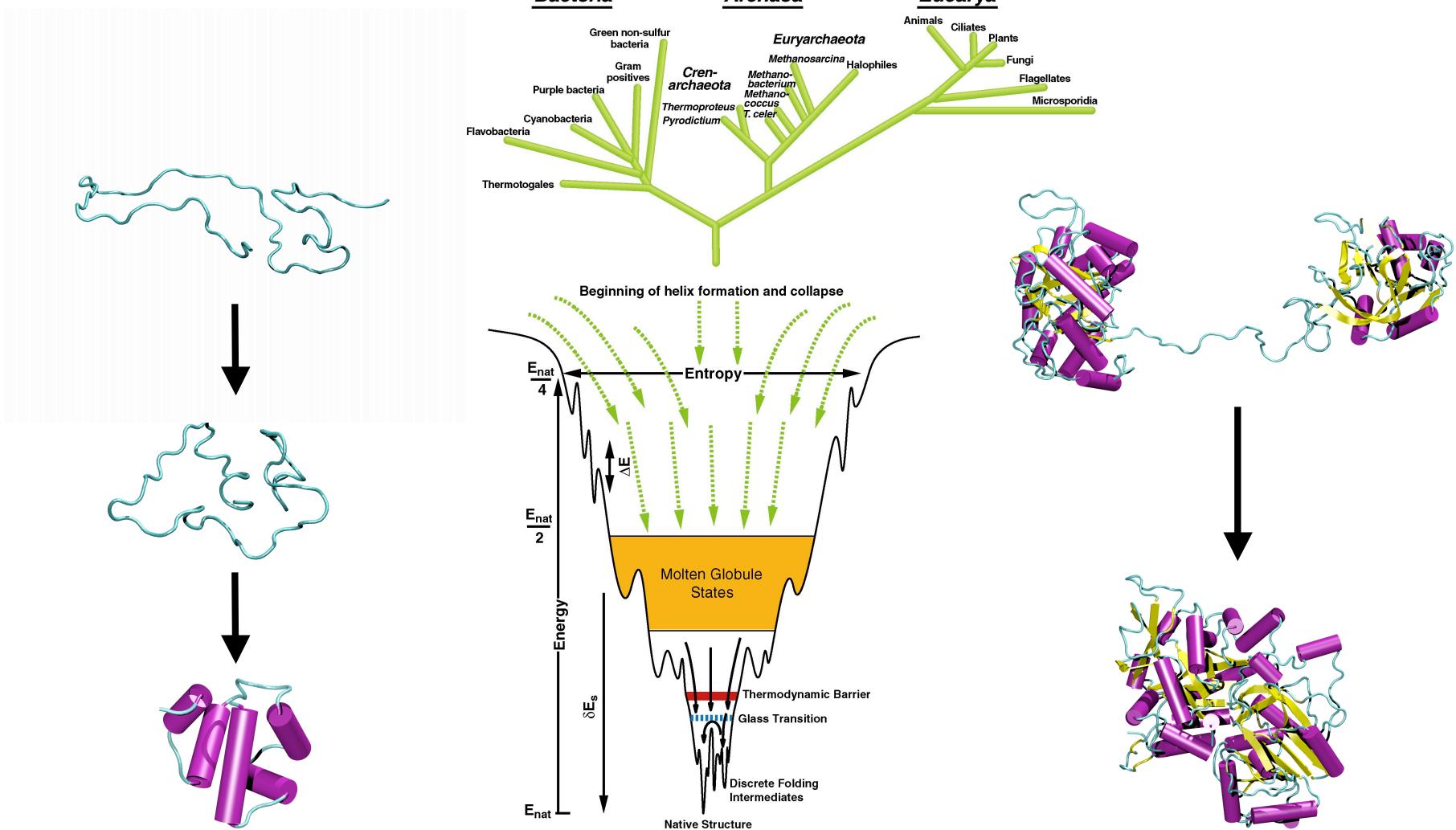


# Bioinformatics – NSF Summer School 2003

## Z. Luthey-Schulten, UIUC



# Sequence-Sequence Alignment

- Smith-Watermann
- Needleman-Wunsch

# Sequence-Structure Alignment

- Threading
- Hidden Markov

# Sequence Alignment & Dynamic Programming

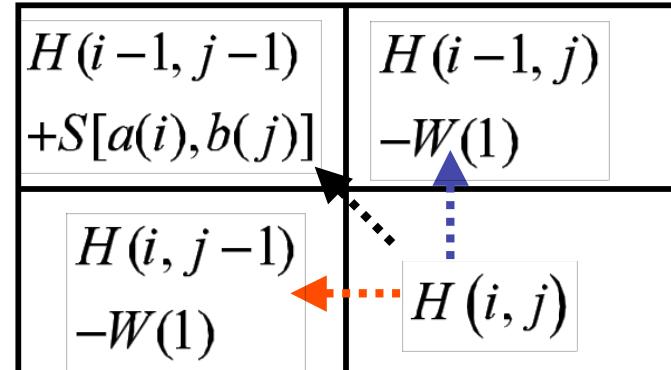
Seq. 1:  $a_1 a_2 a_3 \dots a_4 a_5 \dots a_n$   
 Seq. 2:  $c_1 \dots c_2 c_3 c_4 c_5 \dots c_m$

number of possible alignments:

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

## Smith-Waterman alignment algorithm

$$H(i, j) = \text{MAX} \begin{cases} H(i-1, j-1) + S[a(i), b(j)] \\ H(i, j-k) - W(k), \\ H(i-m, j) - W(m), 0 \end{cases}$$



$S$ : similarity matrix

## Score Matrix H· Traceback

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

AWGHE  
AW--HE

# Smith-Waterman Local Alignment Score Matrix

	H	E	A	G	A	W	G	H	E	E
0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0
A	0	0	0	5	0	5	0	1	0	0
W	0	0	0	3	0	24	16	8	0	0
H	0	13	5	0	0	1	16	22	29	21
E	0	5	20	12	4	0	8	14	22	36
A	0	0	12	25	17	9	1	9	14	28
E	0	0	7	17	22	15	7	1	9	21
				AWGHE						
				AW--HE						

# Blosum 40 Substitution Matrix

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

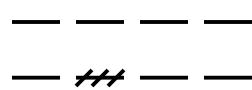
# Protein Structural Relationships

Can protein structural relationships help us to understand evolutionary dynamics?

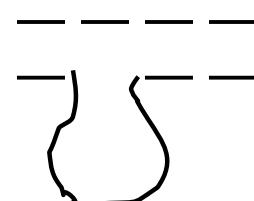
Is there a connection between evolutionary events and changes in protein structure?

What is the effect of gene duplication, horizontal gene transfer, and other evolutionary mechanisms on protein shape?

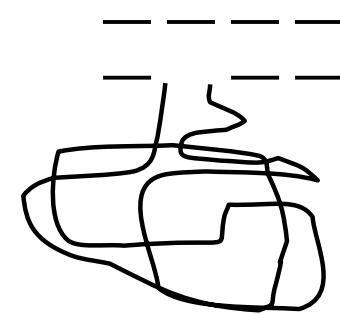
Substitution



Indel



Domain Insertion



# Sequence Alignment & Dynamic Programming

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



number of possible alignments:

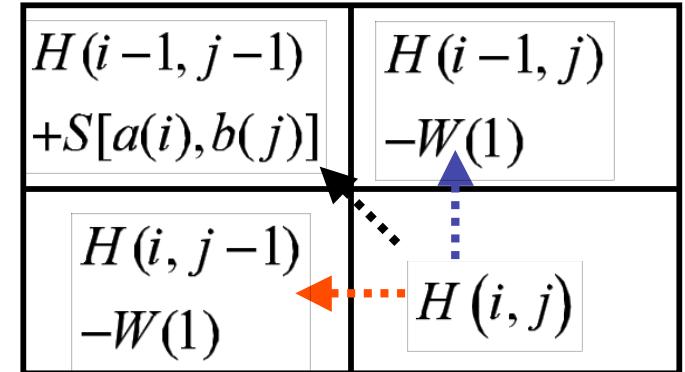
$$= \binom{2n}{n} = 2^{2n} (\sqrt{n\pi})^{-1}$$

## Needleman-Wunsch alignment algorithm

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

S: similarity matrix

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



Score Matrix H: Traceback

??? Tutorial: W=d

# Needleman-Wunsch Global Alignment

Similarity Values

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

Initialization of Gap Penalties

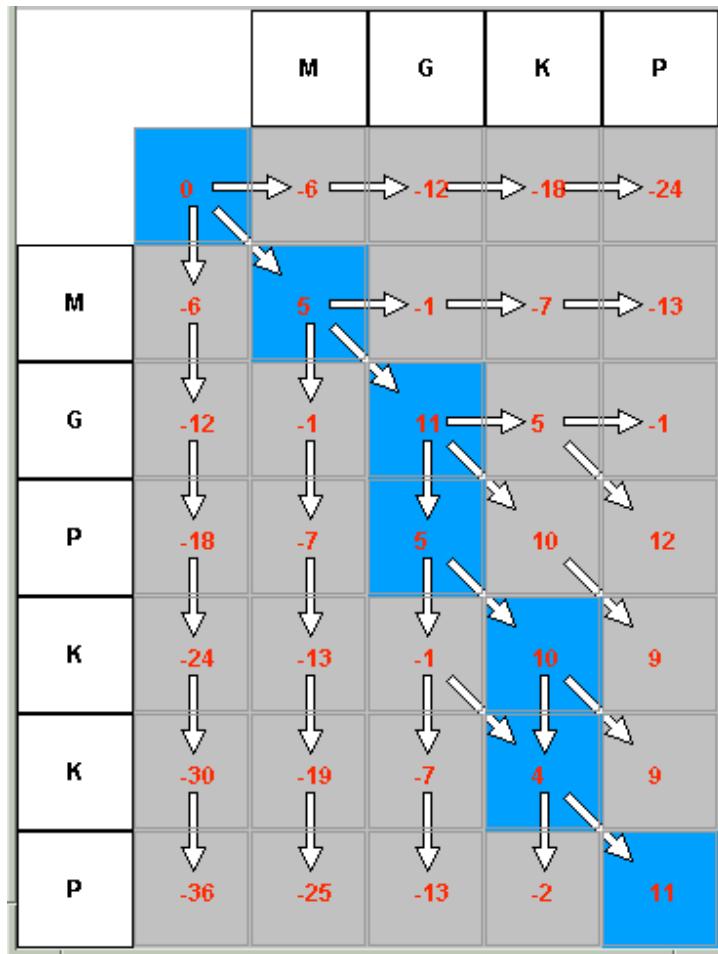
	M	G	K	P	
M	0	-6	-12	-18	-24
G	-6	5	-3	-1	-2
P	-12	-3	6	-2	-2
K	-18	-2	-2	-1	7
K	-24	-1	-2	5	-1
P	-30	-1	-2	5	-1
P	-36	-2	-2	-1	7

# Filling out the Score Matrix H

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

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

# Traceback and Alignment

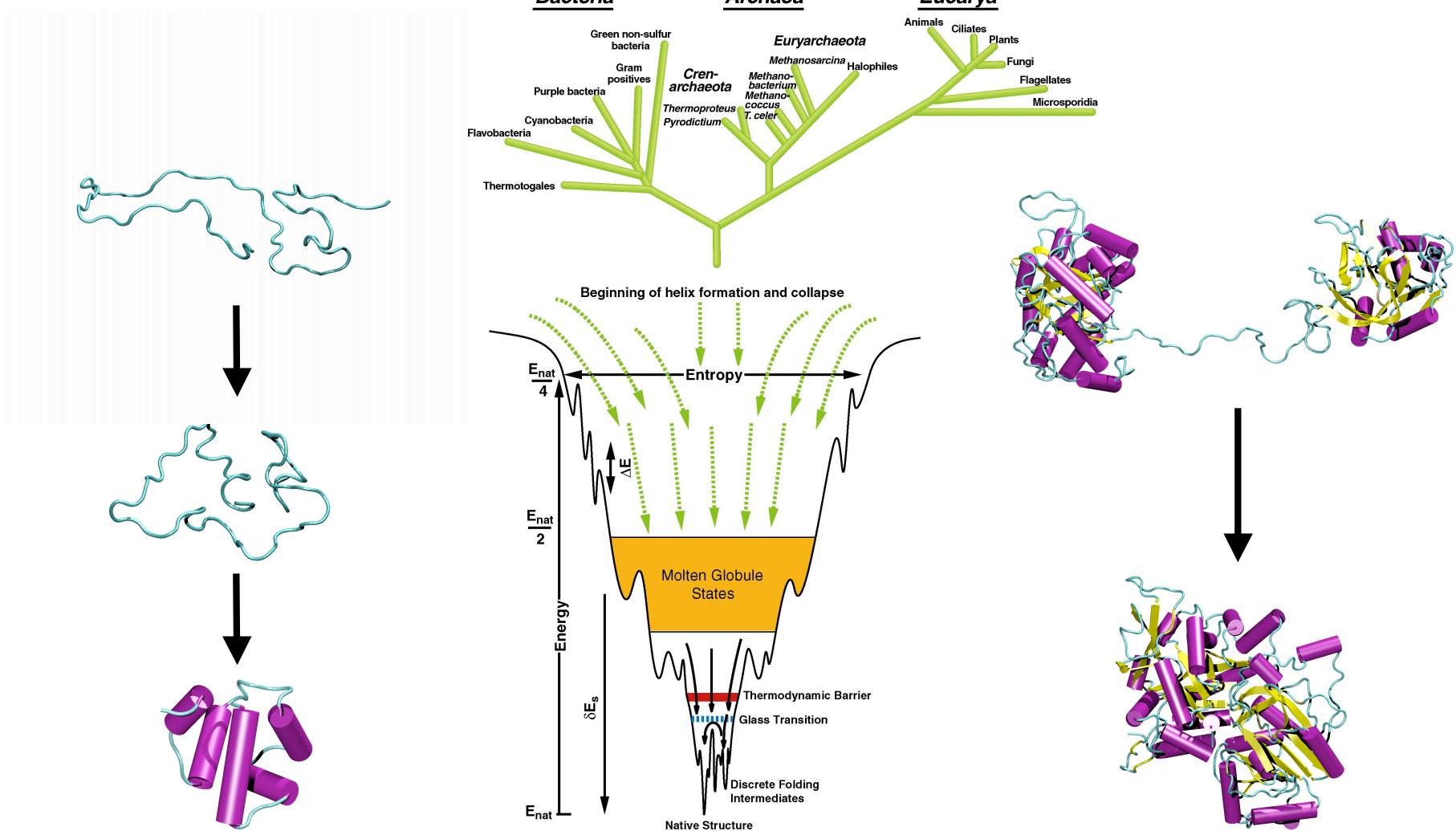


The Alignment

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

Traceback (blue) from optimal score

# Energy Landscape Theory of Structure Prediction



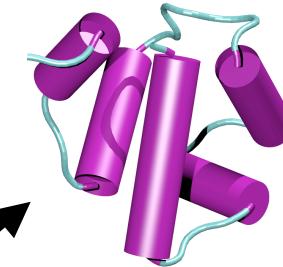
# Protein Structure Prediction

1-D protein sequence

SISSIRVKSKRIQLG....

*Ab Initio* protein folding

3-D protein structure



## Sequence Alignment

Target protein of unknown structure → SISSRVKSKRIQLGLNQAE LAQKV-----GTTQ...

Homologous/analogous protein of known structure → QFANEFKVRRRIKLGYTQTNVGEALAAVHGS...

## Sequence Alignment: the Energy Function

$$E = E_{\text{match}} + E_{\text{gap}} \rightarrow E_{\text{gap}} = ?$$

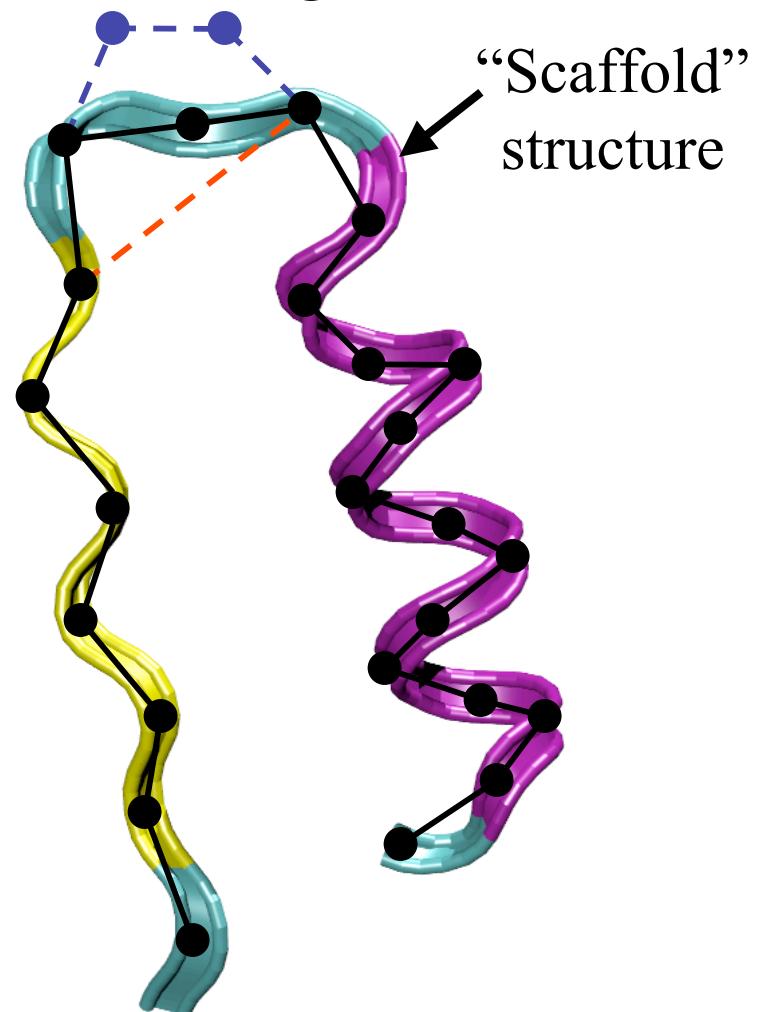
$$E_{\text{match}} =$$

# Threading: Sequence-Structure Alignment

Target sequence

A<sub>1</sub> A<sub>2</sub> A<sub>3</sub> A<sub>4</sub> A<sub>5</sub> ...

threading alignment  
between target and scaffold



## Threading Energy Function

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

$$E_{profile} = \sum_i^n \gamma^{(p)} (A_i, SS_i, SA_i)$$

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

$$E_{gap}(r, l) = \gamma_g \log(P_g)$$

# Gap Penalties

$$E_{gap} = kT \log(P_g)$$

Distribution  
of Gaps

## Sequence-Structure Gap Energy

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

$$P_{insertion}(l) = a_1 \exp(-b_1 l)$$

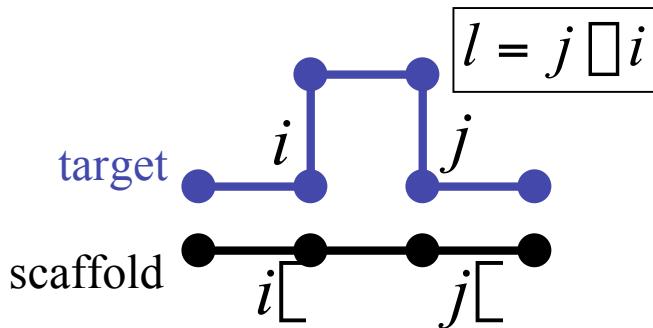
$$P_{deletion}(r) = a_2 \exp\left(-\frac{(r - b_2)^2}{2c_2^2}\right)$$

range  $3.0 \text{ \AA} < r < 7.5 \text{ \AA}$

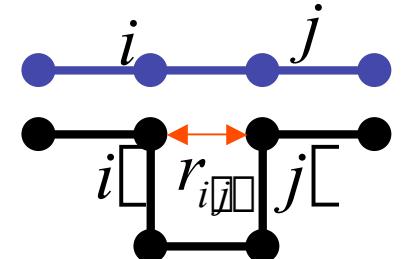
$$P_{bulge}(l, r) = \frac{a_3}{(b_3 l)^{3/2}} \exp\left(-\frac{r^2}{b_3 l}\right)$$

range  $r > 4.0 \text{ \AA}$

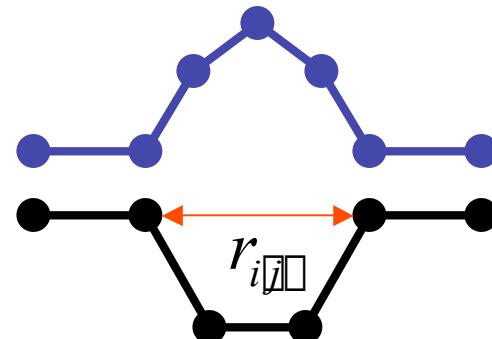
## Insertion



## Deletion



## Bulge



# Similarity Measures

## Sequence Identity

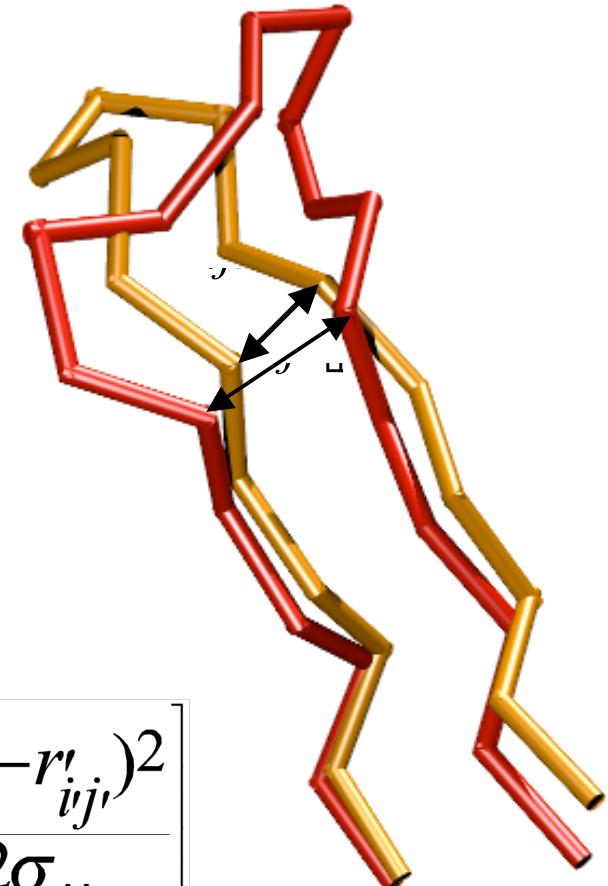
fraction of identically matched residues

$$S = \frac{N_{match}}{N_{sequence\ length}}$$

## Q “Structural Identity”

fraction of native contacts

$$Q = \frac{2}{(N_{ALN}-1)(N_{ALN}-2)} \sum_{i < j-1} \exp \left[ \frac{(r_{ij} - r'_{ij'})^2}{2\sigma_{ij}} \right]$$

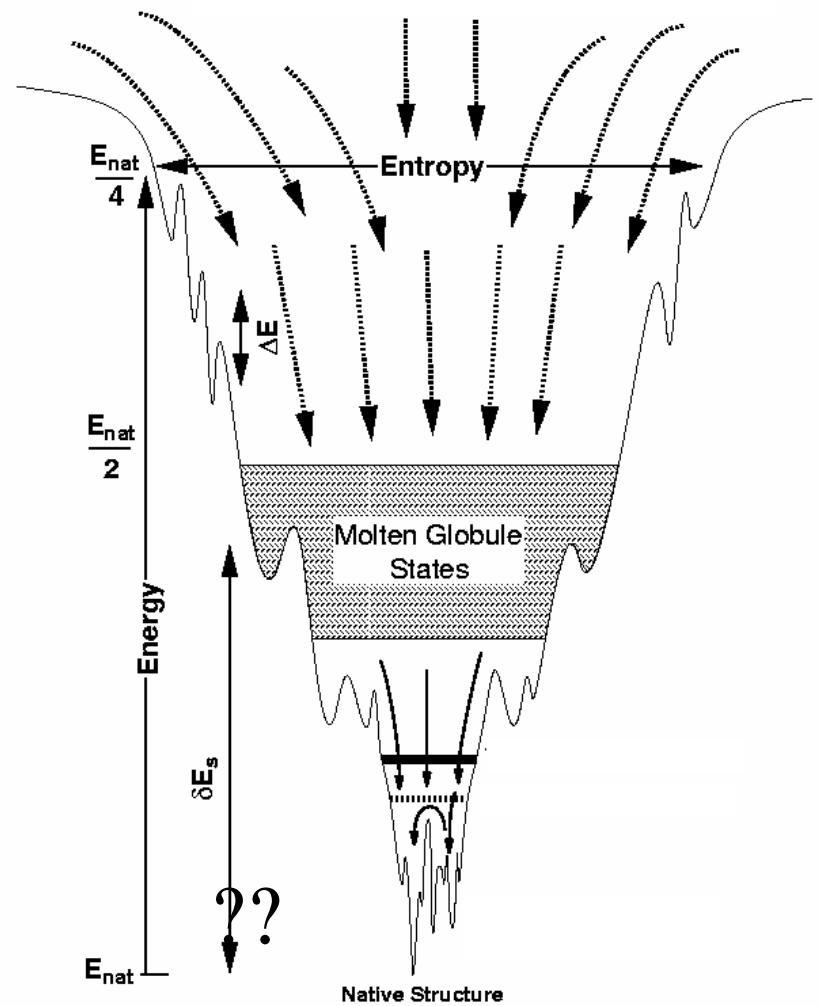
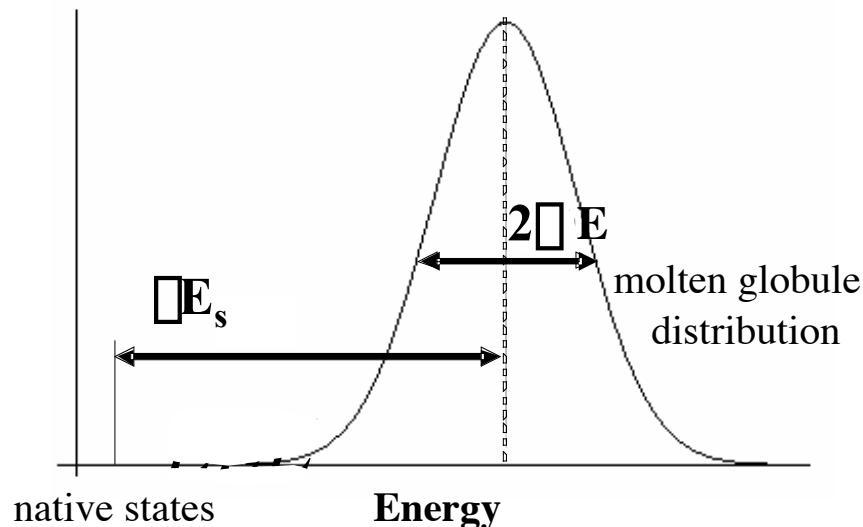


# A summary of Energy Landscape Theory

## Energy Landscape Theory

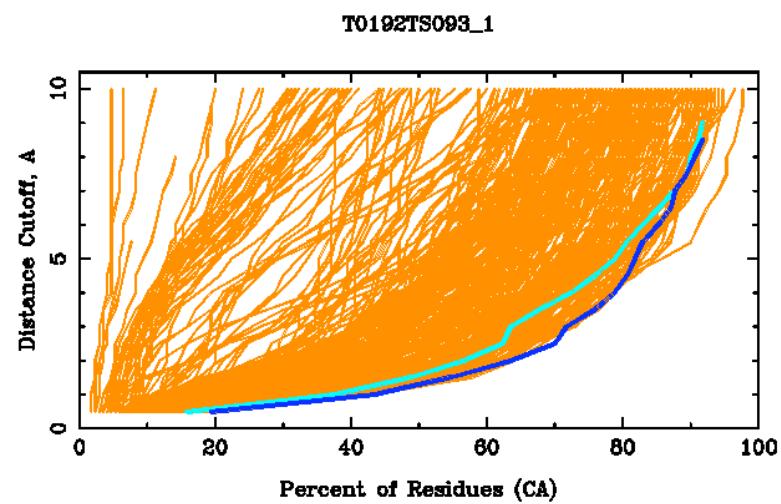
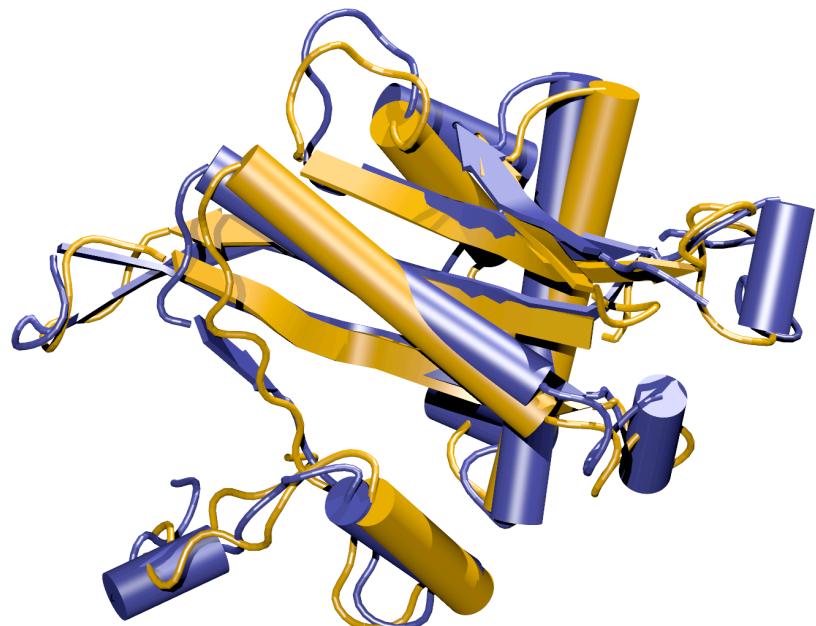
When  $\langle \frac{\partial E_s}{\partial E} \rangle$  is maximum  
the energy landscape is **optimally funneled**.

## Optimization over an Ensemble of Folds



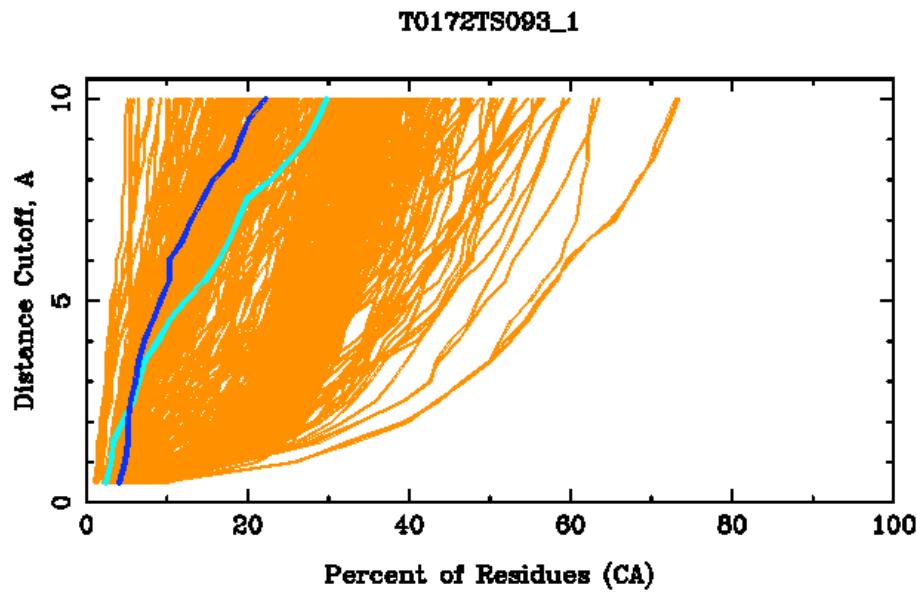
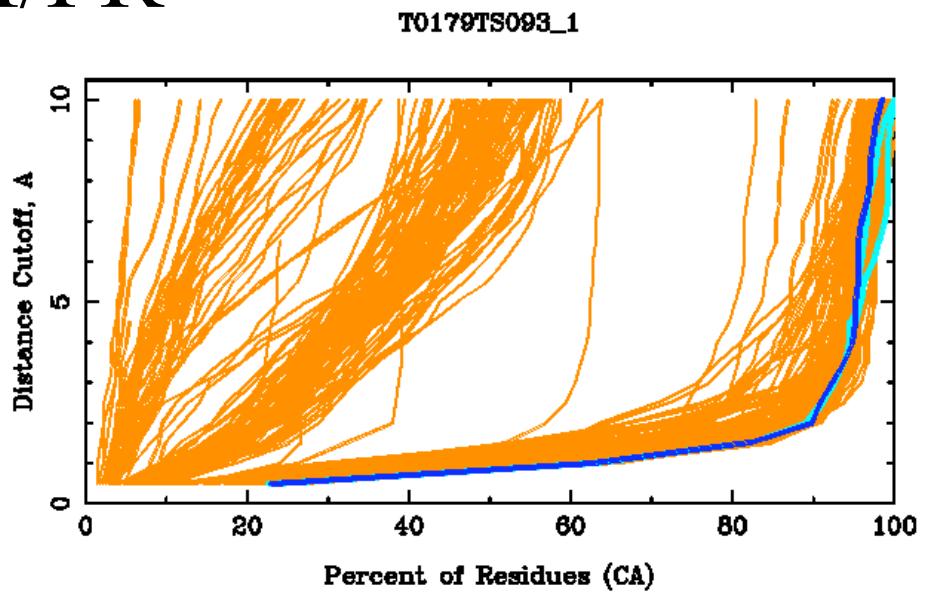
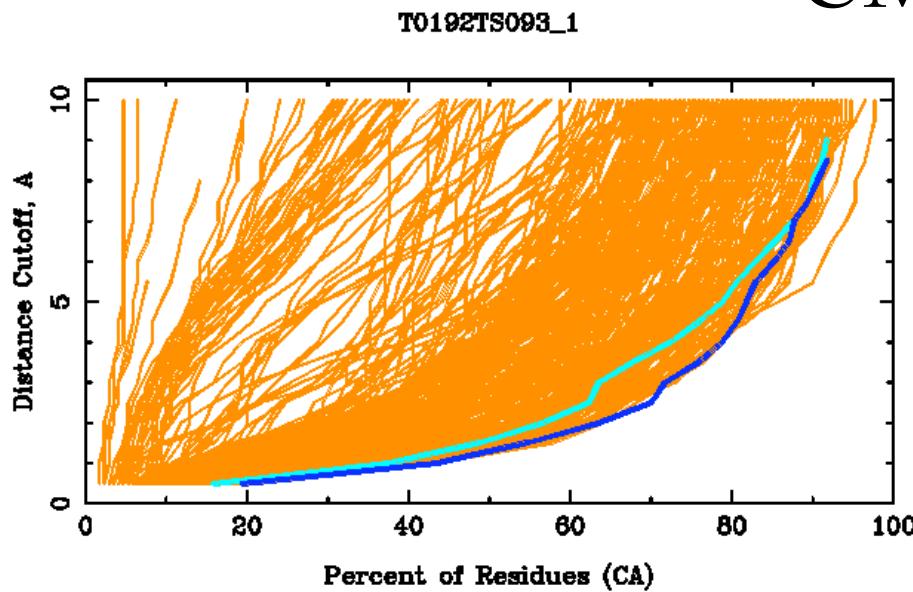
Onuchic , Luthey -Schulten, Wolynes (1997) *Annu . Rev. Phys. Chem.* 48:545 -600.  
Koretke , Luthey -schulten,Wolynes( 1996) *Prot. Sci.* 5:1043

# Homology Modeling - Threading



# Results from CASP5

## CM/FR



The prediction is never better than the scaffold.

Threading Energy function requires improvement.

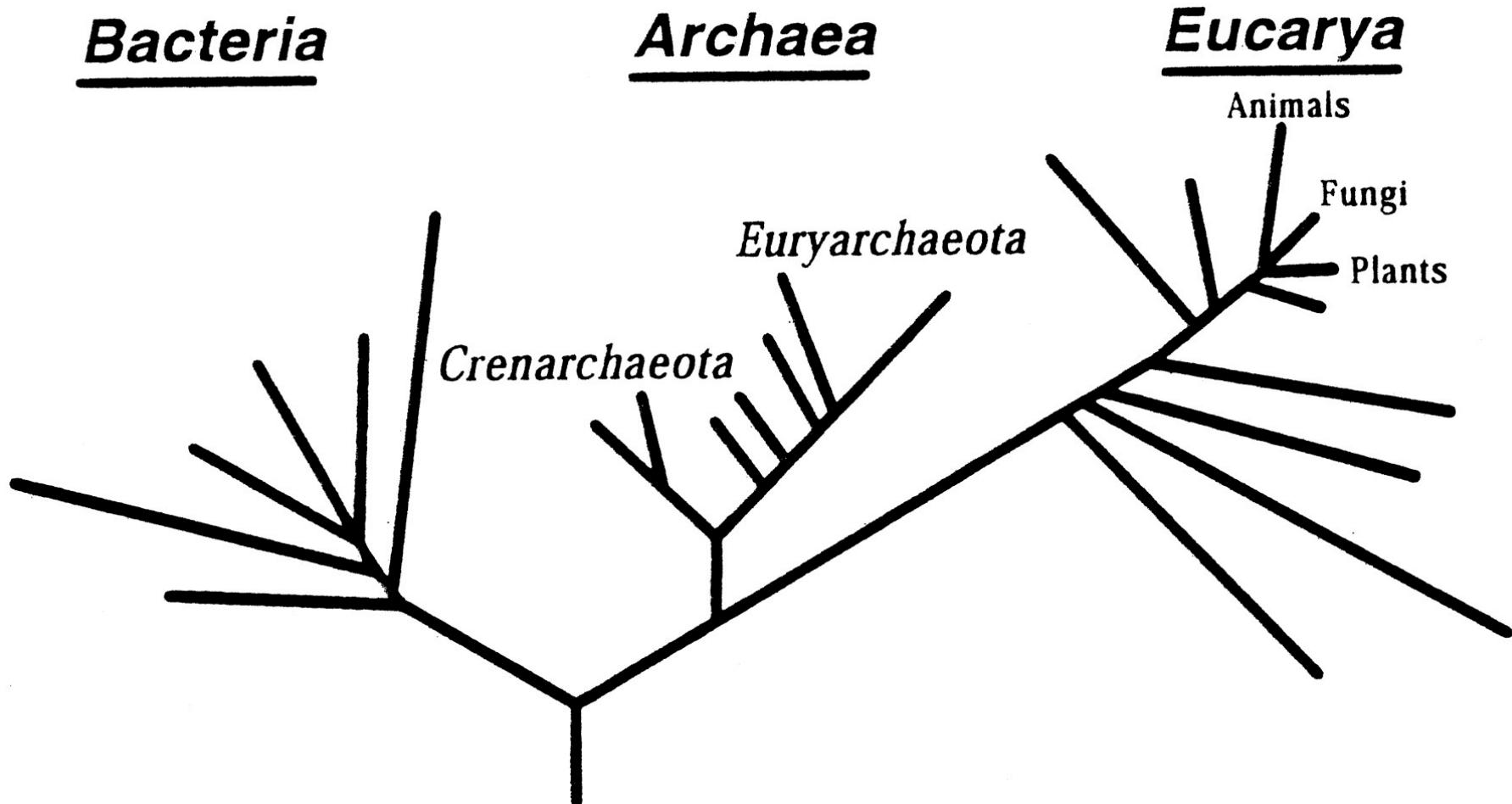
You are now entering the twilight zone  
of sequence identity. We need  
profiles!

Watch for Bioinformants!!!

# Profiles – Evolution Revisited

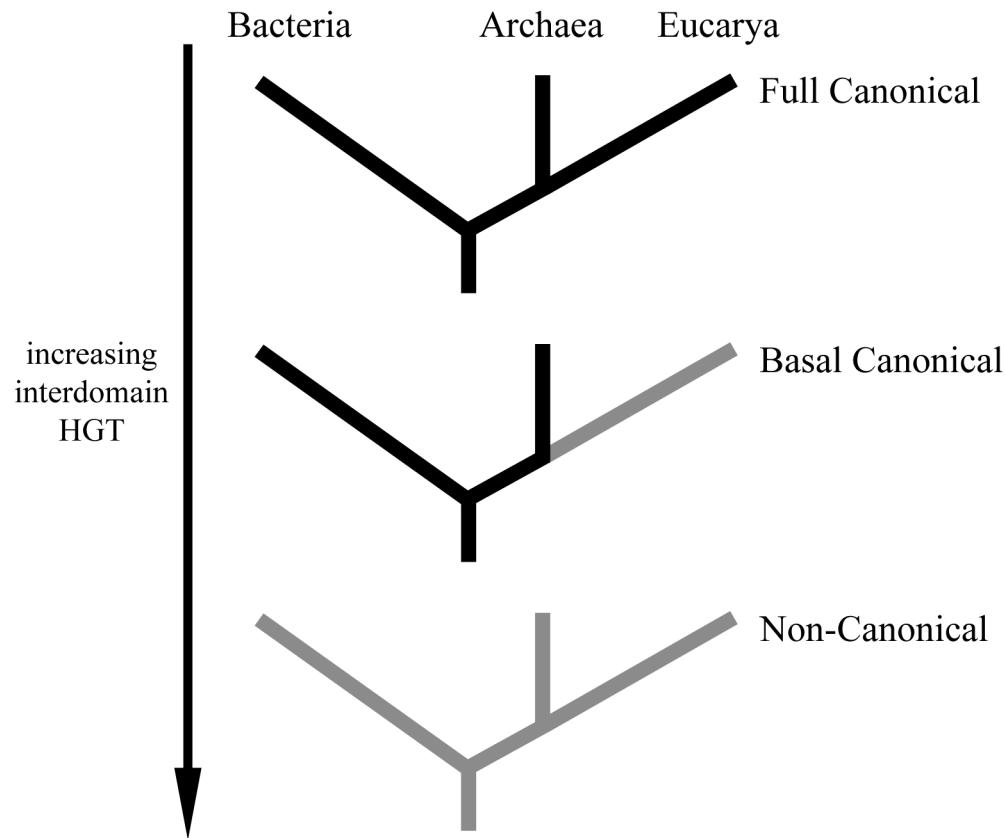
- “What molecular sequences taught us in the 1960’s was that the genealogical history of an organism is written to one extent or another into the sequences of each of its genes, an insight that became the central tenet of a new discipline, molecular evolution”
- Woese (PNAS, 2000) Pauling (1965)

# Universal Tree



The Universal Phylogenetic Tree inferred from comparative analyses of rRNA sequences: Woese(PNAS, 1990)

# Horizontal Gene Transfer

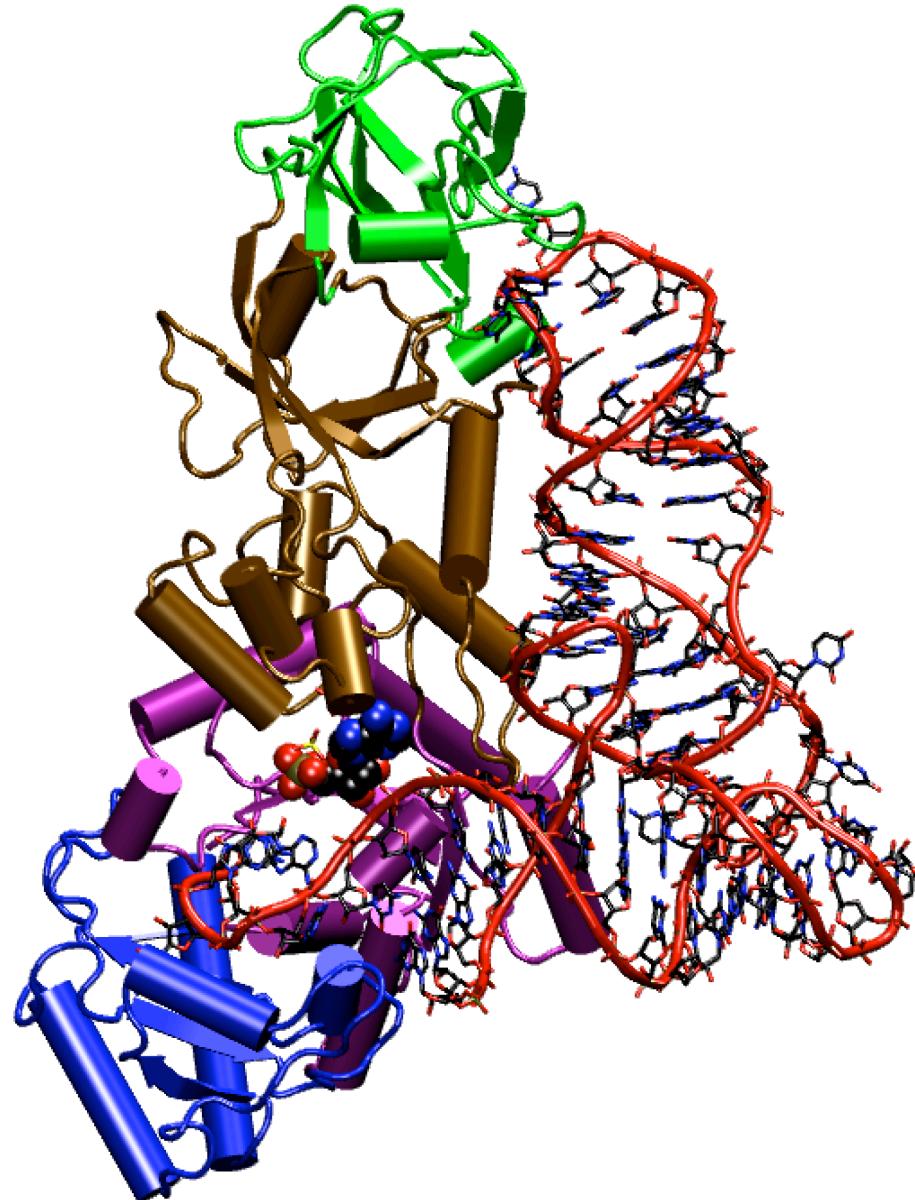


# Multiple Sequence Alignments

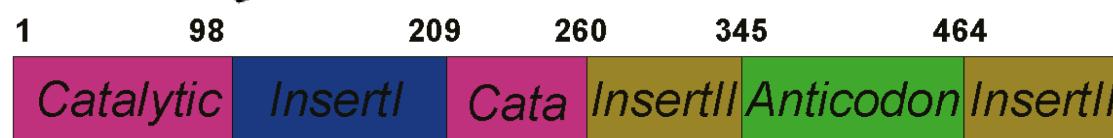
- “The aminoacyl-tRNA synthetases, perhaps better than any other molecules in the cell, eptiomize the current situation and help to under standard (the effects) of HGT” Woese (PNAS, 2000; MMBR 2000)

# Standard Dogma Molecular Biology

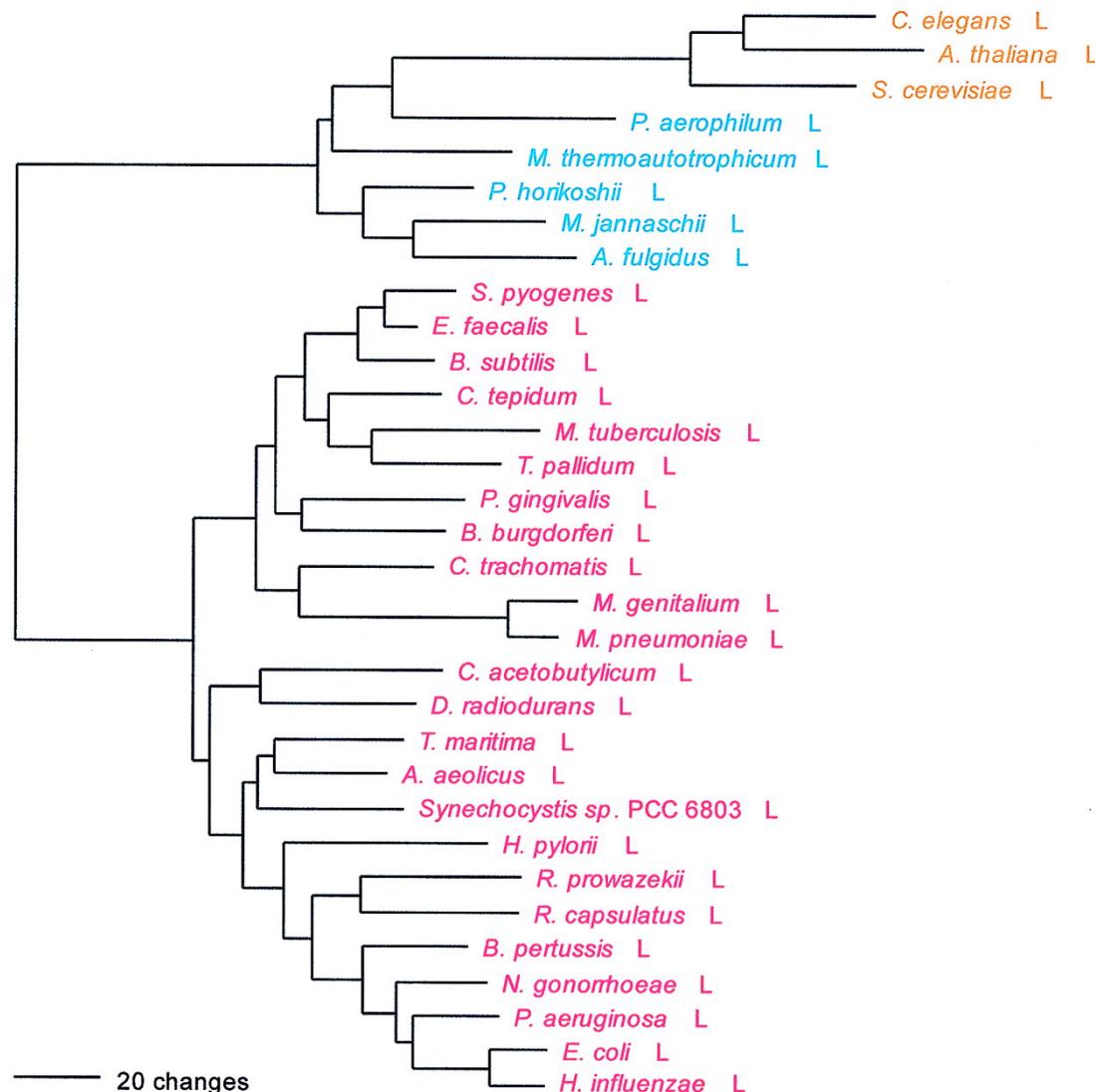
- DNA → RNA → Proteins
- Role of AARS?
- Charging of t-RNA



NCBI 3D

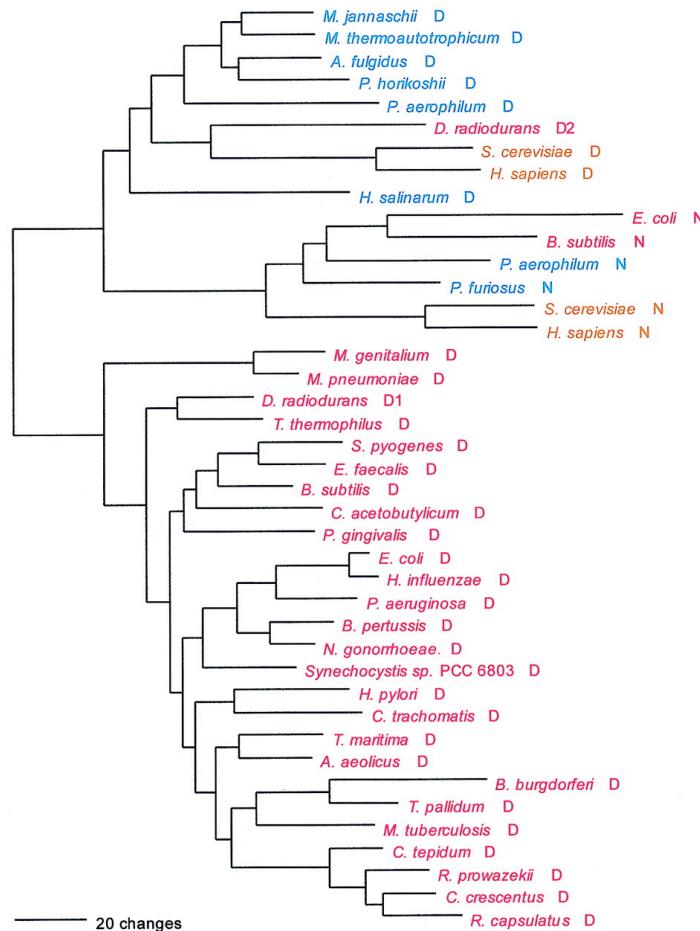


# LeuRS Canonical Tree



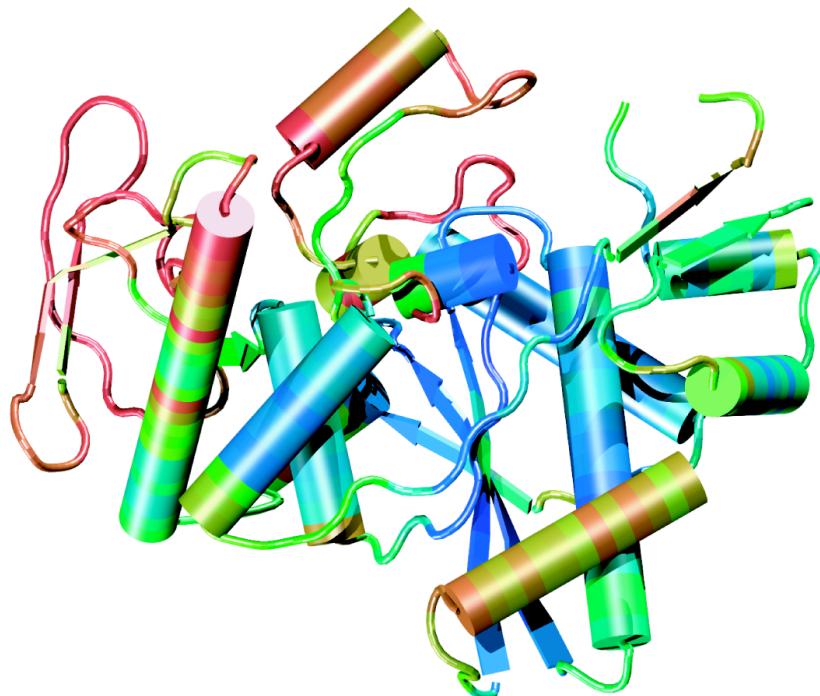
Woese, Olsen (UIUC), Ibba (Panum Inst.), Soll (Yale) *Micro. Mol. Biol. Rev.* March 2000..

# D,N Sequence Phylogenetic Trees



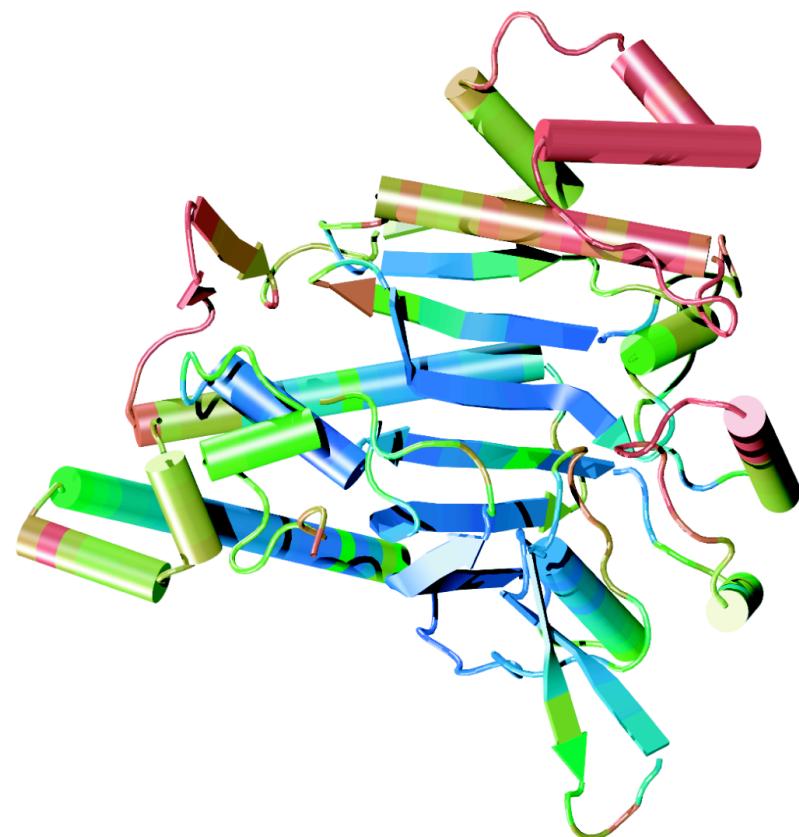
Woese, Olsen (UIUC), Ibba (Panum Inst.), Soll (Yale) *Micro. Mol. Biol. Rev.* March 2000 • •

# Fold Motifs of AARSs



Class I Lysyl-tRNA Synthetase

O'Donoghue and Luthey-Schulten, UIUC 2003

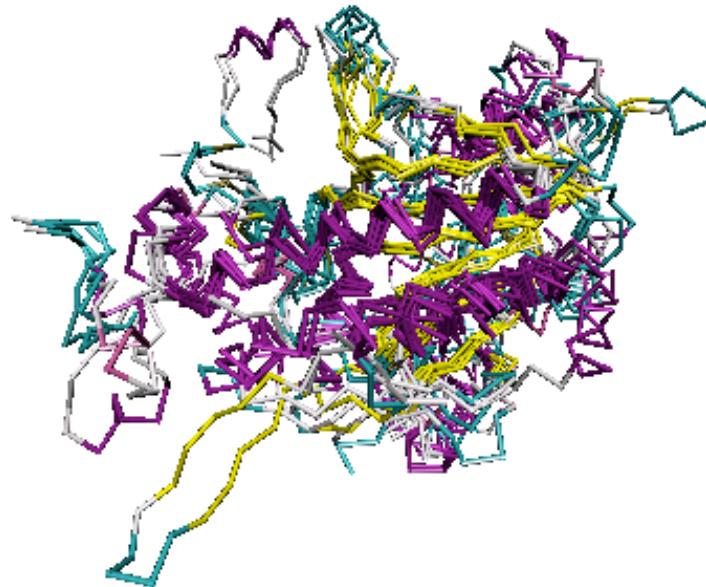


Class II Lysyl-tRNA Synthetase

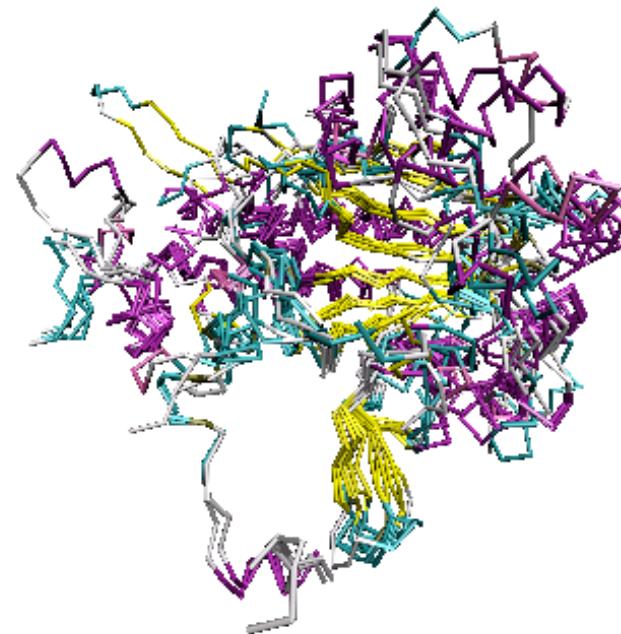
# Structure Conserved More than Sequence

## Structural Overlap of Class II AARS

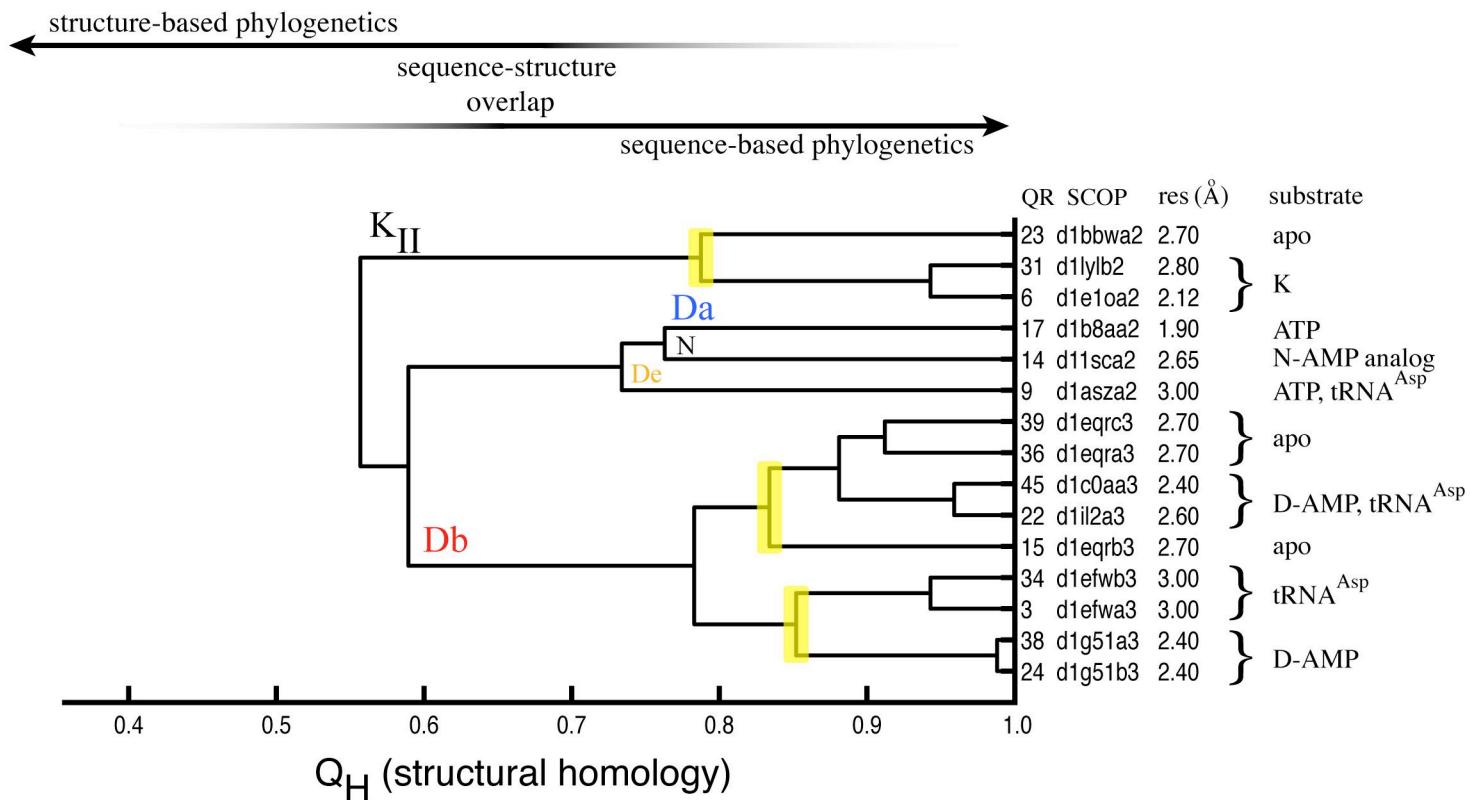
Conserved helices



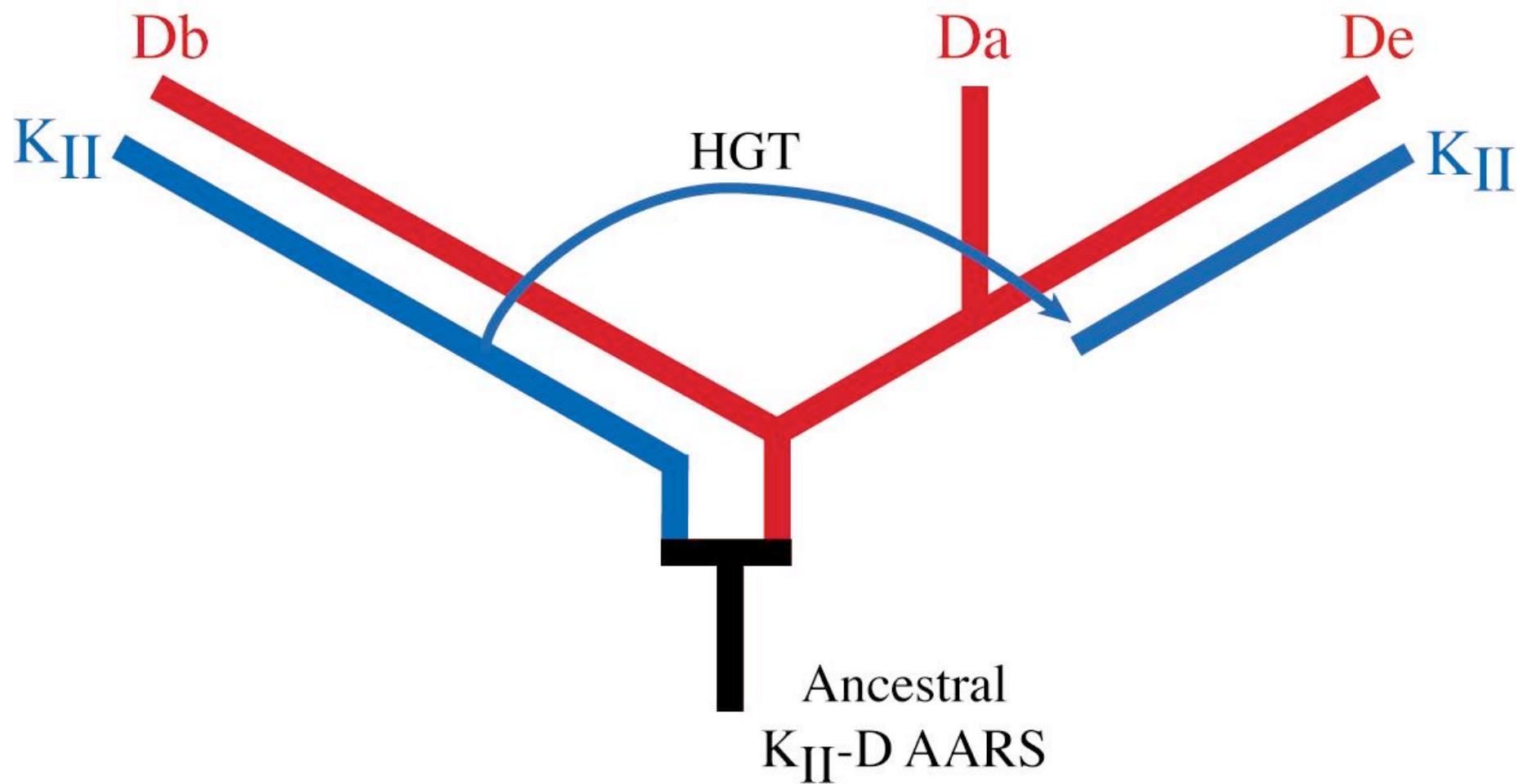
Conserved sheets



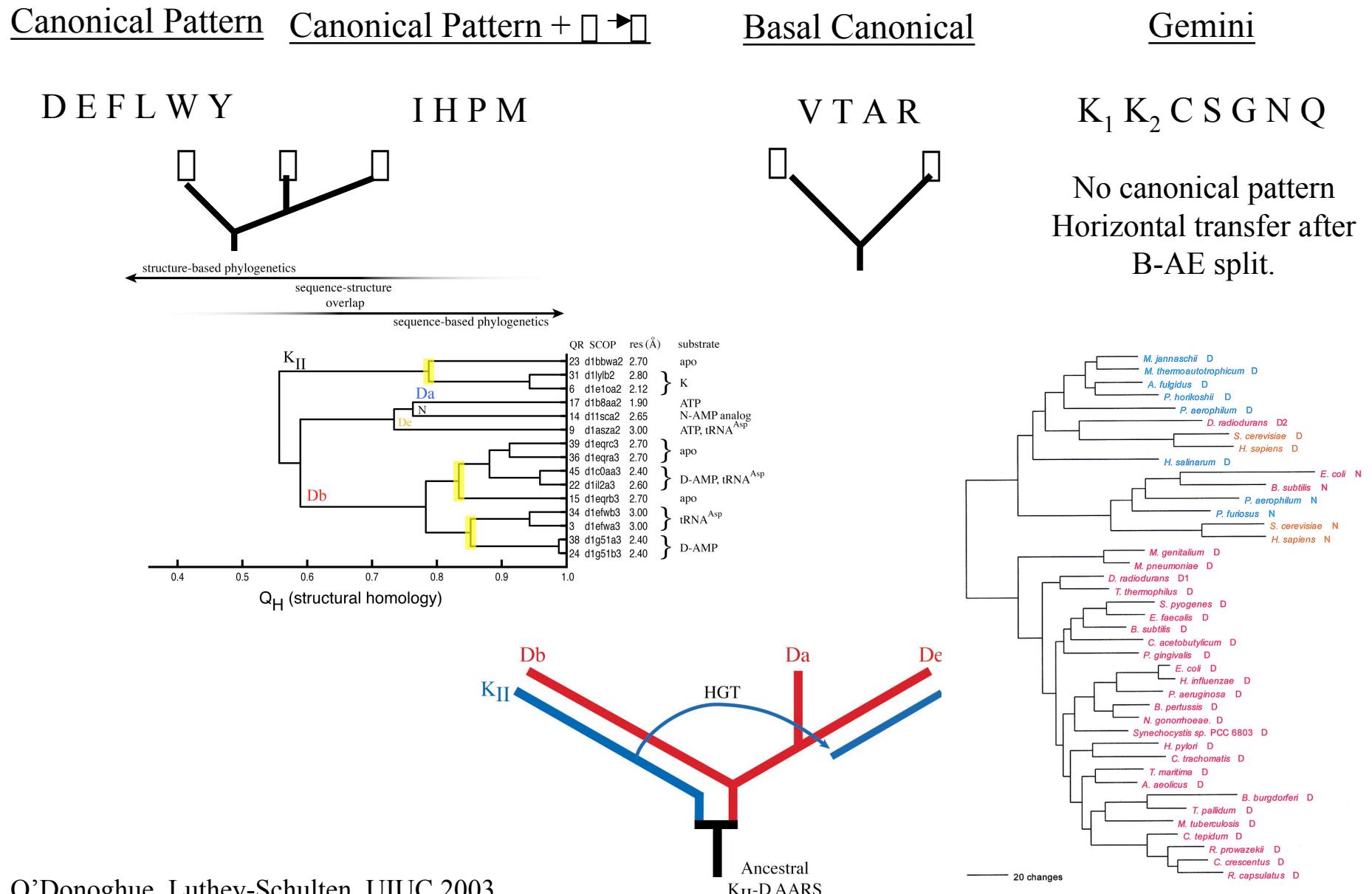
# Subset of Class II Structural Tree



O'Donoghue and Luthey-Schulten, UIUC 2003



# Novel Evolutionary Connections from Sequence and Structure

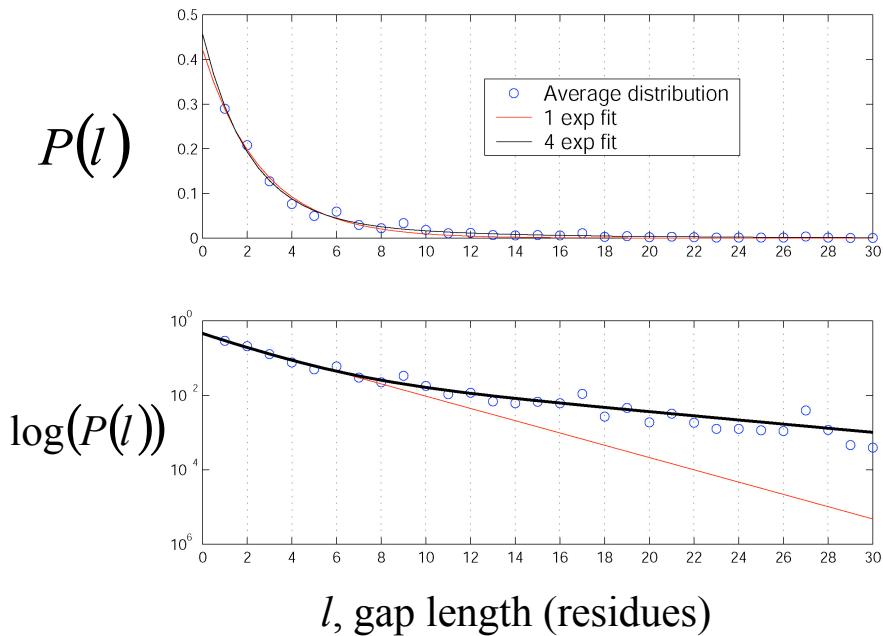


O'Donoghue, Luthey-Schulten, UIUC 2003

Woese, Olsen (UIUC), Ibba (Panum Inst.), Soll (Yale) *Micro. Mol. Biol. Rev.* March 2000..

# Gap Distribution Functions

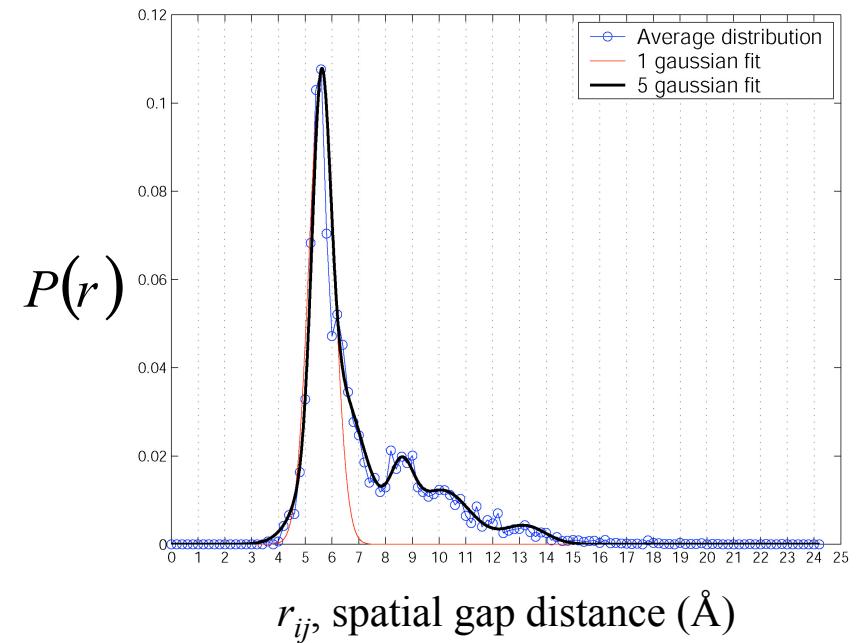
Length Gap Distribution Function



$$P_{insertion}(l) = a_1 * \exp(-b_1 l)$$

$$P_{insertion}(l) = \sum_{i=1}^4 C_i * \exp(-D_i l)$$

Spatial Gap Distribution Function



$$P_{deletion}(r) = a_2 * \exp\left(-\frac{(r - b_2)^2}{2\sigma_2^2}\right)$$

$$P_{deletion}(r) = \sum_{i=1}^5 A_i * \exp\left(-\frac{(r - B_i)^2}{2\theta_i^2}\right)$$

# Structural Alignment Methods

- PDB - Structural Neighbors – CE  
(Bourne)
- Stamp - Russell

# STAMP Multiple Structural Alignments

## 1. Initial Alignment

- Multiple Sequence alignment
- Ridged Body “Scan”

## 2. Refine Initial Alignment & Produce Multiple Structural Alignment

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

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

$d_{ij}$  -- distance between i & j

$S_{ij}$  -- conformational similarity; function of rms 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.

# Multiple Structural Alignments

## STAMP – cont'd

### 2. Refine Initial Alignment & Produce Multiple Structural Alignment

Alignment score:

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

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

$L_p, L_A, L_B$  -- 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.

# Stamp Output/Secondary Structure



# Stamp Output/Clustal Format

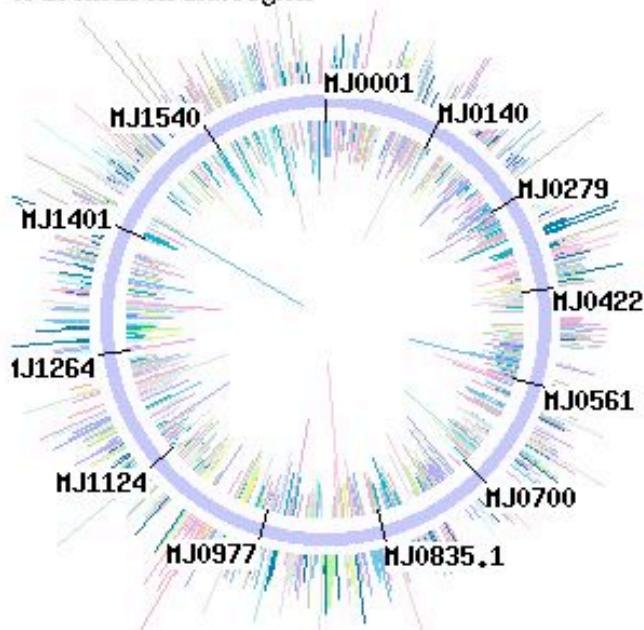
SerRS-T_thermophilus	VGGEENREIKRGGPPEFSFP--P--LDHVALMEKNGWWEPRISQVSGSRSYALKGDLA
ThrRS-E_coli	-----R-DHRKIGKQLDLY-HMQ-EE-APGMVFWHNDGW
ProRS-T_thermophilus	-----KGLTPQSQDFSEWYLEVIQKAELAD-YG--P-VRGTIVVRPYGY
ProRS-M_thermoautotrophicus	-----EFSEWFHNILEEAEIIDQRY--P-VKGMHVWMPHGF
space	-----
SerRS-T_thermophilus	--SGGG-EEEEEEES----SS-----HHHHHHHHHT-B-TTHHHHH-SS---B-THHH
ThrRS-E_coli	-----HHHHHHHHHTT-E-E---TT-STT--EE-HHHH
ProRS-T_thermophilus	-----HHHHHHHHHHHHHHHHHTTSEE-E---S-STT-EEE-HHHH
ProRS-M_thermoautotrophicus	-----HHHHHHHHHHHHHTT-EE----S-STT--EE-HHHH
SerRS-T_thermophilus	LYELALLRFAMDFMARRGFLPMTLPSYAREK-AFLG-TGHFPAYRDQVWAIA----E--
ThrRS-E_coli	TIFRELEVVFVRSKLKEYQYQEVKGPFMMDRV-LWEKT-GHWDNYKDAMFTTS---S-EN
ProRS-T_thermophilus	AIWENIQVQLDRMFKETGHQNAVPLFIPMSFL-----FSPELAVVTHAGGEELE
ProRS-M_thermoautotrophicus	MIRKNLKLRRILD-RDHEEVLFPLLVEDE-LAKEAIHVKGFEDEVYWWVTHGGLSKLQ
space	-----
SerRS-T_thermophilus	HHHHHHHHHHHHHHHTT-EEEE--SEEEHH-HHHH-HT-TTTGGGS-B-T----T--
ThrRS-E_coli	HHHHHHHHHHHHHHHTT-EE---SEEEHH-HHHTT-THHHHGGG--EEE---E-TT
ProRS-T_thermophilus	HHHHHHHHHHHHHHHHHTT-EE---SEESTT-----TT--EEEE-SSSEE
ProRS-M_thermoautotrophicus	HHHHHHHHHHHHHHHTT-TT-EE---SEEEHH-HTTSHHHHHHTTTT--EEEEETTEEEE
SerRS-T_thermophilus	TDLYLTGTAEVVLNALHSGEILPYEALPLRYAGYAPAFRSEA--GSFGKDVRLMRVH-Q
ThrRS-E_coli	REYCIKPMNCPGHVQIFNQGLKSYRDLPLRMAEFGSCHR--NEPS--G-SLHGLMRVR-G
ProRS-T_thermophilus	EPLAVRPTSETVIGYMWSKWIRSWRDLPQLLNQWGNVVRW--E---M-RTRPFLRTSE-
ProRS-M_thermoautotrophicus	RKLALRPTSETVMYPMFALWVRSHTDLPMRFYQVVNTFRY-ET---K-HTRPLIRVREI
space	-----
SerRS-T_thermophilus	SSEEE-S-THHHHHHHHTT-EEEGGG-SEEEEEEEEEE-----S--SSTTTTTTS-S-E
ThrRS-E_coli	EEEEEE-S-SHHHHHHHHHTSS--BTTT-SEEEEE--EEE-----G-G-G-BTTTB-S-E
ProRS-T_thermophilus	EEEEEE-S-SHHHHHHHHHHHH-BGGG--EEEEEEEEEEE-----S-S-BTTTB-SE-
ProRS-M_thermoautotrophicus	EEEEEE-SSSHHHHHHHHHHH--BTTT--EEEEEEEEEEE-----S---BTTTB-SEE

# Examples of Useful Web Tools

- Genomes – Sequence and Gene Information
- Domain Architecture
- Multiple Sequence Alignments
- Phylogenetic Trees
- Structural Databases
- Hidden Markov Methods

## Protein coding genes distribution map

To see map locations of genes, click on a region in the map,  
to zoom in on that region



## Gene Classification based on [COG functional categories](#)

- [Color] Translation, ribosomal structure and biogenesis
- [Color] Transcription
- [Color] DNA replication, recombination and repair
- [Color] Cell division and chromosome partitioning
- [Color] Posttranslational modification, protein turnover
- [Color] Cell envelope biogenesis, outer membrane
- [Color] Cell motility and secretion
- [Color] Inorganic ion transport and metabolism
- [Color] Signal transduction mechanisms
- [Color] Energy production and conversion
- [Color] Carbohydrate transport and metabolism
- [Color] Amino acid transport and metabolism
- [Color] Nucleotide transport and metabolism
- [Color] Coenzyme metabolism
- [Color] Lipid metabolism
- [Color] Secondary metabolites biosynthesis, transport and catabolism
- [Color] General function prediction only
- [Color] Function unknown
- [Color] No COG match

Organism: [Methanocaldococcus jannaschii](#)

Genetic Code: [11](#)

Lineage: Archaea; Euryarchaeota; Methanococci; Methanococcales;  
Methanocaldococcaceae; *Methanocaldococcus*.

## Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*

Bult,C.J., White,O., Olsen,G.J., Zhou,L., Fleischmann,R.D., Sutton,G.G., Blake,J.A., FitzGerald,L.M., Clayton,R.A., Gocayne,J.D., Kerlavage,A.R., Dougherty,B.A., Tomb,J.-F., Adams,M.D., Reich,C.I., Overbeek,R., Kirkness,E.F., Weinstock,K.G., Merrick,J.M., Glodek,A., Scott,J.D., Geoghegan,N.S., Weidman,J.F., Fuhrmann,J.L., Nguyen,D.T., Utterback,T., Kelley,J.M., Peterson,J.D., Sadow,P.W., Hanna,M.C., Cotton,M.D., Hurst,M.A., Roberts,K.M., Kaine,B.B., Borodovsky,M., Klenk,H.P., Fraser,C.M., Smith,H.O., Woese,C.R. and Venter,J.C.

Science 273 (5278), 1058–1073 (1996)

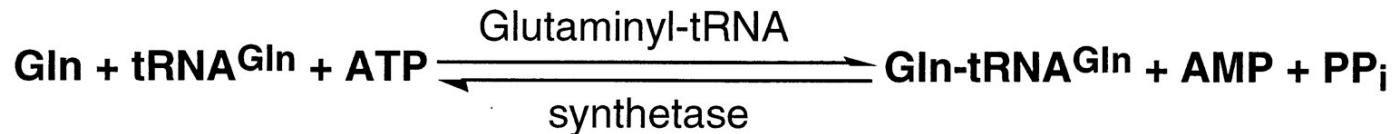
[96337999](#)

NCBI: Genomes

# Charging the tRNA

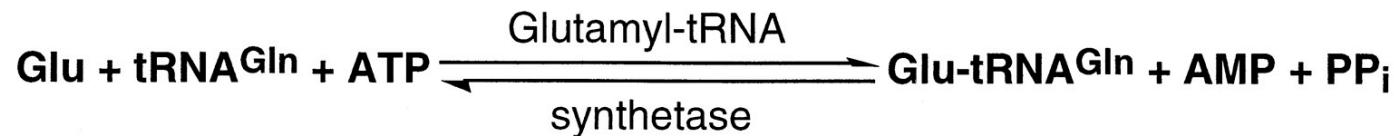
## *Direct acylation*

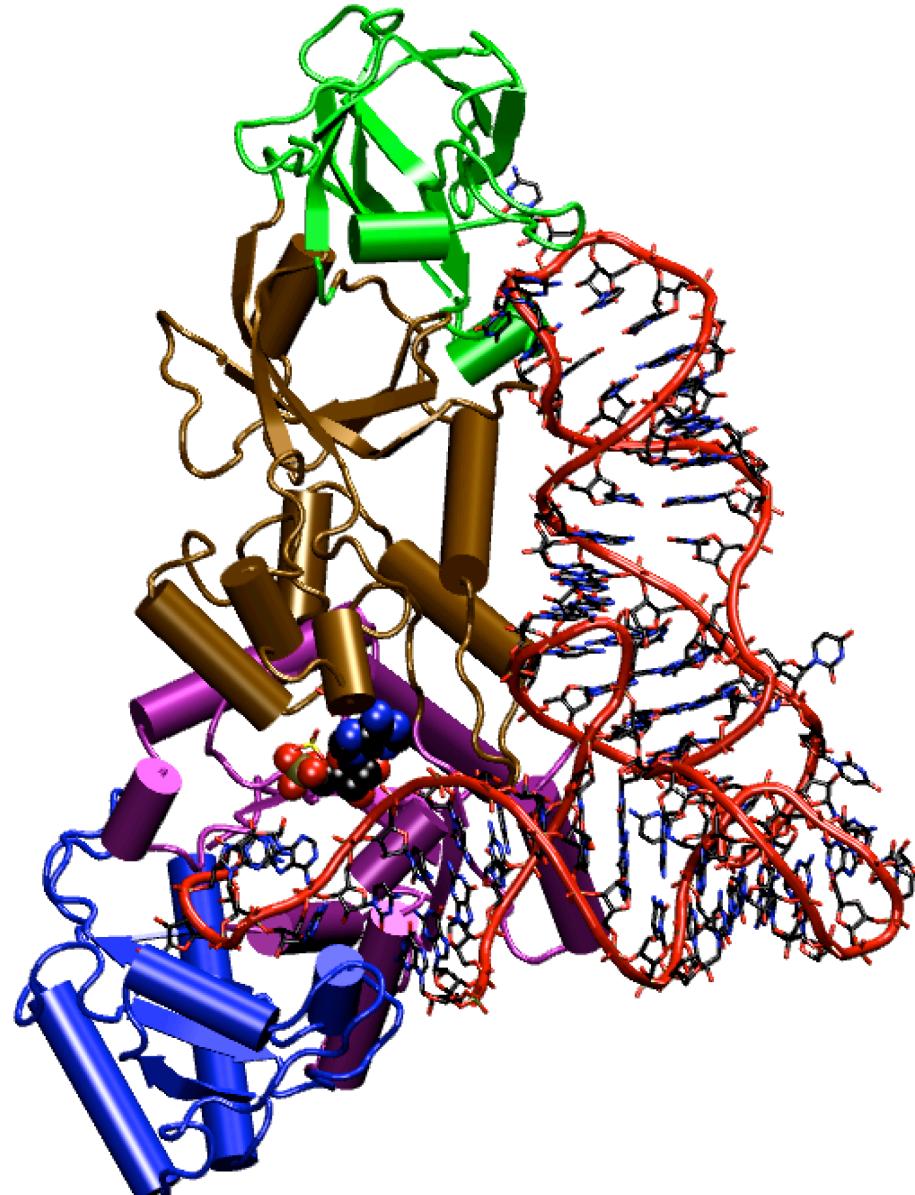
---



## *tRNA-dependent amino acid modification*

---





NCBI 3D

1

98

209

260

345

464

Catalytic

InsertI

Cata

InsertII

Anticodon

InsertII

# Report from SWISS-PROT

## Comments

- **CATALYTIC ACTIVITY:** ATP + L-aspartate + tRNA(Asp) = AMP + diphosphate + L-aspartyl-tRN
- **COFACTOR:** Binds 3 magnesium ions per subunit (*By similarity*).
- **SUBCELLULAR LOCATION:** Cytoplasmic.
- **SIMILARITY:** Belongs to class-II [aminoacyl-tRNA synthetase](#) family.

## Copyright

This SWISS-PROT entry is copyright. It is produced through a collaboration between the Swiss Institute of Bioinformatics a  
outstation – the European Bioinformatics Institute. There are no restrictions on its use by non-profit institutions as long as its  
modified and this statement is not removed. Usage by and for commercial entities requires a license agreement (See  
<http://www.isb-sib.ch/announce/> or send an email to [license@isb-sib.ch](mailto:license@isb-sib.ch))

## Cross-references

EMBL	AB010464; BAA31457.1; -. [ <a href="#">EMBL</a> / <a href="#">GenBank</a> / <a href="#">DDBJ</a> ] [ <a href="#">CodingSequence</a> ]
HSSP	<a href="#">Q52428</a> ; 1B8A. [ <a href="#">HSSP ENTRY</a> / <a href="#">PDB</a> ]
HAMAP	<a href="#">MF_00044</a> ; -; 1. PBIL [ <a href="#">Family</a> / <a href="#">Alignment</a> / <a href="#">Tree</a> ]
InterPro	<a href="#">IPR004523</a> ; AspS_arch. <a href="#">IPR004364</a> ; tRNA-synt_2. <a href="#">IPR002312</a> ; tRNA-synt_asp. <a href="#">IPR004365</a> ; tRNA_anti. <a href="#">IPR006195</a> ; tRNA_ligase_II. <a href="#">Graphical view of domain structure</a> .
Pfam	<a href="#">PF00152</a> ; tRNA-synt_2; 1. <a href="#">PF01336</a> ; tRNA_anti; 1.
PRINTS	<a href="#">PR01042</a> ; TRNASYNTHASP.
TIGRFAMs	<a href="#">TIGR00458</a> ; aspS_arch; 1.
PROSITE	<a href="#">PS50862</a> ; AA_TRNA_LIGASE_II; 1.
ProDom	[ <a href="#">Domain structure</a> / <a href="#">List of seq. sharing at least 1 domain</a> ]
HOBACGEN	[ <a href="#">Family</a> / <a href="#">Alignment</a> / <a href="#">Tree</a> ]
BLOCKS	<a href="#">O24822</a> .
ProtoNet	<a href="#">O24822</a> .
ProtoMap	<a href="#">O24822</a> .

# PFAM Report

## Representative tRNA-synt\_2 family proteins

This family may contain **overlapping domains**, to change the graphical view click [here](#)

[SYDC YEAST](#) [Saccharomyces cerevisiae (baker's yeast)] aspartyl-tRNA synthetase, cytoplasmic (ec 6.1.1.12) (aspartate--tRNA ligase)(asprs)



[SYD CAEEL](#) [Caenorhabditis elegans] aspartyl-tRNA synthetase (ec 6.1.1.12) (aspartate--tRNA ligase)(asprs)



[SYD HUMAN](#) [Homo sapiens (human)] aspartyl-tRNA synthetase (ec 6.1.1.12) (aspartate--tRNA ligase)(asprs)



[SYD RAT](#) [Rattus norvegicus (rat)] aspartyl-tRNA synthetase (ec 6.1.1.12) (aspartate--tRNA ligase)(asprs)



KLTGMAFRUPTRNUSUUD  
KLTGMAFRUPTRNUSUUD  
KLTGMAFR-PTUDUSUUD  
KLDGTSIIRUPTPDUSUUD  
KLNGMAMMRUPTRNUSUUD

# Clustal W

Multiple Sequence Alignment

## CLUSTALW: Multiple Sequence Alignment[\[help\]](#)

---

General Setting Parameters:

Output Format:

Pairwise Alignment:  FAST/APPROXIMATE  SLOW/ACCURATE

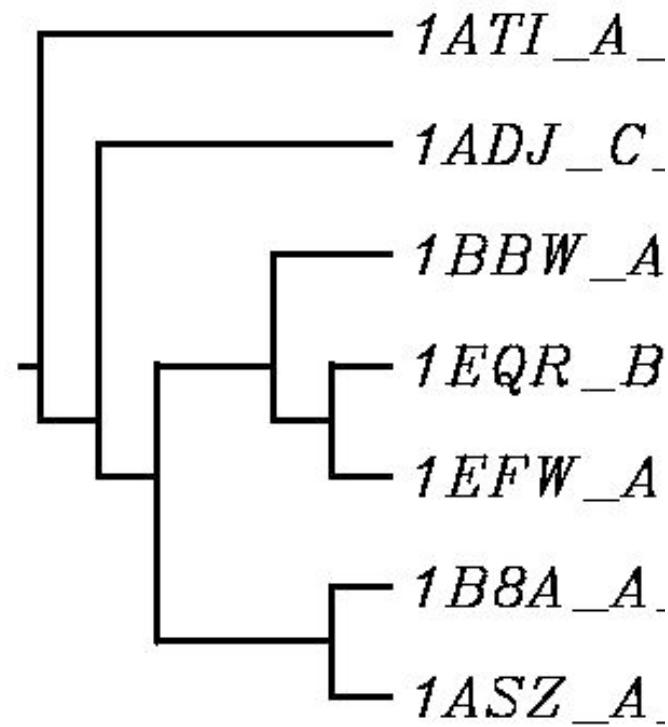
Enter your sequences (with labels) below (copy & paste):  PROTEIN  DNA

Support Formats: FASTA (Pearson), NBRF/PIR, EMBL/Swiss Prot, GDE, CLUSTAL, and GCG/MSF

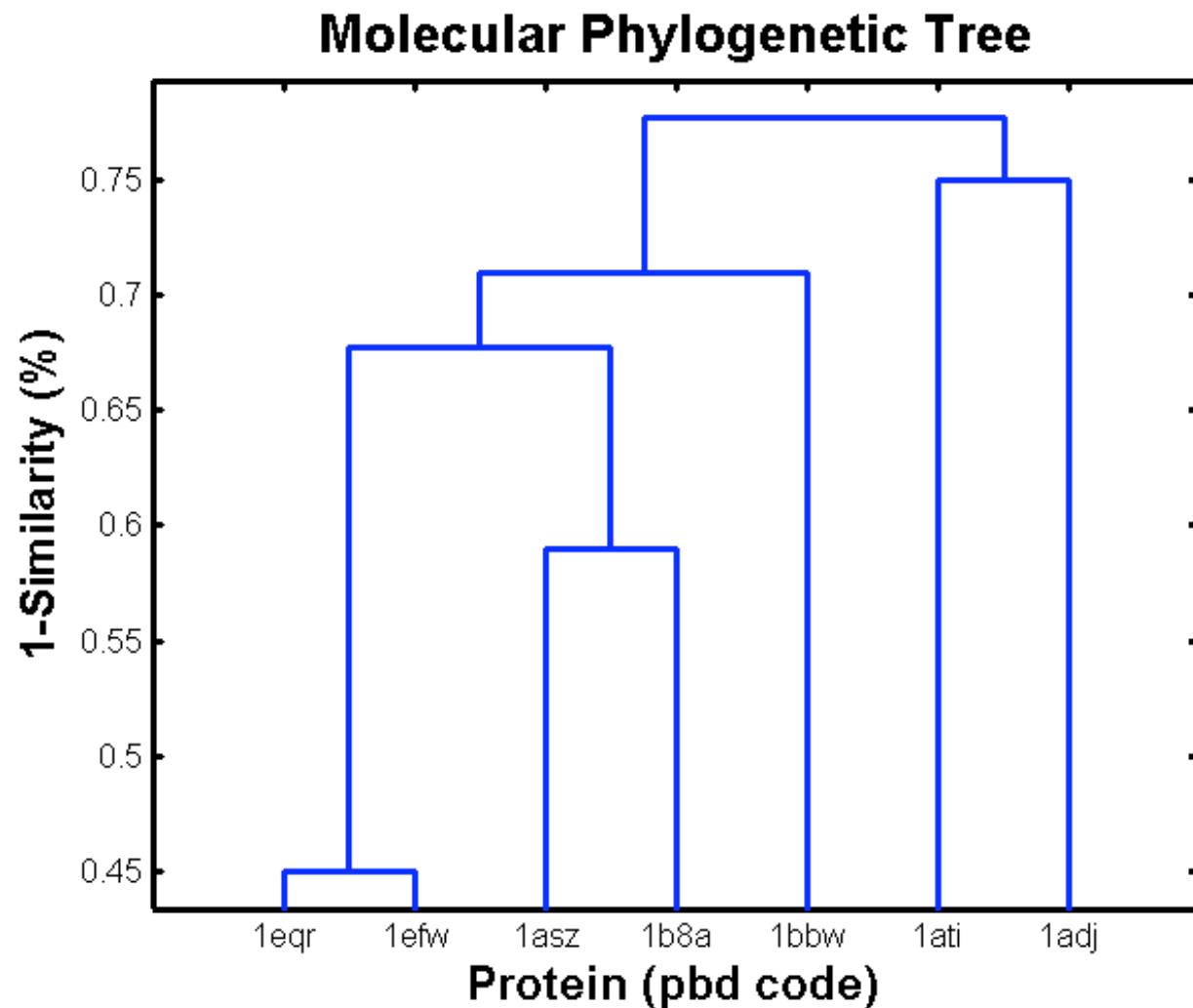
```
>1EQR:B (000:0-000)
VLPLDSNHVNTEEARLKRYLDLRRPEMAQRLKTRAKITSLVRRFMDDHGFLLDIETPMLT
KATPEGARDYLVPNSRVHKGFYALPQS PQLFKQLLMMMSGFDRYYQIVKCFRDEDLRADRQ
PEFTQIDVETS FMTAPQVREVMELVRHLWLEVKGVDLGDFPVMTFAEAERRYGSKPDL
RDESKWA PLWVIDFPMFEDDGEGLTAMHHPFTSPKDMTAAELKAAPENA VANAYDMVIN
```

Or give the file name containing your query

# Sequence Dendrogram from Clustal



# Phylogenetic Tree in Tutorial





## 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. [1i12](#)

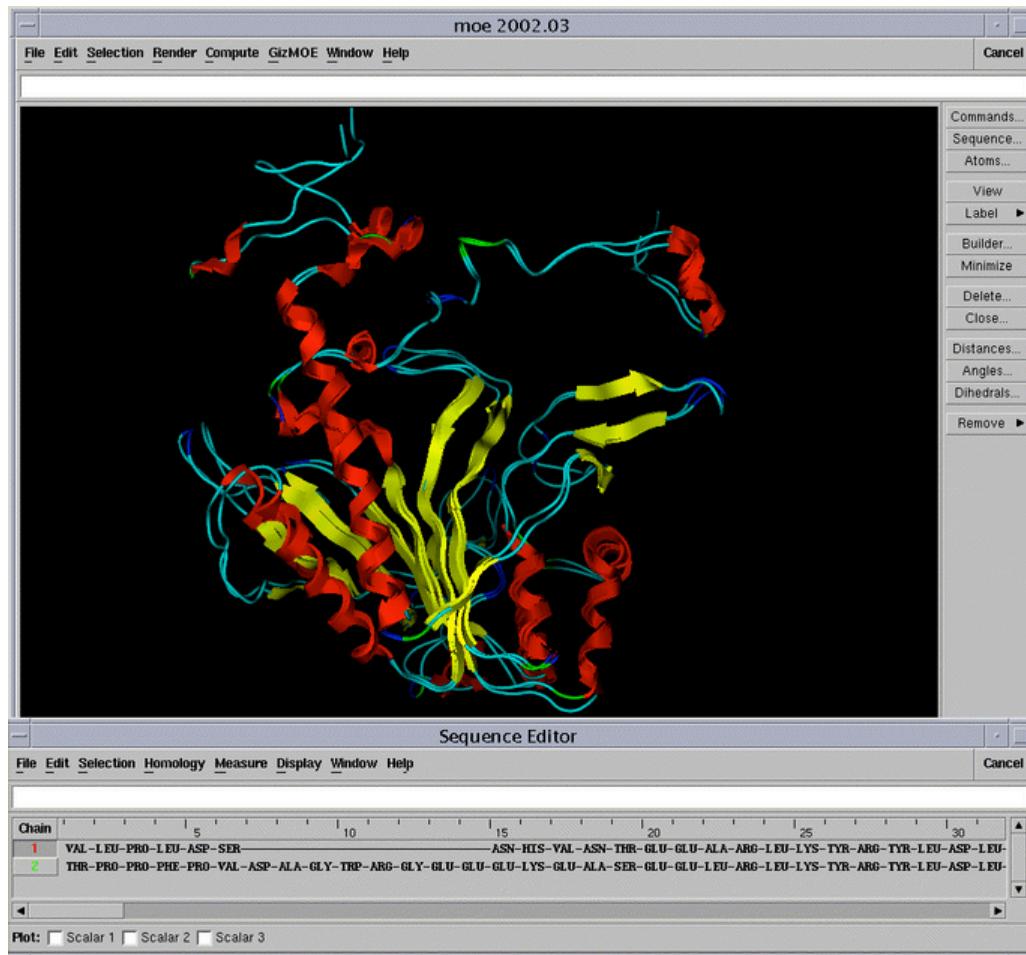
*complexed with 1mg, 5mc, 5mu, amo, h2u, psu, so4*

  1. [region a:1-106](#)
  2. [region b:1001-1106](#)
3. [1eqr](#)

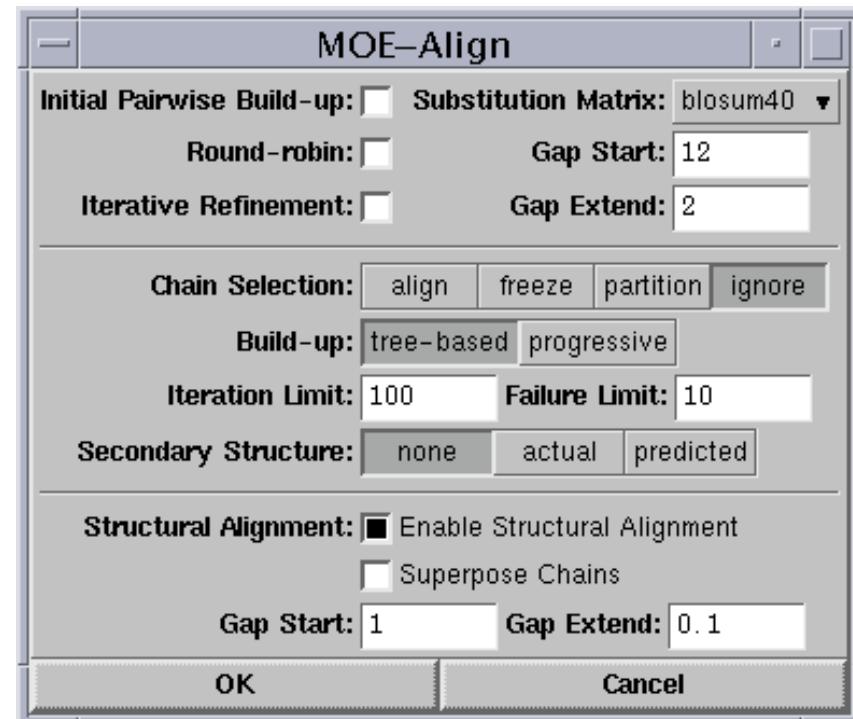
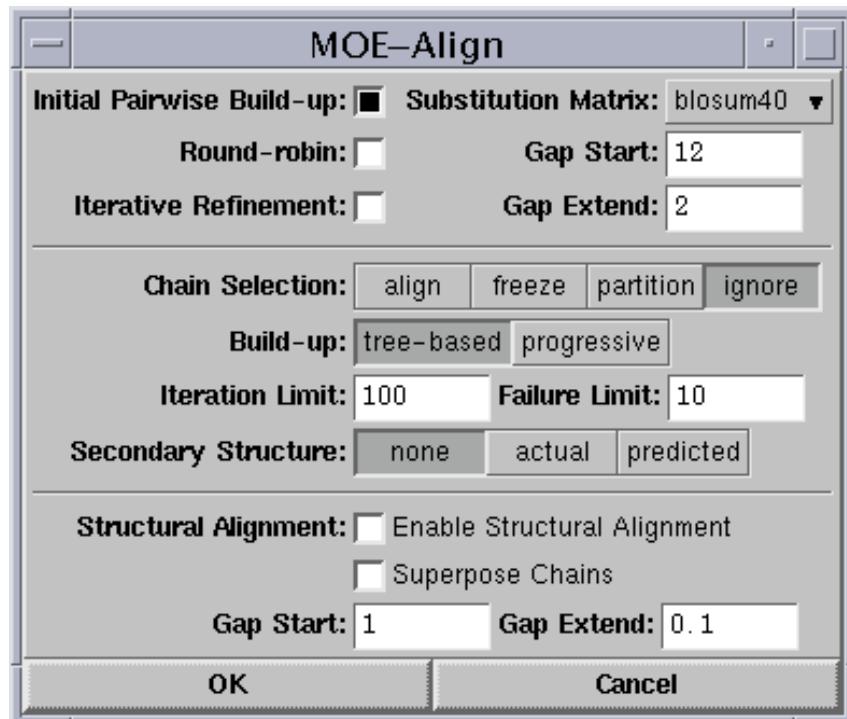
*complexed with mg*

  1. [region a:1-106](#)
  2. [region b:1-106](#)
  3. [region c:1-106](#)

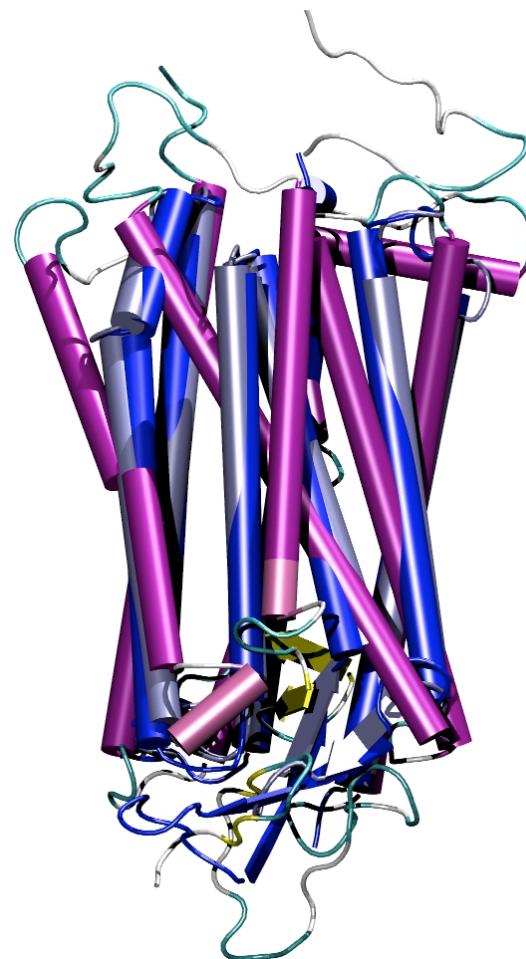
# Alignment in MOE



# Alignment in MOE



# Transmembrane Proteins - HMM



Example Bacteriorhodopsin – Anurag Sethi UIUC

# Stamp Profile

d1l9ha_3	MNGTECPNFYVPPSMKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLWIGFPWFLTLVTVQH
die12a	-----R-ENALLSNSLWNVALAGIILWVVMGRT--IR
d1jgja_1	-----MVGLIULFWLGAIGIANGTLAFACACRD--AG
d1l9ha_3	KKLRTPLNYILNLADLFMIFG-----TTTLYTSLHGTYFY-F-----GPTGCNL
die12a	PG---RPHLIRGATIMIPLSI-SYLGIL-----S---GITYGMDEMPAGHALAKEMVR--SQWG
d1jgja_1	S----GERRYVITLGISGLAA-VHYAVHA-----L--GPGWVP-----EERT--VFVP
d1l9ha_3	EGFFATLGGEIAIW-SLW-----IAIERYVVVCKPMMSNFRFGENHAIIGCWAFTWVHAGCAAPPLVGW
die12a	RYLTWALSTPY-IILA-LGLL-A-----D-----DGGSGFTVIAADEGNCNTG--LA
d1jgja_1	RYEDWILTTPL-IHYF-LGLL-A-----G-----DSREFHIVITENTVVMLAG--FA
d1l9ha_3	SRYIPEGMQCSCCIDYY-TPHEETNNEFVIYMFVVHFIPPLIVIFF-CYS-QLVFTVKEARAAAT
die12a	SA-----M---TTGHL--IFRCAFSAISCA-FPPPPSALVTDW-GASA-S-----
d1jgja_1	SA-----M---VP-H---SERIALPGCHGAV-AEFIGVYYLVGPM-TESA-S-----
d1l9ha_3	TQKAEEKETRIVVIIIVVIAFICHLPLYAGVAF-Y-IFTHQGSD-FGPIFMDIPAFRAK-TSAVYNP
die12a	--SA--GTAEIDDTLRULTVVLWLCGYPIVWAIGVE--G--ALDQSVGATSWAYSWLDFAKYVPR
d1jgja_1	--QRSSGIKSILYURLRNLTVVLWAIYPFWLIGPP--G--ALD-QPTVDVALIVLDLVIKVGFQ
d1l9ha_3	VIVYLM-NNKQFRNCMVTTLCCGKNPLGDST--TVSKTETSQV-APA-----
die12a	FILLRWMAN-----NERT-----VAV-----
d1jgja_1	FIALDA-AA-----

## Building HMM HMM.982259 ..

### Selected Option for HMM Model HMM.982259: build

```
HMMER2.0 [2.2g]
NAME inclustal
LENG 370
ALPH Amino
RF no
CS no
MAP yes
COM /usr/local/bin/hmmbuild /bio/tmp/inclustal.982259.hmm /bio/tmp/inclustal.982259
NSEQ 3
DATE Sun Jun  8 18:12:11 2003
CKSUM 1057
XT -8455 -4 -1000 -1000 -8455 -4 -8455 -4
NULT -4 -8455
NULE 595 -1558 85 338 -294 453 -1158 197 249 902 -1085 -142
HMM   A   C   D   E   F   G   H   I   K   L   M   N
      m->m m->i m->d i->m i->i d->m d->d b->m m->e
      -567 * -1622
1 -1029 -1038 -2200 -1928 -323 -2073 -1373 319 -1471 569 4218 -1777
- -149 -500 233 43 -381 399 106 -626 210 -466 -720 275
- -31 -6105 -7147 -894 -1115 -701 -1378 -567 *
2 -706 -1410 -63 -215 -1846 -1134 -697 -2058 -581 -2198 -1604 3525
- -149 -500 233 43 -381 399 106 -626 210 -466 -720 275
- -31 -6105 -7147 -894 -1115 -701 -1378 * *
3 -855 -1188 -1421 -1605 -2567 3376 -1671 -2629 -1846 -2761 -2202 -1433
- -149 -500 233 43 -381 399 106 -626 210 -466 -720 275
- -31 -6105 -7147 -894 -1115 -701 -1378 * *
4 -101 -603 -1245 -1194 -1643 -916 -1116 -943 -1033 -1432 -944 -909
- -149 -500 233 43 -381 399 106 -626 210 -466 -720 275
- -31 -6105 -7147 -894 -1115 -701 -1378 * *
```

# HMMer Profile-Profile Alignment

d1l9ha\_3 MNGTEGPNFYVPPSNKTGVVRSPFEAPQYY|AEPWQFSMLAAYMFLLIMLGFP|NFLTLYVTVQH  
die12a -----R-ENALLSSSLWVVALAG|ILFVYNGRT--IR  
diat9\_1 -----A-Q-TGRPEWIVLALGTALMGLGTLYF|VKGMG-VSD  
d1jgja\_1 -----MVGL|TLFWLGAIGML|GTLAFAWAGR-D-A-G

d1l9ha\_3 -K|LRTPLNLYILLNLADLFM|FGFET|TLYTSLHGYFV-F-----GPTGCN  
die12a PGE----PRLI|GAT|WIPLE|E-SYLG|LL-----SG|TGME|MPAGHAL|GENVR--SQW  
diat9\_1 P-D----A|SFYAITT|WPAIAF|W|MYL|ML-----LG|YC|TMVPF-----GEQNP--|W  
d1jgja\_1 S-G----ER|YYVTL|EGISGIAA|V|YAVMA-----LG|GWVPV-----AERT--|W

d1l9ha\_3 LEGFFATLGGE|A|W-SL|W|IAIERYVVVCKPMNSNFRFGENHA|MG|WFTWWMA|ACAAPPL|G  
die12a PRY|TWAL|TP|W|LLA-LGLD|---A-----D|---D|G|S|FTV|I|AD|GMCVTG-L-  
diat9\_1 PRYADW|FTTF|L|LLD-L|L|L|---V-----D|---ADQG|V|LA|ADGIM|GTG-L-  
d1jgja\_1 PRY|DW|LTTP|L|WYF-L|GLD|---A-----G|---DSREF|V|VITL|TVV|LAG-F-

d1l9ha\_3 WSRYIPEGMQCSCGIDYY|PHEETNN|E|FVIYMFVVHF|I|PLIV|FF-CYG-QLV-FTVKEAAAA  
die12a A|A|---M-TT|AL|L|R|AAFA|SICA|FF|W|LSALVTD|---AAS|---AS|---  
diat9\_1 VGA|---L-T-KVY|---S|RE|V|W|A|STA|AM|Y|LYV|LFFG|---TSK|---A|---  
d1jgja\_1 AGA|---M-V-P-G|---I|ER|A|L|W|G|AV|AF|G|Y|Y|LVGPM|TES|---AS|---

d1l9ha\_3 TTQ-KAEKEVTR|V|W|VIAF|W|CULPVAGVAF-Y-IFTHQGD|D-FGP|IFMT|PAFFAK|---AVY  
die12a ---SA|---GTA|E|FDTLRV|LT|VVL|W|L|Y|P|W|VA|G|V|---C|---ALV|Q|VGAT|W|A|SVLDVFAK|Y  
diat9\_1 ---ESMRPEVASTFK|L|LRN|T|VVL|W|L|Y|P|W|L|G|SE|---GA|---V|PLN|---T|L|W|VLDV|A|K|Y  
d1jgja\_1 ---Q-RSSG|K|S|---W|LRN|L|TVVL|W|A|I|P|F|W|L|G|P|P|---G|---A|L|---S|P|T|---V|A|L|V|Y|D|W|V|T|K|V|G

d1l9ha\_3 NPV|Y|---M|---W|---FRNCMV|T|L|CCG|K|N|PLG|D|---TT|V|SK|T|E|T|S|Q|V|A|P|A  
die12a F|F|I|L|L|---R|W|A|M|---E|RT|V|---A|V|---  
diat9\_1 F|G|L|L|R|S|R|A|---I|F|---G|---  
d1jgja\_1 F|G|F|A|L|D|---A|A|A|---

# Clustal Profile-Profile Alignment

d1l9ha_3	NNGTEGPNFYVPPFSNKTGVVRSPFEAPQQYLAEPWQFSNLAAYMFLILIGFFPNELTLY
d1e12a	-----R-ENALLSSSWINNAALAGKALIFVWNGR
d1jgja_1	-----EVGLLTLFWLIAIGMLQGTLAFAAAGR
1AT9__BACTERIO	-----XAATGRPEWWLSTGTALCGTLLDENVF
d1l9ha_3	VTVQHKKLRPLNYILLNLAADLFMDFGKTTLYTSLHGYV-F
d1e12a	T-RPGR--RPRRLIGATIPIPLSS-SSYLGLL----S-GLTGGM EMPAGHALA-
d1jgja_1	D-AGS---GEREYYVTLGISGIAA-S-YASHA----L-CIGWVP-----
1AT9__BACTERIO	GNGNSDP--DAFAYAITTPAIAFTTYLALLG----GLTAVPFG-----
d1l9ha_3	--GPTGCNLEGFFATLGGEADW-SLV-LAIERYVVVCKPMNSFRFGENHA-MGDNFT
d1e12a	ENVR--SQWRYITWALITPL-LLA-LG-L-A-----D-----DING-FTV
d1jgja_1	-ERT- -VPRYDWLTPL-SYF-LO-L-A-----G-----DSREFIV
1AT9__BACTERIO	-GEQNPVWVRYADWFTTPLLLDLALLD-----ADQGDLA
d1l9ha_3	WVMIAACAAPPPLVGWSRYIPEGMQCSCGIDYY-.PHEETMNEFVIYMFVVHFIPLIV
d1e12a	IADGMCVTG--LAIA-----M--TTGUL-LRRAFAISCA-FFVIALSAL
d1jgja_1	ITLTVVLAG--FAGA-----M--VP---IERT ALANGAV-AFIGQVYLYL
1AT9__BACTERIO	WADGIMGTG--LVGA-----LTKVYSRIVVMAISTA-AMVYLYVL
d1l9ha_3	FF-CYG-QLVFTVKEAAAATTQKAKEKVTRIVAVIAFLCCLPVAGVAF-Y-IFTHQG
d1e12a	VTDS-AASA-S-----SA--GTAEFDTLRVLTVVLDLVPVWAWGVE--G-
d1jgja_1	VGPM-TESA-S-----QRSSGKSKS-NRLRNLTVVVWAIYPFWLGGPP--G-
1AT9__BACTERIO	FFGCTSKLE-----SMRPEVASTFKELRNNTVVLSHYPRVWLGGSE---G
d1l9ha_3	ID-FGPIFMTPAFFAK-VAVYNPVIYIM-DNKQFRNCMVTLCCGKNPLGDST--TVS
d1e12a	ALQQVGATWAVSVLDVFAKYVFIFILLRWVAN-----NERT--
d1jgja_1	ALI-PTPVALIIVYLDFTVKVGFGLIALDA-AA-----
1AT9__BACTERIO	AGIVPLNLTLDLVLDVAKVGFGLLRLSRAIFG-----EAEAP
d1l9ha_3	KTETSQV-APA
d1e12a	-----VAV-
d1jgja_1	-----
1AT9__BACTERIO	EPSADGAAATS

# Structure Prediction Modeller

## 6.2/Hmmer



Sethi and Luthey-Schulten, UIUC 2003

Modeller 6.2 A. Sali, et al.

# Acknowledgements

- Felix Autenrieth
- Barry Isralewitz
- Patrick O'Donoghue
- Taras Pogorelov
- Anurag Sethi