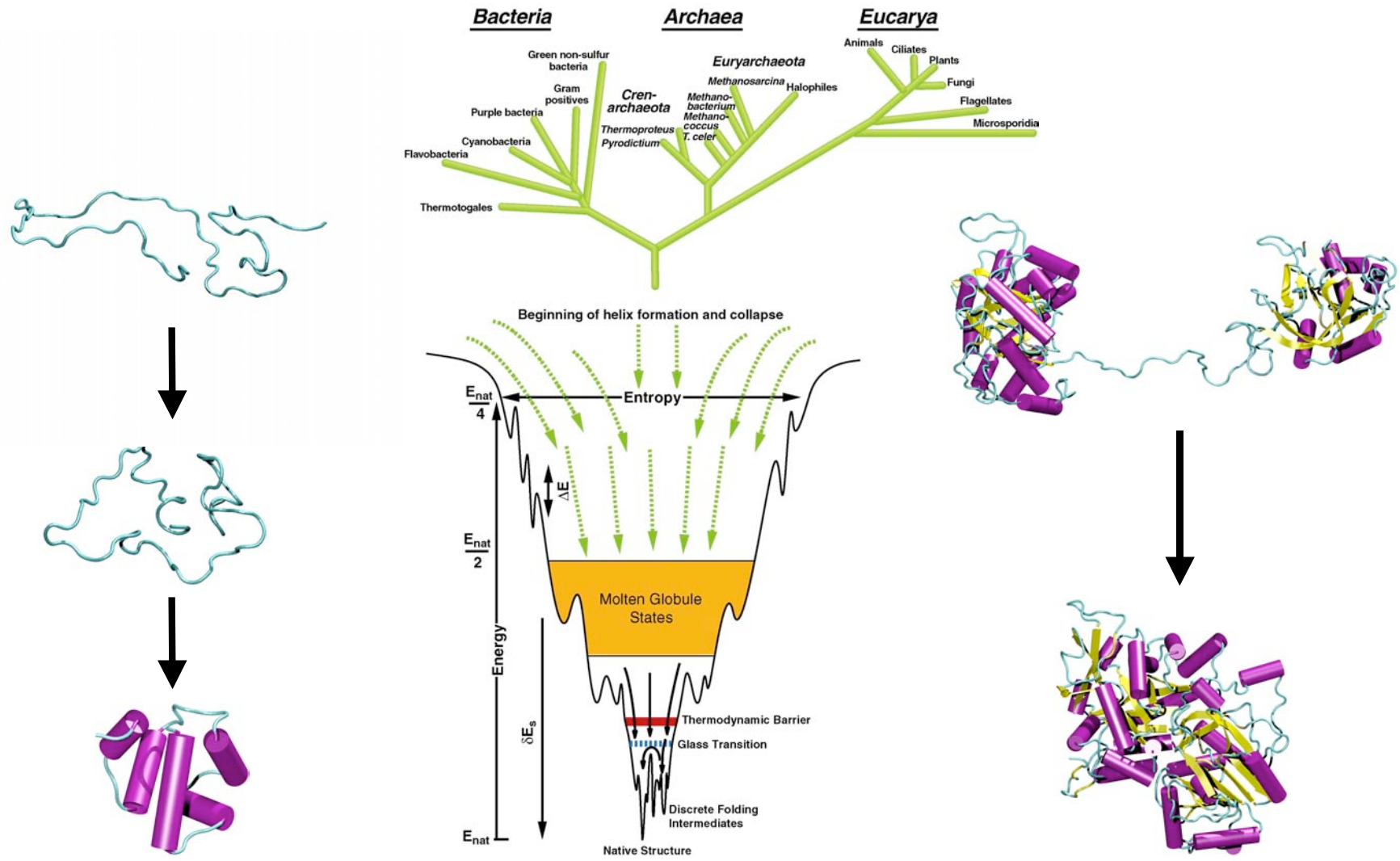


Bioinformatics I - Sequence and Structure Alignments

Z. Luthey-Schulten, UIUC



Perth, 2004

Sequence-Sequence Alignment (P)

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

Sequence-Structure Alignment (MS)

- Threading Profile 1: $A_1 A_2 A_3 - - A_4 A_5 \dots A_n$
- Hidden Markov Profile 2: $C_1 - C_2 C_3 C_4 C_5 - \dots C_m$

Structure-Structure Alignment (MS)

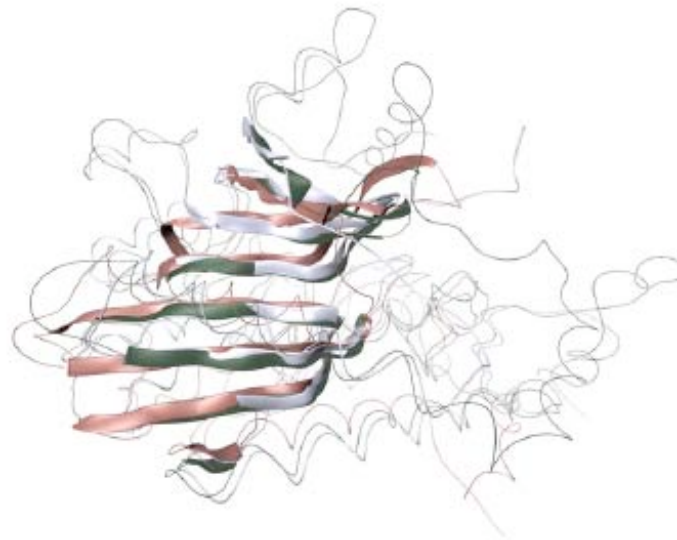
- STAMP - Barton and Russell
- CE - Bourne et al.

Sequence Database Searches (MS)

- Blast and Psi-Blast

University of Illinois at Urbana-Champaign
Luthey-Schulten Group
Theoretical and Computational Biophysics Group
Summer School 2004 - University of Western Australia, Perth

Sequence Alignment Algorithms



Rommie Amaro
Felix Autenrieth
Brijeet Dhaliwal
Barry Isralewitz

Zaida Luthey-Schulten
Anurag Sethi
Taras Pogorelov

June 2004

Sequence Alignment & Dynamic Programming

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

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

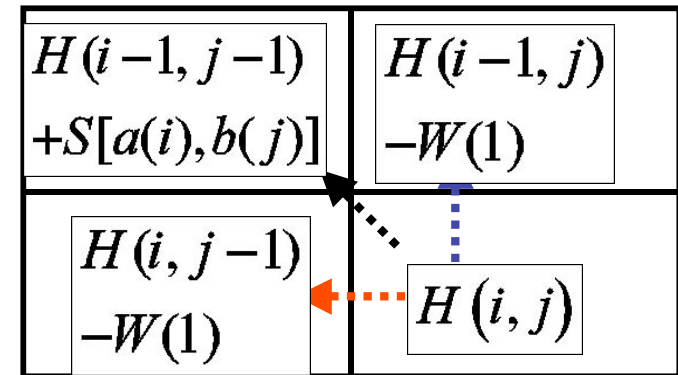


number of possible alignments:

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

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

AWGHE

AW--HE

	H	E	A	G	A	W	G	H	E	E
	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

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
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	0	3	0	24	16	8	0
H	0	13	5	0	0	1	16	22	29	13
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

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	

Sequence Alignment & Dynamic Programming

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



number of possible alignments:

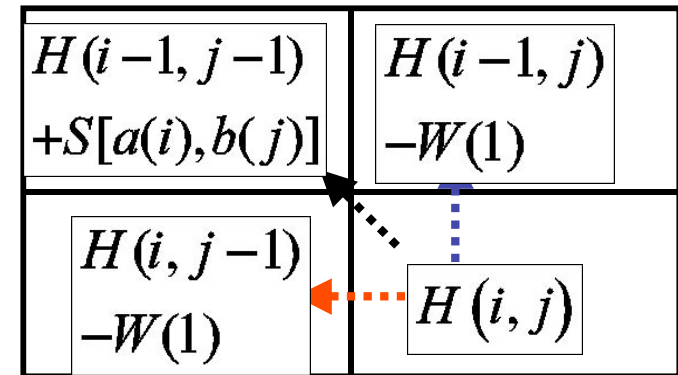
$$= \binom{2n}{n} = 2^{2n} (\sqrt{n\pi})^{-1}$$

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 : 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



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

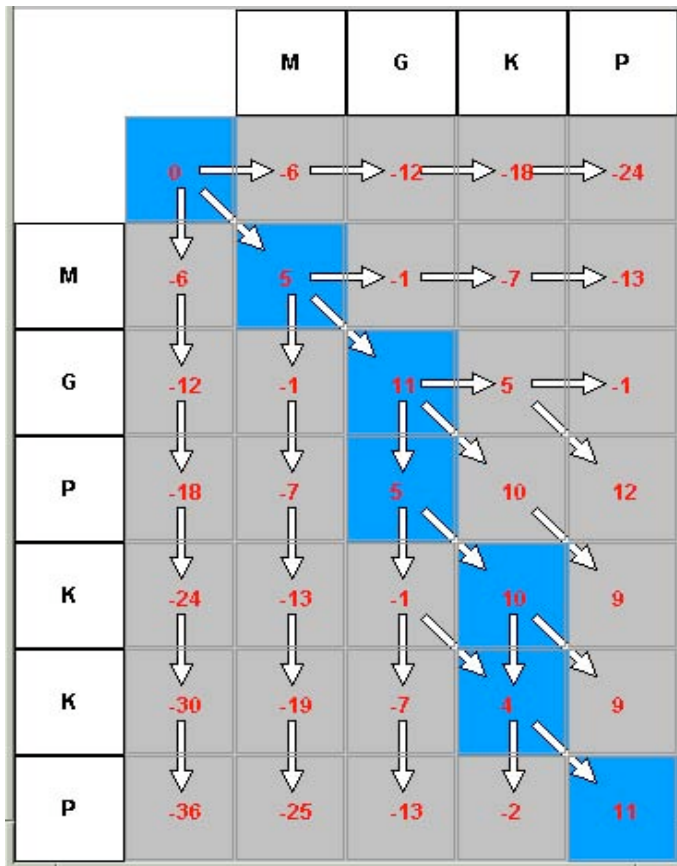
		M	G	K	P
	0	-6	-12	-18	-24
M	-6	5	-3	-1	-2
G	-12	-3	6	-2	-2
P	-18	-2	-2	-1	7
K	-24	-1	-2	5	-1
K	-30	-1	-2	5	-1
P	-36	-2	-2	-1	7

Filling out the Score Matrix H

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

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

Traceback and Alignment

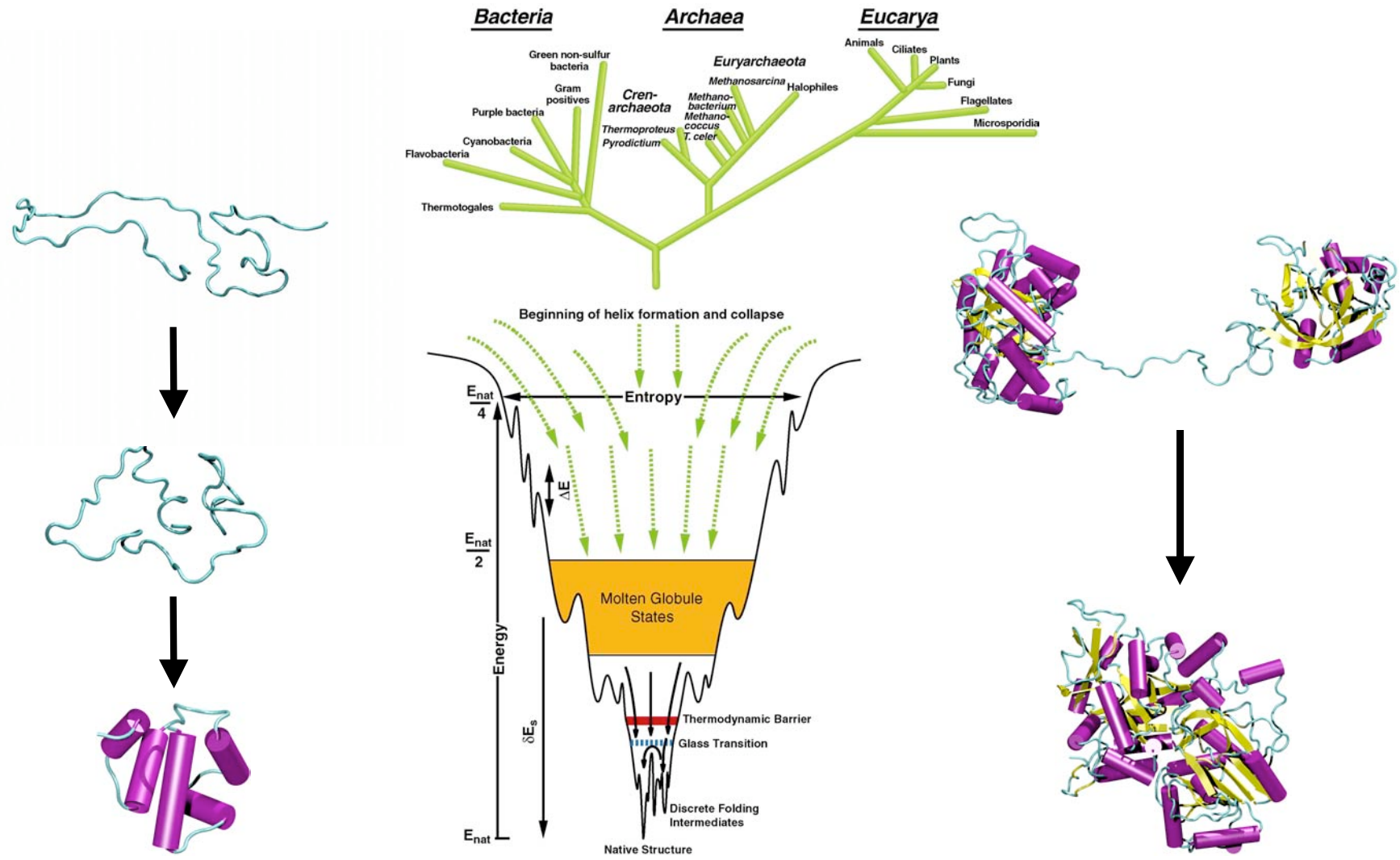


The Alignment

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

Traceback (blue) from optimal score

Energy Landscape Theory of Structure Prediction



Protein Structure Prediction

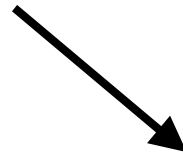
1-D protein sequence

SISSIRVKSKRIQLG....

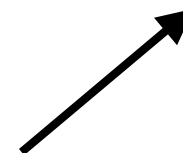
3-D protein structure



Ab Initio protein folding



Seq-Str Alignment



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

Homologous/analogous protein → QFANEFKVRRIKLGYTQTNVGEALAAVHGS...

of known structure

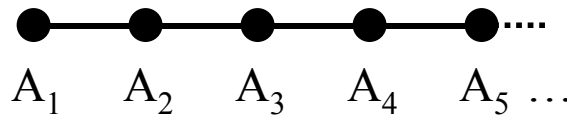
Sequence -Structure Alignment: the Energy Function

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

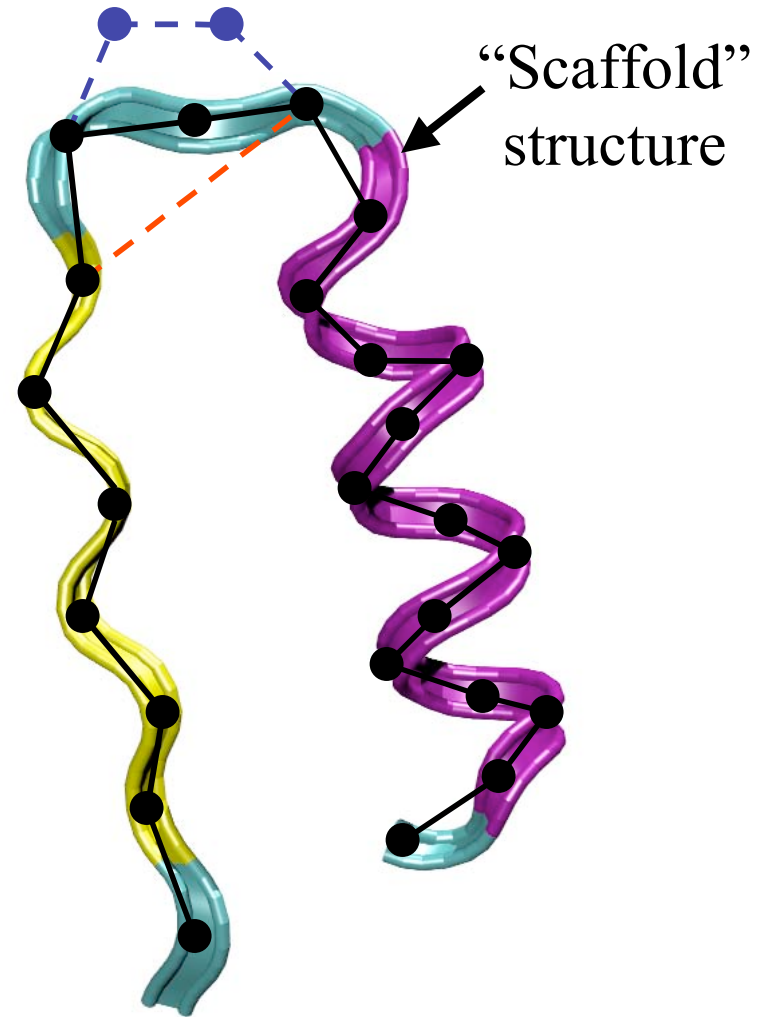
$$E_{\text{match}} =$$

Threading: Sequence-Structure Alignment

Target sequence



threading alignment
between target and scaffold



Threading Energy Function

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

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

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

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

Gap Penalties

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

Distribution
of Gaps

Sequence-Structure Gap Energy

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

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

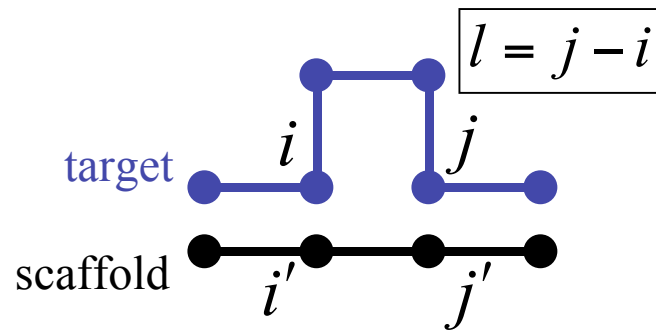
$$P_{deletion}(r) = a_2 * \exp\left(-\frac{(r - b_2)^2}{2\sigma_2^2}\right)$$

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

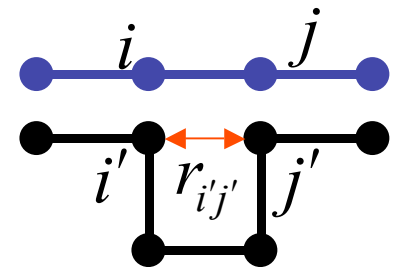
$$P_{bulge}(l, r) = \frac{a_3}{(\sigma_3 l)^{3/2}} * r^2 * \exp\left(-c_3 l - \frac{r^2}{\sigma_3 l}\right)$$

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

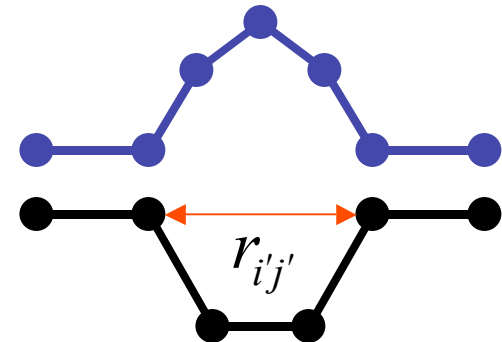
Insertion



Deletion



Bulge



Similarity Measures

Sequence Identity

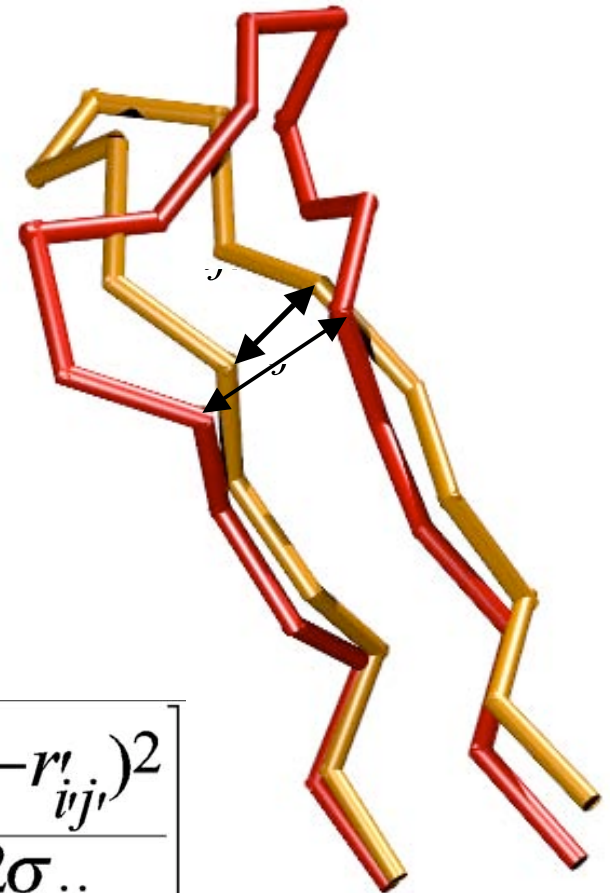
fraction of identically
matched residues

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

Q “Structural Identity”

fraction of native contacts

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

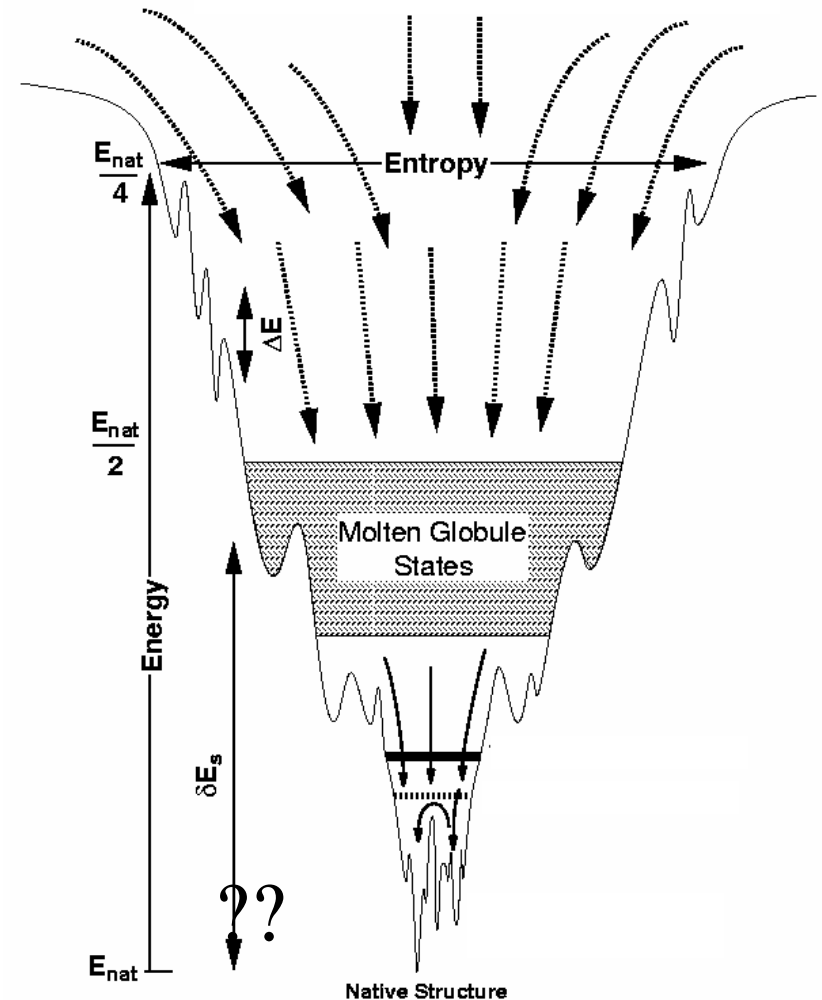
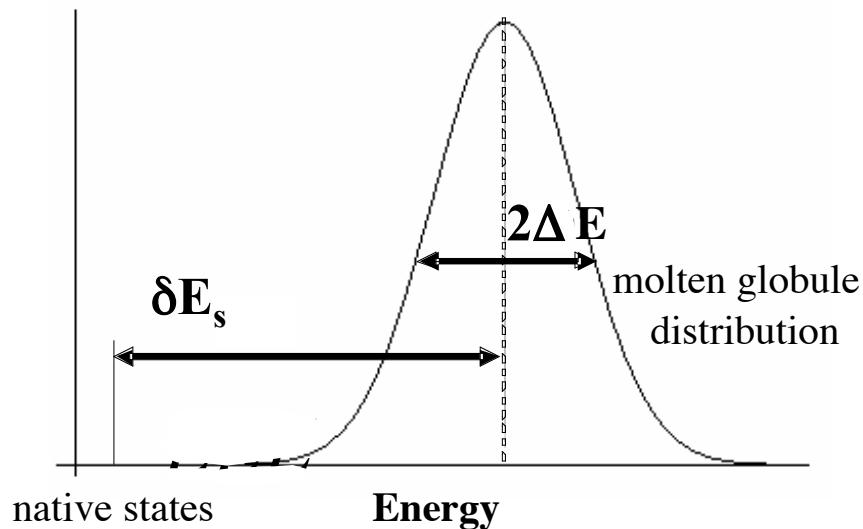


A summary of Energy Landscape Theory

Energy Landscape Theory

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

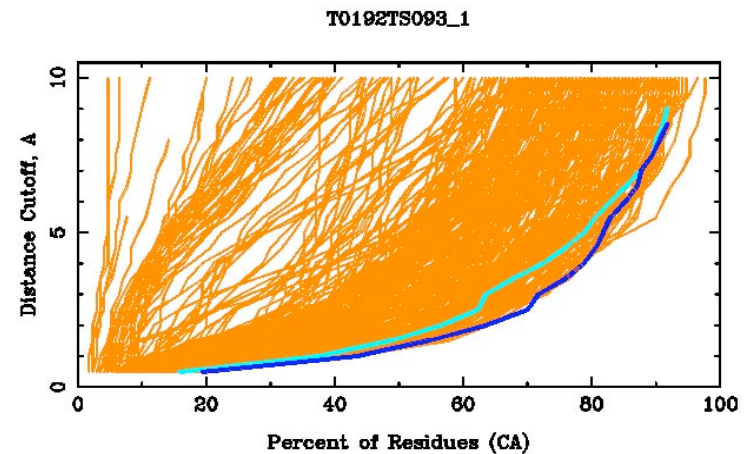
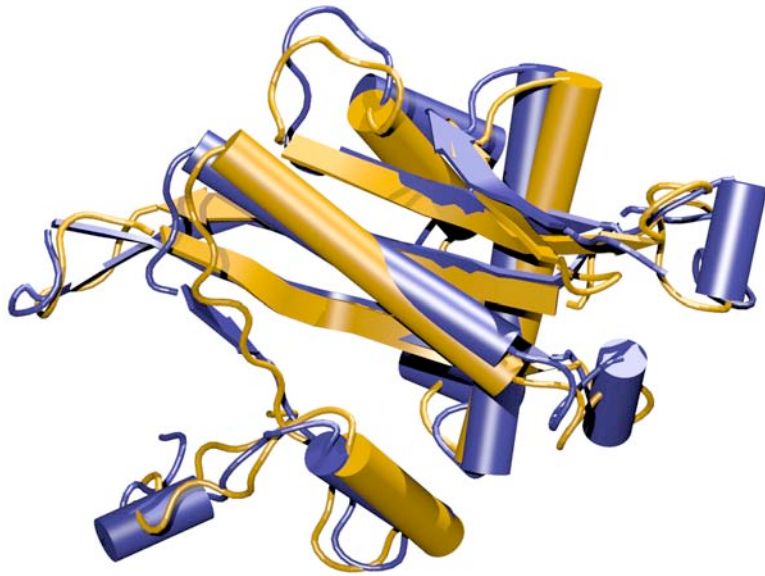
Optimization over an Ensemble of Folds



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

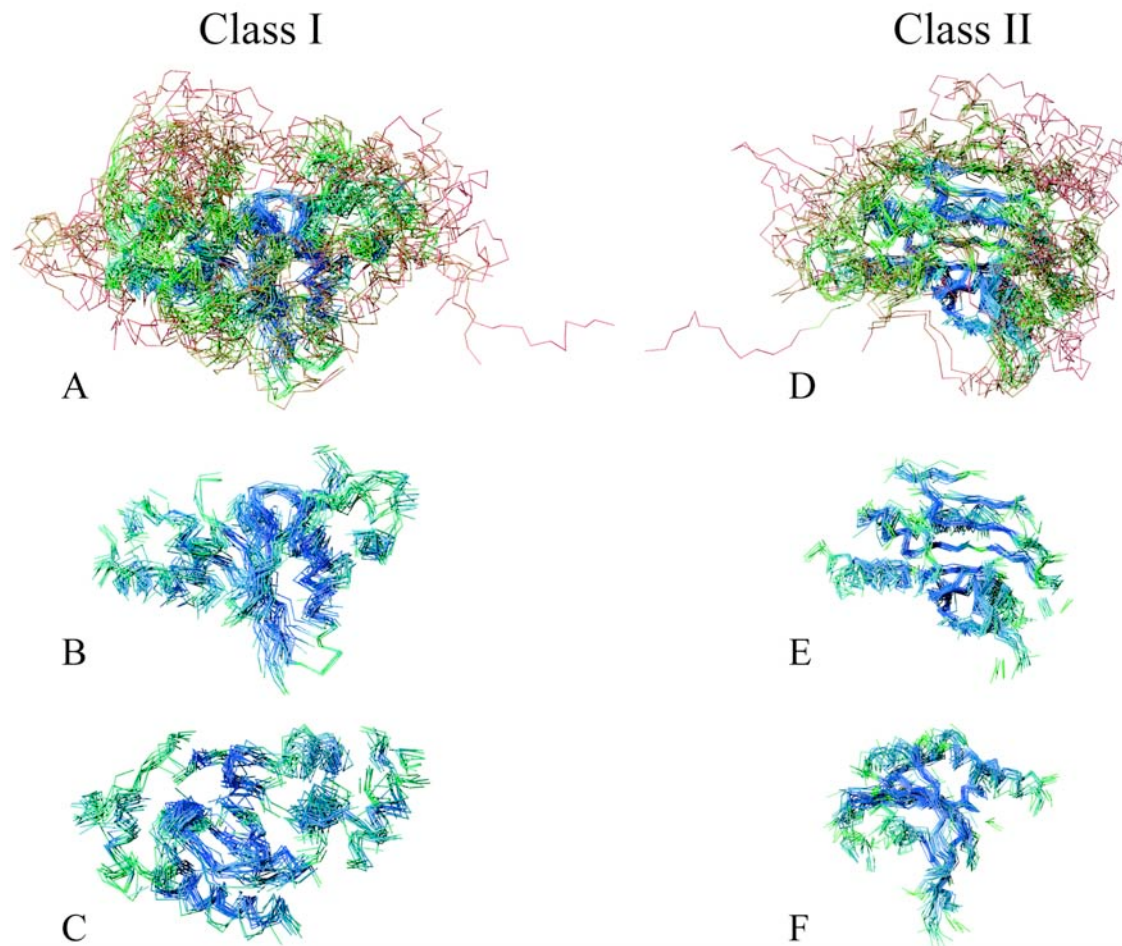
Homology Modeling - Threading

Single Sequence to Single Structure



Profile - Multiple Structural Alignments

Representative Profile of AARS Family



STAMP - Multiple Structural Alignments

1. Initial Alignment Inputs

- Multiple Sequence alignment
- Ridged Body “Scan”

2. Refine Initial Alignment & Produce Multiple Structural Alignment

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

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 between 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:

$$S_C = \frac{S_P}{L_P} \frac{L_P - i_A}{L_A} \frac{L_P - i_B}{L_B}$$
$$S_P = \sum_{aln.path} P_{ij}$$

L_P, L_A, L_B – length of alignment, sequence A, sequence B

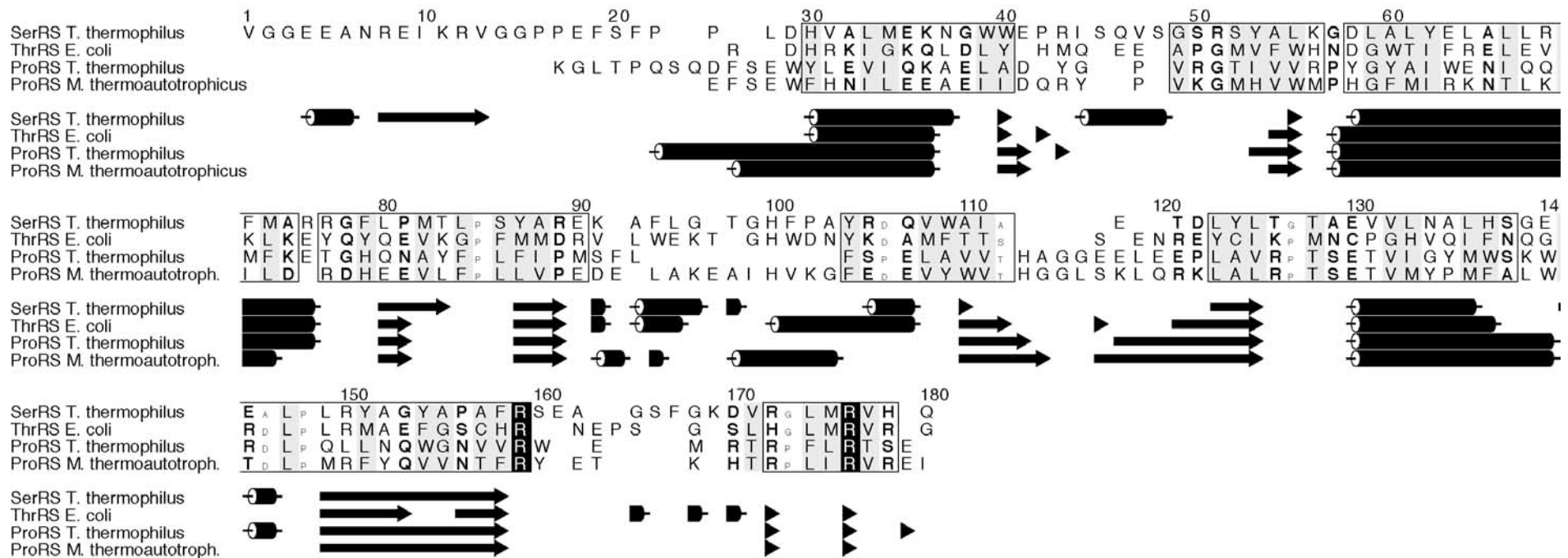
i_A, i_B – length of gaps in A and B.

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.

Variation in Secondary Structure

STAMP Output



Stamp Output/Clustal Format

```

SerRS-T_thermophilus      VGGEENREIKRVGGPPEFSFP--P--LDHVALMEKNGWWEPRISQVSGSRSYALKGDLA
ThrRS-E_coli              -----R--DHRKIGKQLDLY-HMQ-EE-APGMVFWHNDGW
ProRS-T_thermophilus      -----KGLTPQSQDFSEWYLEVIQKAEAD-YG--P-VRGTIVVRPYGY
ProRS-M_thermoautotrophic -----EFSEWFHNILEEAEIIDQRY--P-VKGMHVWMPHGF
space                     -----
SerRS-T_thermophilus      --SGGG-EEEEEESS-----SS-----HHHHHHHHT-B-TTHHHHH-SS---B-THHH
ThrRS-E_coli              -----HHHHHHHHT-E-E---TT-STT--EE-HHHH
ProRS-T_thermophilus      -----HHHHHHHHHHHHHHHTTSEE-E---S-STT-EEE-HHHH
ProRS-M_thermoautotrophic -----HHHHHHHHHHHHTT-EE-----S-STT--EE-HHHH

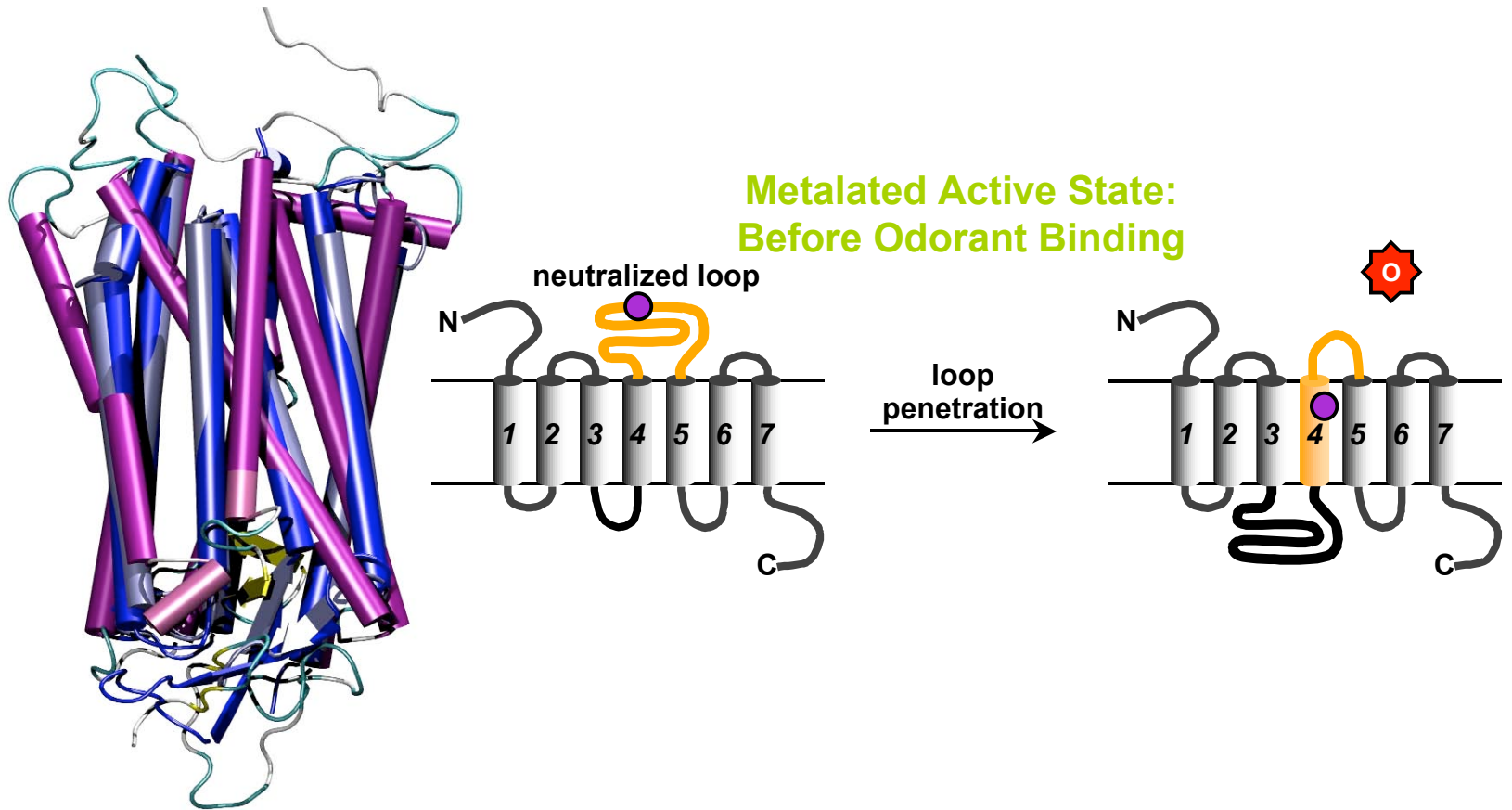
SerRS-T_thermophilus      LYELALLRFAMDFMARRGFLPMTLPSYAREK-AFLG-TGHFPAYRDQVWAI-----E--
ThrRS-E_coli              TIFRELEVFRSKLKEYQYQEVKGPFFMDRV-LWEKT-GHWDNYKDAMFTTS---S-EN
ProRS-T_thermophilus      AIWENIQQLDRMFKETGHQNAFYPLFIPMSFL-----FSPELAVVTHAGGEELE
ProRS-M_thermoautotrophic MIRKNTLKILRRILD-RDHEEVLFPLLPEDE-LAKEAIHVKGFEDEVYVWTHGGLSKLQ
space                     -----
SerRS-T_thermophilus      HHHHHHHHHHHHHHHHTT-EEEE--SEEEHH-HHHH-HT-TTTGGGS-B-T----T--
ThrRS-E_coli              HHHHHHHHHHHHHHHHTT-EE----SEEEHH-HHHT-THHHHGGG--EEE---E-TT
ProRS-T_thermophilus      HHHHHHHHHHHHHHHHTT-EE----SEESTT-----TT--EEEE-SSSEEE
ProRS-M_thermoautotrophic HHHHHHHHHHHHHHTT-TT-EE----SEEEHH-HTTSHHHHHHTTTT--EEEEETTEEE

SerRS-T_thermophilus      TDLYLTGTAEVVLNALHSGEILPYEALPLRYAGYAPAFRSEA--GSFGKDVRGLMRVH-Q
ThrRS-E_coli              REYCIKPMNCPGHVQIFNQGLKSYRDLPLRMAEFGSCHR--NEPS--G-SLHGLMRVR-G
ProRS-T_thermophilus      EPLAVRPTSETVIGYMWSKWIRSWRDLPQLLNQWGNVVRW--E----M-RTRPFLRTSE-
ProRS-M_thermoautotrophic RKLALRPTSETVMYPMFALWVRSHDLPMPFYQVVNTFRY-ET----K-HTRPLIRVREI
space                     -----
SerRS-T_thermophilus      SSEE-S-THHHHHHHTTT-EEEGG-SEEEEEEEEE-----S--SSTTTTTTS-S-E
ThrRS-E_coli              EEEEE-S-SHHHHHHHTSS--BTTT-SEEE--EEE-----G--G-G-BTTTB-S-E
ProRS-T_thermophilus      EEEEE-S-SHHHHHHHHHH--BGGG--EEEEEEEE-----S-S-BTTTB-SE-
ProRS-M_thermoautotrophic EEEEE-SSSHHHHHHHHH--BTTT--EEEEEEEE-----S--BTTTB-SEE

```

From multiple sequence alignment compute position probabilities for amino acids and gaps!!!!

Hidden Markov Models of Transmembrane Proteins



Bacteriorhodopsin/Rhodopsins

Olfactory Receptor/Bovine Rhodopsin

J. Wang, Z. Luthey-Schulten, K. Suslick (2003) *PNAS* **100**(6):3035-9

Stamp Profile

```

d1l9ha_3  MNGTEGPNFYVPFSNKTGVVRSPPFEAPQYYLAEPWQFSMLAAYNFLLGGFPNFLTLYVTVQH
d1e12a    -----R-ENALLSSSLWNVVALAGAILFVNGRT--IR
d1jgja_1  -----MVGLTFLFWGAGGGGTLAFAAGRD--AG

d1l9ha_3  KKLRTPLNYILNLAAADLFMFGSTTTTLYTSLHGYYFV-F-----GPTGCNL
d1e12a    PG---RPRLINGATPIPLSS-SSYLG-L----S--GTVGMENPAGHALAEMVR--SQWG
d1jgja_1  S----GERRYETLGGISGAA-VNYAVAA-----L--GGMVPP-----ERT--VFVP

d1l9ha_3  EGFFATGGEEAEW-SLESLAIERYVVVCKPMSNFRFGENHAIMGSEFTWVBAISCAAPPLVGW
d1e12a    RYTWALSTPN-ILLA-LGLL-A-----D---DAGSFTVIAADLCTGTG--LA
d1jgja_1  RYDWTITTP-ISYF-LGLL-A-----G---DSREFIVITNTVNMAG--FA

d1l9ha_3  SRYIPEGMQCSCGIDYYTPHEETNNEFVIYMFVVHFIPLIVFF-CYK-QLVFTVKEAAAAT
d1e12a    SA-----M--TTEEL--FRNAFVNSCA-PPNALSALVTDW-SASA-S-----
d1jgja_1  SA-----M--VP--SERIALGGAGAV-APIGAYYLVGPM-TESA-S-----

d1l9ha_3  TQKAEKETRVVIVVAFVCLPYNAGVAF-Y-IFTHQGID-FGPIMIPAFPAK-TAVYNP
d1e12a    --SA--GTAEEDTLRLTVVLLGYPPIVWANGVE--GL-ALQGVGATWAYSVLDFAKYVFN
d1jgja_1  --QRSSGKSLYRLRLNLTVVLLAIYPFVWLGGPP--GN-ALDPTVDVALIVLDVVKVGF

d1l9ha_3  VIYVM-WNKQFRNCMVTTLCCKGNPLGDST--TVSKTETSQV-APA-----
d1e12a    FLLRWGAN-----NERT-----VAV-----
d1jgja_1  FALDA-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      *      *

```

Protein X : A B - B A
Protein Y : A - - B A
Protein Z : A A B A A

State π : $\mathbf{M}_1$ $\mathbf{M}_2$ $\mathbf{I}_2$ $\mathbf{M}_3$ $\mathbf{M}_4$
 $\mathbf{D}_1$

$\mathbf{M}_i$ – i^{th} Match State
 $\mathbf{I}_i$ – i^{th} Insert State
 $\mathbf{D}_i$ – i^{th} Delete State

$$\begin{aligned}
 P(x_j, t) &= P\left(x_j \mid \pi_j = t\right) \times P\left(\pi_j = t\right) \\
 &= e\left(x_j \mid \pi_j = t\right) \times a\left(\pi_j = t \mid \pi_{j-1} = s\right)
 \end{aligned}$$

New protein aligned to profile with Verterbi
 (Dynamic Programming) algorithm -
 Maximum probability path through state
 transitions.

State transition Probabilities (a)

i	$\mathbf{M}_i \longrightarrow \mathbf{M}_{i+1}$	$\mathbf{M}_i \longrightarrow \mathbf{D}_i$	$\mathbf{M}_i \longrightarrow \mathbf{I}_i$
1	0.67	0.33	0
2	0.67	0	0.33
3	1	0	0

**Position dependent amino acid
(Emission) Probabilities (e) - PSSM**

$i\text{-}M$	$e(\mathbf{A} \mathbf{M})$	$e(\mathbf{B} \mathbf{M})$
1	1	0
2	0.5	0.5
3	0.33	0.67
4	1	0

Amino acid probabilities at insert states is background
 probability of occurrence of the corresponding amino acid.

$$P(\mathbf{A}|\mathbf{I}) = 0.72$$

$$P(\mathbf{B}|\mathbf{I}) = 0.28$$

Leads to affine gap penalty.

$$P(-|\mathbf{D}) = 1.$$

HMMer Profile-Profile Alignment

```

d1l9ha_3 MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYAEPWQFSMLAAYMFLLIMLGFPNFLTLYVTVQH
d1e12a -----R-ENALLSSSLWNVALAGILFYNGRT--IR
d1at9_1 -----A--Q--TGRPEVIWLALGTALMGLGTLTYFVKGMG-VSD
d1jgja_1 -----MVGLTLFWLGAIGMLCTLAFAWAGRD-A-G

d1l9ha_3 -KRLRTPLNVIILLNLAADLFMFGDTTLYTSLHGYFYF-----GPTGCN
d1e12a PG-----PRLNGATPIPL-----SYLGLL-----SGTGMEMPAGHALGEMVR--SQW
d1at9_1 P-D---AIFYAITTTPAIAFMYLML-----LGYGTNVPP-----GEQNP--W
d1jgja_1 S-G---ERYVVTLEGISGIAA-VYAVMA-----LGGVVPV-----AERT--W

d1l9ha_3 LEGFFATLGGEAW-SL--SAIERVYVVCKPMSNFRFGENHAMGFTWVMAACAAPPLG
d1e12a RYRTWALTPM--LLA-LGL--A-----D---DGGFTVIADGMCVTG-L-
d1at9_1 RYADWFTTPL--LLD-L-L--V-----D---ADQGLA--ADGINGTG-L-
d1jgja_1 PRYDWLTTPL--GYF-LGL--A-----G---DSREFIVITLTVVLAG-F-

d1l9ha_3 WSRYPGEMQCSCGIDYYPHEETNNEFVIYMFVVHFIPLIVFFCYG-QLV-FTVKEAAAA
d1e12a A--A-----M-TTAL--LRFAFAISCA-FF--LSALVTD--AAS--AS-----
d1at9_1 VGA-----L-T-KVY--SRVGAISTA-AMVLYVLFFG--TSK--A-----
d1jgja_1 AGA-----M-V-P-G--IERALGAGAV-AFIGVYVLVGPM-TES--AS-----

d1l9ha_3 TTQ-KAEKEVTRV--VIAF--CLPAGVAF-Y-IFTHQC-D-FGPIFMTPAFFAK--AVY
d1e12a ---SA--GTAEFDTLRVLTVVLMCYPTVWAQVE--G--ALVQVGATWASVLDVFAKYV
d1at9_1 ---ESMRPEVASTFKLRNTVVLSYPTVWLQSE--GA--V-PLN--TVL--VLDV--AKVG
d1jgja_1 ---Q-RSSGKSK--RLRNLTVVLAIYPTVWLQGP--G--AL--PT--VALISYLDVTKVG

d1l9ha_3 NPVY--M--NKIFRNCMVTTLCCKGNPLGDS--TTVSKTETSQVAPA
d1e12a F--LL--RW--AM-----ERTV-----AV-
d1at9_1 FGLLLRSRA-I-F-----G-----
d1jgja_1 FGFALD--A-A-A-----

```

Clustal Profile-Profile Alignment

```

d1l9ha_3      MNGTEGPNFYVPFSNKTGVVRSPPFEAPQYYLAEPWQFSNLAAYMFLLIAGFPNLTLY
die12a        -----R-ENALLSSEWNNALAGELFLFVAGR
d1jgja_1      -----VGLLFWLAIGMLGTLAFAAGR
1AT9__BACTERIO -----XAGTGRPEWWSLTALGGLTLEVV

d1l9ha_3      VTVQHKLRTPLNYYILLNLADLFHFGSSSTLYTSLNGYV-F-----
die12a        T--RPG---RPRLIGATIPLESSYLGLL----S--GLTGMEMPAGHALA
d1jgja_1      D--AGS---GERYYVTLCISGIAAAYAA----L--GCGWVP-----
1AT9__BACTERIO GNGSDP--DAFYAITTPAIAFTYLLGLL-----GLTVVFG-----

d1l9ha_3      ---GPTGCNLEGFFATLGGEAW-SL--LAIERVVVCKPMSNFRFGENHAMGFT
die12a        ENVR--SQWRYTWTALTP--LLA-LGL-A-----D--DCTFTV
d1jgja_1      -ERT--GVPRYDWLTPL--GYF-LGL-A-----G--DSREFIV
1AT9__BACTERIO -GEQNPWRYADWFTTPLLDLALLD-----ADQCIA

d1l9ha_3      WVMAACAAPPLVGSRYIPEGMQCSCGIDYY--PHEETNNEFVIYMFVVHFIPLIV
die12a        IADGMCVTG--LAA-----M--TTGL--LRRAF AISCA-FFGLSAL
d1jgja_1      ITLTVVLAG--FAGA-----M--VP--IERALGAV-AFIGYYL
1AT9__BACTERIO IADGIMGTG--LVGA-----LTKVYSRVVAISTA-AMYYLYVL

d1l9ha_3      FF-CYG-QLVFTVKEAAAATTQKAEKEVTRVSVIAFCLPAGVAF-Y-IFTHQG
die12a        VTD--AASA-S-----SA--GTAEFDTLRVLTVVLLVFWAGVE--G-
d1jgja_1      VGPM-TESA-S-----QRSSGKSLRLRLTVVLAINPFWLGPP--G-
1AT9__BACTERIO FFGTSKE-----SMRPEVASTFKLRNTVVLSDFVWL GSE---G

d1l9ha_3      D-FGPIFMTPAFFAK--AVYNPVYVM--NKQFRNCMVTTLCCGKNPLGDST--TVS
die12a        ALQVGATWASVLDVFAKYVF--LLRWAN-----NERT--
d1jgja_1      AL--PTWVALIYLDVTKVGFGLALDA-AA-----
1AT9__BACTERIO AGVPLNLTLLVLDVAKVGFGLLLRSRAIFG-----EAEAP

d1l9ha_3      KTETSQV-APA
die12a        -----VAV-
d1jgja_1      -----
1AT9__BACTERIO EPSADGAAATS

```

Refine Structure Prediction with Modeller 6.2

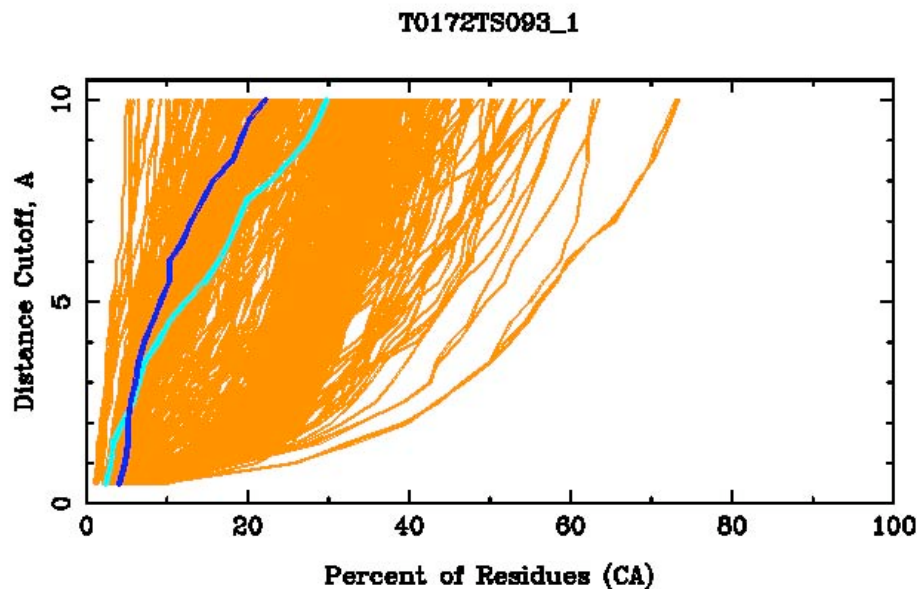
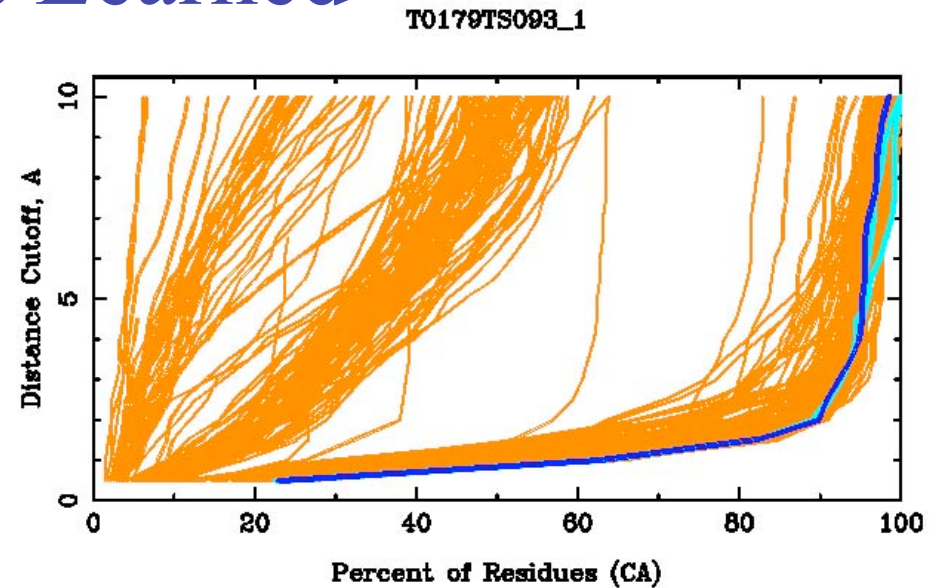
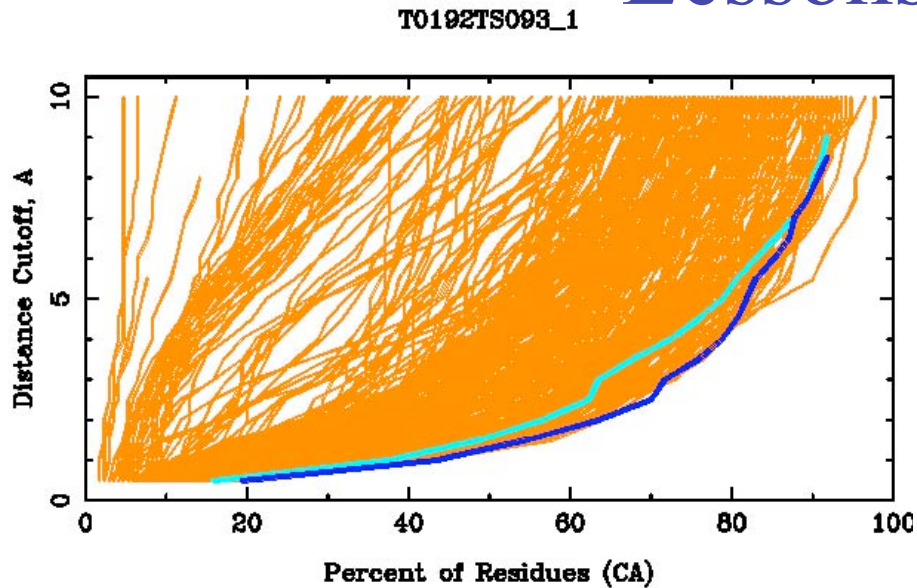


Sethi and Luthey-Schulten, UIUC 2003

Modeller 6.2 A. Sali, et al.

CM/Fold Recognition Results from CASP5

Lessons Learned



The prediction is never better than the scaffold.

Threading Energy Function and Profiles need improvement.

We need non-redundant, evolutionary profiles! True representative sets of protein sequences and structures from which to draw correct statistical inferences. Structure more conserved than sequence!!!! You are now entering the twilight zone of sequence identity.

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)

Acknowledgements

Patrick O'Donoghue

Rommie Amaro

Anurag Sethi

John Eargle

Corey Hardin

Michael Baym

Michael Januszczak

Felix Autenrieth

Taras Pogorelov

Graphics Programmers VMD

John Stone, Dan Wright, John Eargle

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