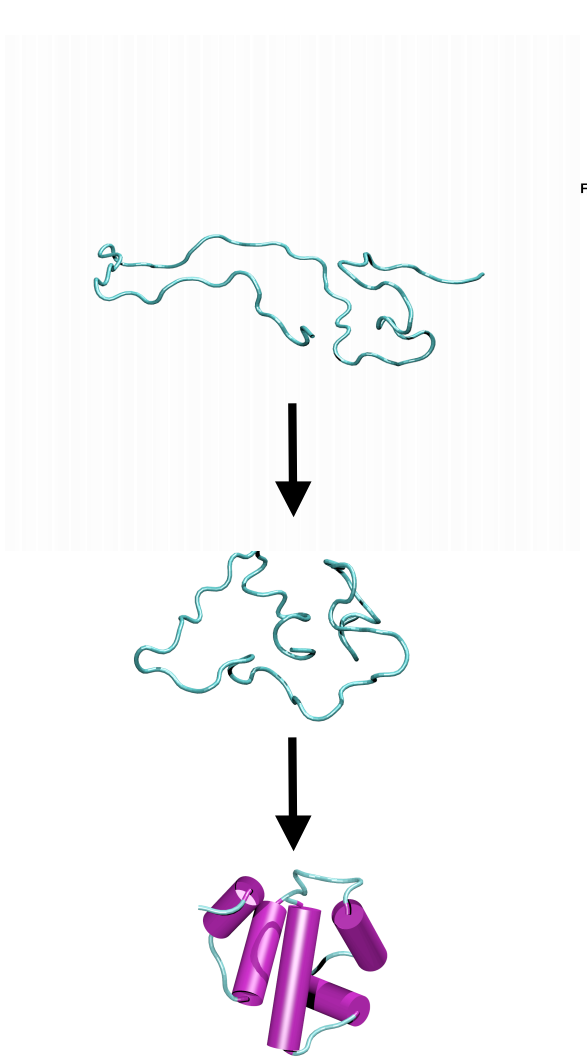
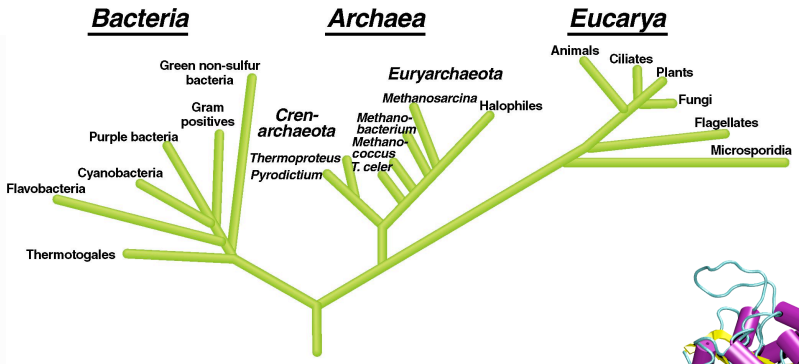
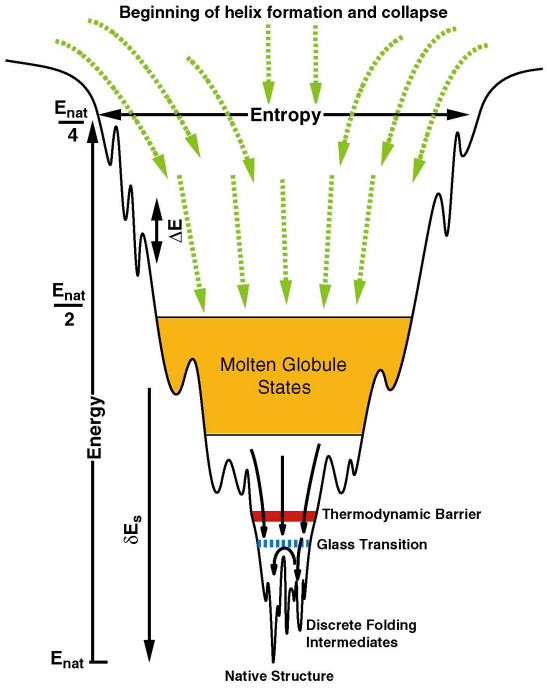
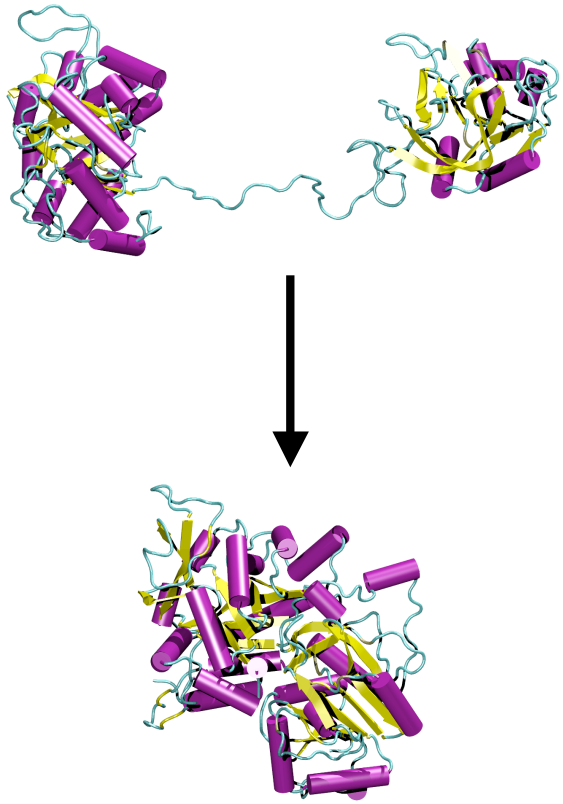
Bioinformatics – NSF Summer School 2003
Z. Luthey-Schulten, UIUC









Sequence-Sequence Alignment

- Smith-Watermann
- Needleman-Wunsch

Sequence-Structure Alignment

- Threading
- Hidden Markov

Sequence Alignment & Dynamic Programming

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

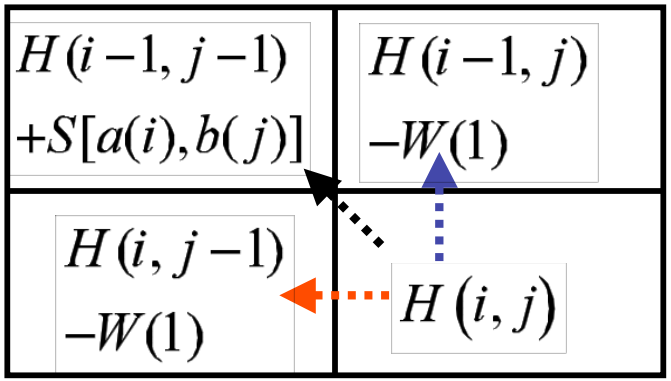
Seq. 2: c_1 - $c_2\ c_3\ c_4\ c_5$ - $\ldots c_m$

number of possible alignments:

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

Smith-Waterman alignment algorithm

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



S : similarity matrix

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

Score Matrix H: Traceback

AWGHE
AW--HE

		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	0
A	0	0	0	5	0	5	0	1	0	0	0
W	0	0	0	0	3	0	24	16	8	0	0
H	0	13	5	0	0	1	16	22	29	21	13
E	0	5	20	12	4	0	8	14	22	36	28
A	0	0	12	25	17	9	1	9	14	28	35
E	0	0	7	17	22	15	7	1	9	21	35

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	0
A	0	0	0	5	0	5	0	1	0	0	0
W	0	0	0	0	3	0	24	16	8	0	0
H	0	13	5	0	0	1	16	22	29	21	13
E	0	5	20	12	4	0	8	14	22	36	28
A	0	0	12	25	17	9	1	9	14	28	35
E	0	0	7	17	22	15	7	1	9	21	35

AWGHE

AW--HE

Blosum 40 Substitution Matrix

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

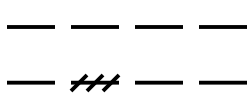
Protein Structural Relationships

Can protein structural relationships help us to understand evolutionary dynamics?

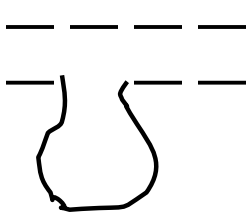
Is there a connection between evolutionary events and changes in protein structure?

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

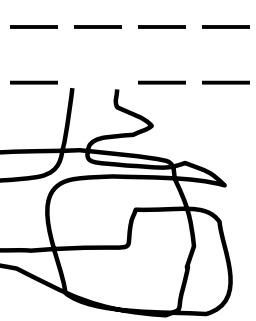
Substitution



Indel



Domain Insertion



Sequence Alignment & Dynamic Programming

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

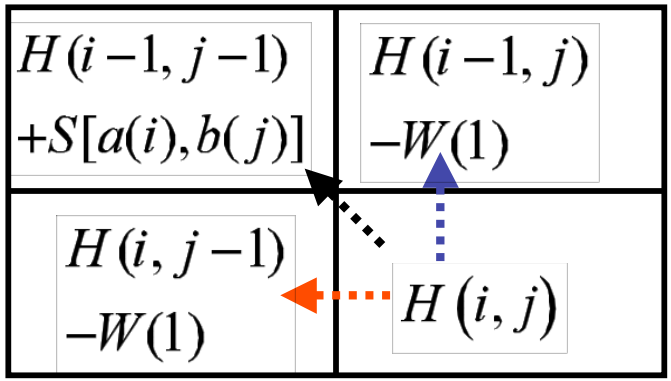
Seq. 2: c_1 - $c_2\ c_3\ c_4\ c_5$ - $\ldots c_m$

number of possible alignments:

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

Needleman-Wunsch alignment algorithm

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



Score Matrix H: Traceback

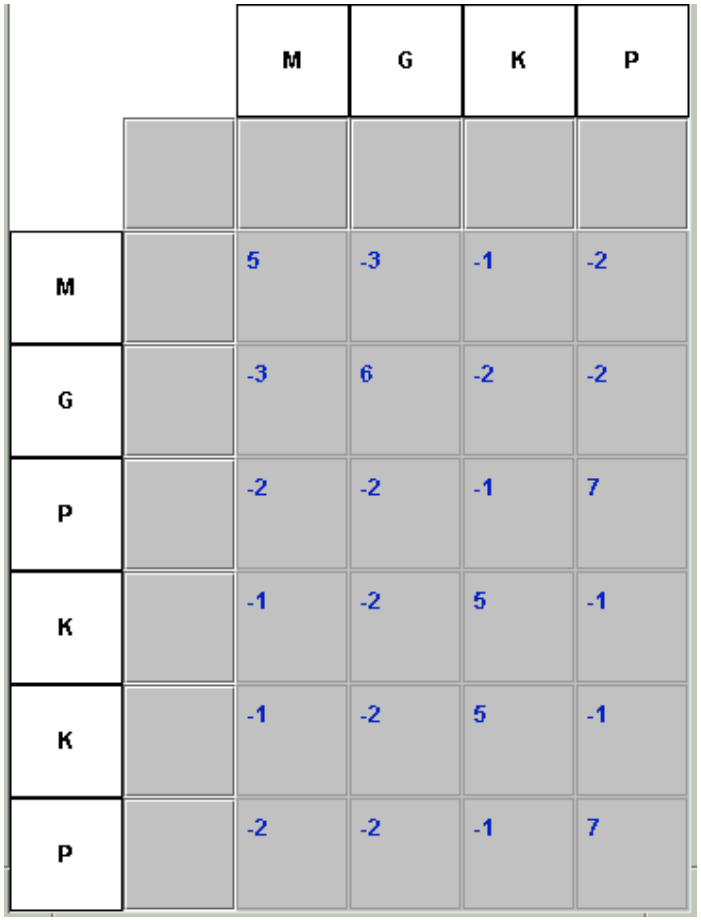
S : similarity matrix

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

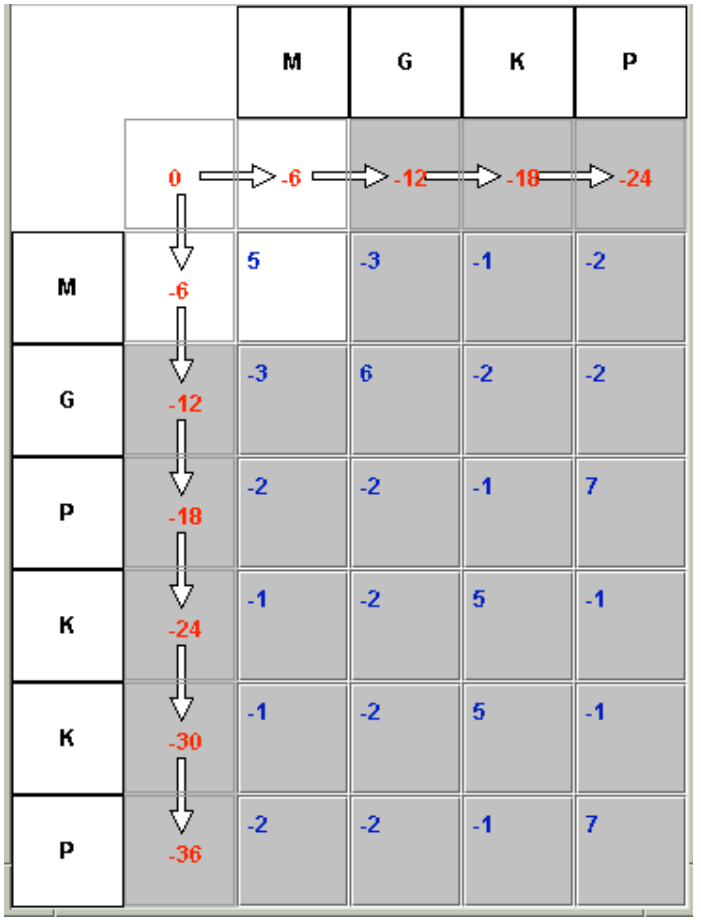
??? Tutorial: W=d

Needleman-Wunsch Global Alignment

Similarity Values

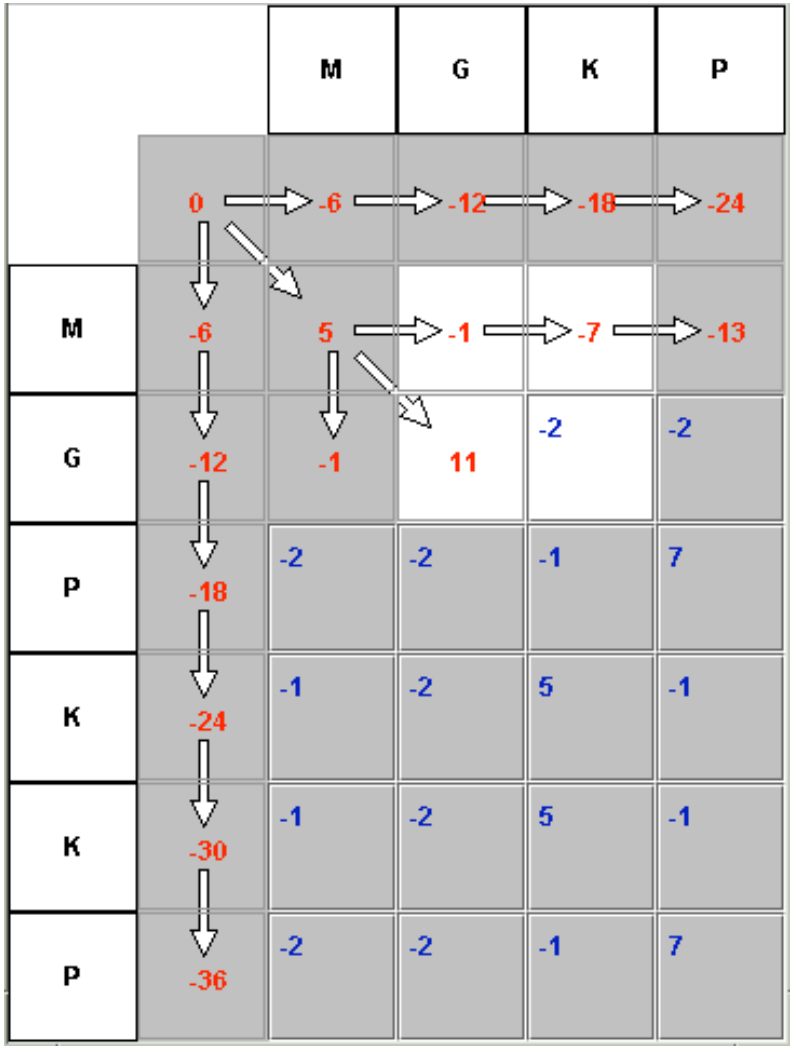


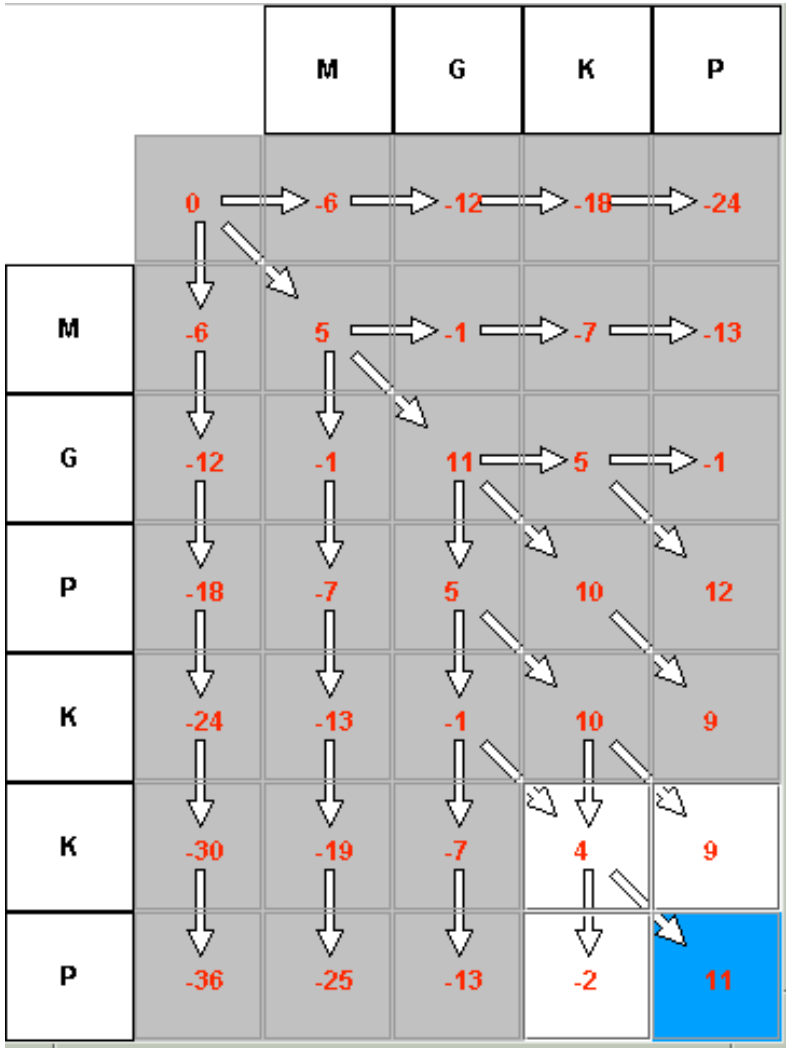
Initialization of Gap Penalties



http://www.dkfz-heidelberg.de/tbi/bioinfo/PracticalSection/AliApplet/index.html

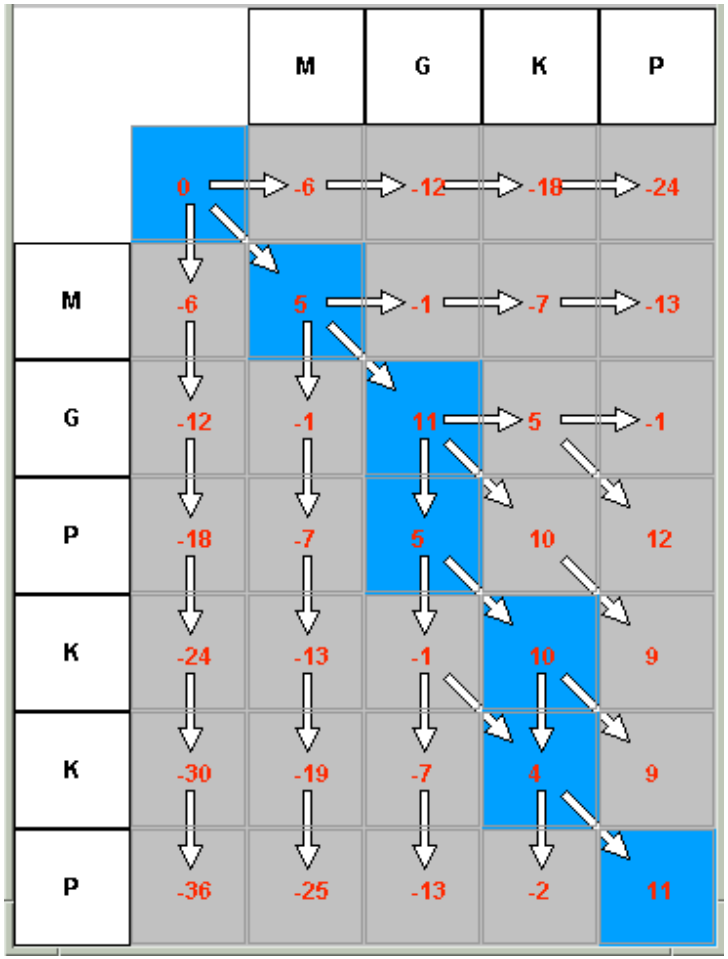
Filling out the Score Matrix H





http://www.dkfz-heidelberg.de/tbi/bioinfo/PracticalSection/AliApplet/index.html

Traceback and Alignment



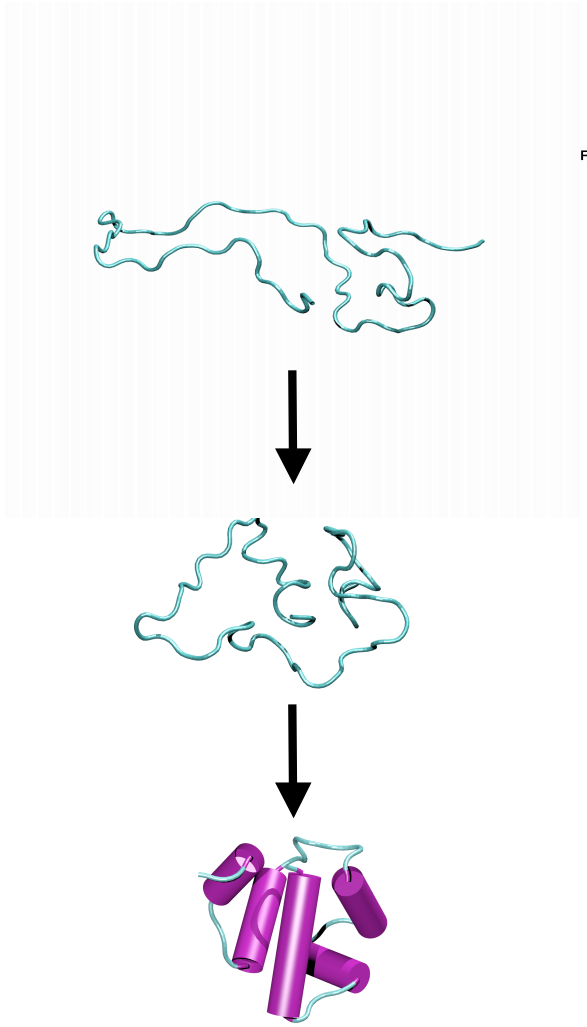
The Alignment

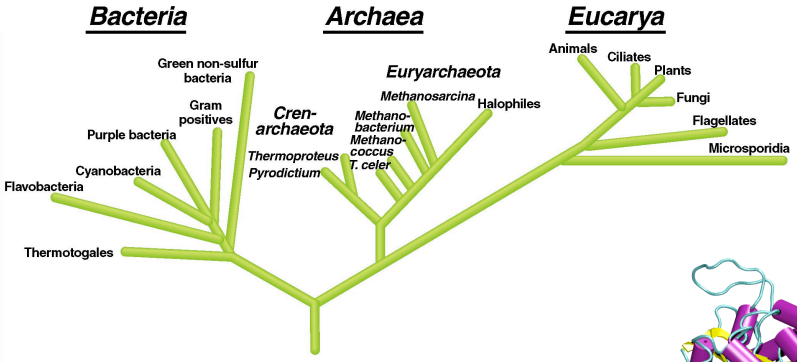
M	G	-	K	-	P
:	:		:		:
M	G	P	K	K	P

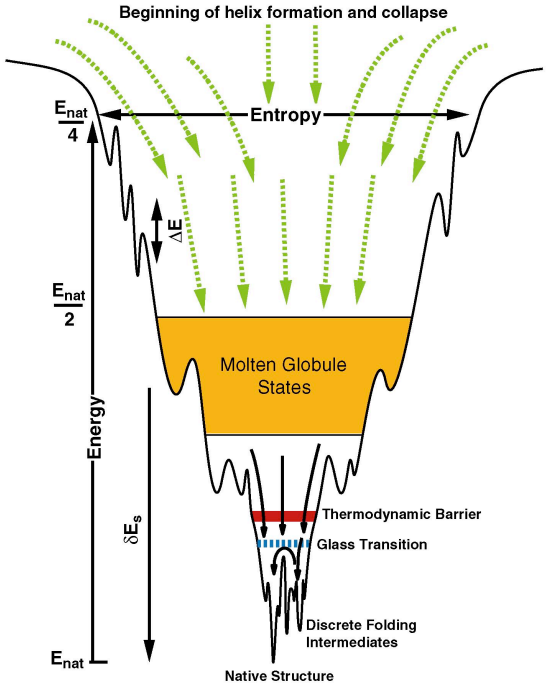
Traceback (blue) from optimal score

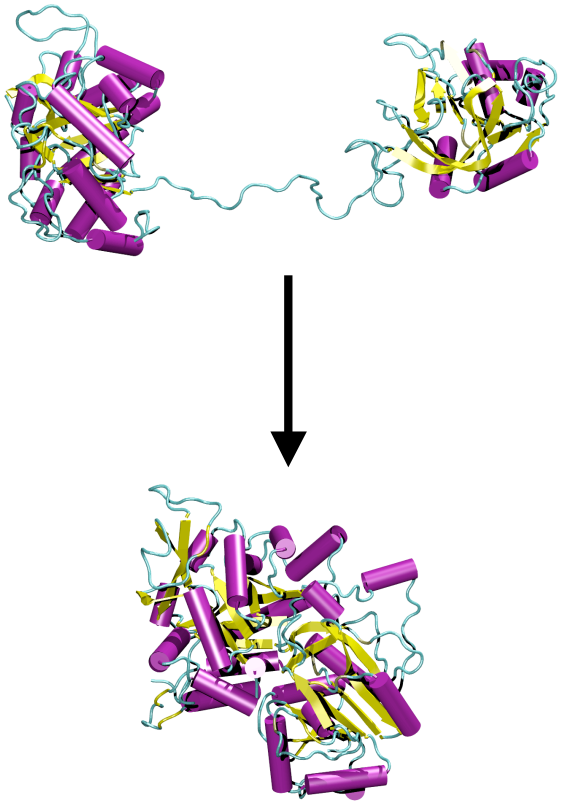
http://www.dkfz-heidelberg.de/tbi/bioinfo/PracticalSection/AliApplet/index.html

Energy Landscape Theory of Structure Prediction









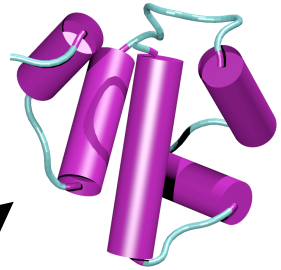
Protein Structure Prediction

1-D protein sequence

SISSIRVKSKRIQLG….

Ab Initio protein folding

3-D protein structure



Sequence Alignment

Target protein of unknown structure → SISSRVKSKRIQLGLNQAELAQKV------GTTQ···

Homologous/analogous protein of **known** structure → QFANEFKVRRIKLGYTQTNVGEALAAVHGS···

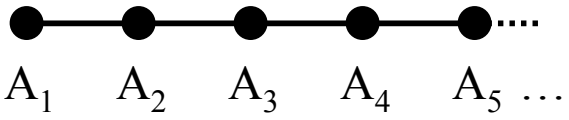
Sequence Alignment: the Energy Function

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

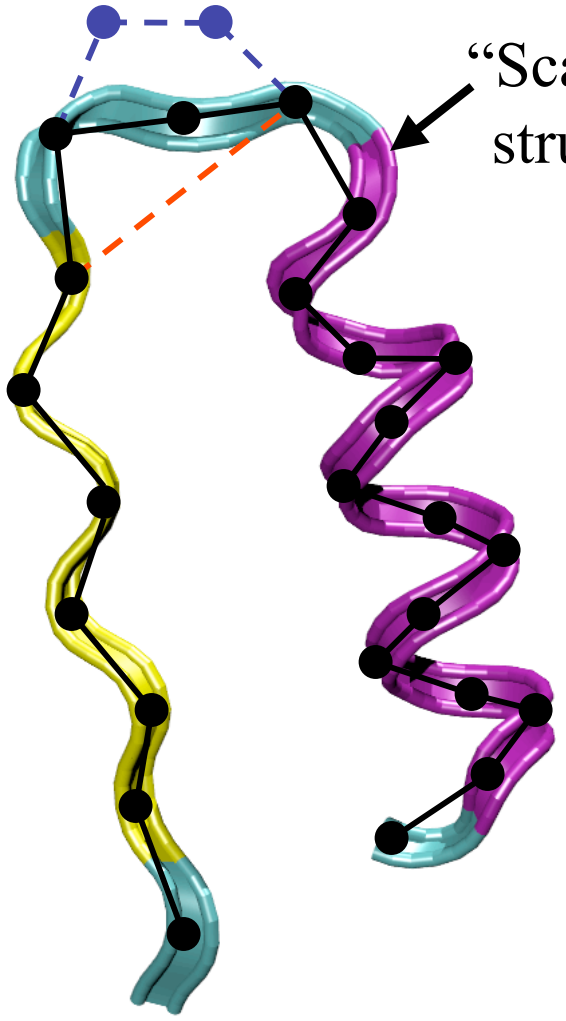
$E_{match} =$

Threading: Sequence-Structure Alignment

Target sequence



threading alignment between target and scaffold



Threading Energy Function

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

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

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

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

R. Goldstein, Z. Luthey-Schulten, P. Wolynes (1992, PNAS)

Gap Penalties

Distribution of Gaps

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

Sequence-Structure Gap Energy

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

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

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

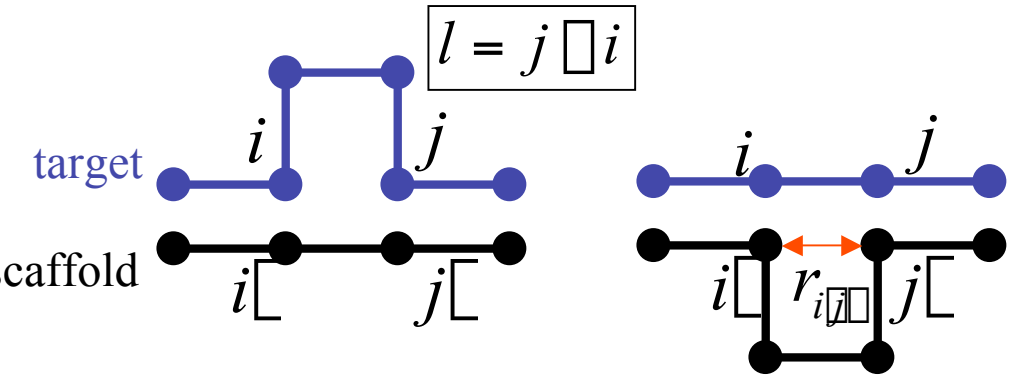
range: $3.0\,\text{Å} < r < 7.5\,\text{Å}$

$$P_{bulge}(l, r) = \frac{a_3}{(\sigma_3 l)^{3/2}} r^2 \exp\left[-c_3 l - \frac{r^2}{\sigma_3 l}\right]$$

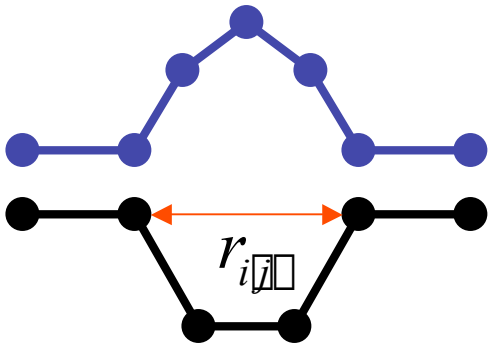
range: $r > 4.0\,\text{Å}$

Insertion

Deletion



Bulge



R. Goldstein, Z. Luthey-Schulten, P. Wolynes (1994) *Proc 27th Annu Hawaii Int Conf Sys Sci*:306.

Similarity Measures

Sequence Identity

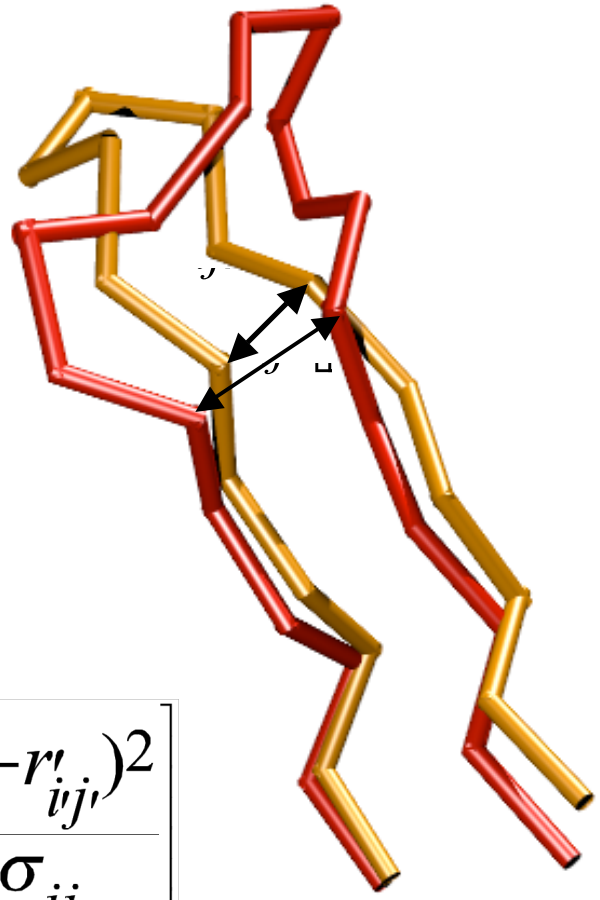
fraction of identically matched residues

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

Q "Structural Identity"

fraction of native contacts

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

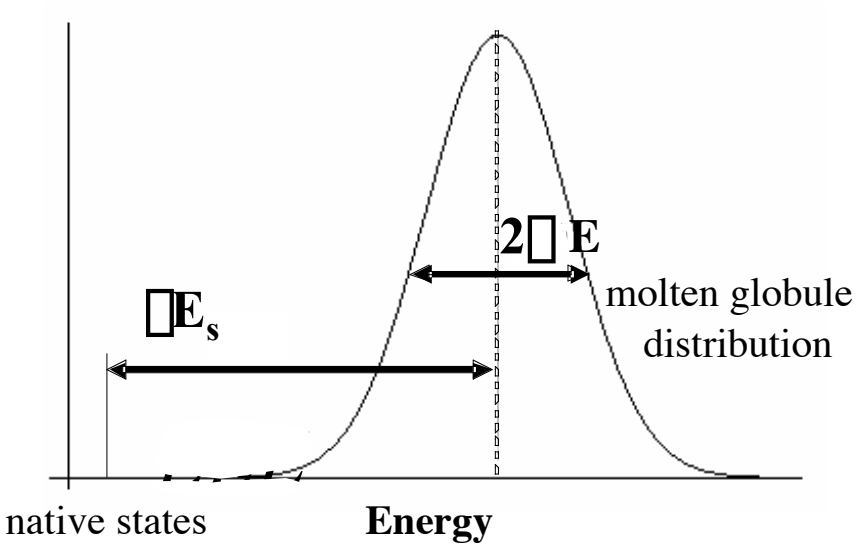


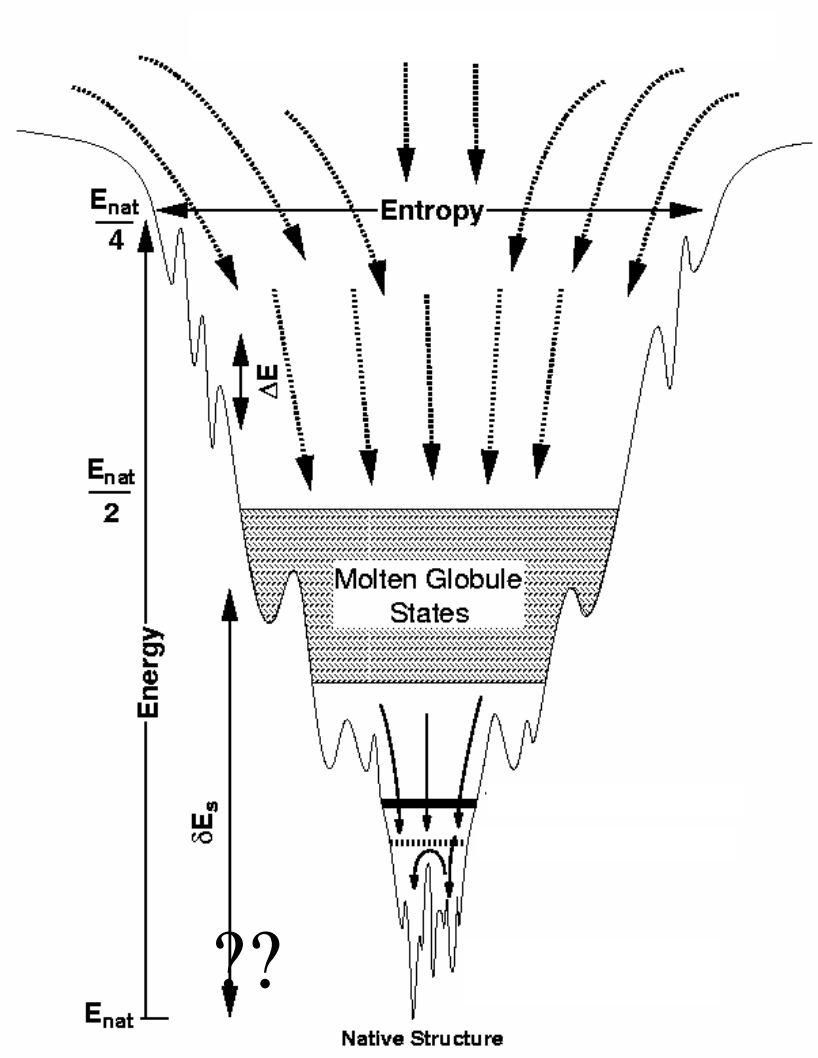
A summary of Energy Landscape Theory

Energy Landscape Theory

When $\langle \delta E_s / \delta E \rangle$ is maximum
the energy landscape is **optimally funneled.**

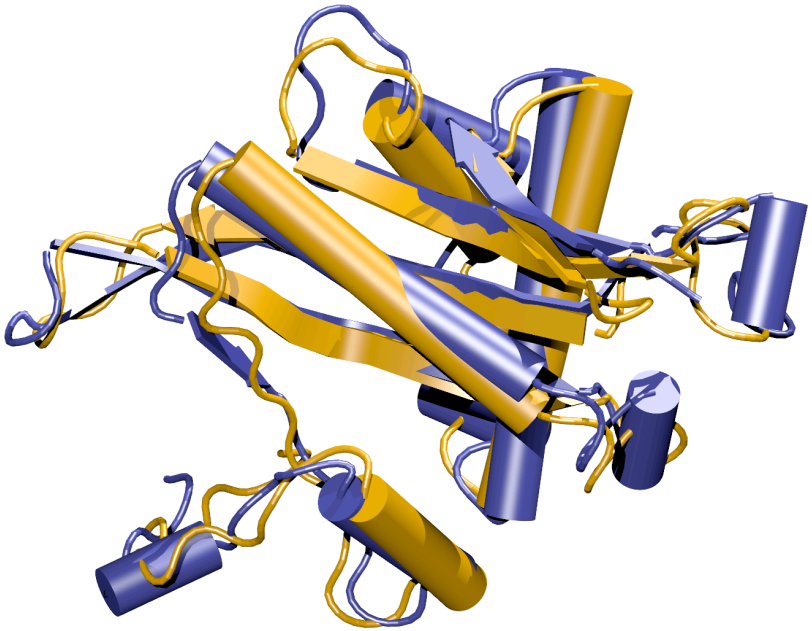
Optimization over an Ensemble of Folds

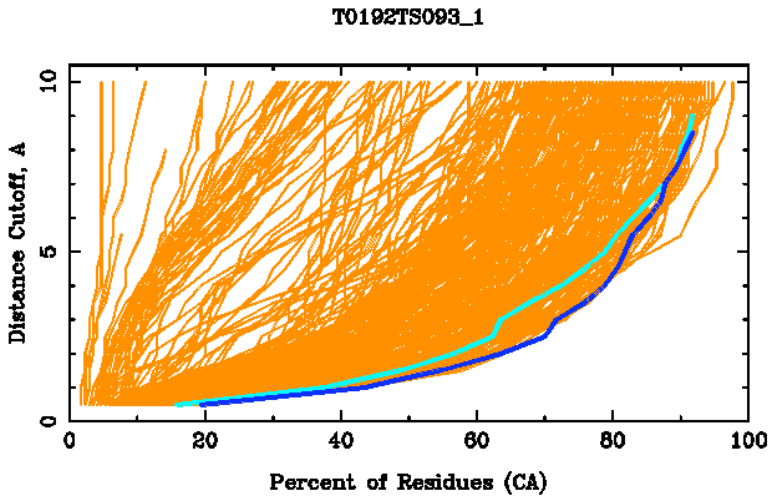




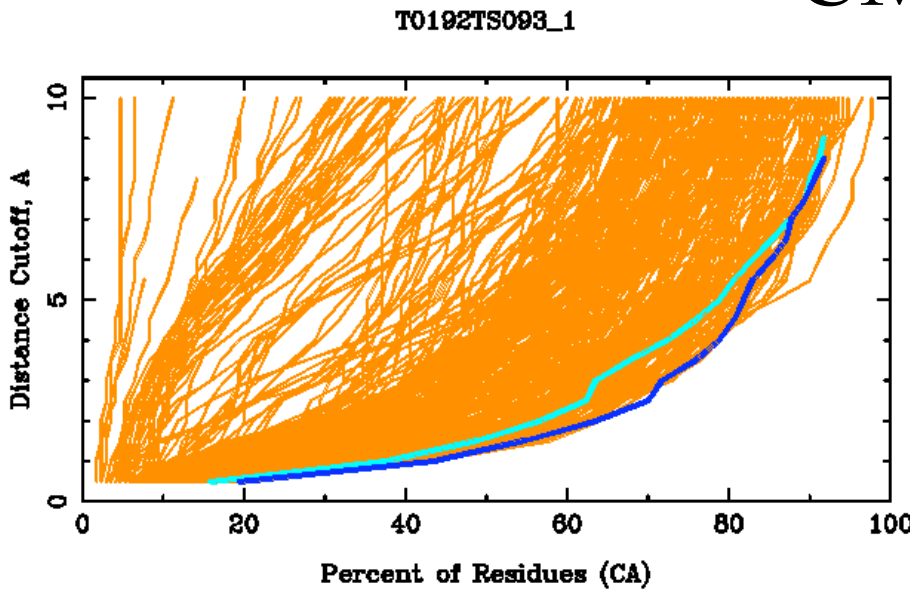
Onuchic , Luthey -Schulten, Wolynes (**1997**) *Annu . Rev. Phys. Chem.* 48:545 -600.
Koretke , Luthey -schulten,Wolynes(**1996)** *Prot. Sci***.** 5:1043

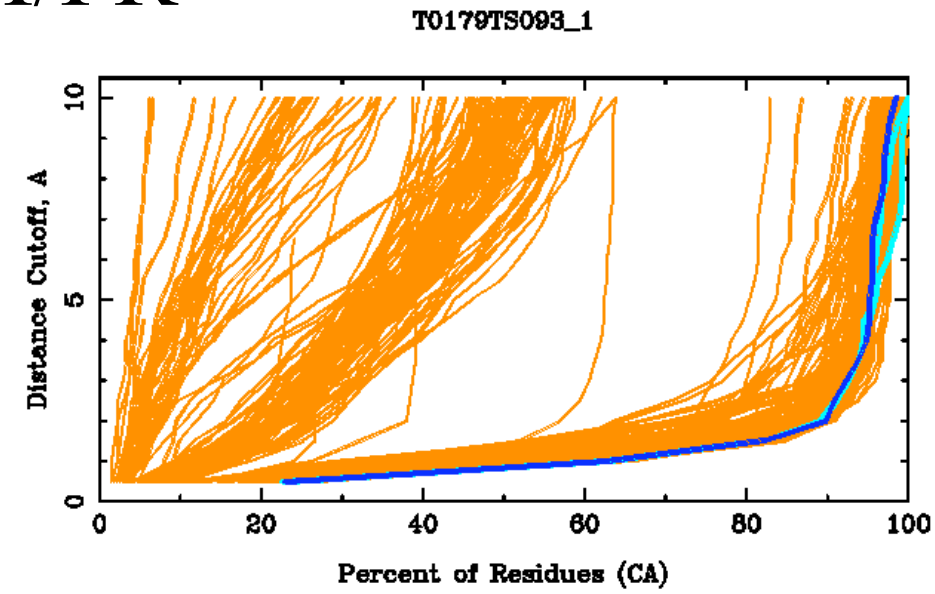
Homology Modeling - Threading

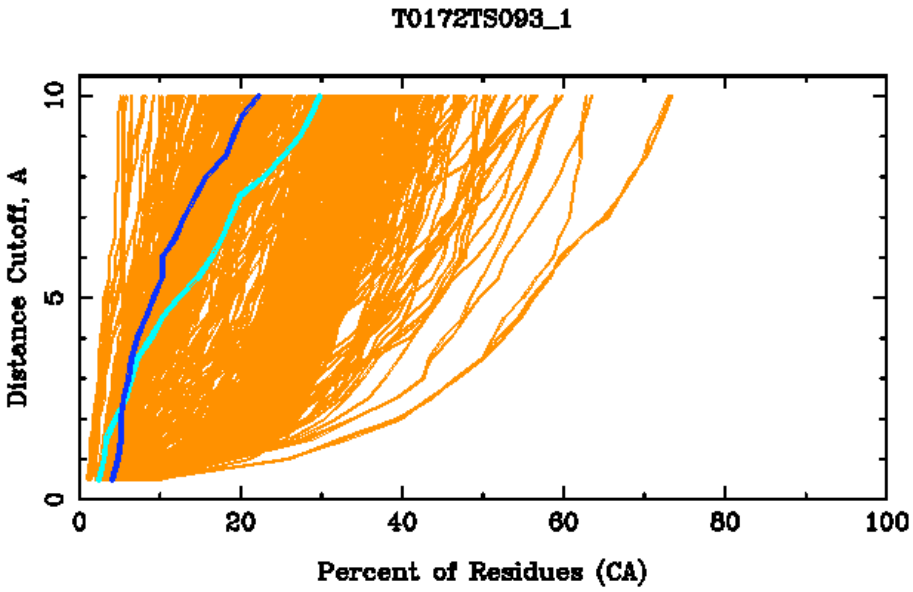




Results from CASP5
CM/FR







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!

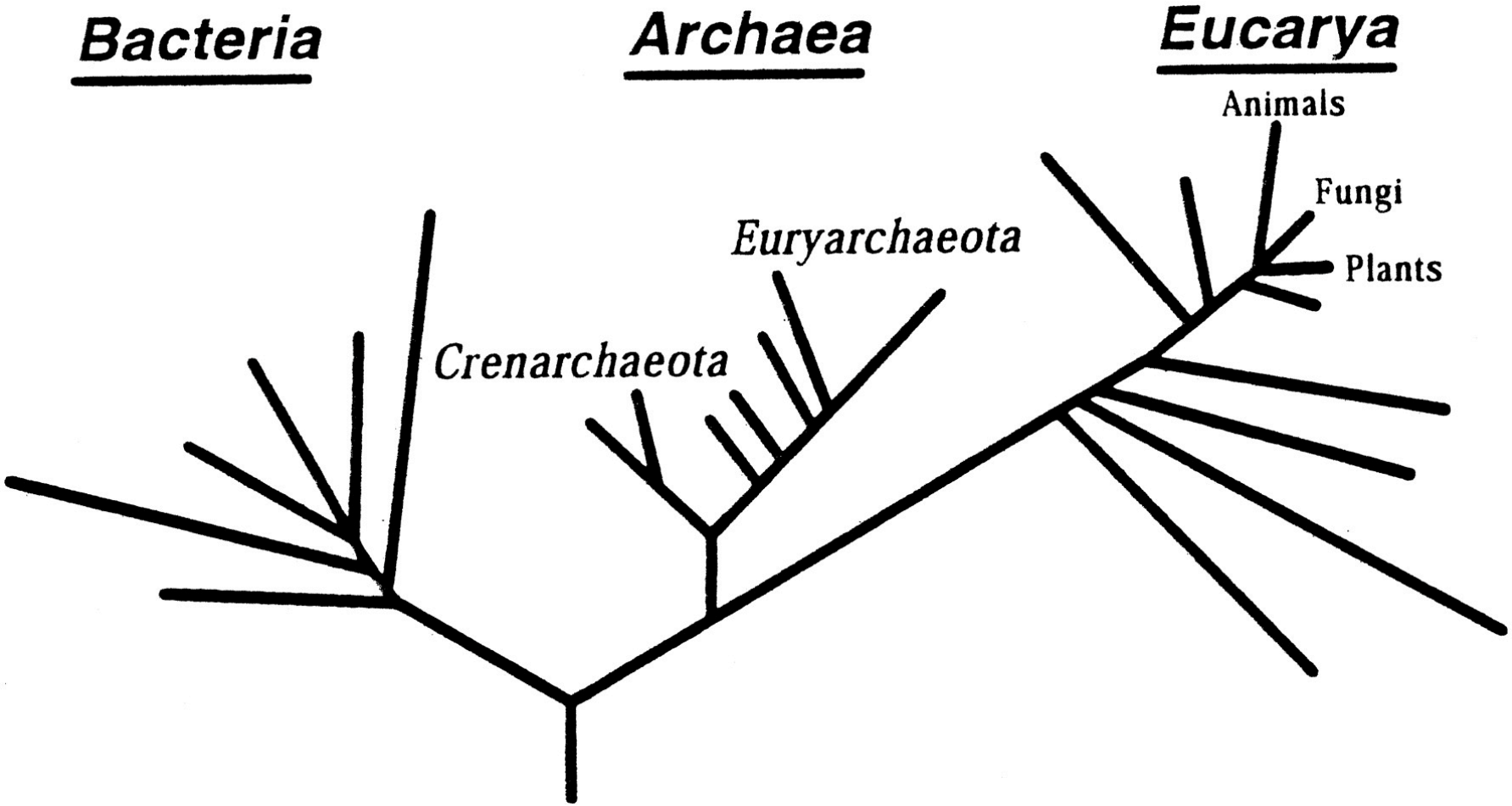
Watch for Bioinformants!!!

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"

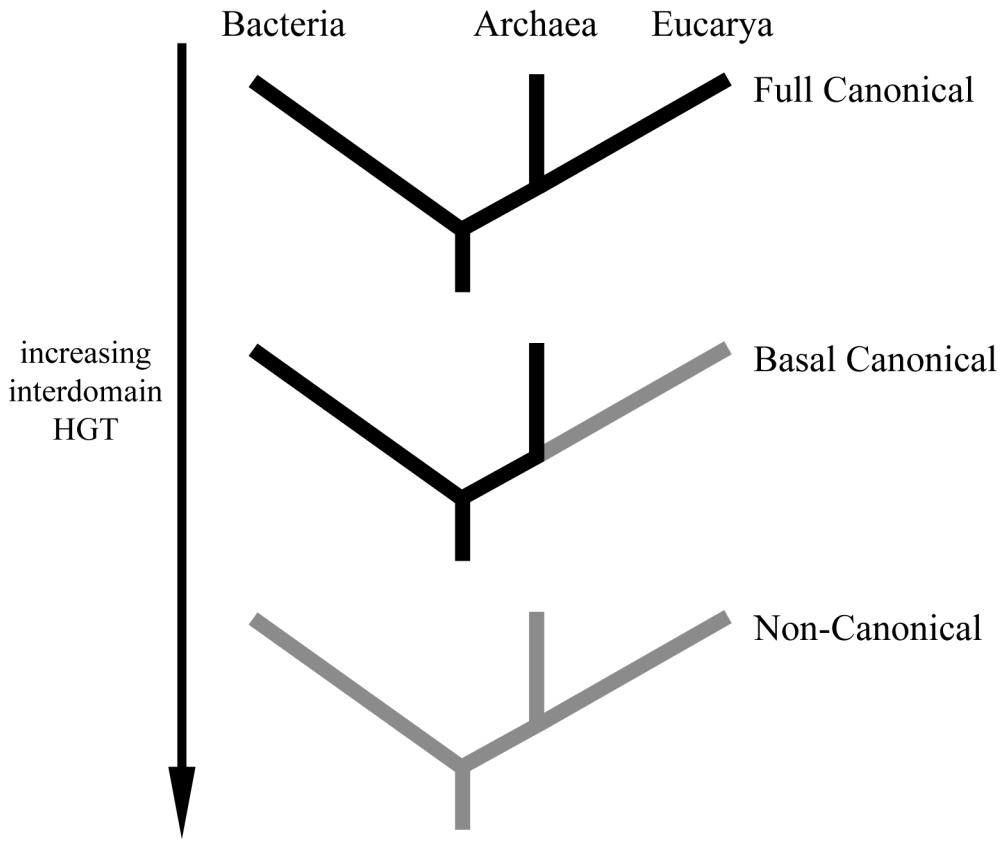
- Woese (PNAS, 2000) Pauling (1965)

Universal Tree



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

Horizontal Gene Transfer



O'Donoghue and Luthey-Schulten, UIUC 2003

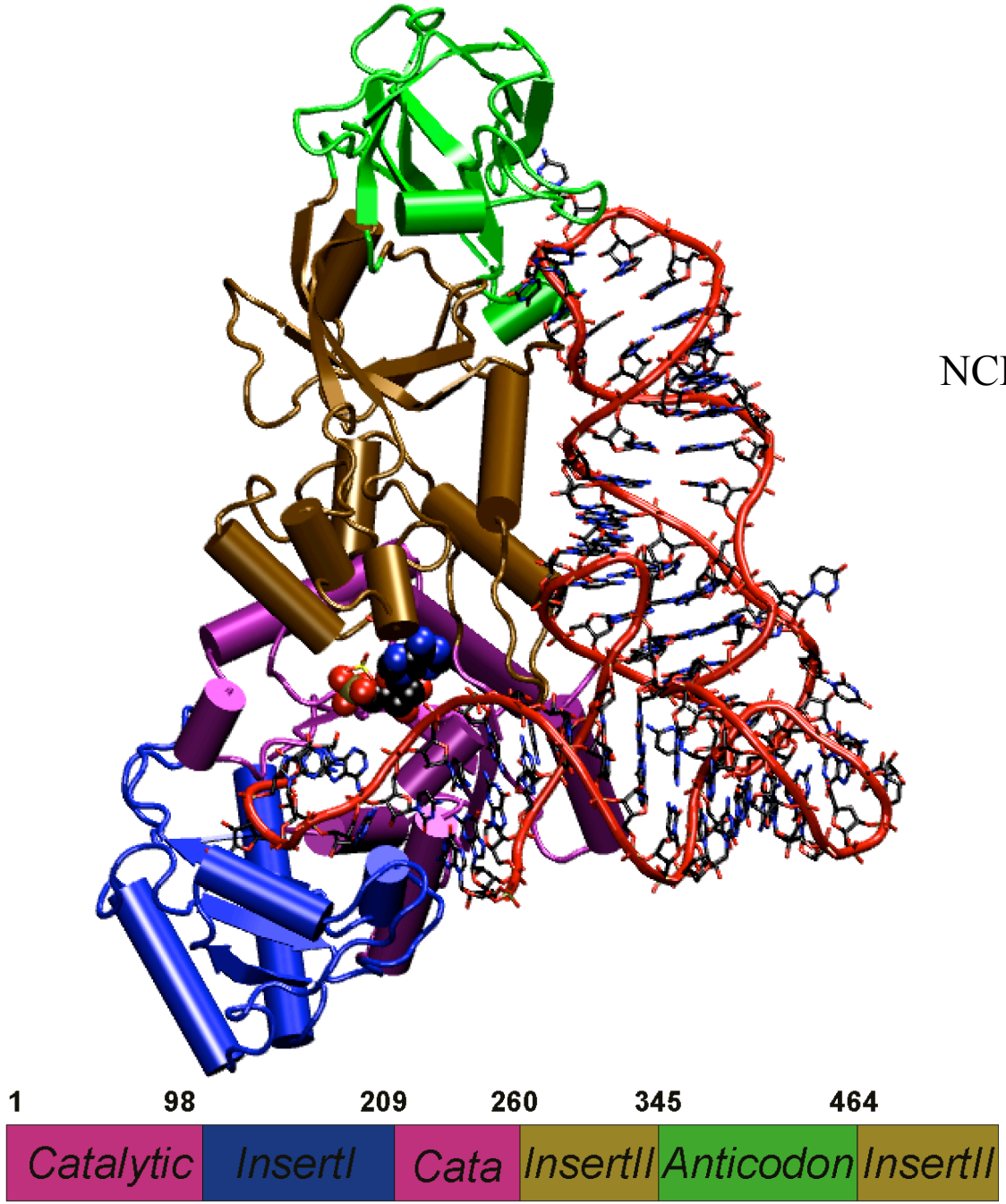
Multiple Sequence Alignments

- “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

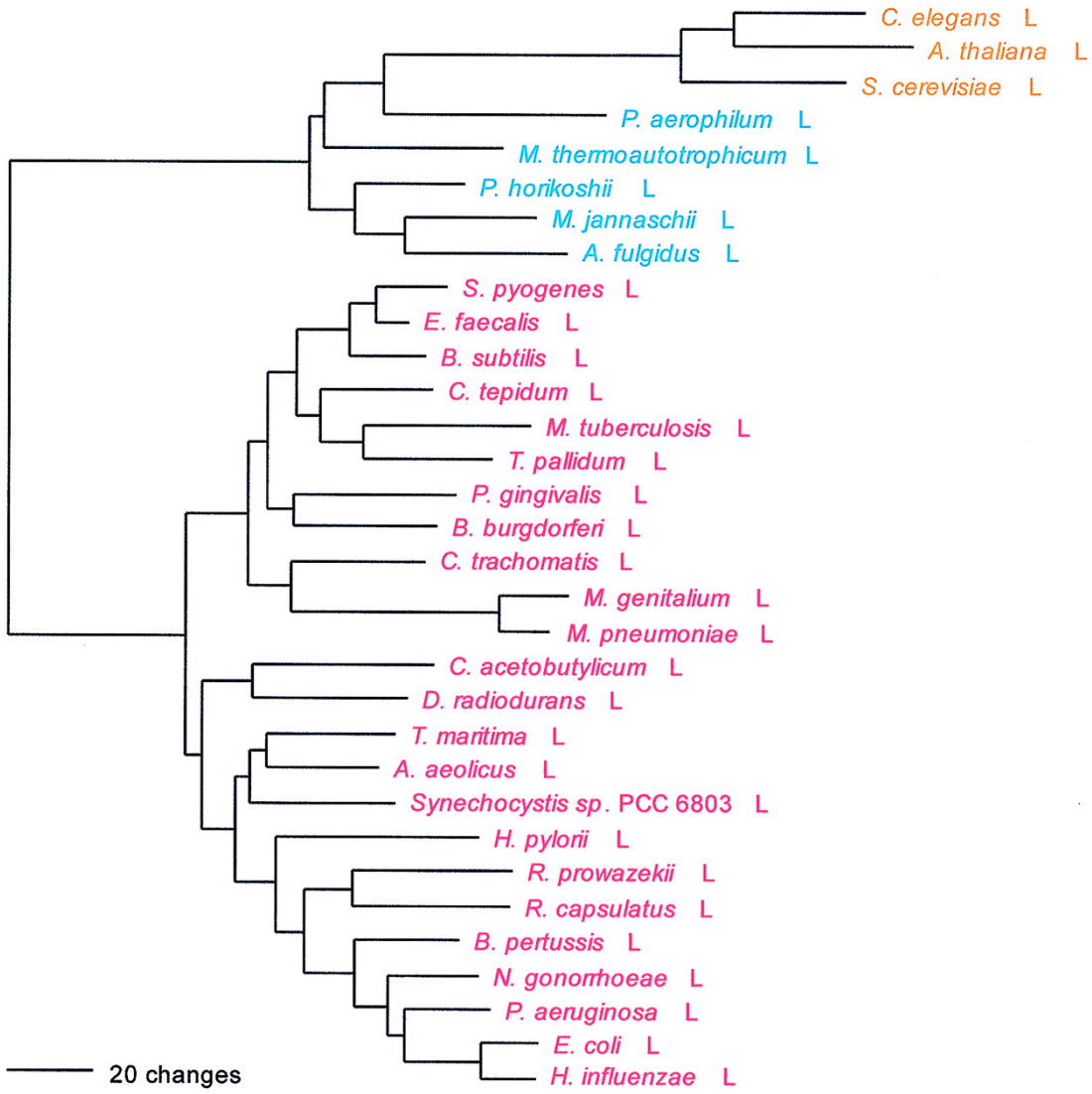
- DNA → RNA → Proteins

- Role of AARS?
- Charging of t-RNA


1
98
209
260
345
464
Catalytic
InsertI
Cata
InsertII
Anticodon
InsertII

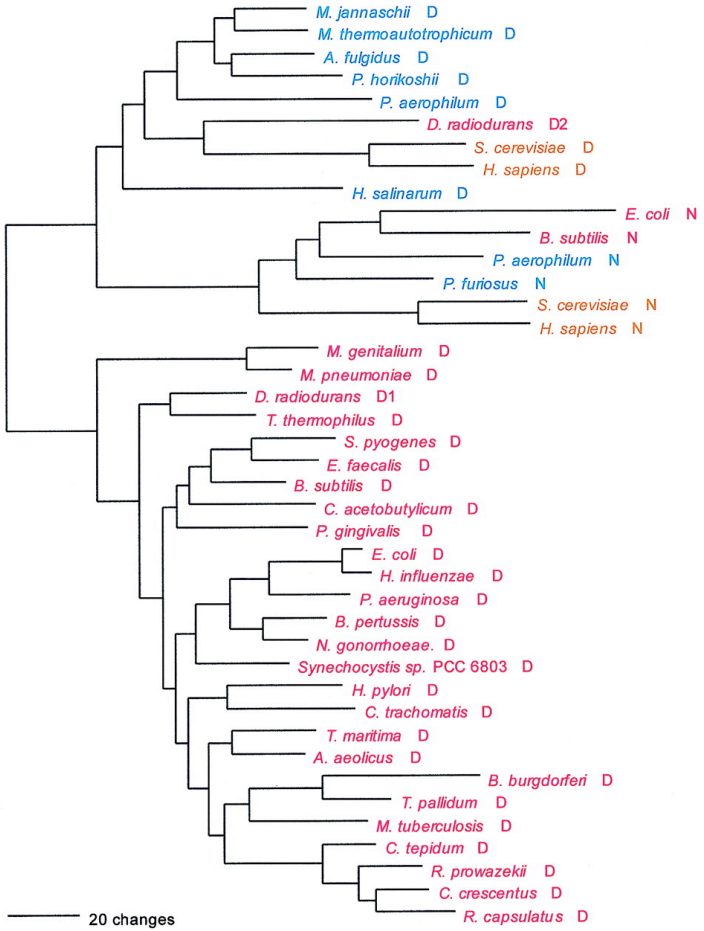
NCBI 3D

LeuRS Canonical Tree



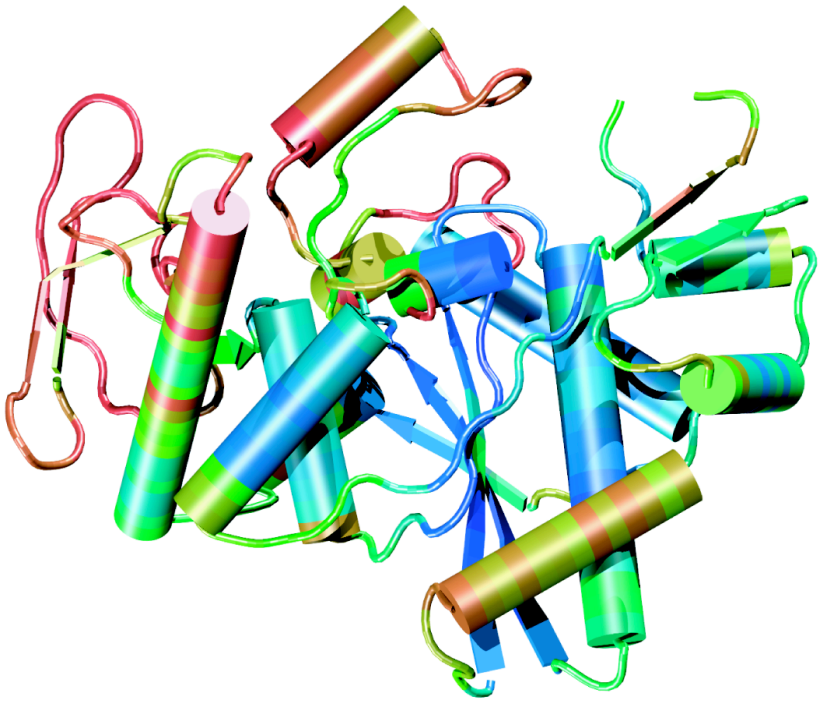
Woese, Olsen (UIUC), Ibba (Panum Inst.), Soll (Yale) *Micro. Mol. Biol. Rev.* March 2000..

D,N Sequence Phylogenetic Trees

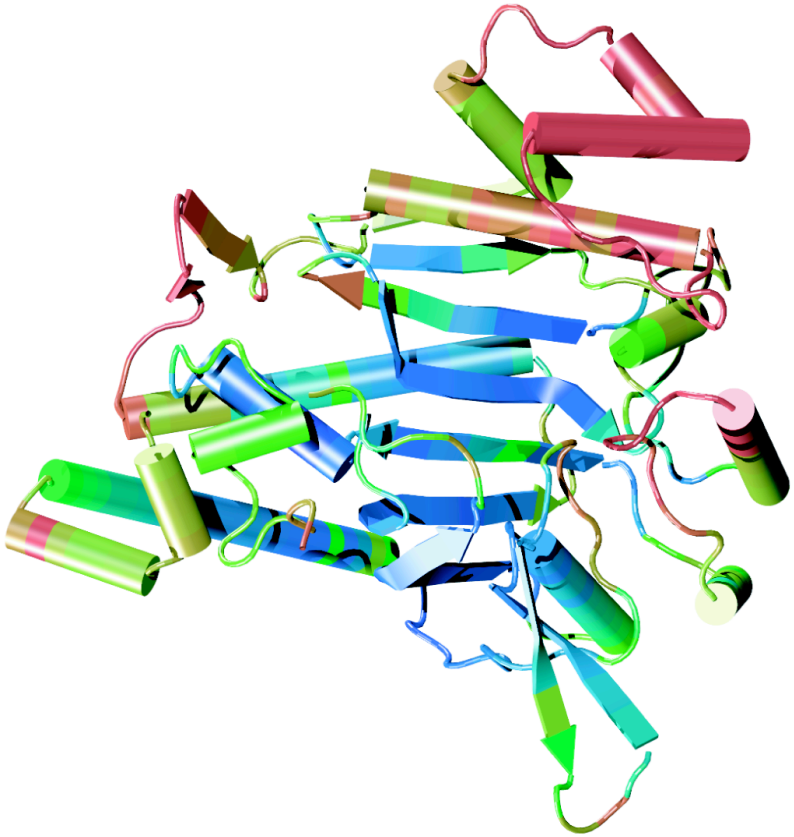


Woese, Olsen (UIUC), Ibba (Panum Inst.), Soll (Yale) *Micro. Mol. Biol. Rev.* March 2000

Fold Motifs of AARSs



Class I Lysyl-tRNA Synthetase



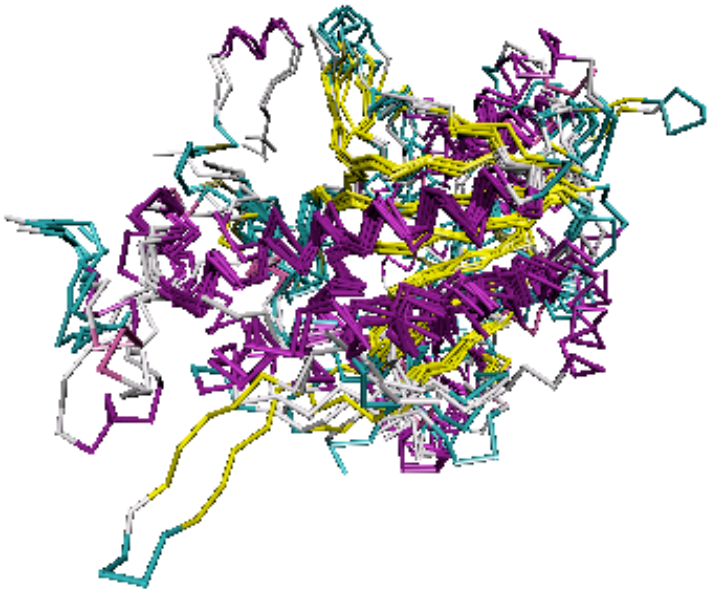
Class II Lysyl-tRNA Synthetase

O'Donoghue and Luthey-Schulten, UIUC 2003

Structure Conserved More than Sequence

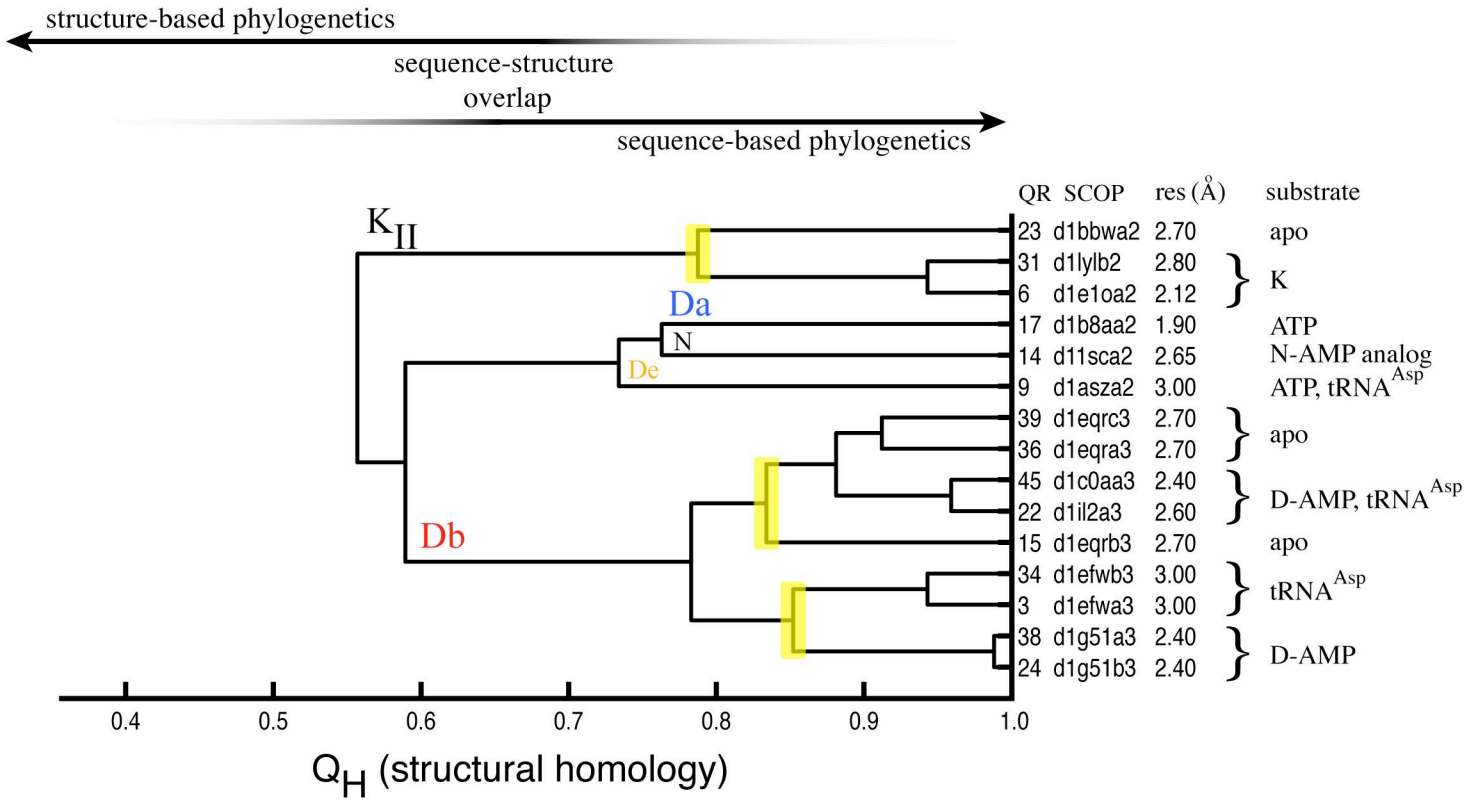
Structural Overlap of Class II AARS

Conserved helices

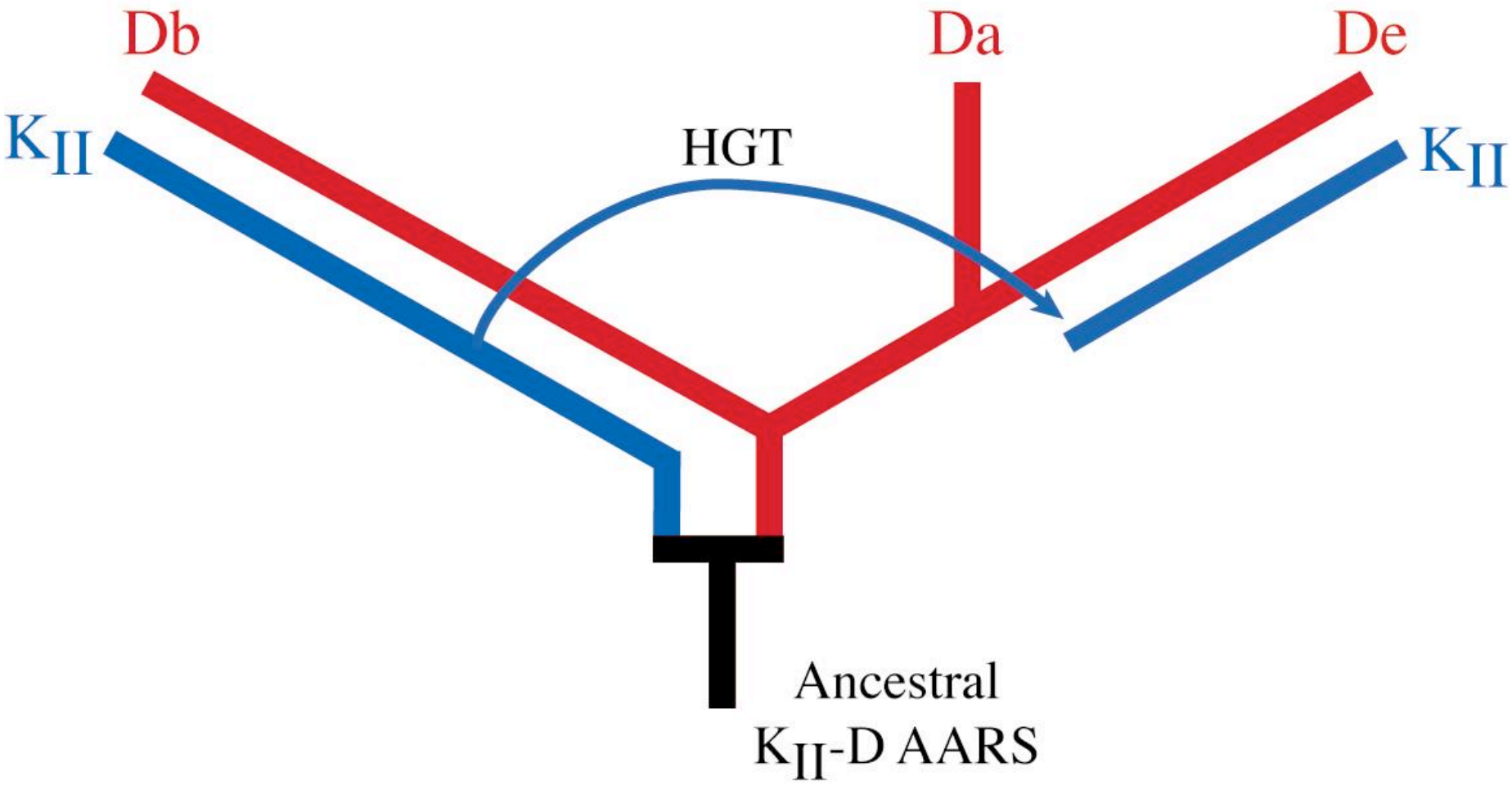


Conserved sheets

Subset of Class II Structural Tree

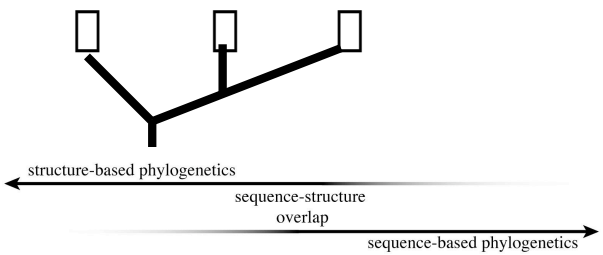


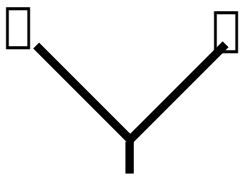
O'Donoghue and Luthey-Schulten, UIUC 2003


Db
Da
De
K_{II}
HGT
K_{II}
Ancestral
K_{II}-D AARS

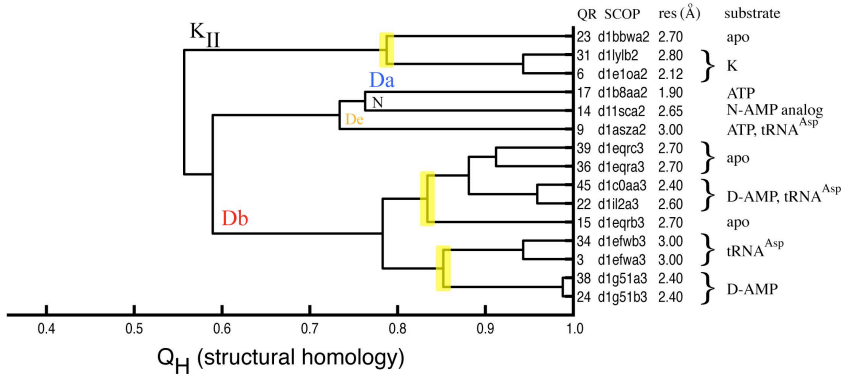
Novel Evolutionary Connections from Sequence and Structure

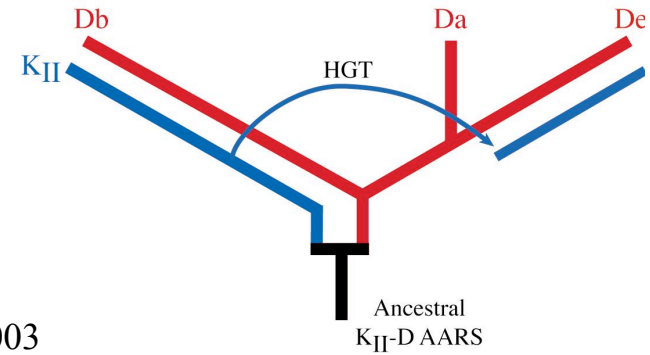
Canonical Pattern	Canonical Pattern + □→□	Basal Canonical	Gemini
D E F L W Y	I H P M	V T A R	K_1 K_2 C S G N Q

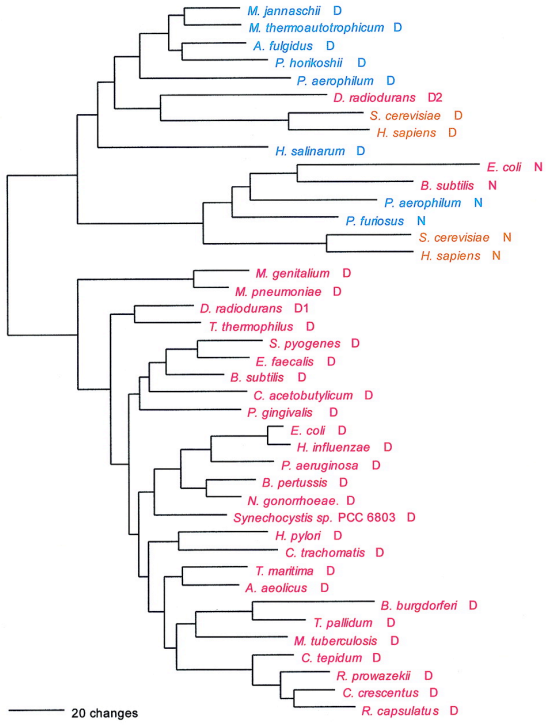




No canonical pattern
Horizontal transfer after
B-AE split.







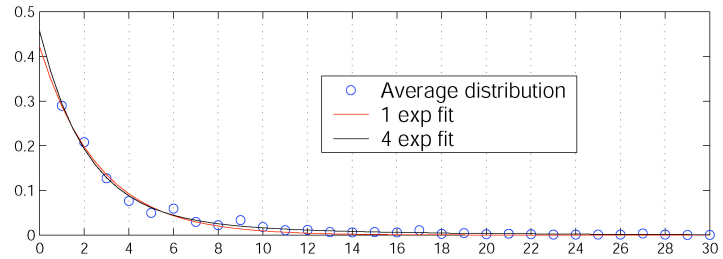
O'Donoghue, Luthey-Schulten, UIUC 2003

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

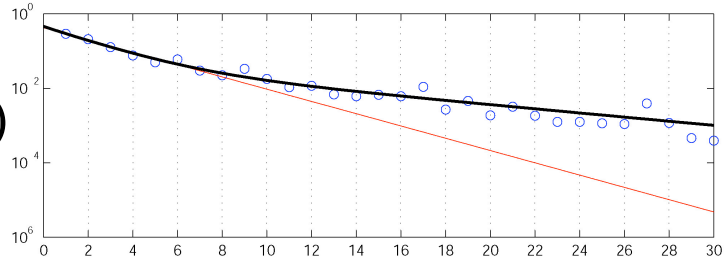
Gap Distribution Functions

Length Gap Distribution Function

$P(l)$



$\log(P(l))$



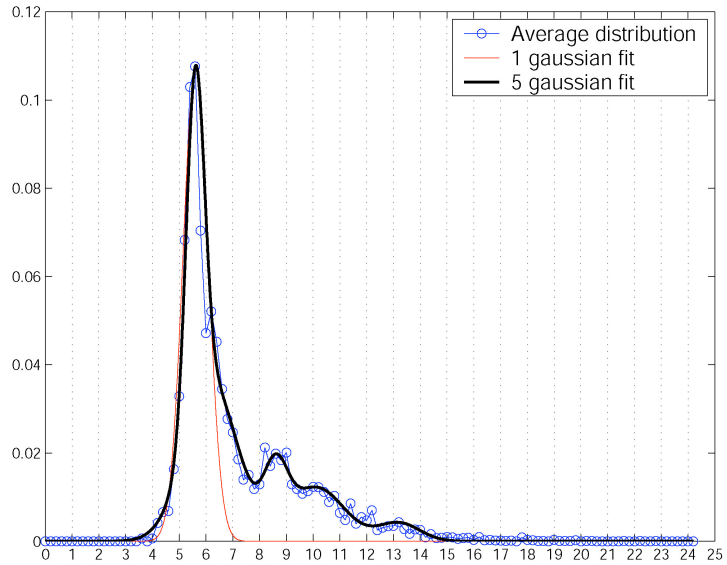
l, gap length (residues)

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

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

Spatial Gap Distribution Funciton

$P(r)$



r_{ij}, spatial gap distance (Å)

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

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

B. Qian & R. Goldstein. (2001) *Proteins* **45**:102.

Structural Alignment Methods

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

Multiple Structural Alignments

STAMP

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.

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

Multiple Structural Alignments

STAMP – cont'd

2. Refine Initial Alignment & Produce Multiple Structural Alignment

Alignment score:

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

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

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

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

Multiple Alignment:

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

Stamp Output/Secondary Structure

O'Donoghue and Luthey-Schulten, UIUC 2003

Stamp Output/Clustal Format

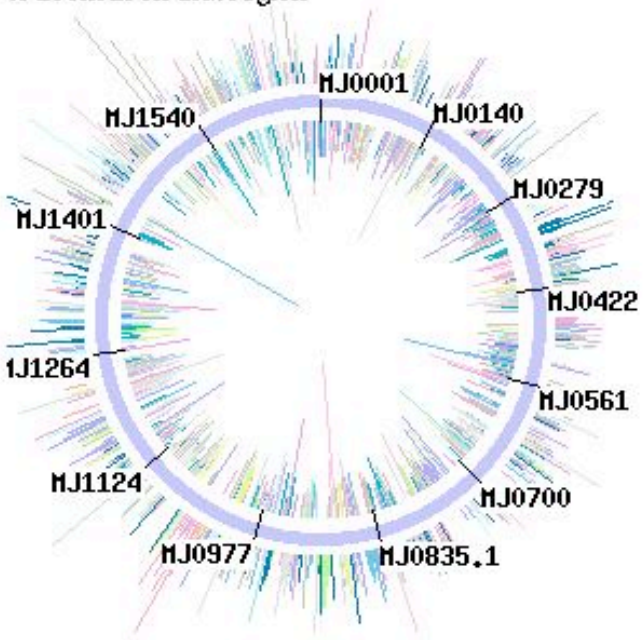
```
SerRS-T_thermophilus          VGGEEANREIKRVGGPPEFSFP--P--LDHVALMEKNGWWEPRISQVSGSRSYALKGDLA
ThrRS-E_coli                  -------------------------R--DHRKIGKQLDLY-HMQ-EE-APGMVFWHNDGW
ProRS-T_thermophilus          ----------------KGLTPQSQDFSEWYLEVIQKAELAD-YG--P-VRGTIVVRPYGY
ProRS-M_thermoautotrophicus   ------------------------EFSEWFHNILEEAEIIDQRY--P-VKGMHVWMPHGF
space                         ------------------------------------------------------------
SerRS-T_thermophilus          --SGGG-EEEEEES-----SS--------HHHHHHHHT-B-TTHHHHH-SS---B-THHH
ThrRS-E_coli                  ----------------------------HHHHHHHTT-E-E---TT-STT--EE-HHHH
ProRS-T_thermophilus          --------------------HHHHHHHHHHHHHHHHTTSEE-E---S-STT-EEE-HHHH
ProRS-M_thermoautotrophicus   ------------------------HHHHHHHHHHHTT-EE-----S-STT--EE-HHHH

SerRS-T_thermophilus          LYELALLRFAMDFMARRGFLPMTLPSYAREK-AFLG-TGHFPAYRDQVWAIA-----E--
ThrRS-E_coli                  TIFRELEVFVRSKLKEYQYQEVKGPFMMDRV-LWEKT-GHWDNYKDAMFTTS----S-EN
ProRS-T_thermophilus          AIWENIQQVLDRMFKETGHQNAYFPLFIPMSFL----------FSPELAVVTHAGGEELE
ProRS-M_thermoautotrophicus   MIRKNTLKILRRILD-RDHEEVLFPLLVPEDE-LAKEAIHVKGFEDEVYWVTHGGLSKLQ
space                         ------------------------------------------------------------
SerRS-T_thermophilus          HHHHHHHHHHHHHHHHTT-EEEE--SEEEHH-HHHH-HT-TTTTGGGS-B-T-----T--
ThrRS-E_coli                  HHHHHHHHHHHHHHHHTT-EE----SEEEHH-HHHTT-THHHHHGGG--EEE----E-TT
ProRS-T_thermophilus          HHHHHHHHHHHHHHHHTT-EE----SEEESTT-------------TT--EEEE-SSSEEE
ProRS-M_thermoautotrophicus   HHHHHHHHHHHHHHT-TT-EE----SEEEHHH-HTTSHHHHHHTTTT--EEEEETTEEEE

SerRS-T_thermophilus          TDLYLTGTAEVVLNALHSGEILPYEALPLRYAGYAPAFRSEA--GSFGKDVRGLMRVH-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-THHHHHHHTTT-EEEGGG-SEEEEEEEEE------S--SSTTTTTTS-S-E
ThrRS-E_coli                  EEEEE-S-SHHHHHHHHTSS--BTTT-SEEEEE--EEE------G--G-G-BTTTB-S-E
ProRS-T_thermophilus          EEEEE-S-SHHHHHHHHHHH--BGGG--EEEEEEEEEE---------S-S-BTTTB-SE-
ProRS-M_thermoautotrophicus   EEEEE-SSSHHHHHHHHHHH--BTTT--EEEEEEEEEE---------S---BTTTB-SEE
```

O'Donoghue and Luthey-Schulten, UIUC 2003

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

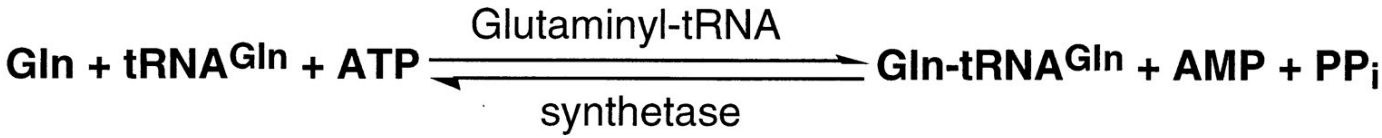
- Translation, ribosomal structure and biogenesis
- Transcription
- DNA replication, recombination and repair
- Cell division and chromosome partitioning
- Posttranslational modification, protein turnover
- Cell envelope biogenesis, outer membrane
- Cell motility and secretion
- Inorganic ion transport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

Organism: Methanocaldococcus jannaschii
Genetic Code: 11
Lineage: Archaea; Euryarchaeota; Methanococci; Methanococcales; Methanocaldococcaceae; Methanocaldococcus.

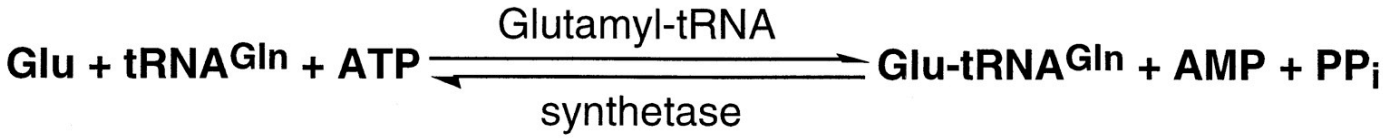
Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii
Bult,C.J., White,O., Olsen,G.J., Zhou,L., Fleischmann,R.D., Sutton,G.G., Blake,J.A., FitzGerald,L.M., Clayton,R.A., Gocayne,J.D., Kerlavage,A.R., Dougherty,B.A., Tomb,J.-F., Adams,M.D., Reich,C.I., Overbeek,R., Kirkness,E.F., Weinstock,K.G., Merrick,J.M., Glodek,A., Scott,J.D., Geoghagen,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

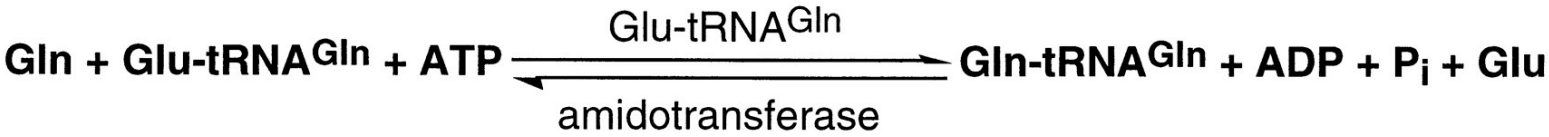
Charging the tRNA

Direct acylation

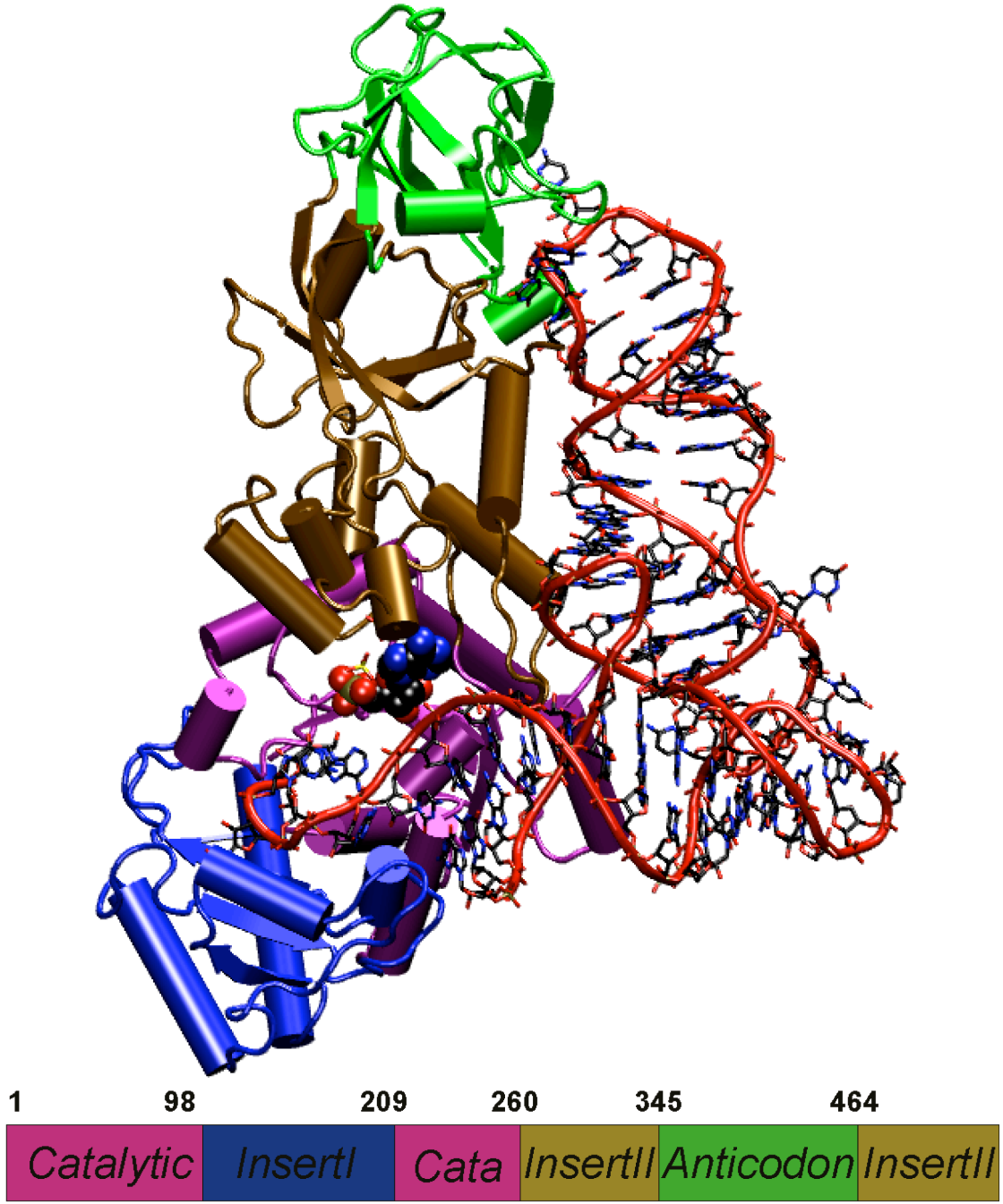


tRNA-dependent amino acid modification





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


1
98
209
260
345
464
Catalytic
InsertI
Cata
InsertII
Anticodon
InsertII

NCBI 3D

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

Cross-references

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

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--trnaligase) (asprs)

[556 residues]

SYD_CAEEL [Caenorhabditis elegans] aspartyl-trna synthetase (ec 6.1.1.12) (aspartate--trna ligase)(asprs)

[531 residues]

SYD_HUMAN [Homo sapiens (human)] aspartyl-trna synthetase (ec 6.1.1.12) (aspartate--trna ligase)(asprs)

[501 residues]

SYD_RAT [Rattus norvegicus (rat)] aspartyl-trna synthetase (ec 6.1.1.12) (aspartate--trna ligase)(asprs)

[501 residues]

CLUSTALW: Multiple Sequence Alignment[help]

General Setting Parameters:

Output Format: CLUSTAL

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

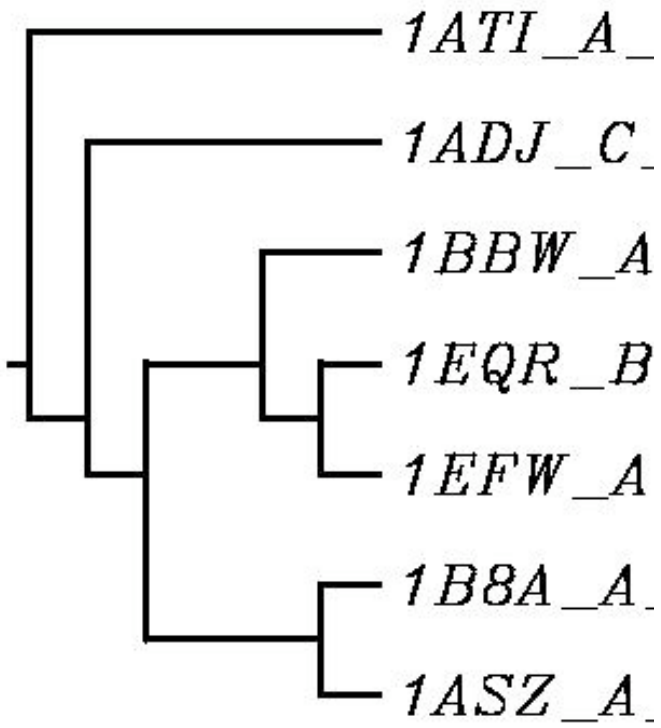
```
>1EQR:B(000:0-000)
VLPLDSNHVNTEEARLKYRYLDLRRPEMAQRLKTRAKITSLVRRFMDDHGFLDIETPMLT
KATPEGARDYLVPSRVHKGKFYALPQSPQLFKQLLMMSGFDRYYQIVKCFRDEDLRADRQ
PEFTQIDVETSFMTAPQVREVMEALVRHLWLEVKGVDLGDFPVMTFAEAERRYGSDKPDL
RDESKWAPLWVIDFPMFEDDGEGGLTAMHHPFTSPKDMTAAELKAAPENAVANAYDMVIN
```

Or give the file name containing your query

Browse...

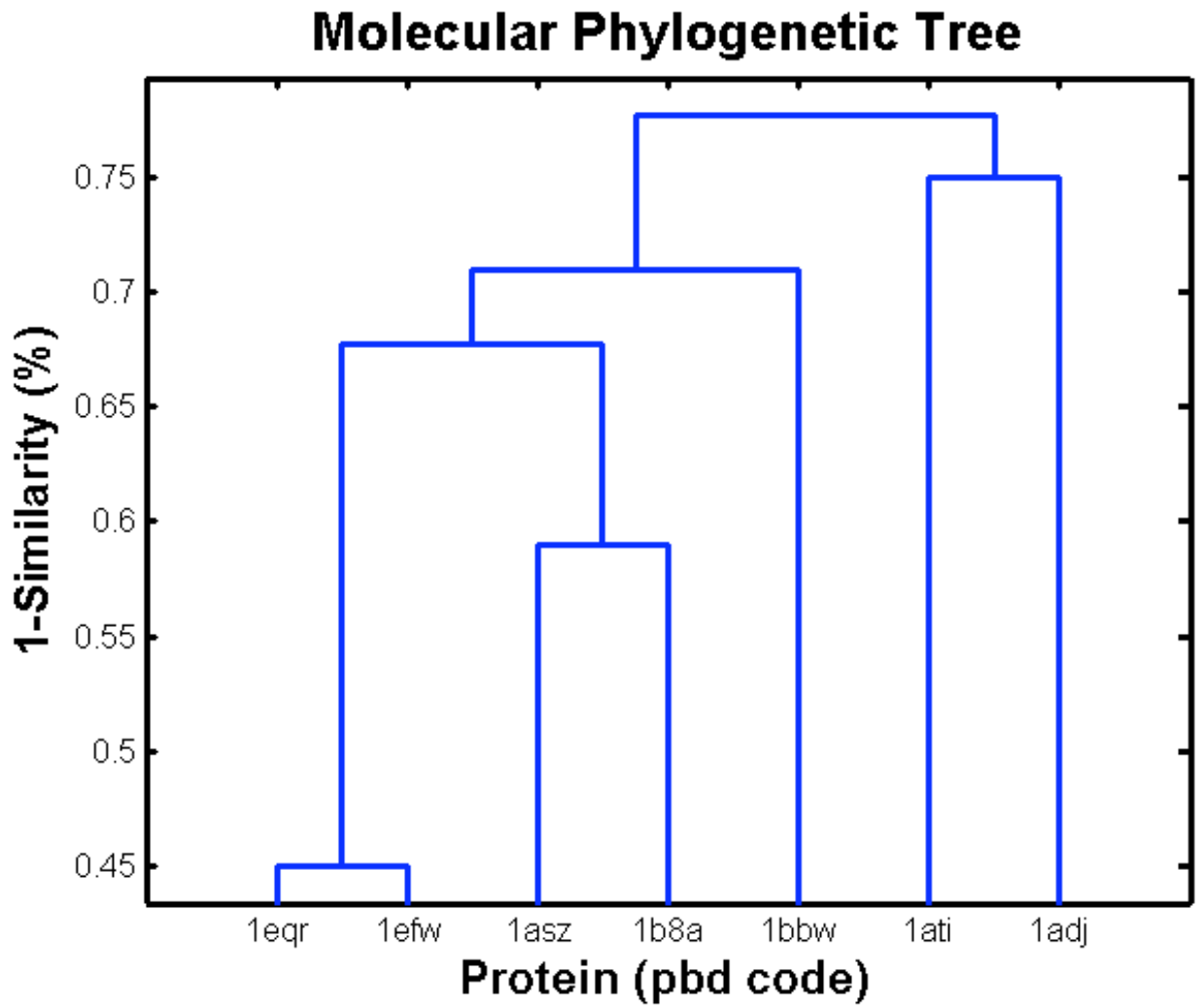
Execute Multiple Alignment | Reset

Sequence Dendrogram from Clustal



Luthey-Schulten, UIUC 2003

Phylogenetic Tree in Tutorial



Pogorelov and Luthey-Schulten, UIUC 2003

Structural Classification of Proteins

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

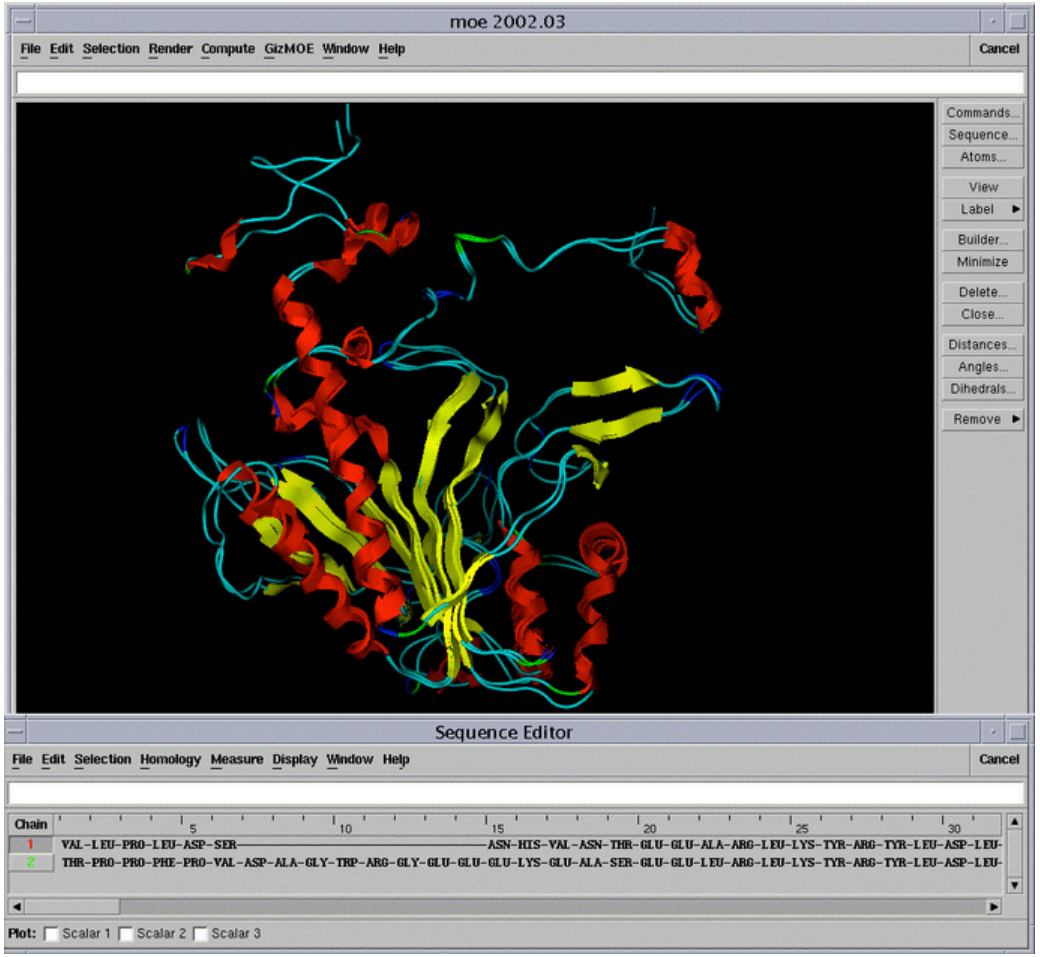
Lineage:

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

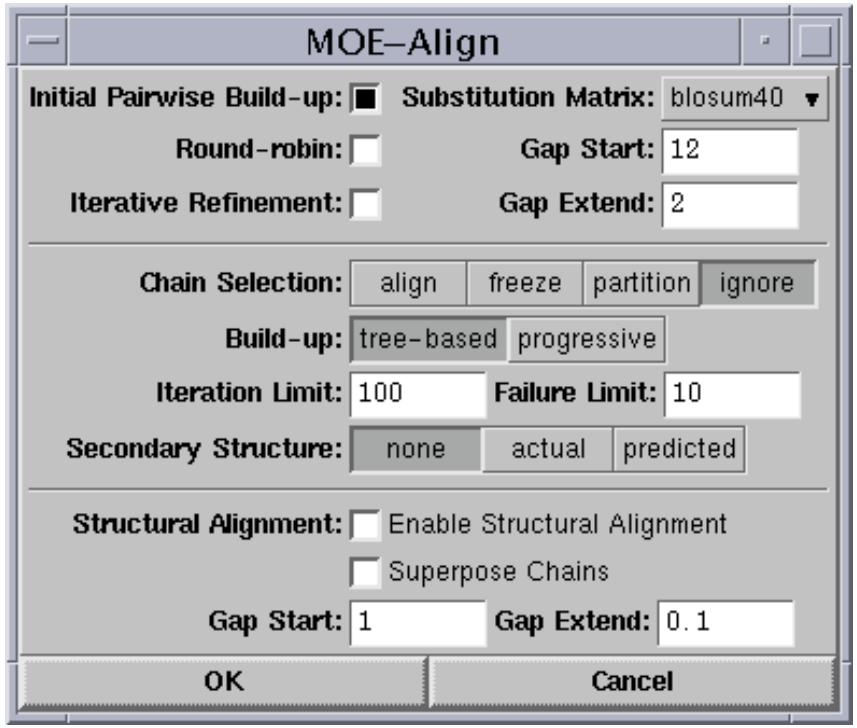
PDB Entry Domains:

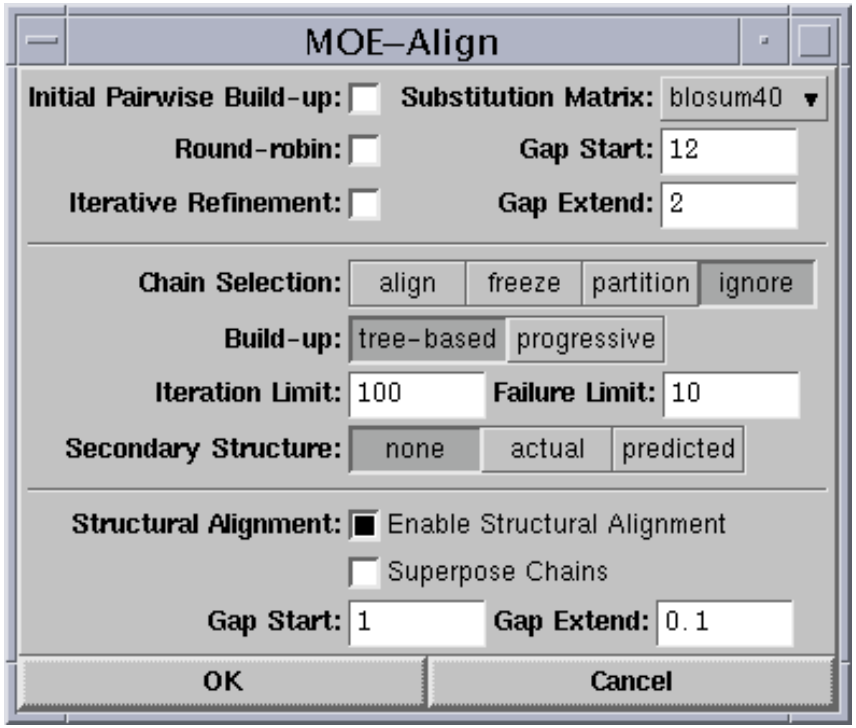
1. 1c0a
   1. region a:1-106
2. 1il2
   complexed with 1mg, 5mc, 5mu, amo, h2u, psu, so4
   1. region a:1-106
   2. region b:1001-1106
3. 1eqr
   complexed with mg
   1. region a:1-106
   2. region b:1-106
   3. region c:1-106

Alignment in MOE

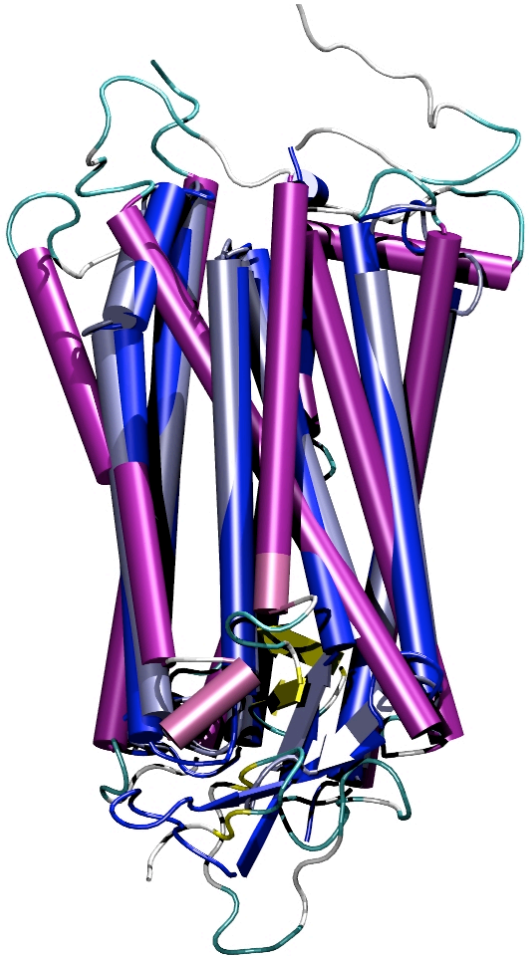


Alignment in MOE



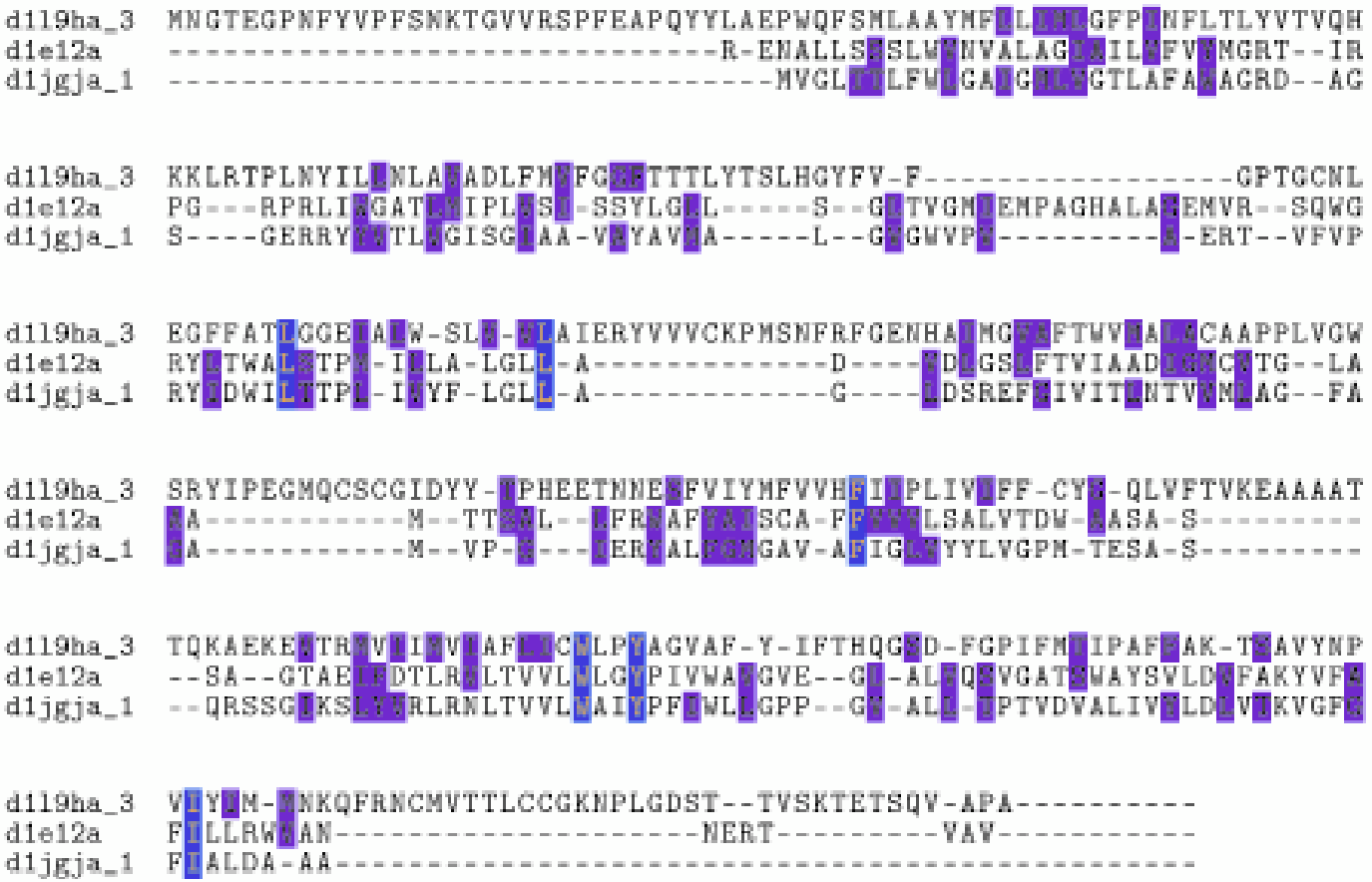


Transmembrane Proteins - HMM



Example Bacteriorhodpsin – Anurag Sethi UIUC

Stamp Profile

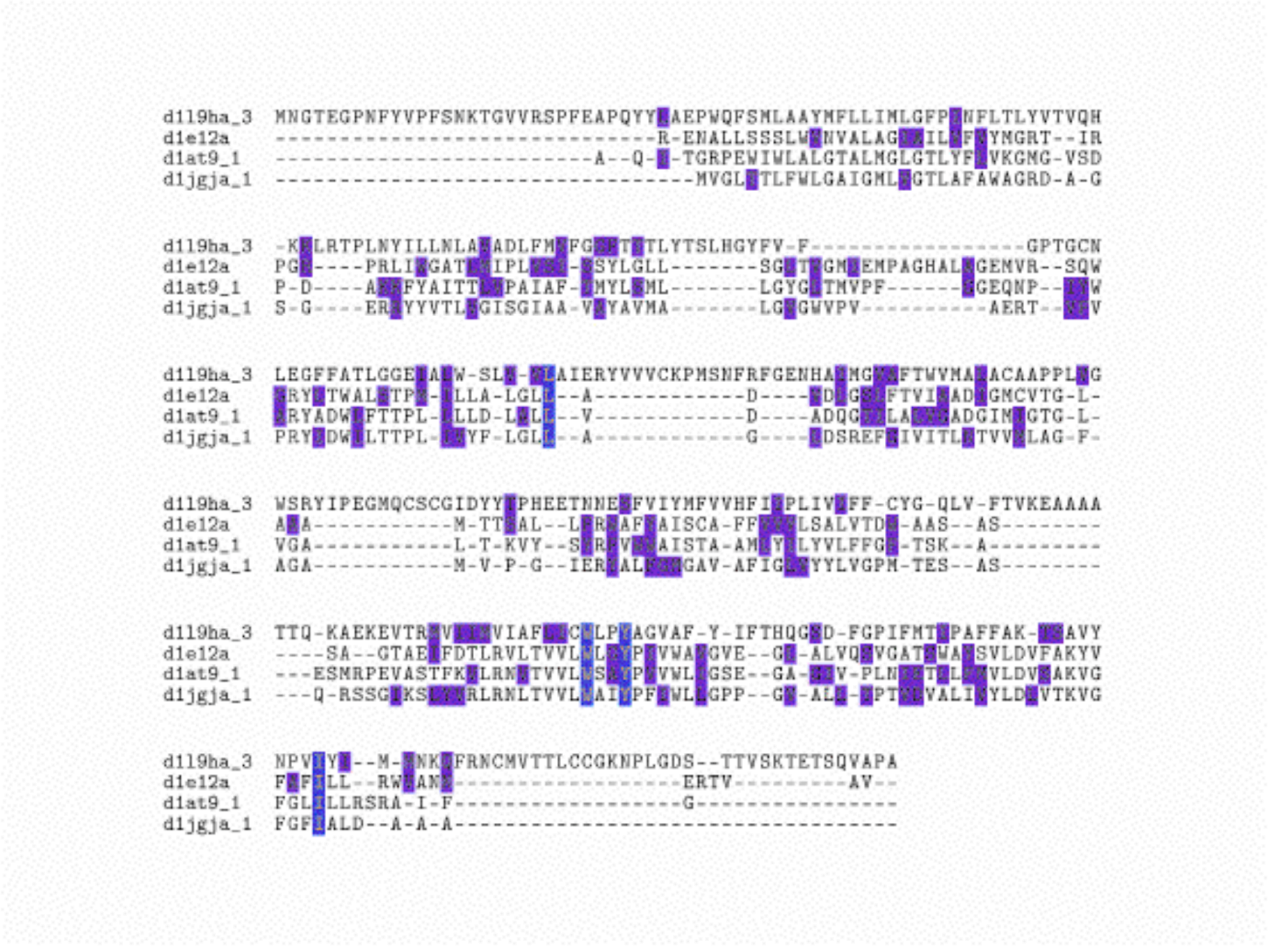


Sethi and Luthey-Schulten, UIUC 2003

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

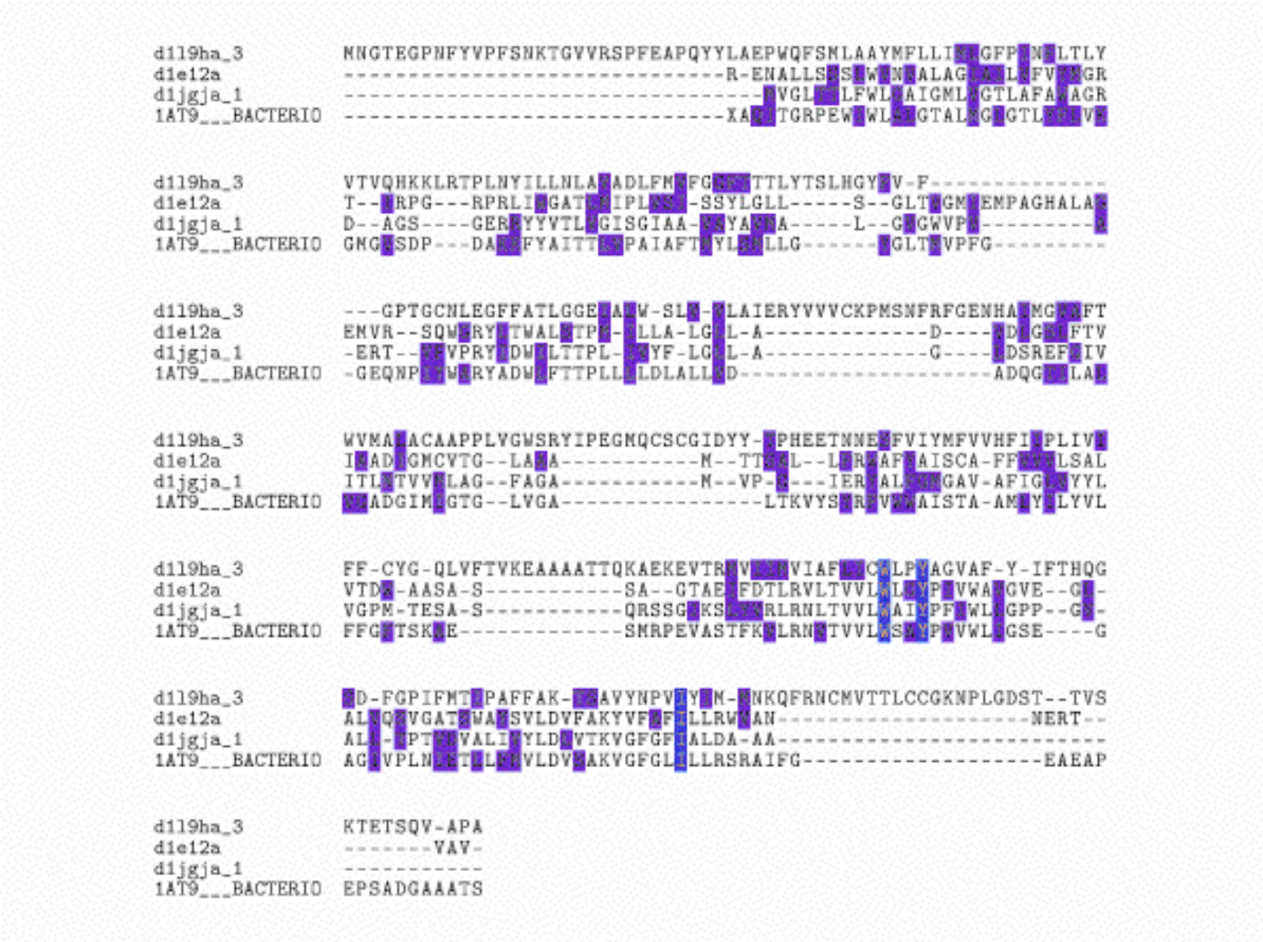
```
HMMER2.0  [2.2g]
NAME  inclustal
LENG  370
ALPH  Amino
RF    no
CS    no
MAP   yes
COM   /usr/local/bin/hmmbuild /bio/tmp/inclustal.982259.hmm /bio/tmp/inclustal.982259
NSEQ  3
DATE  Sun Jun  8 18:12:11 2003
CKSUM 1057
XT      -8455     -4  -1000  -1000  -8455     -4  -8455     -4
NULT      -4  -8455
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



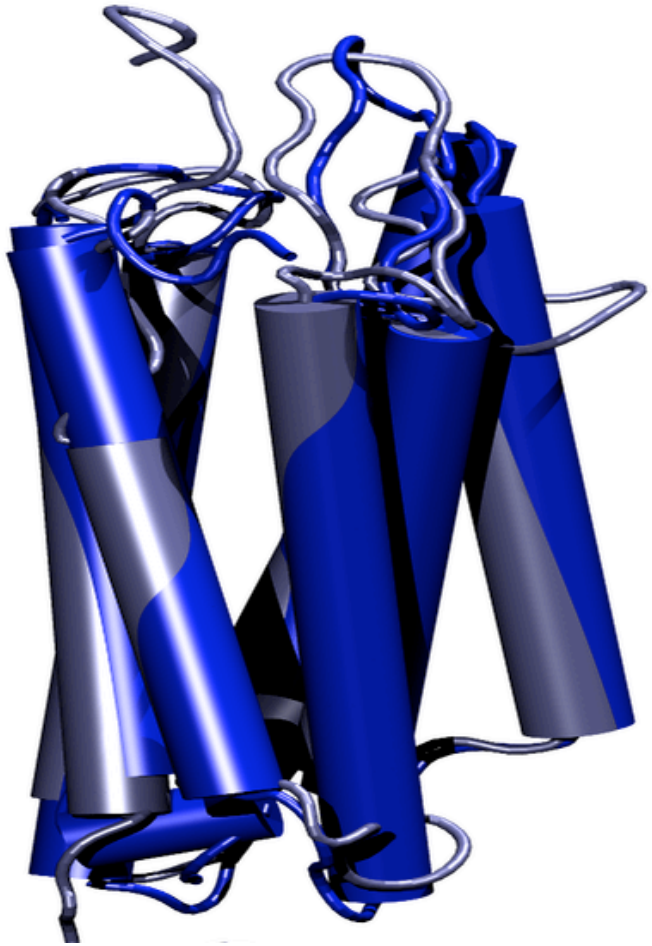
Sethi and Luthey-Schulten, UIUC 2003

Clustal Profile-Profile Alignment



Sethi and Luthey-Schulten, 2003

Structure Prediction Modeller 6.2/Hmmer



Sethi and Luthey-Schulten, UIUC 2003

Modeller 6.2 A. Sali, et al.

Acknowledgements

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