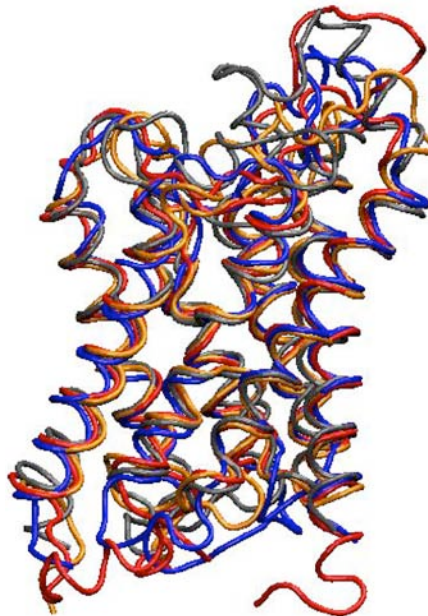


Sequence and Structure Alignment

Z. Luthey-Schulten, UIUC

Lake Tahoe, May 2005



Multiple Sequence Alignment

F... To... H...

Align Molecules...
FASTA
Highlight PDB
Pairwise RMSD
Sequence Display

```
1fqy -----KLFWRVVAEFLATLTFVFTISIGSAL-GF-KY---PVGNNQTAVQDNVKS LAPGLS IATLAQS-VGHISGAHLNPAVTLG LLLSCQISIF-RAI
1j4n MASEFKKLFWRVVAEFLAMILFIFLISIGSAL-GF-HYPIKSNQT-TGAVQDNVKS LAPGLS IATLAQSVGH-ISGAHLNPAVTLG LLLSCQ-I SVLRAI
1lda -----TLKGQCIAEFLGTGLLIFFGVGCVA-ALKVA-----G-A-SFGQWEISVINGLGVAMAIYLT A-GVSGAHLNPAVTLALWLF A-CFDKRV
1rc2 -----MFRKLA AECFGTFWLVFGGCGSAVLA-AG-----FPE-LGIGFAGVALAPGLTVLTMFAVVG-HISGGHFNPAVTIG LWAGG-RFPAKEV
```

Sequence-Sequence Alignment

- 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

- Threading
- Hidden Markov, Clustal

Structure-Structure Alignment

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

Sequence Database Searches

- Blast and Psi-Blast

Sequence-Sequence Alignment

- Smith-Watermann

Profile 1: $A_1 A_2 A_3 - - A_4 A_5 \dots A_n$

- Needleman-Wunsch

Profile 2: $C_1 - C_2 C_3 C_4 C_5 - \dots C_m$

Sequence-Structure Alignment

- Threading
- Hidden Markov, Clustal

Structure-Structure Alignment

- STAMP - Barton and Russell **SCOP, Astral**
- CE - Bourne et al. **PDB**

Sequence Database Searches

- Blast and Psi-Blast **NCBI** **Swiss Prot**

Search for



Swiss-Prot
Protein knowledgebase
TrEMBL
Computer-annotated supplement to Swiss-Prot



The [UniProt Knowledgebase](#) consists of:

- **Swiss-Prot**; a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases [[More details](#) / [References](#) / [Linking to Swiss-Prot](#) / [User manual](#) / [Recent changes](#) / [Commercial users](#) / [Disclaimer](#)].
- **TrEMBL**; a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot.

These databases are developed by the Swiss-Prot groups [at SIB](#) and [at EBI](#).

UniProt Release 3.2 consists of:

Swiss-Prot Release 45.2 of 23-Nov-2004: 164201 entries ([More statistics](#))

TrEMBL Release 28.2 of 23-Nov-2004: 1503829 entries ([More statistics](#))

> *Swiss-Prot headlines*

Major update of *C.elegans* entries (Read [more...](#))

 ExpASY Home page	Site Map	Search ExpASY	Contact us	Swiss-Prot					
Hosted by NCSC US	Mirror sites:	Australia	Bolivia	Brazil <small>new</small>	Canada	China	Korea	Switzerland	Taiwan
Search		<input type="text" value="Swiss-Prot/TrEMBL"/>	for	<input type="text" value="aqp"/>	<input type="button" value="Go"/>	<input type="button" value="Clear"/>			

Search in Swiss-Prot and TrEMBL for: aqp

Swiss-Prot Release 45.2 of 23-Nov-2004

TrEMBL Release 28.2 of 23-Nov-2004

-
- Number of sequences found in [Swiss-Prot](#)₍₈₉₎ and [TrEMBL](#)₍₁₂₂₎: **211**
 - Note that the selected sequences can be saved to a file to be later retrieved; to do so, go to the [bottom](#) of this page.
 - For more directed searches, you can use the Sequence Retrieval System [SRS](#).
-

Search in Swiss-Prot: There are matches to 89 out of 164201 entries

[AQPI_BOVIN \(P47865\)](#)

Aquaporin-CHIP (Water channel protein for red blood cells and kidney proximal tubule) (Aquaporin 1) (Water channel protein CHIP29). {GENE: Name=AQP1} - Bos taurus (Bovine)

[AQPI_HUMAN \(P29972\)](#)

Aquaporin-CHIP (Water channel protein for red blood cells and kidney proximal tubule) (Aquaporin 1) (AQP-1) (Urine water channel). {GENE: Name=AQP1; Synonyms=CHIP28} - Homo sapiens (Human)

[AQPI_MOUSE \(Q02013\)](#)

Aquaporin-CHIP (Water channel protein for red blood cells and kidney proximal tubule) (Aquaporin 1) (Early response protein DER2). {GENE: Name=Aqp1} - Mus musculus (Mouse)

Search for

NiceProt View of Swiss-Prot: P47865

[\[Entry info\]](#)
[\[Name and origin\]](#)
[\[References\]](#)
[\[Comments\]](#)
[\[Cross-references\]](#)
[\[Keywords\]](#)
[\[Features\]](#)
[\[Sequence\]](#)
[\[Tools\]](#)

Note: most headings are clickable, even if they don't appear as links. They link to the user manual or other documents.

Entry information	
Entry name	AQP1_BOVIN
Primary accession number	P47865
Secondary accession numbers	None
Entered in Swiss-Prot in	Release 33, February 1996
Sequence was last modified in	Release 44, July 2004
Annotations were last modified in	Release 45, October 2004
Name and origin of the protein	
Protein name	Aquaporin-CHIP
Synonyms	Water channel protein for red blood cells and kidney proximal tubule Aquaporin 1 Water channel protein CHIP29
Gene name	Name: AQP1
From	Bos taurus (Bovine) [TaxID: 9913]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Cetartiodactyla; Ruminantia; Pecora; Bovidae; Bovinae; Bos.
References	
[1]	SEQUENCE FROM NUCLEIC ACID. TISSUE=Ocular ciliary epithelium;

Sequence information

Length: **270 AA** Molecular weight: **28669 Da** CRC64: **F3ECAD45DCCDB309** [This is a checksum on the sequence]

```
10      20      30      40      50      60
ASEFKKLFW RAVVAEFLAM ILFIFISIGS ALGFHYPIKS NQTTGAVQDN VKVSLAFGLS
70      80      90     100     110     120
IATLAQSVGH ISGAHLNPAV TLGLLLSCQI SVLRAIMYII AQCVGAIIVAT AILSGITSSL
130     140     150     160     170     180
PDNSLGLNAL APGVNSGQGL GIEIIGTLQL VLCVLATDR RRRDLGGSGP LAIGFSVALG
190     200     210     220     230     240
HLLAIDYTCG GINPARSFGS SVITHNFQDH WIFWVGPFIG AALAVLIYDF ILAPRSSDLT
250     260     270
DRVKVWTSQ VEEYDL DADD INSRVEMKPK
```

P47865 in [FASTA format](#)

The screenshot shows a web browser window with the URL `http://us.expasy.org/cgi-bin/get-sprot-fasta?P47865`. The browser's address bar and search bar are visible. Below the browser window, the text content of the page is displayed, including a description of the protein and its amino acid sequence. An arrow points from the text 'P47865 in FASTA format' in the top image to the browser window. Another arrow points from the word 'cut' to the protein sequence in the browser window.

```
>sp|P47865|AQP1_BOVIN Aquaporin-CHIP (Water channel protein for red blood cells and kidney proximal tubule) (Aquaporin 1) (Water channel protein CHIP29) - Bos taurus (Bovine).
ASEFKKLFWRAVVAEFLAMILFIFISIGSALGFHYPIKSNQTTGAVQDNVKVSLAFGLS
IATLAQSVGHISGAHLNPAVTLGLLLSCQISVLRAIMYIIAQCVGAIIVATAILSGITSSL
PDNSLGLNALAPGVNSGQGLGIEIIGTLQLVLCVLATDRRRDLGGSGPLAIGFSVALG
HLLAIDYTCGGINPARSFGSSVITHNFQDHWIFWVGPFIGAALAVLIYDFILAPRSSDLT
DRVKVWTSQVEEYDL DADD INSRVEMKPK
```

cut

[Search](#)

```
ASEFKKLFWRVAVVAEFLAMILFIFISIGSALGFHYPIKSNQTTGAVQDNVKVSLAFGLS
IATLAQSVGHISGAHLNPAVTLGLLLSCQISVLRAIMYIIAQCVGAIVATAILSGITSSL
PDNSLGLNALAPGVNSGQGLGIEIGTLQLVLCVLATTD RRRRLGCGPLAIGFSVALG
HLLAIDYTGCGINPARSFGSSVITHNFQDHWIFWVGPFIGAALAVLIYDFILAPRSSDLT
DRVKVVWTSQGVEEYDL DADDINSRVEMKPK
```

paste

[Set subsequence](#) From: To: [Choose database](#) [Do CD-Search](#) Now: or **Options** for advanced blasting[Limit by entrez query](#) or select from: [Composition-based statistics](#) [Choose filter](#) Low complexity Mask for lookup table only Mask lower case[Expect](#) [Word Size](#) [Matrix](#) Gap Costs Choice of
substitution matrix
and gap penalty

Your request has been successfully submitted and put into the Blast Queue.

Query = (270 letters)

Putative conserved domains have been detected, click on the image below for detailed results.



The request ID is

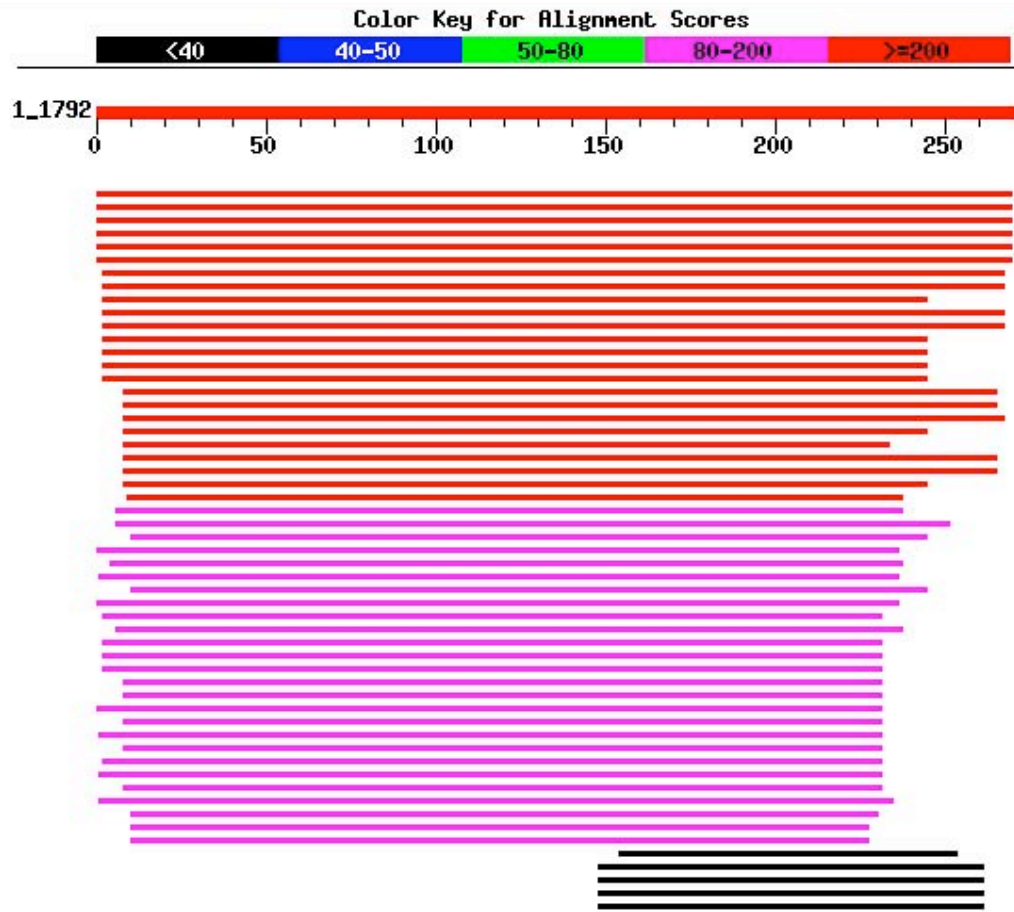
or

The results are estimated to be ready in 28 seconds but may be done sooner.

Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.



Distribution of 164 Blast Hits on the Query Sequence

Mouse-over to show define and scores. Click to show alignments



Sequences producing significant alignments:			score	n	
			(bits)	Value	
gi 1351965 sp P47865 AQP1_BOVIN	Aquaporin-CHIP (Water chann...	530	e-150	G	
gi 3023310 sp P56401 AQP1_SHEEP	Aquaporin-CHIP (Water chann...	521	e-148		
gi 267412 sp P29972 AQP1_HUMAN	Aquaporin-CHIP (Water channe...	481	e-136	G	
gi 543832 sp O02013 AQP1_MOUSE	Aquaporin-CHIP (Water channe...	477	e-134	G	
gi 47117785 sp P29975 AQP1_RAT	Aquaporin-CHIP (Water channe...	474	e-134	G	
gi 1703359 sp P50501 AOPA_RANES	Aquaporin FA-CHIP	431	e-121		
gi 730026 sp O06019 MIP_RANPI	Lens fiber major intrinsic pr...	238	1e-62		
gi 127102 sp P06624 MIP_BOVIN	Lens fiber major intrinsic pr...	236	3e-62	G	
gi 728874 sp P41181 AQP2_HUMAN	Aquaporin-CD (AQP-CD) (Water...	235	7e-62	G	
gi 127106 sp P09011 MIP_RAT	Lens fiber major intrinsic prot...	233	5e-61	G	
gi 47117800 sp P51180 MIP_MOUSE	Lens fiber major intrinsic ...	231	1e-60	G	
gi 266537 sp P30301 MIP_HUMAN	Lens fiber major intrinsic pr...	231	1e-60	G	
gi 3913084 sp O62735 AQP2_SHEEP	Aquaporin-CD (AQP-CD) (Wate...	231	2e-60		
gi 23503041 sp P56402 AQP2_MOUSE	Aquaporin-CD (AQP-CD) (Wat...	228	9e-60	G	
gi 461529 sp P34080 AQP2_RAT	Aquaporin-CD (AQP-CD) (Water c...	225	8e-59	G	
gi 1351967 sp P47863 AQP4_RAT	Aquaporin 4 (WCH4) (Mercurial...	222	8e-58	G	
gi 47116232 sp O923J4 AQP4_DIPME	Aquaporin 4	219	4e-57		
gi 1703358 sp P55064 AQP5_HUMAN	Aquaporin 5	219	5e-57	G	
gi 2506859 sp P55087 AQP4_HUMAN	Aquaporin 4 (WCH4) (Mercuri...	218	2e-56	G	
gi 7387547 sp O9WTY4 AQP5_MOUSE	Aquaporin 5	218	2e-56		
gi 7387545 sp O77750 AQP4_BOVIN	Aquaporin 4 (WCH4) (Mercuri...	217	3e-56		
gi 47117859 sp P55088 AQP4_MOUSE	Aquaporin 4 (WCH4) (Mercur...	216	4e-56	G	
gi 1351968 sp P47864 AQP5_RAT	Aquaporin 5	215	8e-56	G	
gi 32469581 sp O9NHW7 AQP_AEDAE	Aquaporin AQP Ae.a	201	1e-51		
gi 32469580 sp O25074 AQP_HAEIE	Aquaporin (Water channel 1)...	192	5e-49		
gi 2497939 sp O13520 AQP6_HUMAN	Aquaporin 6 (Aquaporin-2 li...	192	9e-49	G	
gi 47115531 sp O9WTY0 AQP6_RAT	Aquaporin 6	191	2e-48	G	
gi 21431896 sp P43287 PI22_ARATH	Aquaporin PIP2.2 (Plasma m...	189	5e-48	G	
gi 32469582 sp O9V527 AQP_DROME	Aquaporin	188	1e-47	G	
gi 1175013 sp P43286 PI21_ARATH	Aquaporin PIP2.1 (Plasma me...	187	2e-47	G	
gi 47115796 sp O8C4A0 AQP6_MOUSE	Aquaporin 6	185	1e-46	G	
gi 267136 sp P30302 PI23_ARATH	Aquaporin PIP2.3 (Plasma mem...	184	1e-46	G	
gi 32363439 sp O92V07 PI26_ARATH	Probable aquaporin PIP2.6 ...	184	1e-46	G	

Final Result: Sequence Alignment - **Approximate**

 >[gi|46395801|sp|Q88F17|AQPZ_PSEPK](#)  Aquaporin Z
Length = 230

Score = 119 bits (299), Expect = 6e-27

Identities = 70/186 (37%), Positives = 105/186 (56%), Gaps = 12/186 (6%)

```
Query: 53  VSLAFGLSIATLAQSVGHISGAHLNPAVTLGLLLSCQISVLRAIMYIIAQCVGAIVATAI 112
          V+ AFGL++ T+A ++GHISG HLNPAV+ GL++ + + Y+IAQ +GAI+A +
Sbjct: 40  VAFAFGLTVLTMFAIGHISGCHLNPAVVSFGLVVGGRFPKELLPYVIAQVIGAILAAGV 99

Query: 113 LSGITSSLP--DNSLGL--NALAP----GVNSGQGLGIEIIGTLQLVLCVLATTD RRRRD 164
          + I S + S GL N A G G G E++ T ++ ++ TD R
Sbjct: 100 IYLIASGKAGFELSAGLASNGYADHSPGGYTLGAGFVSEVVMTAMFLVVIMGATDARAP- 158

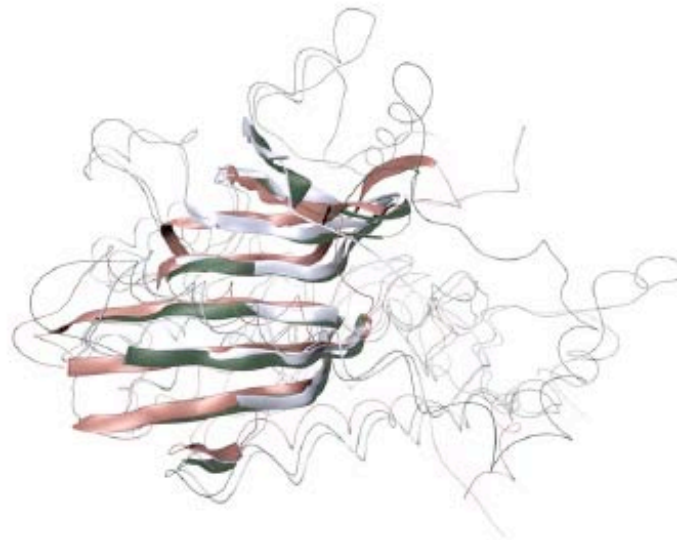
Query: 165 LGGSGPLAIGFSVALGHLLAIDYTGCGINPARSFGSSVITHNF--QDHWIFWVGPFIGAA 222
          G P+AIG ++ L HL++I T +NPARS G ++ + Q W+FWV P IGAA
Sbjct: 159 -AGFAPIAIGLALTLIHLISIPVTNTSVNPARSTGPALFVGGWALQQLWLFWVAPLIGAA 217

Query: 223 LAVLIY 228
          + +Y
Sbjct: 218 IGGALY 223
```

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

*Tutorial for the
material of this
lecture available*

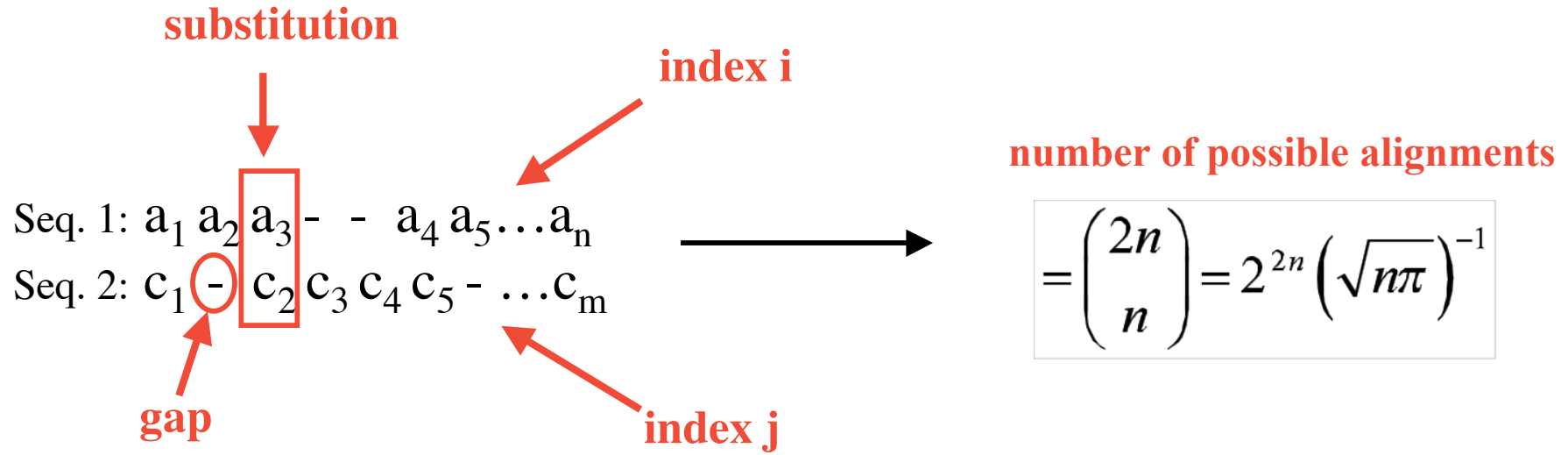


Rommie Amaro
Felix Autenrieth
Brijeet Dhaliwal
Barry Isralewitz

Zaida Luthey-Schulten
Anurag Sethi
Taras Pogorelov

June 2004

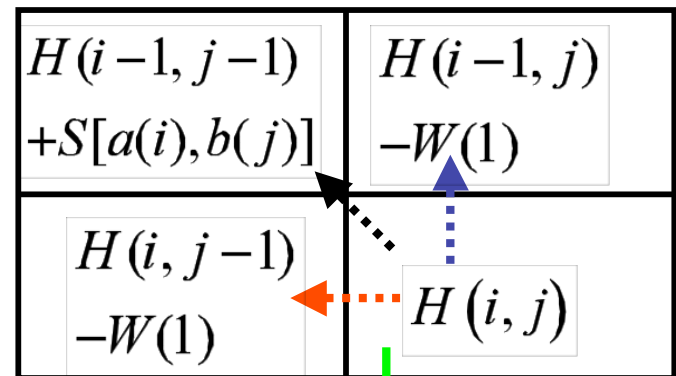
Sequence Alignment & Dynamic Programming



Smith-Waterman alignment algorithm

objective function

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



substitution matrix

gap penalty

traceback defined through choice of maximum

Blosum 40 Substitution Matrix

AA not resolved

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

Amino Acid Three Letter and One Letter Code

Amino acid	Three letter code	One letter code
alanine	ala	A
arginine	arg	R
asparagine	asn	N
aspartic acid	asp	D
asparagine or aspartic acid	asx	B
cysteine	cys	C
glutamic acid	glu	E
glutamine	gln	Q
glutamine or glutamic acid	glx	Z
glycine	gly	G
histidine	his	H
isoleucine	ile	I
leucine	leu	L
lysine	lys	K
methionine	met	M
phenylalanine	phe	F
proline	pro	P
serine	ser	S
threonine	thr	T
tryptophan	try	W
tyrosine	tyr	Y
valine	val	V

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:

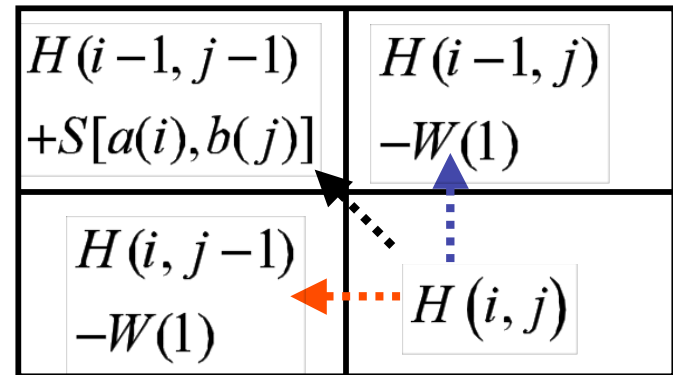
$$= \binom{2n}{n} = 2^{2n} (\sqrt{n\pi})^{-1}$$

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	
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	-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	-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	-2	-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

gap penalty $W = -6$

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

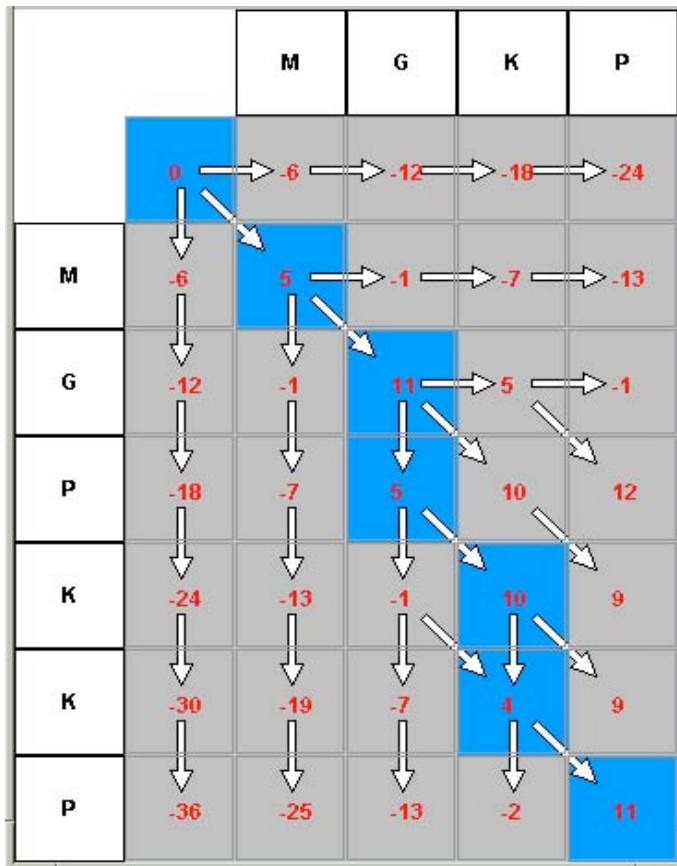
		M	G	K	P	
		0	-6	-12	-18	-24
M		-6	5	-3	-1	-2
G		-12	-3	6	-2	-2
P		-18	-2	-2	-1	7
K		-24	-1	-2	5	-1
K		-30	-1	-2	5	-1
P		-36	-2	-2	-1	7

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

Protein Structure Prediction

1-D protein sequence

SISSIRVKS KRIQLG...



3-D protein structure



Homology Modeling/ FR

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

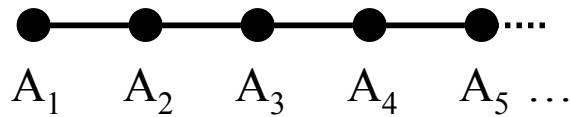
Target Sequence

SISSRVKSKRIQLGLNQAELAQKV-----GTTQ...
QFANEFKVRRIKLGYTQ-----TNVGEALAAVHGS...

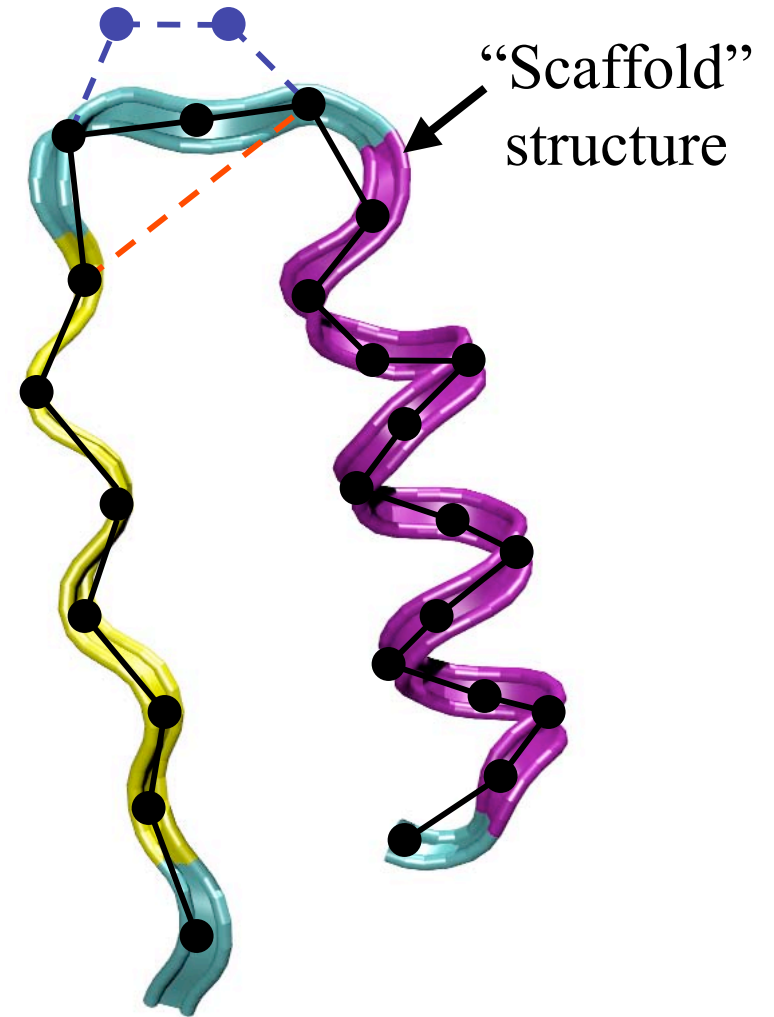
Known structure(s)

Sequence-Structure Alignment

Target sequence



Alignment between
target(s) and scaffold(s)



1. Energy Based Threading*

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

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

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

2. Sequence – Structure Profile Alignments

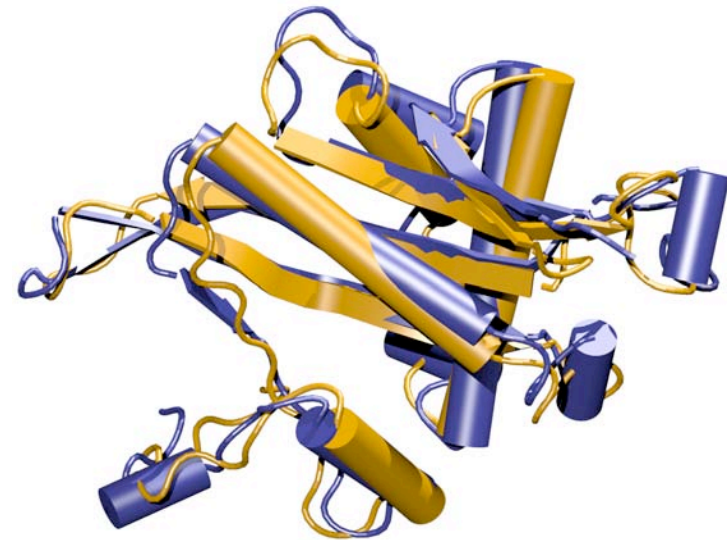
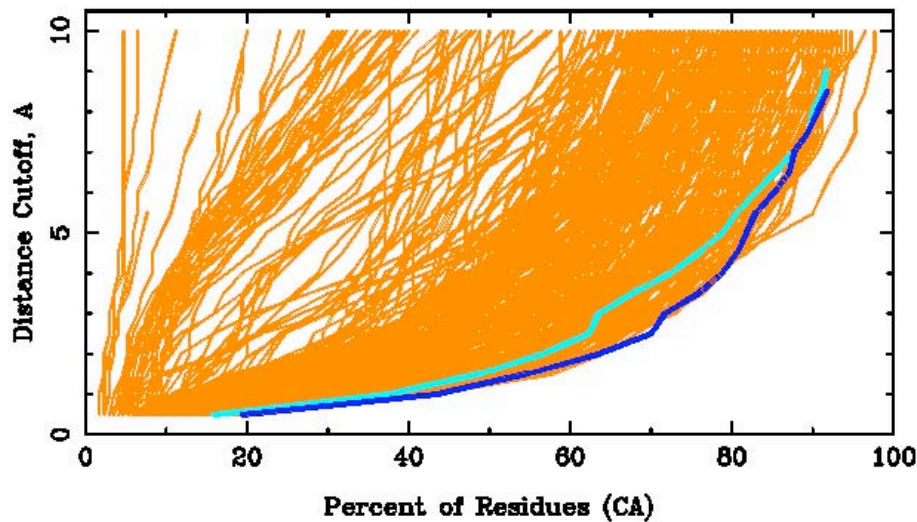
Clustal, Hidden Markov (HMMER, PSSM)
with position dependent gap penalties

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

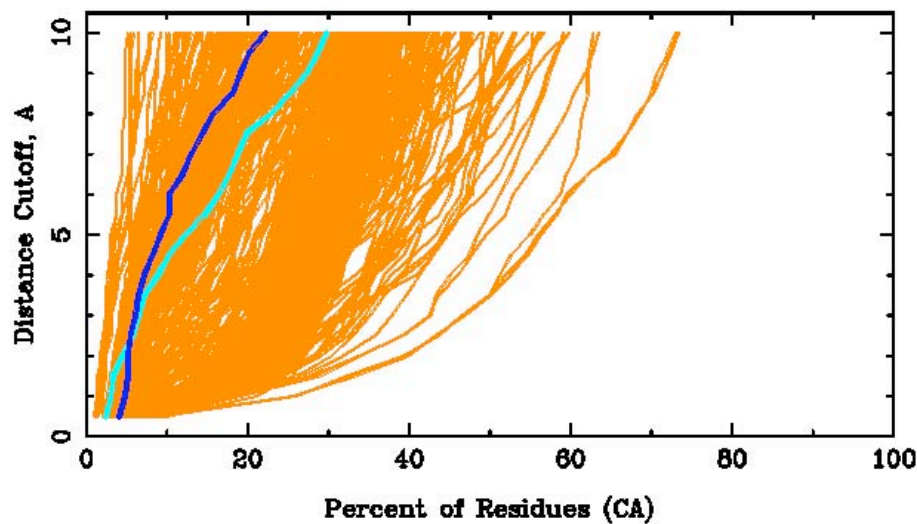
CM/Fold Recognition Results from CASP5

Lessons Learned

T0192TS093_1



T0172TS093_1



The prediction is never better than the scaffold.

Threading Energy Function and Profiles need improvement.

Structural Profiles

1. Structure more conserved than sequences!!! Similar structures at the Family and Superfamily levels.

Add more structural information

2. Which structures and sequences to include? Evolution, QR redundancy

Structural Domains

Structural Classification of Proteins

































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

Lineage:

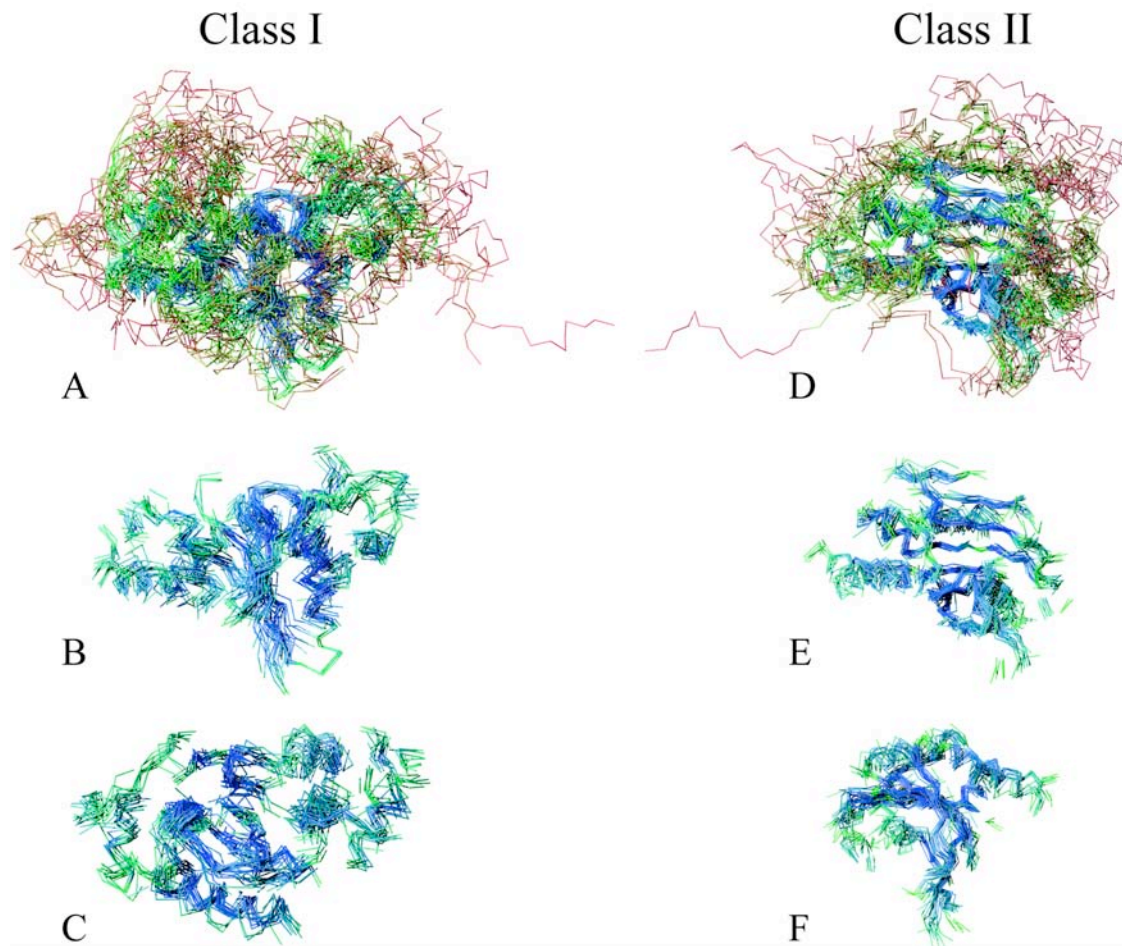
1. Root: [scop](#)
2. Class: [All beta proteins](#)
3. Fold: [OB-fold](#)
barrel, closed or partly opened n=5, S=10 or S=8; greek-key
4. Superfamily: [Nucleic acid-binding proteins](#)
5. Family: [Anticodon-binding domain](#)
barrel, closed; n=5, S=10
6. Protein: Aspartyl-tRNA synthetase (AspRS)
this 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. [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](#)   

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      -----KGLTPQSQDFSEWYLEVIQKAEALAD-YG--P-VRGTIVVRPYGY
ProRS-M_thermoautotrophic space -----EFSEWFHNILEEAEIIDQRY--P-VKGMHVWMPHGF
SerRS-T_thermophilus      --SGGG-EEEEEEES-----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 space -----HHHHHHHHHHHHTT-EE-----S-STT--EE-HHHH

SerRS-T_thermophilus      LYELALLRFAMDFMARRGFLPMTLPSYAREK-AFLG-TGHFPAYRDQVWAI-----E--
ThrRS-E_coli              TIFRELEV FVRSKLKEYQYQEVKGPFFMDRV-LWEKT-GHWDNYKDAMFTTS----S-EN
ProRS-T_thermophilus      AIWENIQQVLDRMFKETGHQNAFYPLFIPMSFL-----FSPELAVVTHAGGEELE
ProRS-M_thermoautotrophic space -----MIRKNTLKILRRILD-RDHEEVLFPLLVPEDE-LAKEAIHVKGFEDEVYVWVTHGGLSKLQ
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 space -----HHHHHHHHHHHHHTT-TT-EE----SEEEHH-HHTSHHHHHHTTTT--EEEEETEEEE

SerRS-T_thermophilus      TDLYLTGTAEVVLNALHSGEILPYEALPLRYAGYAPAFRSEA--GSFGKDVRGLMRVH-Q
ThrRS-E_coli              REYCIKPMNCPGHVQIFNQGLKSYRDLPLRMAEFGSCHR--NEPS--G-SLHGLMRVR-G
ProRS-T_thermophilus      EPLAVRPTSETVIGYMWSKWIRSWRDLPLLNQWGNVVRW--E----M-RTRPFLRTSE-
ProRS-M_thermoautotrophic space -----RKLALRPTSETVMYPMFALWVRSHDLPMPFYQVVNTFRY-ET----K-HTRPLIRVREI
SerRS-T_thermophilus      SEEEE-S-THHHHHHHTTT-EEEGGG-SEEEEEEEEE-----S--SSTTTTTTS-S-E
ThrRS-E_coli              EEEEE-S-SHHHHHHHTSS--BTTT-SEEEEE--EEE-----G--G-G-BTTTB-S-E
ProRS-T_thermophilus      EEEEE-S-SHHHHHHHHHH--BGGG--EEEEEEEE-----S-S-BTTTB-SE-
ProRS-M_thermoautotrophic space -----EEEE-SSSHHHHHHHHH--BTTT--EEEEEEEE-----S--BTTTB-SEE
```

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

PSSM-based approach

I. Construction of Profile

	1	2	3	4	5
Sequence 1	–	B	B	–	C
Sequence 2	C	C	–	–	C
Sequence 3	C	B	C	C	B

Multiple Sequence Alignment



Position Specific Amino Acid Probabilities

j	1	2	3	4	5
P(C _j)	1	0.33	0.5	1	0.67
P(B _j)	0	0.67	0.5	0	0.33

Position specific score for aligning i^{th} residue of S to j^{th} position of profile

$$Sc(S_i^j) = \log(P(S_i^j)/P(S_i^{rnd}))$$

II Database search

Align every sequence in the database to the profile using Dynamic Programming algorithm.

Sequence represented by $S(S_1, S_2, \dots, S_{n_{res}})$

Progressive alignment score = $H(i, j)$

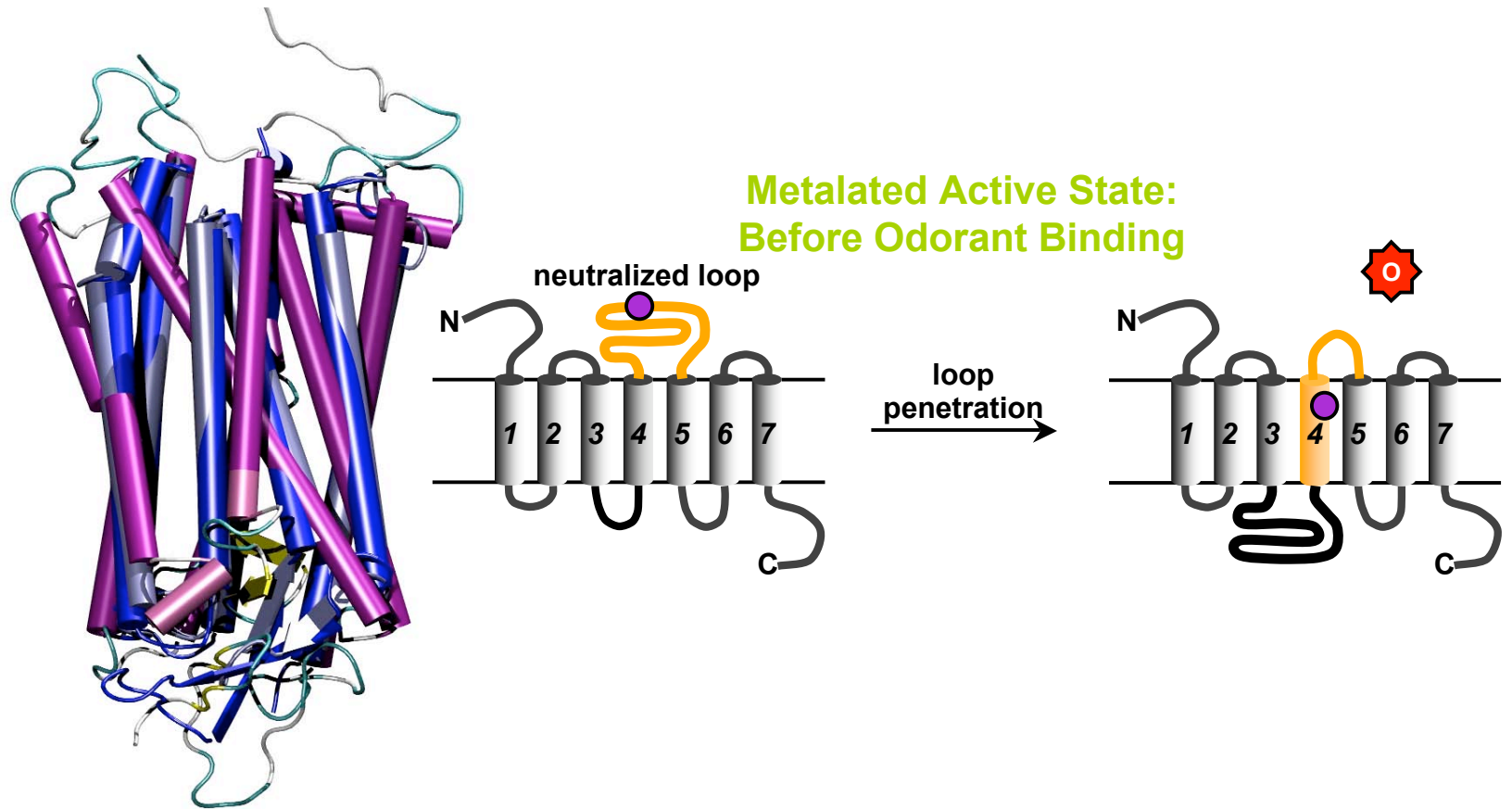
$$H(0, 0) = 0, H(i, 0) = i \times \delta, H(0, j) = j \times \delta.$$

$$H(i, j) = \max \begin{cases} H(i-1, j-1) + Sc(S_i^j) \\ H(i-1, j) + \delta \\ H(i, j-1) + \delta \end{cases}$$

for $j=1, 2, \dots, m_{aln}$ and $i=1, 2, \dots, N_{res}$

Traceback gives the optimal alignment of the sequence S to profile.

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

```
d119ha_3 MNGTEGPNFYVPFSNKTGVVRSPPFEAPQYYLAEPWQFSMLAAYNF L GFP NFLTLYVTVQH
d1e12a -----R-ENALLS SLW HVALAG IL FV NGRT--IR
d1jgja_1 -----MVGL LFW GA G GTLAFA AGRD--AG

d119ha_3 KKLRTPLNYIL NLA ADLFM FG TTTLYTSLHG YFV-F-----GPTGCNL
d1e12a PG---RPRLI GAT IPL S -SSYL G L-----S--G TVGM EMPAGHALA EMVR--SQWG
d1jgja_1 S----GERRY T L G I S G AA-V YAV A-----L--G GWVP -----ERT--VFVP

d119ha_3 EGF FAT GGE A W-SL E -LA IERYVVVCKPMSNFRFGENHA MG FTWV A CAAPPLVGW
d1e12a RY TWAL STP I LA-LGLL -A-----D----D GS FTVIAAD CTG--LA
d1jgja_1 RY DWIL TTP I SYF-LGLL -A-----G----SREF I VIT NTV M AG--FA

d119ha_3 SRYIPEGMQCSCGIDYY -PHEETNNE FVIYMFVVF I I PLIV FF-CY -QLVFTVKEAAAAT
d1e12a SA-----M--TT E L --FRN A F SCA -P S L S A LV TDW -ASA-S-----
d1jgja_1 SA-----M--VP - - - - - E R A L G A V - A I G Y Y L V G P M - T E S A - S -----

d119ha_3 TQKAEKE TR V I V A F C L P V A G V A F - Y - I F T H Q G D - F G P I F M I P A F A K - T A V Y N P
d1e12a --SA--GTA E D T L R L T V V L L G V P I V W A G V E --G -AL Q V G A T W A Y S V L D F A K Y V F
d1jgja_1 --QRSSG K S Y R L R N L T V V L A I P F W L G P P --G -AL - P T V D V A L I V L D V K V G F

d119ha_3 V I Y M - N K Q F R N C M V T T L C C G K N P L G D S T -- T V S K T E T S Q V - A P A -----
d1e12a F L L R W A N -----NERT-----VAV-----
d1jgja_1 F A L D A - A A -----
```

Clustal Profile-Profile Alignment

```
d119ha_3      NNGTEGPNFYVPPFSNKTGVVRSPPFEAPQYYLAEPWQFSNLAAYMFLLEGGFMMLTLY
die12a      -----R-ENALLSSEWENALAGLFLFVGR
d1jgja_1     -----VGLLFWLAIGMLGTLAFAAGR
IAT9__BACTERIO -----XALTRPEWVWLTALGGLTLEVV

d119ha_3      VTVQHKLRTPLNYYILLNLAADLFNFGSSDTTLYTSLNGYV-F-----
die12a      T--RPG---RPRLIGATIPLE---SSYLGLL-----S--GLTGMEMPAGHALA
d1jgja_1     D--AGS---GERYYVTLGISGIAA---YAA---L--GCWVP-----
IAT9__BACTERIO GNGSDP---DA---FYAITT---PAIAFTVLLG-----GLTVPFG-----

d119ha_3      ---GPTGCNLEGGFATLGGEA---W-SL---LAIERYVVVCKPMSNFRFGENHAMG---FT
die12a      ENVR--SQWRYTWALTP---LLA-LG---L-A-----D---DGLFTV
d1jgja_1     -ERT--FVPRYDWTTP---GYF-LG---L-A-----G---DSREFIV
IAT9__BACTERIO -GEQNP---WRYADWFTTPLL---LDLALLD-----ADQQLA

d119ha_3      WVMAACAAPPLVGSRYIPEGNQCSGIDYY---PHEETMNEFVIYMFVVHFIPLIV
die12a      IADGMCVTG--LA---A-----M--T---L---L---RFAF---AISCA--FF---LSAL
d1jgja_1     ITLTVVLAG--FAGA-----M--VP---I---ERAL---GAV-AFIG---YYL
IAT9__BACTERIO ADGIMGTG--LVGA-----LTKVYSR---VAISTA-AM---LYVL

d119ha_3      FF-CYG-QLVFTVKEAAAATTQKAKEVTRV---VIAF---CALPAGVAF-Y-IFTHQG
die12a      VTD--AASA-S-----SA--GTAEFDTLRVLTVVLLV---VWA---GVE--G---
d1jgja_1     VGPM-TESA-S-----QRSSGK---RLRNLTVVLAIPF---WL---GPP--G---
IAT9__BACTERIO FFGTSK---E-----SMRPEVASTFKLRN---TVVLS---FVWL---GSE---G

d119ha_3      D-FGPIFMTPAFFAK---AVYNPV---Y---M---NKQFRNCMVTTLCCGKNPLGDST--TVS
die12a      ALQVGATVA---SVLDVFAKYVFF---LLRW---AH-----NERT--
d1jgja_1     AL---PT---VALI---YLD---VTKVGFGF---ALDA-AA-----
IAT9__BACTERIO AGVPLN---L---VLDY---AKVGFGL---LLRSRAIFG-----EAEAP

d119ha_3      KTETSQV-APA
die12a      -----VAV-
d1jgja_1     -----
IAT9__BACTERIO EPSADGAAATS
```

Refine Structure Prediction with Modeller 6.2



Sethi and Luthey-Schulten, UIUC 2003

Modeller 6.2 A. Sali, et al.

Comparative Modeling Tools in SwissProt - SwissModel

Databases	Tools and software packages
<ul style="list-style-type: none">● Swiss-Prot and TrEMBL - Protein knowledgebase● PROSITE - Protein families and domains● SWISS-2DPAGE - Two-dimensional polyacrylamide gel electrophoresis● ENZYME - Enzyme nomenclature● SWISS-3DIMAGE - 3D images of proteins and other biological macromolecules● SWISS-MODEL Repository - Automatically generated protein models ● GermOnLine - Knowledgebase on germ cell differentiation● Ashbya Genome Database● Links to many other molecular biology databases	<ul style="list-style-type: none">● Proteomics and sequence analysis tools<ul style="list-style-type: none">○ Proteomics [Aldente (PMF) ^{new}, PeptideMass, ...]○ DNA -> Protein [Translate]○ Similarity searches [BLAST]○ Pattern and profile searches [ScanProsite]○ Post-translational modification and topology prediction○ Primary structure analysis [ProtParam, pI/MW, ProtScale]○ Secondary and tertiary structure prediction [SWISS-MODEL, Swiss-PdbViewer]○ Alignment [T-COFFEE, SIM]○ Biological text analysis● ImageMaster / Melanie - Software for 2-D PAGE analysis● MSight - Mass Spectrometry Imager● Roche Applied Science's Biochemical Pathways

Structure Prediction Resources - on line

http://cgat.ukm.my/spores/Predictory/structure_prediction.html

Fold Recognition / Secondary Structure Prediction	
WWW Servers	
3D-PSSM	A Fast, Web-based Method for Protein Fold Recognition using 1D and 3D Sequence Profiles coupled with Secondary Structure and Solvation Potential Information. Biomolecular Modelling Group, Imperial Cancer Research Fund, UK
BCM PSSP	Protein secondary structure prediction, Baylor College of Medicine, USA.
bioinbgu	Fold-recognition services based on Sequence-Derived Properties - Computer Science Dept., Ben Gurion University, Israel
CODA	A combined algorithm for predicting protein loops. CODA can predict fragments of length 3-8 using the consensus methodology. Biochemistry Dept. Cambridge, UK
CONSENSUS SECONDARY STRUCTURE	Submits input sequence to multiple servers for consensus
NNPREDICT	Protein secondary structure prediction. Dept of Cellular & Molecular Pharmacology, UCSF, USA
PREDATOR	Protein secondary structure prediction from single sequence or a set of sequences, EMBL-Heidelberg, Germany
Pred2ary	Secondary structure and class prediction server, Cohen Lab, Dept of Cellular & Molecular Pharmacology, UCSF, USA
PredictProtein	PredictProtein is a service for sequence analysis, and structure prediction. Mirrors - America, n Asia, Australia, Europe
PROF	B. Rost : PROFfile based neural network prediction
PROF	R.D. King :Cascaded multiple classifiers for protein secondary structure prediction. - Neural Networks - Bayesian classifiers - Linear discrimination - Quadratic discrimination, Computational Biology Group, Univ. of Wales, Aberystwyth, UK.
PSCAN	A combined sequence structure profile fold recognition method - alignment of 2 sequences.
PSI Pred /Gen Threader	Protein structure prediction server (V2.0) using PSIPRED - a highly accurate secondary structure prediction method, GenTHREADER - a new sequence profile based fold recognition method & MEMSAT2 - a widely used transmembrane topology prediction method, David Jones et al, Brunel Univ. UK
RPFOLD	This is a simple method based on amino acid properties.
SAM-T99/T98	Hidden Markov Models applications, Computational Biology, UCSC, USA
SAWTED	SAWTED stands for Structure Assignment With Text Description. It is a method to improve the coverage of the detection of remote homologues of known structure by sequence searches d (e.g. PSI-BLAST) and fold recognition programs.
ssPsi	A fold recognition/homology modelling server that uses predicted secondary structure and evolutionary information.
UCLA-DOE	Protein fold recognition by threading. UCLA-DOE, USA

Cont'd - Web Resources

Ab - Initio Methods

WWW Servers

ELAN-PROT	Uses an elastic net combinatorial optimization to determine the lowest-energy conformation for a residue-residue energy function. This energy function incorporates information from secondary structure prediction methods, and also adds a term for any known non-local residue contacts. Side-chain detail is presently being added to the protein model. Side chains will be added to the predicted C _α trace by adding a full backbone and then using SCWRL.
Isites	Sequences are converted to sequence profiles using PSI-BLAST. Fragments of structure are predicted using the I-sites Library, which also produces a moveset for the next step. Rosetta uses the I-sites fragments to do a Monte Carlo Fragment Insertion conformational search for pieces of length 36-50. These pieces are spliced together using a genetic algorithm, to create the coordinates for the complete sequence.
PETRA	An <i>ab-initio</i> protein fragment prediction method, Biochemistry Dept., Cambridge, UK

Model Building

WWW Servers

AL2TS	Translates sequence-structure alignment into tertiary structure. PredictionCenter, Lawrence Livermore Natl. Lab.
-----------------------	--

Software / Programs

Homology	Builds models from manual alignment and assignment of SCRs, loops and side chains (rotamers).
MODELLER	MODELLER-4: Homology protein structure modelling by satisfaction of spatial restraints, Andrej Sali Lab, Rockefeller University, USA.