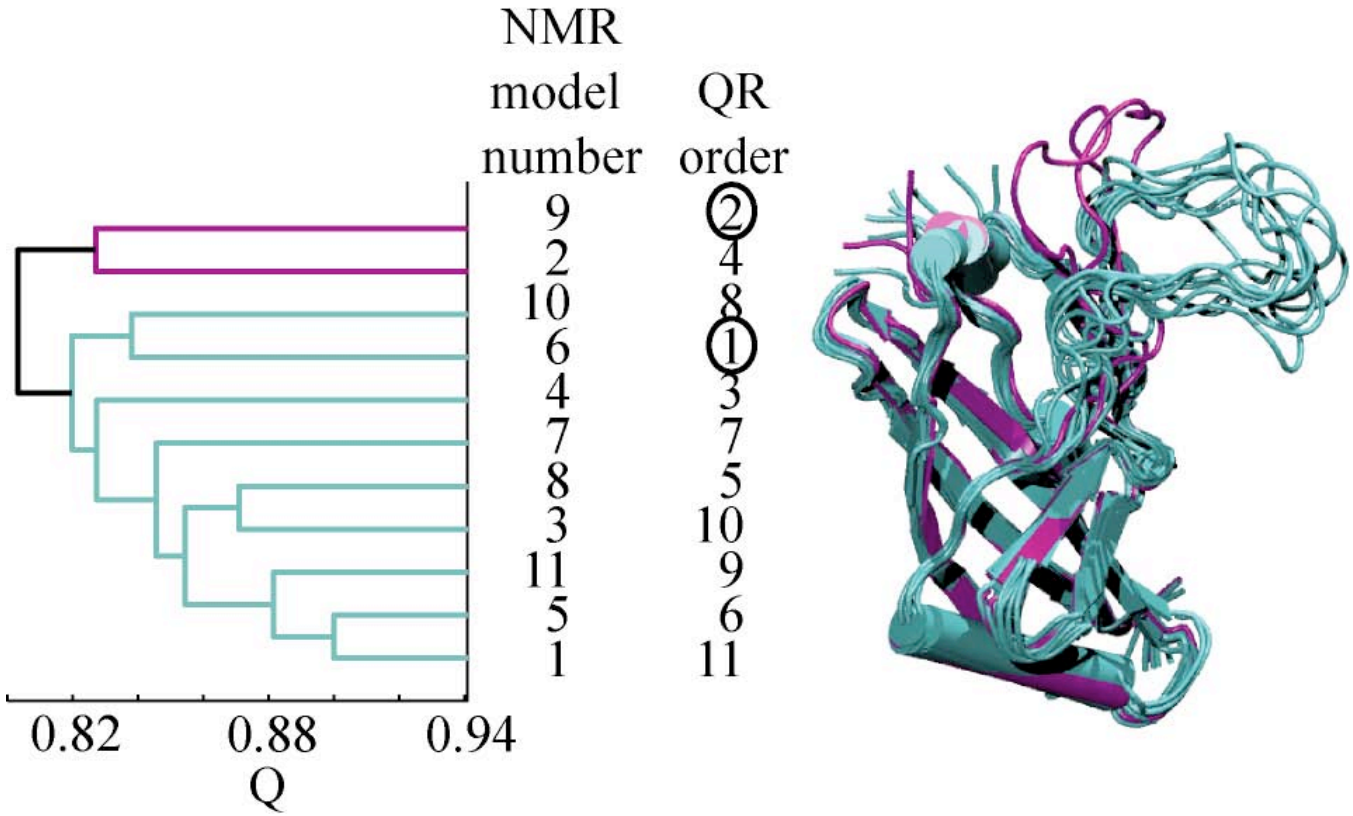



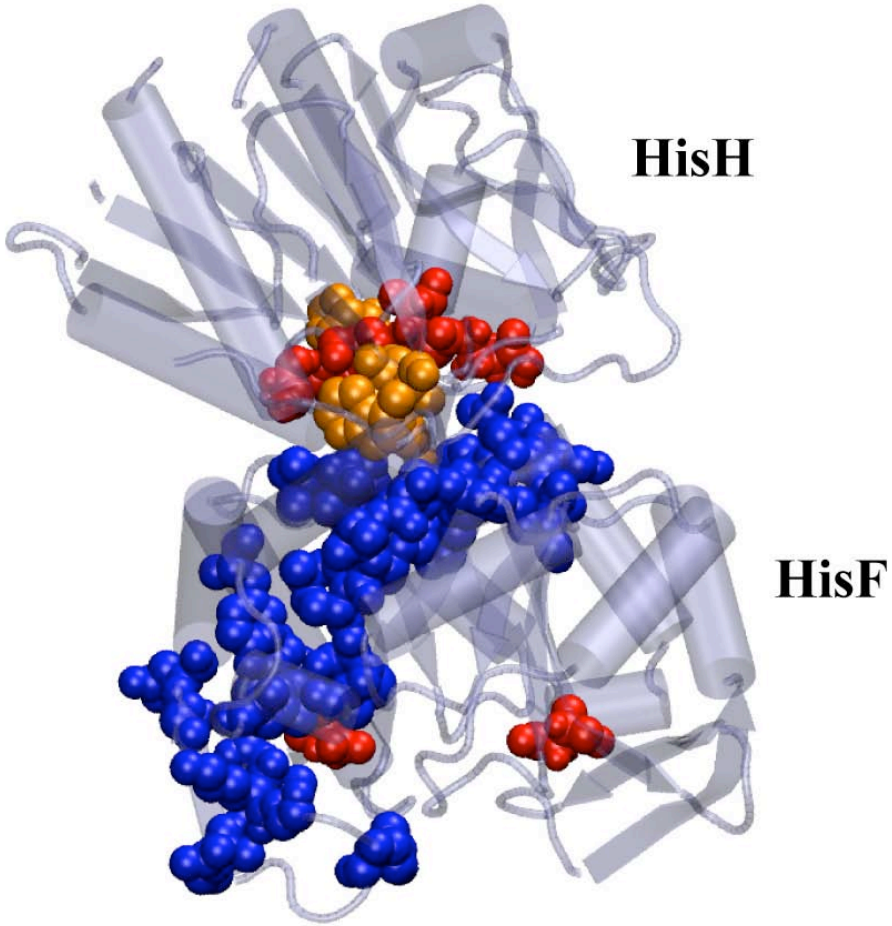

If the sequences well represent the
evolutionary history of the protein family,
a factor of 10 to 100 less information is required.

# QR factorization of an ensemble of NMR structures



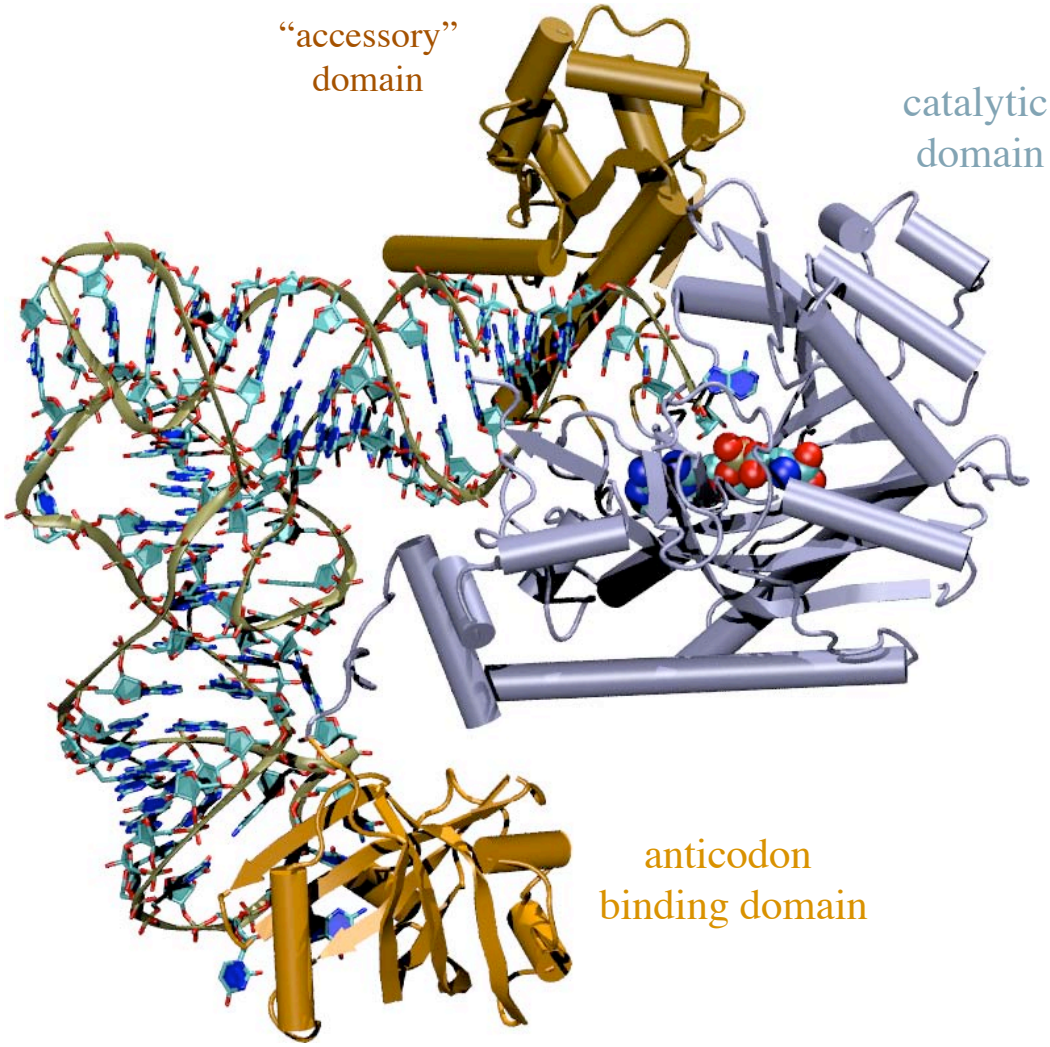

The QR algorithm can be applied to conformational, evolutionary ensembles or both simultaneously.

# Evolutionary Structure/Sequence Profiles Suggest Reaction Pathway



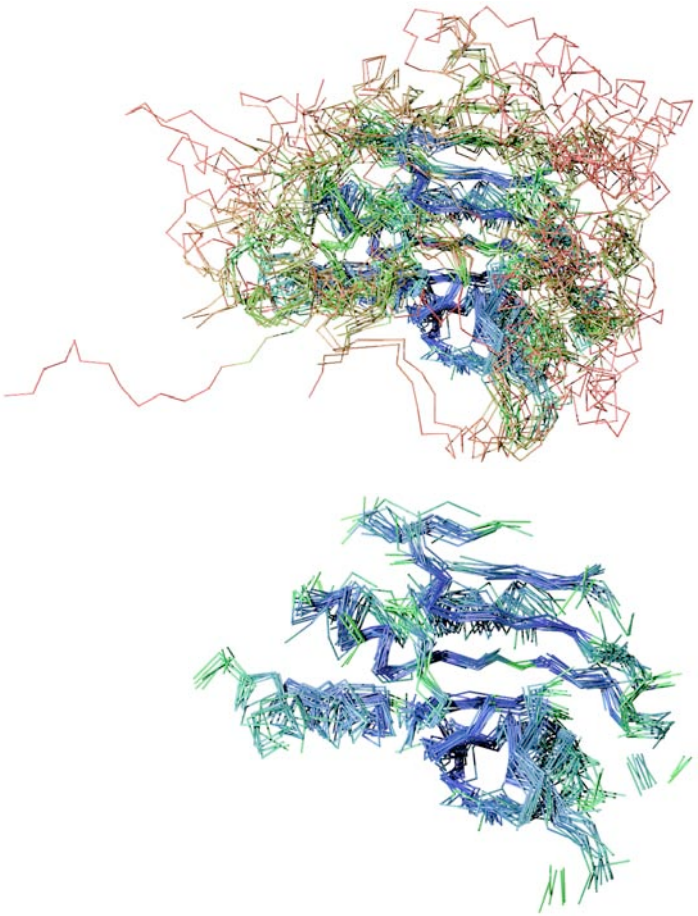

R. Amaro and Z. Schulten, *MD Simulations of Substrate Channeling*, Chemical Physics Special Issue, 2004 (in press). *FE Landscapes of Ammonia Channeling*, PNAS 2003

# Domain Structure in AspRS

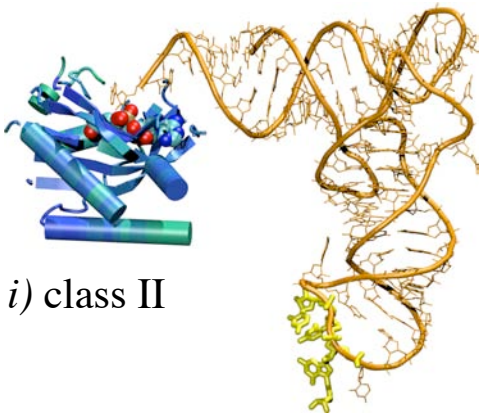



bacterial type aspartyl-tRNA synthetase
*E. coli*, homodimer

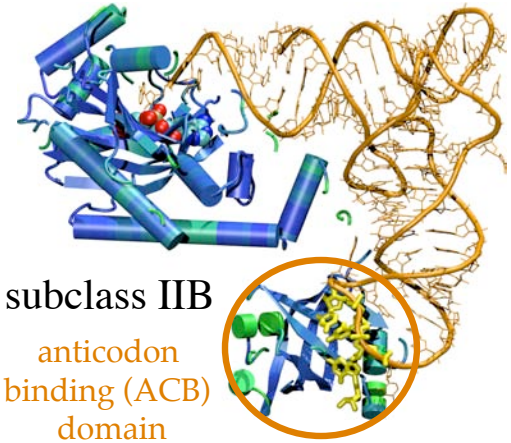
catalytic domains
class II AARSs



# Evolution of Structure and Function in AspRS

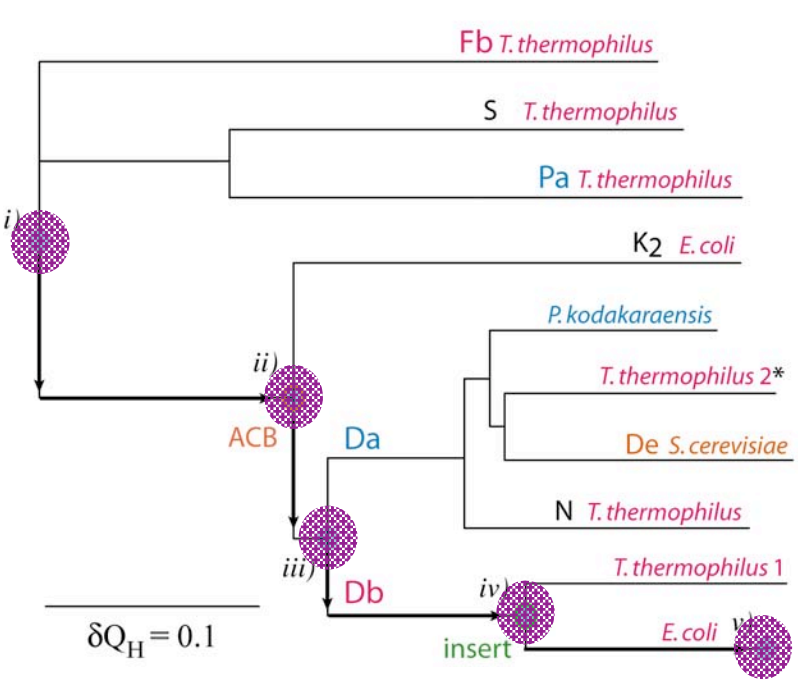


*i)* class II

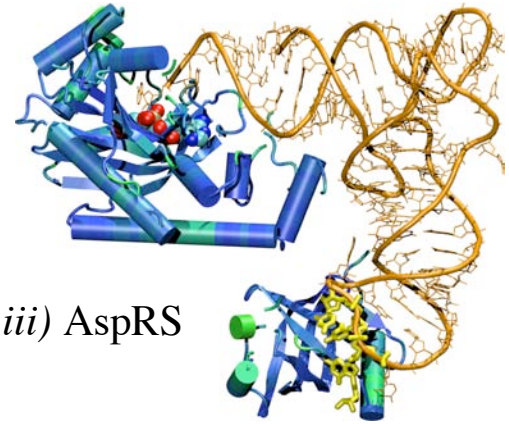


*ii)* subclass IIB

anticodon binding (ACB) domain

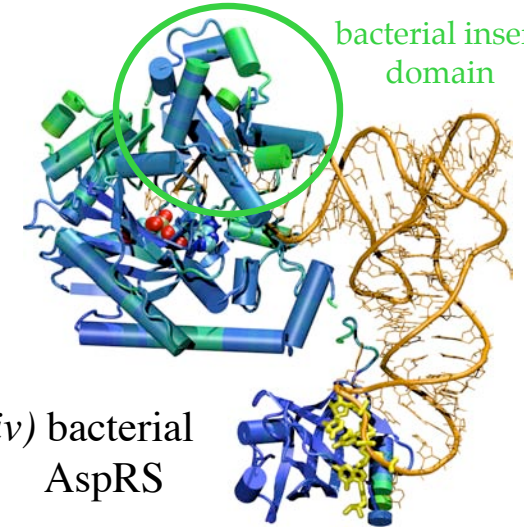



| SCOP | QR order |
|---|---|
| d1b70a_ | (1) |
| d1serb2 | (3) |
| d1h4sb2 | 6 |
| d1bbua2 | 4 (2) |
| d1b8ab2 | 9 5 4 |
| d1n9wb2 | 10 7 6 |
| d1asza2 | 5 (3) 3 |
| d11sca2 | 7 4 (2) |
| d1efwa3 | 8 6 5 |
| d1c0aa3 | (2) (1) (1) |

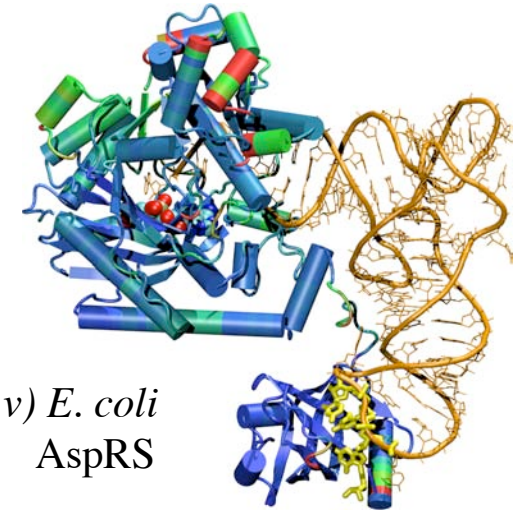


*iii)* AspRS

bacterial insert domain



*iv)* bacterial AspRS



*v) E. coli* AspRS

# Summary

## Evolutionary information is encoded in protein structure.

Protein structure allows investigation of evolutionary events that pre-date the origin of species.

Accounting for gaps is critical for comparing homologous structures.

Sequence and structure can be combined to give a unified phylogenetic framework.

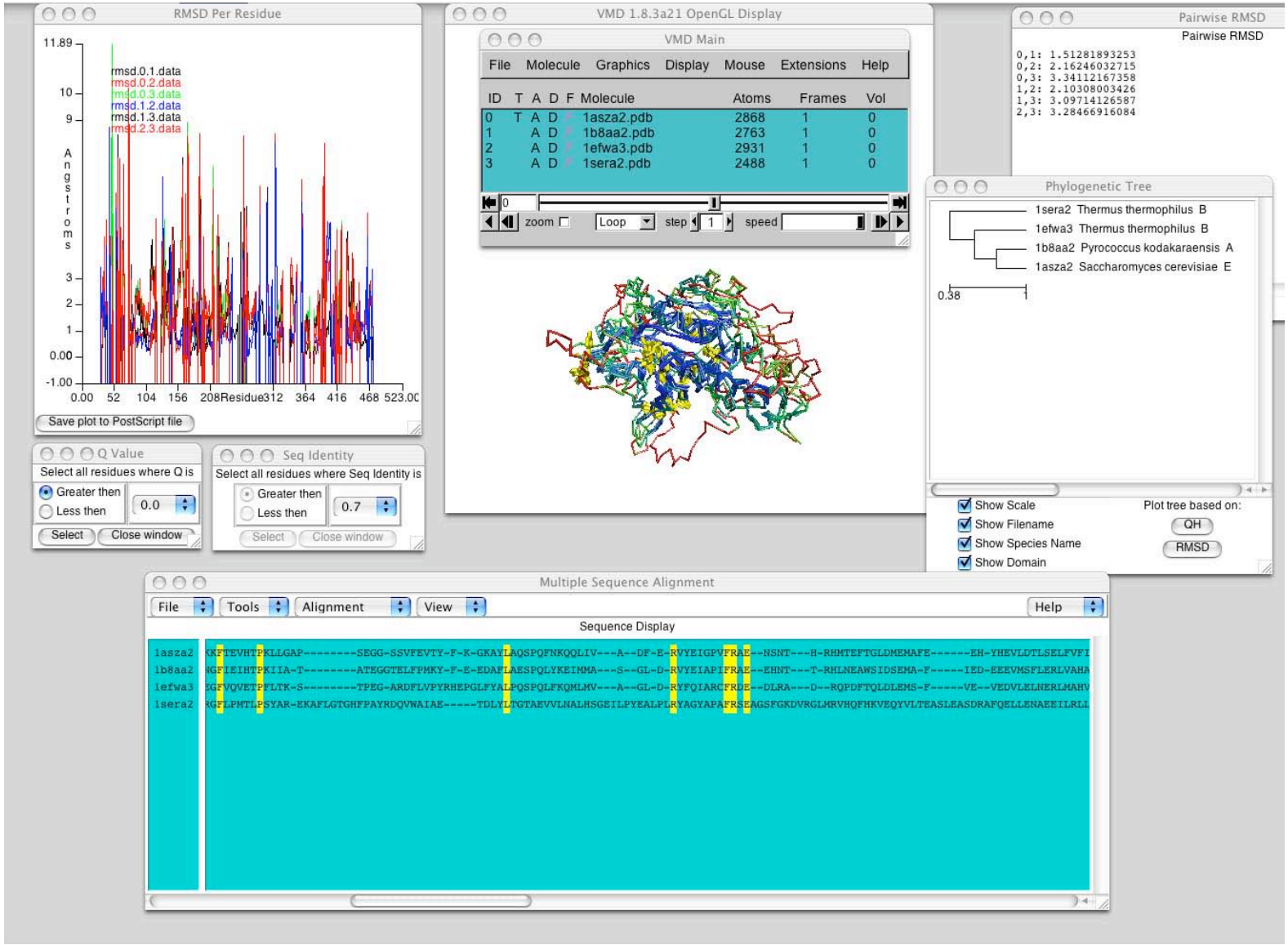
## The QR factorization provides evolutionary profiles (EPs).

By spanning the evolutionary space with a small number of representative sequences
EPs outperform traditional profiles.

> 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.

The QR algorithm can be applied to conformational, evolutionary ensembles
or both simultaneously.

# Multiseq in VMD: Merging the sequence and structure worlds




Brijeet Dhaliwal, John Eargle, John Stone, Dan Wright

# 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