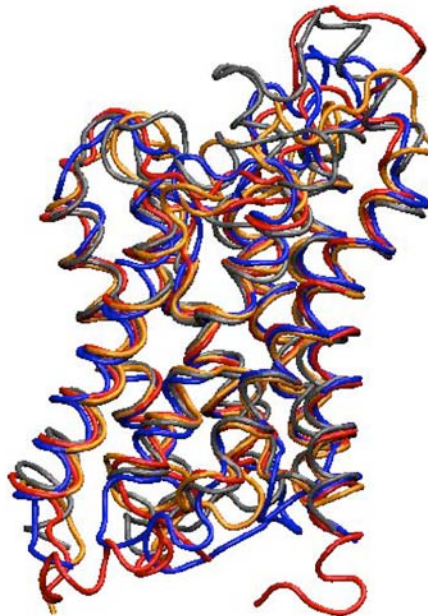


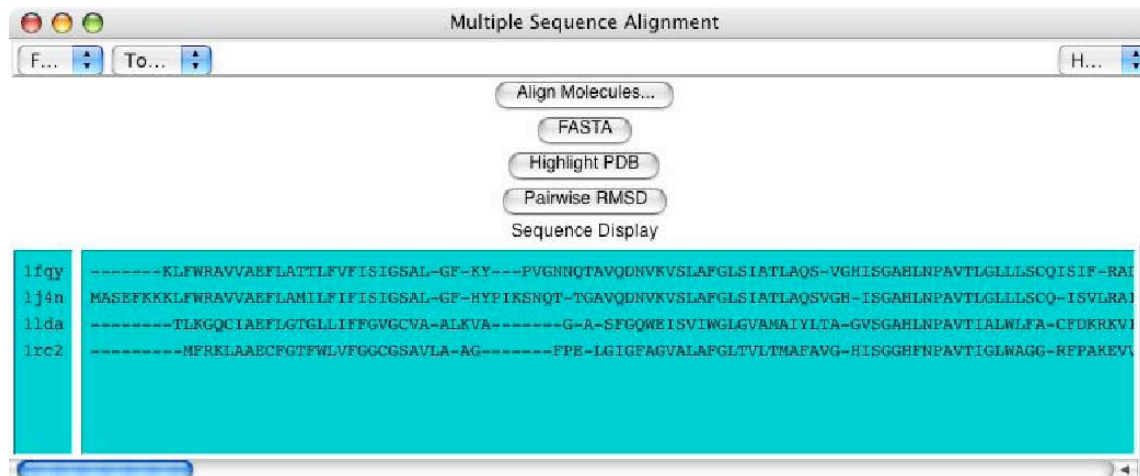
# Sequence and Structure Alignment

Z. Luthey-Schulten, UIUC

San Francisco, 2005



VMD 1.83



Multiple Sequence Alignment

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

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

```
1fqy -----KLFWRVVAEFLATLTFVFTSIGSAL-GF-KY---PVGNNQTAVQDNVKS LAPGLS IATLAQS-VGHISGAHLNPAVTLG LLLSCQISIF-RAI
1j4n MASEFKKLFWRVVAEFLAMILFIFLISIGSAL-GF-HYPIKSNQT-TGAVQDNVKS LAPGLS IATLAQSVGH-ISGAHLNPAVTLG LLLSCQ-I S VLRAI
1lda -----TLKGQCIAEFLGTGLLIFFGVGCVA-ALKVA-----G-A-SFGQWEISVINGLGVAMAIYLTA-GVSGAHLNPAVTLALWLF A-CFDKRV I
1rc2 -----MFRKLA AECFGTFWLVFGGCGSAVLA-AG-----FPE-LGIGFAGVALAPGLTVLTMFAVVG-HISGGHFNPAVTIGLWAGG-RFP AKEV I
```



# 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...](#))

 <a href="#">ExPASy Home page</a>	<a href="#">Site Map</a>	<a href="#">Search ExPASy</a>	<a href="#">Contact us</a>	<a href="#">Swiss-Prot</a>					
<a href="#">Hosted by NCSC US</a>	Mirror sites:	<a href="#">Australia</a>	<a href="#">Bolivia</a>	<a href="#">Brazil <small>new</small></a>	<a href="#">Canada</a>	<a href="#">China</a>	<a href="#">Korea</a>	<a href="#">Switzerland</a>	<a href="#">Taiwan</a>
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](#)<sub>(89)</sub> and [TrEMBL](#)<sub>(122)</sub>: **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

#### [AQP1\\_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)

#### [AQP1\\_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)

#### [AQP1\\_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	<b>AQP1_BOVIN</b>
Primary accession number	<b>P47865</b>
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	<b>Aquaporin-CHIP</b>
Synonyms	<b>Water channel protein for red blood cells and kidney proximal tubule Aquaporin 1 Water channel protein CHIP29</b>
Gene name	<b>Name: AQP1</b>
From	<b>Bos taurus (Bovine)</b> [TaxID: 9913]
Taxonomy	<b>Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Cetartiodactyla; Ruminantia; Pecora; Bovidae; Bovinae; Bos.</b>

## 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
ASEFKKKLFW 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 following text is displayed:

```
>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).
ASEFKKKLFWRAVVAEFLAMILFIFISIGSALGFHYPIKSNQTTGAVQDNVKVSLAFGLS
IATLAQSVGHISGAHLNPAVTLGLLLSCQISVLRAIMYIIAQCVGAIIVATAILSGITSSL
PDNSLGLNALAPGVNSGQGLGIEIIGTLQLVLCVLATDRRRDLGGSGPLAIGFSVALG
HLLAIDYTCGGINPARSFGSSVITHNFQDHWIFWVGPFIGAALAVLIYDFILAPRSSDLT
DRVKVWTSQVEEYDL DADD INSRVEMKPK
```

cut



[Search](#)

```
ASEFKKLFWRVAVVAEFLAMILFIFISIGSALGFHYPIKSNQTTGAVQDNVKVSLAFGLS
IATLAQSVGHISGAHLNPAVTLGLLSQCISVLRIMYIIAQCVGAIVATAILSGITSSL
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



Nucleotide

Protein

Translations

# formatting BLAST

Retrieve results for an RID

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

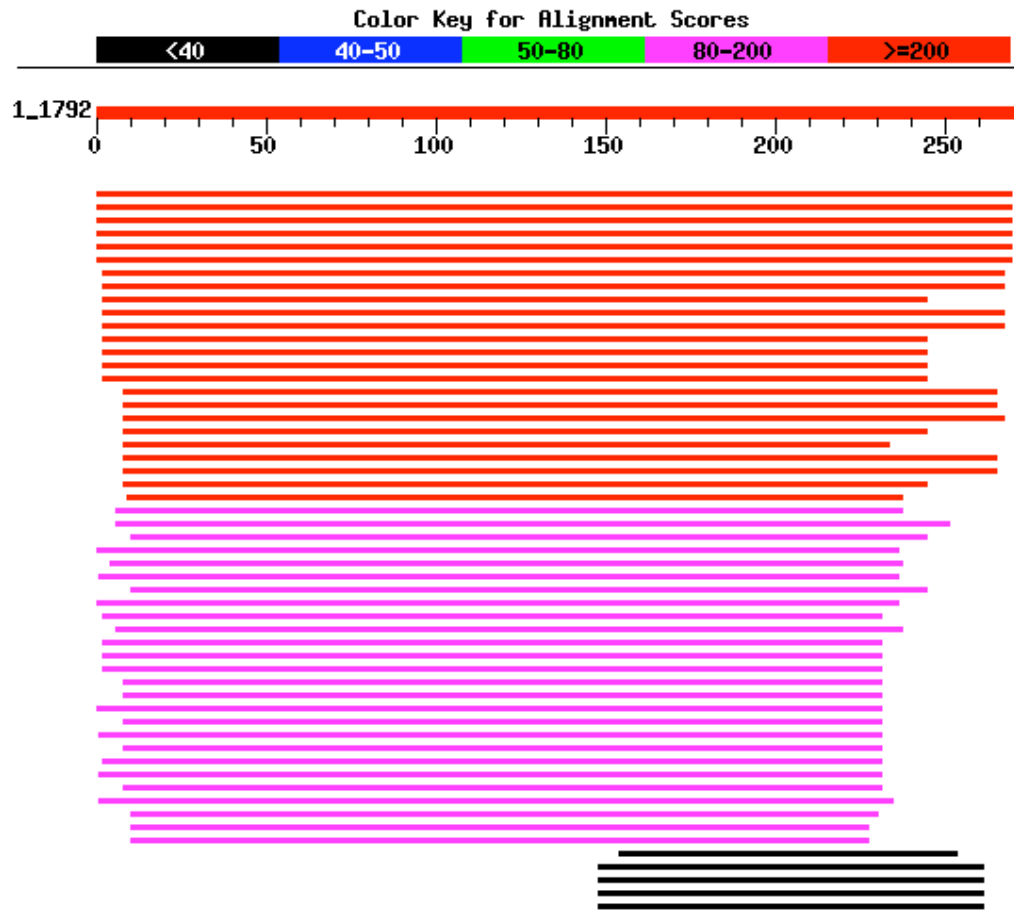
or

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

## 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 VAFAFGLTVLTMAFAIGHISGCHLNPAVVSFGLVVGGRFPAKELLPYVIAQVIGAILAAGV 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

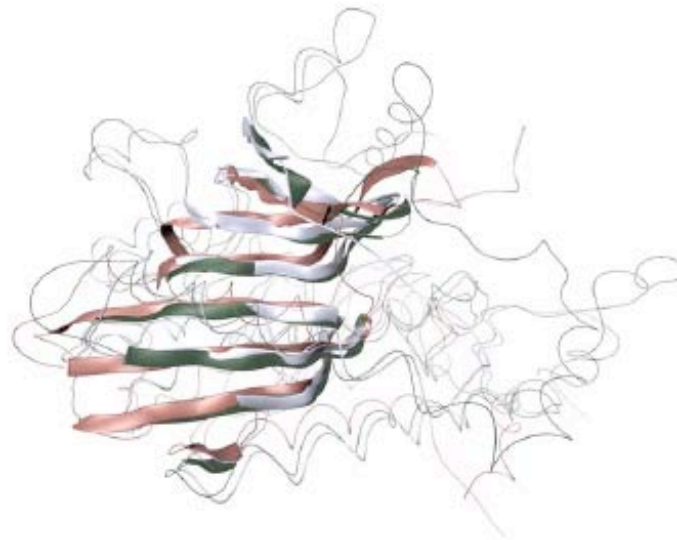
Sbjct: 218 IGGALY 223

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*



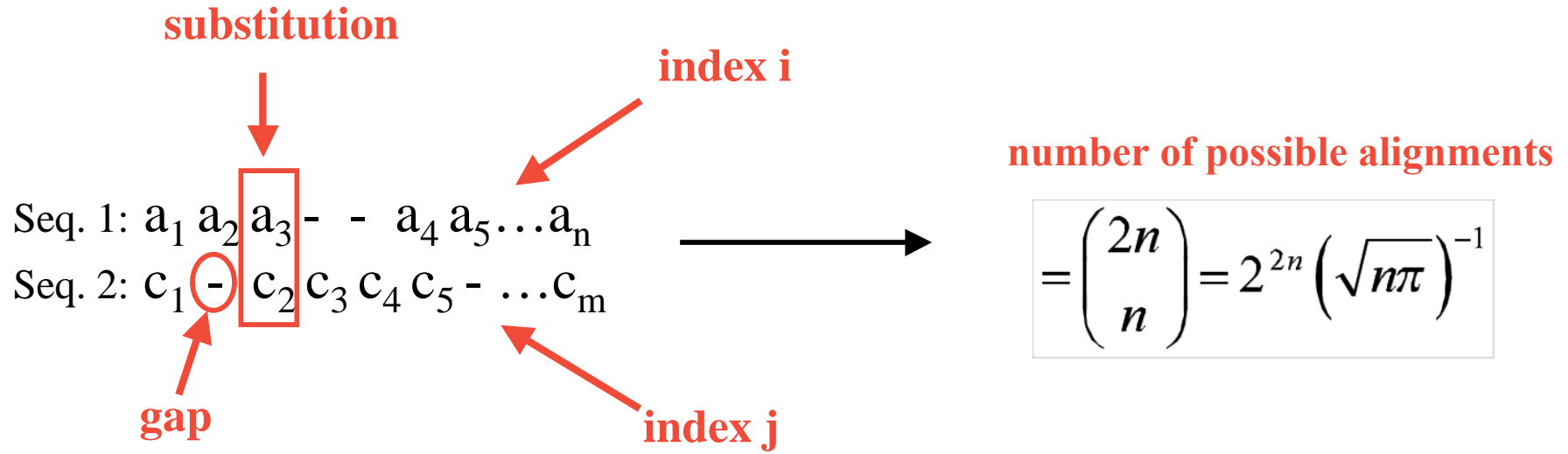
Rommie Amaro  
Felix Autenrieth  
Brijeet Dhaliwal  
Barry Isralewitz

Zaida Luthey-Schulten  
Anurag Sethi  
Taras Pogorelov

June 2004



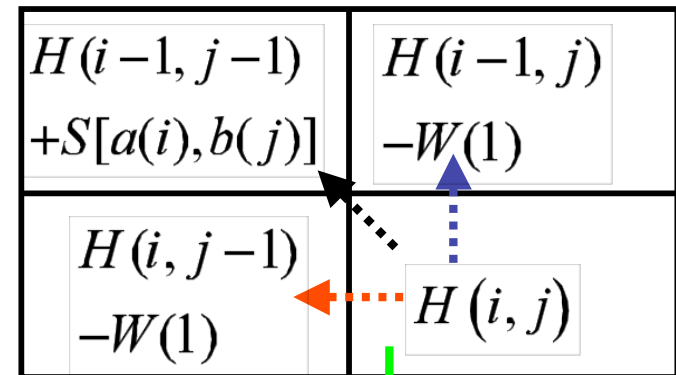
# Sequence Alignment & Dynamic Programming



## 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:

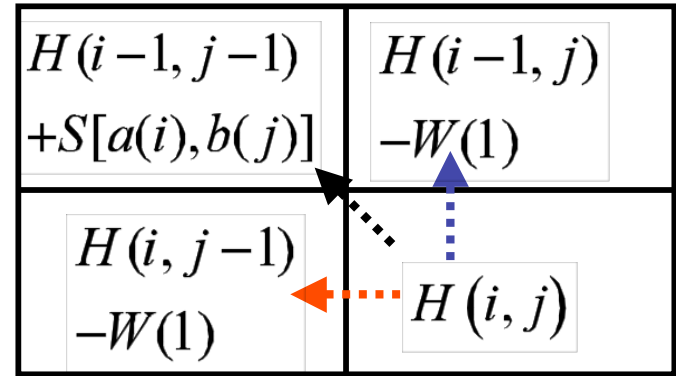
$$= \binom{2n}{n} = 2^{2n} (\sqrt{n\pi})^{-1}$$

## 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	-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	-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	-2	-1	-2	-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	0	-1	-2	0	0	-2	-1	-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			
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

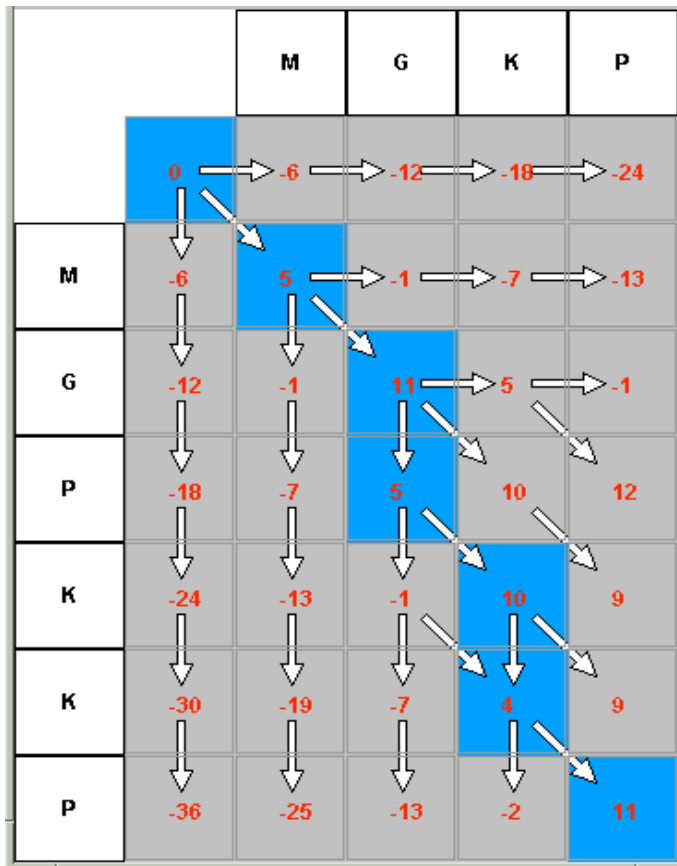
# Filling out the Score Matrix H

	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

<b>M</b>	G	-	K	-	P
:	:		:		:
<b>M</b>	G	P	K	K	P

Traceback (blue) from optimal score

# Protein Structure Prediction

1-D protein sequence

SISSIRVKS KRIQLG...

Homology Modeling/ FR

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

Target Sequence

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

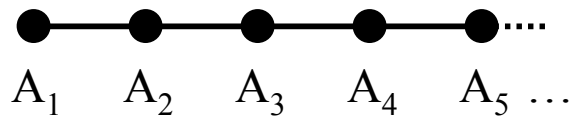
Known structure(s)

3-D protein structure

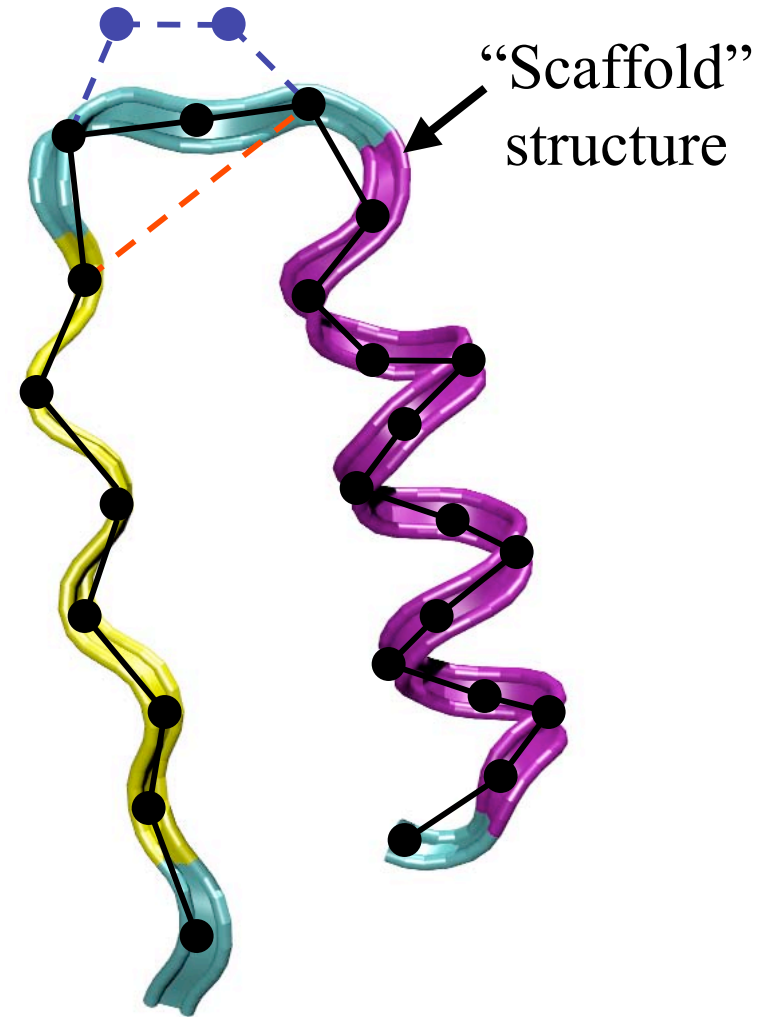


# Sequence-Structure Alignment

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

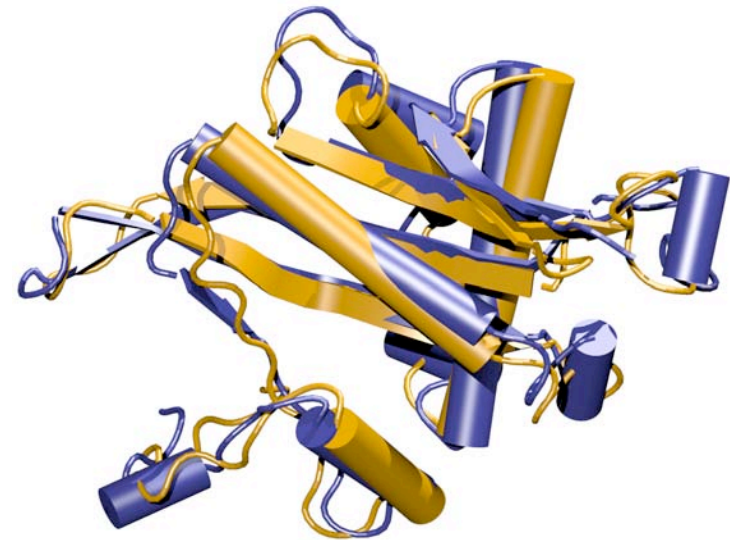
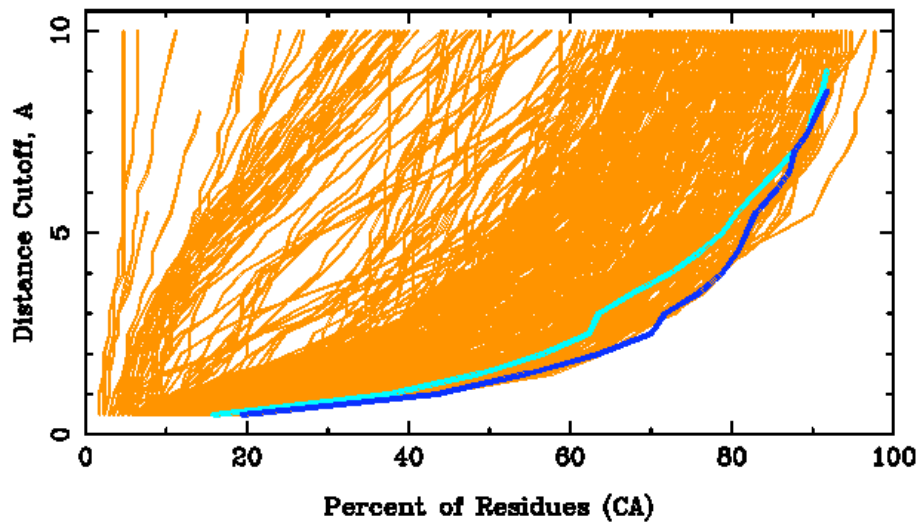
Clustal, Hidden Markov (HMMER, PSSM)  
with position dependent gap penalties

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

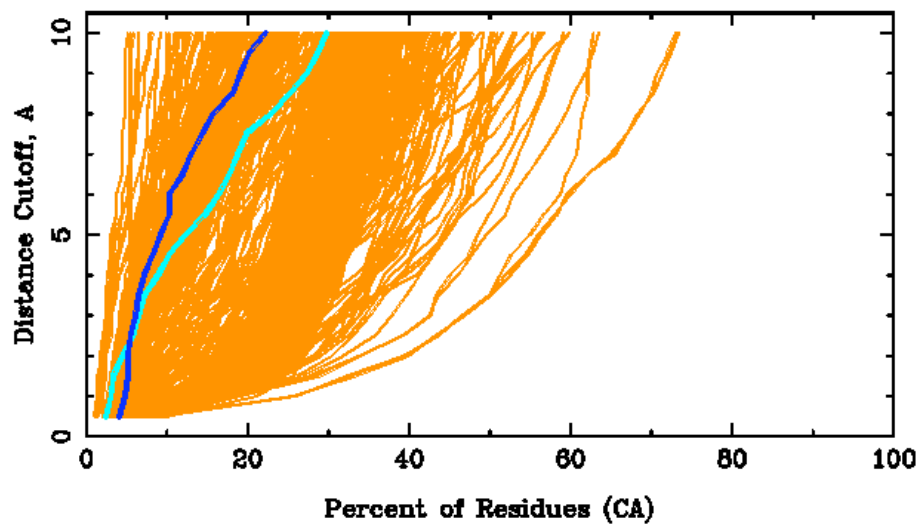
# CM/Fold Recognition Results from CASP5

## 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:

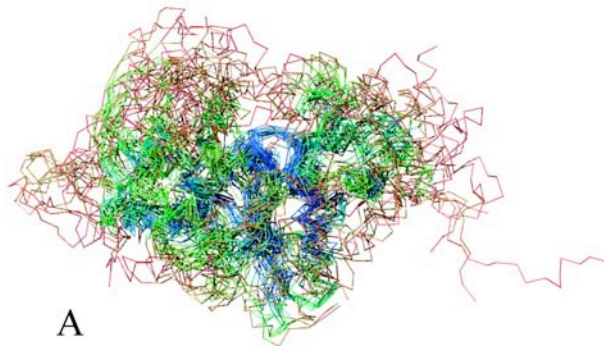
1. [1c0a](#)    
  1. [region a:1-106](#)   
2. [1il2](#)      
*complexed with 1mg, 5mc, 5mu, amo, h2u, psu, so4*
  1. [region a:1-106](#)   
  2. [region b:1001-1106](#)   
3. [1eqr](#)      
*complexed with mg*
  1. [region a:1-106](#)   
  2. [region b:1-106](#)   
  3. [region c:1-106](#)   



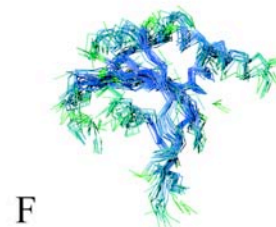
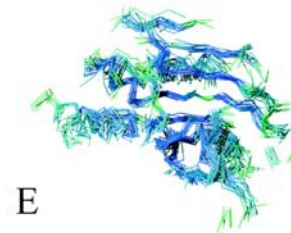
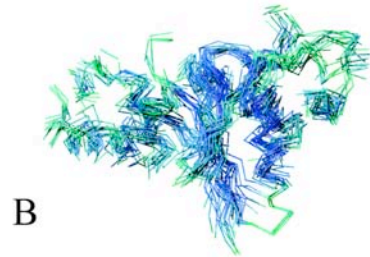
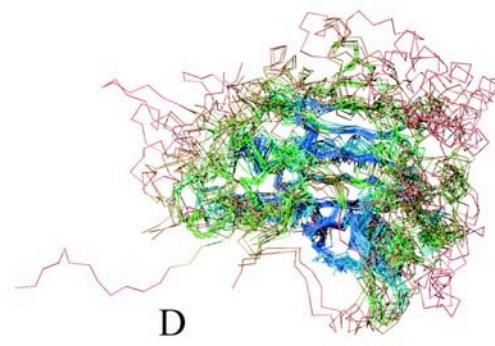
# Profile - Multiple Structural Alignments

Representative Profile of AARS Family  
Catalytic Domain

Class I



Class II



# 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_thermoautotrophicus
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_thermoautotrophicus
                           -----HHHHHHHHHHHHTT-EE-----S-STT--EE-HHHH

SerRS-T_thermophilus     LYELALLRFAMDFMARRGFLPMTLPSYAREK-AFLG-TGHFPAYRDQVWAI-----E--
ThrRS-E_coli              TIFRELEVFVRSKLKEYQYQEVKGPFFMMDRV-LWEKT-GHWDNYKDAMFTTS----S-EN
ProRS-T_thermophilus     AIWENIQQVLDRMFKETGHQNAFYPLFIPMSFL-----FSPELAVVTHAGGEELE
ProRS-M_thermoautotrophicus
space                    MIRKNTLKILRRILD-RDHEEVLFPLLVPEDE-LAKEAIVKGFEDVYVWVTHGGLSKLQ
SerRS-T_thermophilus     HHHHHHHHHHHHHHHHTT-EEEE--SEEEHH-HHHH-HT-TTTGGGS-B-T-----T--
ThrRS-E_coli              HHHHHHHHHHHHHHHHTT-EE----SEEEHH-HHHTT-THHHHGGG--EEE----E-TT
ProRS-T_thermophilus     HHHHHHHHHHHHHHHHTT-EE----SEESTT-----TT--EEEE-SSSEEE
ProRS-M_thermoautotrophicus
                           HHHHHHHHHHHHHHTT-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
space                    RKLALRPTSETVMYPMFALWVRSHDLPMPRFYQVVNTFRY-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_thermoautotrophicus
                           EEEEE-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
<b>Sequence 1</b>	-	B	B	-	C
<b>Sequence 2</b>	C	C	-	-	C
<b>Sequence 3</b>	C	B	C	C	B

Multiple Sequence Alignment



Position Specific Amino Acid Probabilities

j	1	2	3	4	5
P(C <sub>j</sub> )	1	0.33	0.5	1	0.67
P(B <sub>j</sub> )	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** =  $H(i, j)$

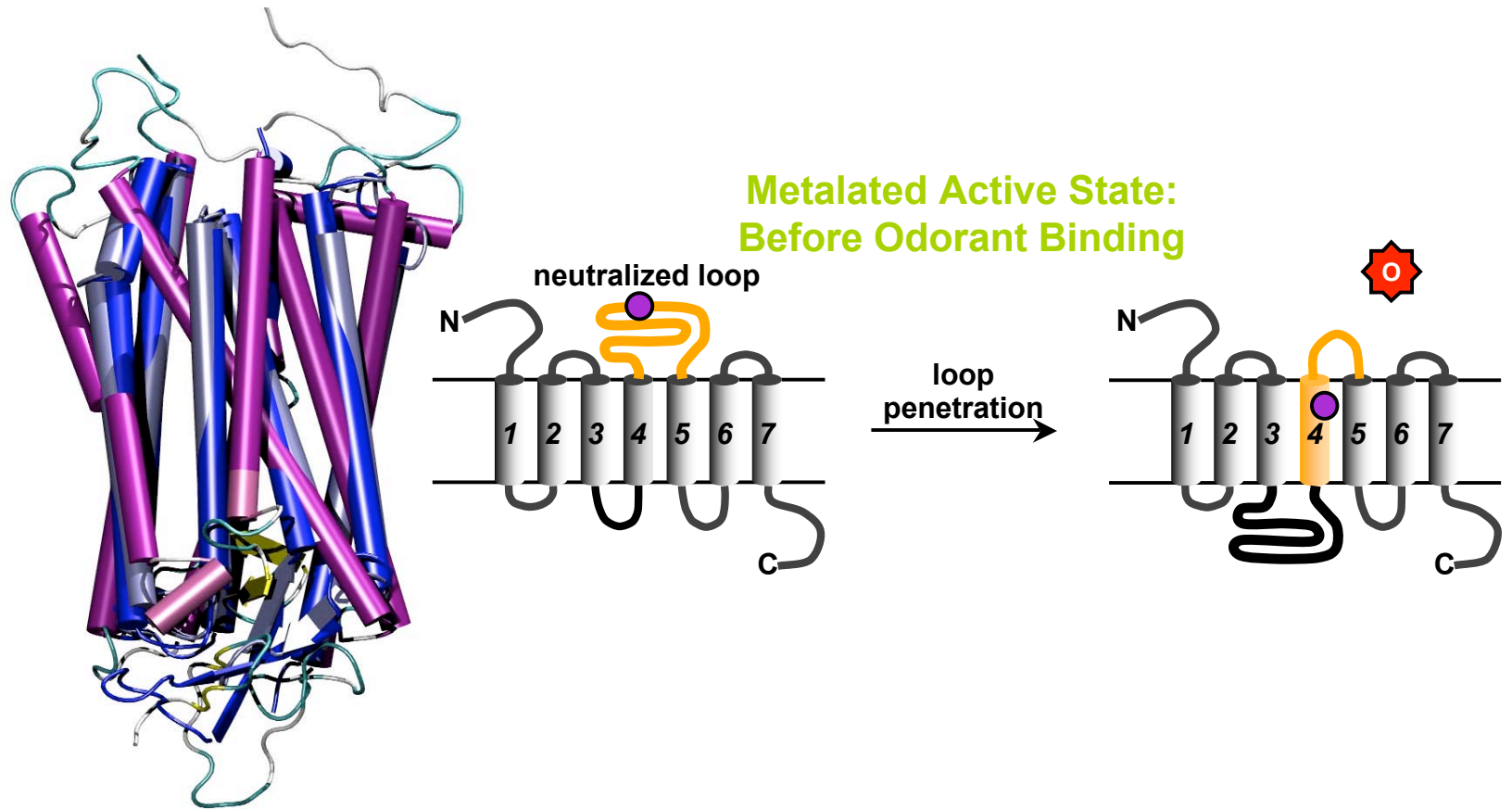
$$H(0, 0) = 0, H(i, 0) = i \times \delta, H(0, j) = j \times \delta.$$

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

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.

# HMM / Clustal 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 - 3 Structures

```
d119ha_3 MNGTEGPNFYVPFSNKTGVVRSPPFEAPQYYLAEPWQFSMLAAYNF L L GFP NFLTLYVTVQH
d1e12a -----R-ENALLS SLW HVALAG IL FV NGRT--IR
d1jgja_1 -----MVGL LFW GA G G TLAFA 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 TL GISG AA-V YAV A-----L--G GWVP -----ERT--VFVP
```

```
d119ha_3 EGF FAT GGE A W-SL - 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 YF-LGLL -A-----G----SREF IVIT NTV M AG--FA
```

```
d119ha_3 SRYIPEGMQCSCGIDYY -PHEETNNE FVIYMFVVF I I PLIV FF-CY -QLVFTVKEAAAAT
d1e12a A-----M--TT L -FRNAP SCA-F S L S ALVTDW -ASA-S-----
d1jgja_1 A-----M--VP - -ERIAL GAV-AE IG Y YLVGPM-TE SA-S-----
```

```
d119ha_3 TQKAEKE TR V I V A F C L P VAGVAF-Y-IFTHQG D-FGPIFM I PAF AK-T AVYNP
d1e12a --SA--GTAEL DTLR LTVVLL GYPIVWA GVE--G -AL Q V GAT WAYSVLD FAKYVF
d1jgja_1 --QRSSG KS RLRNLTVVLL A I P F WL GPP--G -AL -PTVDVALIV LD V KVGF
```

```
d119ha_3 V Y M -NKQFRNCMVTTLCGGKNPLGDST--TVSKTETSQV-APA-----
d1e12a F LLRW AN-----NERT-----VAV-----
d1jgja_1 F ALDA-AA-----
```



# Clustal Profile-Profile Alignment

Profile 1 Structures  
Profile 2 Sequence

```

d119ha_3      NNGTEGPNFYVFPFSNKTGVVRSPPFEAPQYYLAEPWQFSMLAAYMFLLI  G  G  P  N  L  T  L  Y
d1e12a      -----R-ENALLS  S  W  N  A  L  A  G  L  L  F  V  A  G  R
d1jgja_1      -----V  G  L  L  F  W  L  A  I  G  M  L  G  T  L  A  F  A  A  G  R
1AT9___BACTERIO -----X  A  T  G  R  P  E  W  W  L  S  G  T  A  L  G  G  T  L  S  P  V  E

d119ha_3      VTVQHKKLRTPLNHYILLNLA  A  D  L  F  M  F  G  S  S  T  T  L  Y  T  S  L  N  G  Y  V  F  -----
d1e12a      T--R  P  G  ---R  P  R  L  I  G  A  T  F  I  P  L  S  S  S  Y  L  G  L  L  ---S  ---G  L  T  G  M  E  M  P  A  G  H  A  L  A
d1jgja_1      D--A  G  S  ---G  E  R  F  Y  Y  V  T  L  G  I  S  G  I  A  A  S  Y  A  S  T  A  ---L  ---G  W  V  P  F  -----
1AT9___BACTERIO G  N  G  S  D  P  ---D  A  S  F  Y  A  I  T  T  P  A  I  A  F  T  Y  L  S  L  L  G  ---G  L  T  V  P  F  G  -----

d119ha_3      ---G  P  T  G  C  N  L  E  G  F  F  A  T  L  G  G  E  A  W  S  L  G  L  A  I  E  R  Y  V  V  V  C  K  P  M  S  N  F  R  F  G  E  N  H  A  M  G  F  T
d1e12a      E  M  V  R  ---S  Q  W  R  Y  T  W  A  L  T  P  S  L  L  A  L  G  L  A  ---D  ---D  G  G  F  T  V
d1jgja_1      -E  R  T  ---V  P  R  Y  D  W  L  T  T  P  L  Y  F  L  G  L  A  ---G  ---D  S  R  E  F  I  V
1AT9___BACTERIO -G  E  Q  N  P  W  R  Y  A  D  W  F  T  T  P  L  L  D  L  L  A  L  L  D  ---A  D  Q  G  L  A

d119ha_3      W  V  M  A  A  C  A  A  P  P  L  V  G  M  S  R  Y  I  P  E  G  M  Q  C  S  C  G  I  D  Y  Y  -P  H  E  E  T  N  N  E  F  V  I  Y  M  F  V  V  H  F  I  P  L  I  V
d1e12a      I  A  D  G  M  C  V  T  G  ---L  A  A  ---M  ---T  T  L  ---L  R  R  A  F  A  I  S  C  A  -P  F  S  L  S  A  L
d1jgja_1      I  T  L  T  V  V  L  A  G  ---F  A  G  A  ---M  ---V  P  S  ---I  E  R  A  L  S  G  A  V  -A  F  I  G  S  Y  Y  L
1AT9___BACTERIO S  A  D  G  I  M  G  T  C  ---L  V  G  A  ---L  T  K  V  Y  S  R  V  S  A  I  S  T  A  -A  M  Y  L  Y  V  L

d119ha_3      F  F  -C  Y  G  -Q  L  V  F  T  V  K  E  A  A  A  A  T  T  Q  K  A  E  K  E  V  T  R  V  V  V  V  I  A  F  C  L  P  V  A  G  V  A  F  -Y  -I  F  T  H  Q
d1e12a      V  T  D  S  -A  A  S  A  -S  ---S  A  ---G  T  A  E  F  D  T  L  R  V  L  T  V  V  L  L  S  Y  P  V  W  A  G  V  E  ---G  ---
d1jgja_1      V  G  P  M  -T  E  S  A  -S  ---Q  R  S  S  G  K  S  S  R  L  R  N  L  T  V  V  L  A  I  Y  P  F  W  L  G  P  P  ---G  ---
1AT9___BACTERIO F  F  G  T  S  K  E  ---S  M  R  P  E  V  A  S  T  F  K  L  R  N  T  V  V  L  S  Y  P  V  W  L  G  S  E  ---G

d119ha_3      D  -F  G  P  I  F  M  T  P  A  F  F  A  K  -G  A  V  Y  N  P  V  Y  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
d1e12a      A  L  Q  V  G  A  T  W  A  S  V  L  D  V  F  A  K  Y  V  F  F  L  L  R  V  A  N  ---N  E  R  T  ---
d1jgja_1      A  L  -P  T  S  V  A  L  I  Y  L  D  V  T  K  V  G  F  G  F  A  L  D  A  -A  A  ---
1AT9___BACTERIO A  G  V  P  L  N  E  T  L  S  V  L  D  V  A  K  V  G  F  G  L  L  L  R  S  R  A  I  F  G  ---E  A  E  A  P

d119ha_3      K  T  E  T  S  Q  V  -A  P  A
d1e12a      -----V  A  V
d1jgja_1      -----
1AT9___BACTERIO E  P  S  A  D  G  A  A  A  T  S
    
```

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

# Structure Prediction Resources - on line

[http://cgat.ukm.my/spores/Predictory/structure\\_prediction.html](http://cgat.ukm.my/spores/Predictory/structure_prediction.html)

Fold Recognition / Secondary Structure Prediction	
WWW Servers	
<a href="#">3D-PSSM</a>	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
<a href="#">BCM PSSP</a>	Protein secondary structure prediction, Baylor College of Medicine, USA.
<a href="#">bioinbgu</a>	Fold-recognition services based on Sequence-Derived Properties - Computer Science Dept., Ben Gurion University, Israel
<a href="#">CODA</a>	A combined algorithm for predicting protein loops. CODA can predict fragments of length 3-8 using the consensus methodology. Biochemistry Dept. Cambridge, UK
<a href="#">CONSENSUS SECONDARY STRUCTURE</a>	Submits input sequence to multiple servers for consensus
<a href="#">NNPREDICT</a>	Protein secondary structure prediction. Dept of Cellular & Molecular Pharmacology, UCSF, USA
<a href="#">PREDATOR</a>	Protein secondary structure prediction from single sequence or a set of sequences, EMBL-Heidelberg, Germany
<a href="#">Pred2ary</a>	Secondary structure and class prediction server, Cohen Lab, Dept of Cellular & Molecular Pharmacology, UCSF, USA
<a href="#">PredictProtein</a>	PredictProtein is a service for sequence analysis, and structure prediction. Mirrors - America, Asia, Australia, Europe
<a href="#">PROF</a>	B. Rost : PROFfile based neural network prediction
<a href="#">PROF</a>	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.
<a href="#">PSCAN</a>	A combined sequence structure profile fold recognition method - alignment of 2 sequences.
<a href="#">PSI Pred /Gen Threader</a>	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
<a href="#">RPFOLD</a>	This is a simple method based on amino acid properties.
<a href="#">SAM-T99/T98</a>	Hidden Markov Models applications, Computational Biology, UCSC, USA
<a href="#">SAWTED</a>	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 (e.g. PSI-BLAST) and fold recognition programs.
<a href="#">ssPsi</a>	A fold recognition/homology modelling server that uses predicted secondary structure and evolutionary information.
<a href="#">UCLA-DOE</a>	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<sub>α</sub> 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 *ab-initio* 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](#)

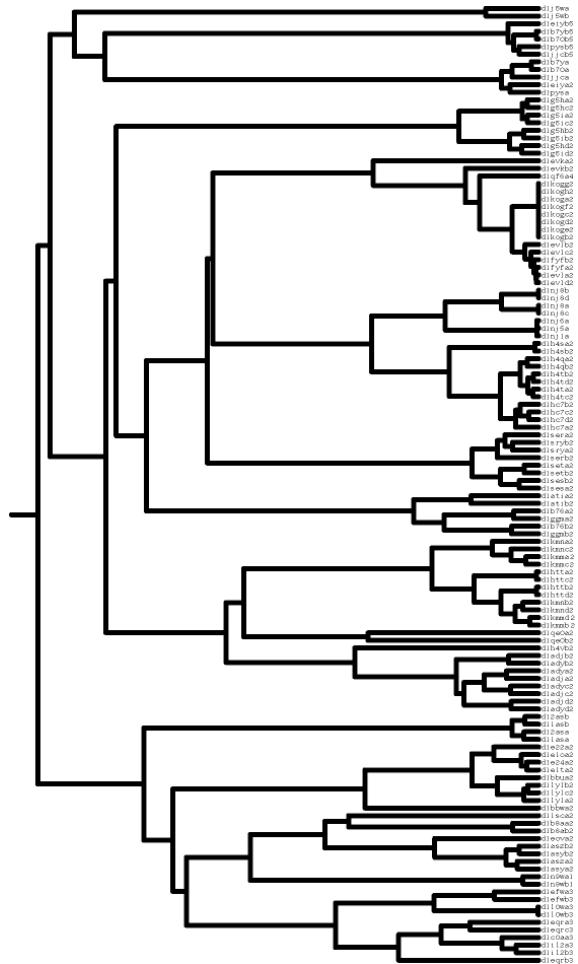
Builds models from manual alignment and assignment of SCRs, loops and side chains (rotamers).

#### [MODELLER](#)

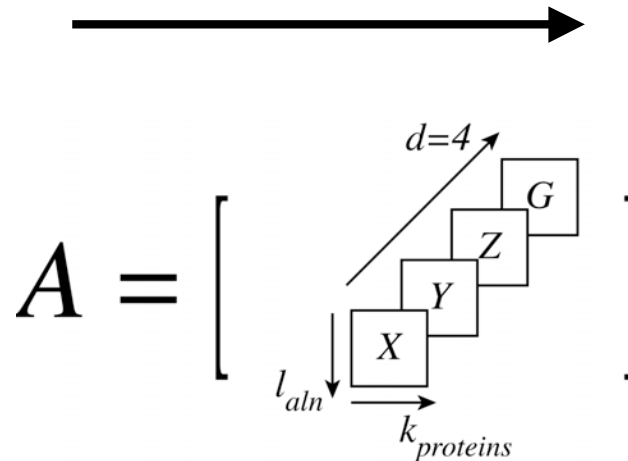
MODELLER-4: Homology protein structure modelling by satisfaction of spatial restraints, Andrej Sali Lab, Rockefeller University, USA.

# Non-redundant Representative Sets of Structures

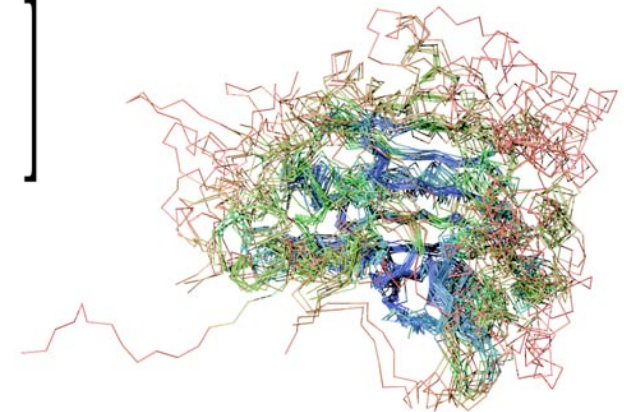
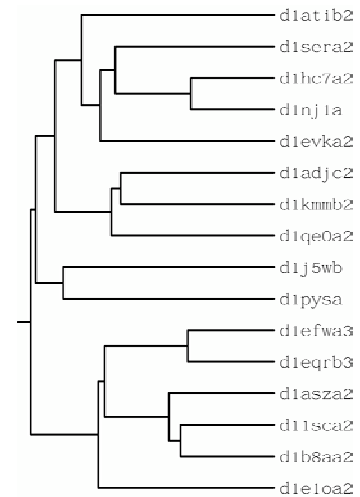
Too much information  
129 Structures



Multidimensional QR factorization  
of alignment matrix,  $A$ .



Economy of information  
16 representatives



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

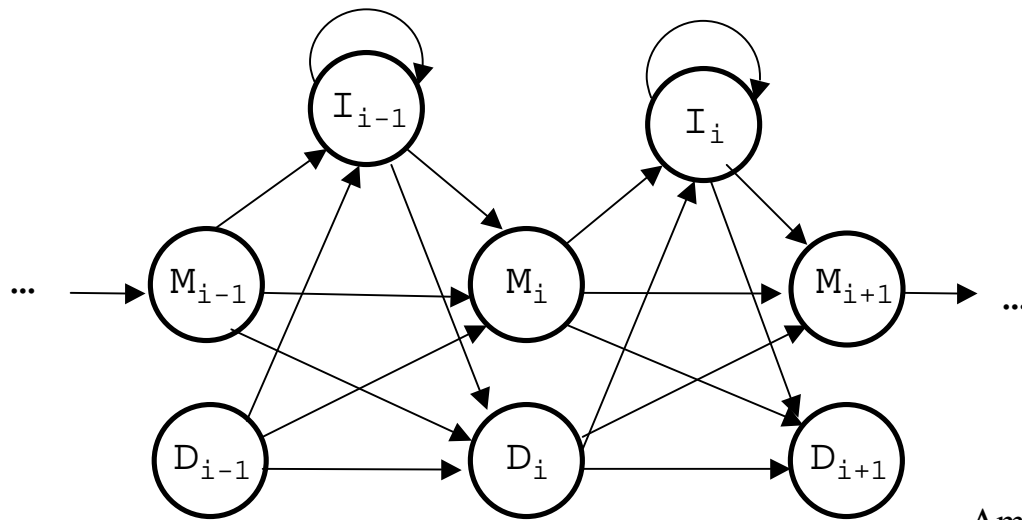
P. O'Donoghue and Z. Luthey-Schulten (2004) *J. Mol. Biol.*, in press.

# HMM-based approach

		<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>
<b>Sequence 1</b>		-	B	B	-	C
<b>Sequence 2</b>		C	C	-	-	C
<b>Sequence 3</b>		C	B	C	C	B
<b>State</b>		$M_1$	$M_2$	$M_3$	$I_3$	$M_4$

State transition Probabilities (ST)

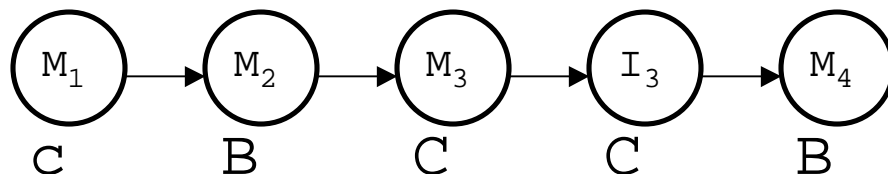
$i$	$M_i \rightarrow M_{i+1}$	$M_i \rightarrow D_{i+1}$	$M_i \rightarrow I_i$
1	1	0	0
2	0.67	0.33	0
3	0.5	0	0.5



Position specific amino acid Probabilities

$i$	$C(M_i)$	$B(M_i)$
1	1	0
2	0.33	0.67
3	0.5	0.5
4	0.67	0.33

State Path of Sequence 3



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

$$P(C|I) = 7/11 = 0.64$$

$$P(B|I) = 4/11 = 0.36$$

Leads to affine gap penalty.

$$P(-|D) = 1.$$



# Generation of HMM Profile

Goal: To select right model and assign match states to selected columns in MSA.

The **model (M)** is constructed so that it generates MSA with the *highest probability*.

The **probability** of finding a **single aligned sequence** A ( $A_1, A_2, \dots, A_N$ ) is:

$$P(A|\pi, M) = \prod_{i=1}^N P(A_i(\pi_i)) \quad P(\pi|M) = \prod_{i=1}^{N-1} P(\pi_i \rightarrow \pi_{i+1})$$

$$P(A, \pi|M) = \prod_{i=0}^{N-1} P(\pi_i \rightarrow \pi_{i+1})P(A_{i+1}(\pi_{i+1}))$$

$$= P(A_{0 \rightarrow i}, \pi_{0 \rightarrow i}|M)P(A_{i+1 \rightarrow j}, \pi_{i+1 \rightarrow j}|M)P(A_{j+1 \rightarrow N}, \pi_{j+1 \rightarrow N}|M)$$

The **probability** of obtaining the **MSA** from the profile is:

$$P(MSA, \pi|M) = \prod_{k=1}^{N_{seq}} P(A^k, \pi^k|M) \quad \text{where } A^k \text{ represents the } k^{th} \text{ aligned sequence.}$$

$$= \prod_{k=1}^{N_{seq}} P(A_{0 \rightarrow i}^k, \pi_{0 \rightarrow i}^k|M) \times P(A_{i+1 \rightarrow j}^k, \pi_{i+1 \rightarrow j}^k|M) \times P(A_{j+1 \rightarrow N}^k, \pi_{j+1 \rightarrow N}^k|M)$$

The **profile** is constructed using dynamic programming.

$$\log P_j = \max_{0 \leq i \leq j-1} \log P_{ij} = \max_{0 \leq i \leq j-1} (\log P_i + ST_{ij} + AM_j + AI_{ij})$$

$P_j$  = Highest probability of obtaining MSA till  $j^{th}$  column such that  $j$  is a match state.

$P_{ij}$  = Highest probability of obtaining MSA such that  $j$  is a match state and  $i$  is the previous match state.

where  $ST_{ij}$  represents the probabilities of all state transitions.

$AI_{ij}$  the probability of all amino acids in insert states between  $i^{th}$  and  $j^{th}$  columns.

$AM_j$  represents the probabilities of all amino acids in  $j^{th}$  column.



# HMM-based approach

New sequence aligned using dynamic programming.

$$H_j^M(i) = \log\left(\frac{aa_{M_j}(x_i)}{aa^{rnd}(x_i)}\right) + \max \begin{cases} H_{j-1}^M(i-1) + \log(ST_{M_{j-1}M_j}) \\ H_{j-1}^I(i-1) + \log(ST_{I_{j-1}M_j}) \\ H_{j-1}^D(i-1) + \log(ST_{D_{j-1}M_j}) \end{cases}$$

$$H_j^I(i) = \log\left(\frac{aa_{I_j}(x_i)}{aa^{rnd}(x_i)}\right) + \max \begin{cases} H_j^M(i-1) + \log(ST_{M_jI_j}) \\ H_j^I(i-1) + \log(ST_{I_jI_j}) \\ H_j^D(i-1) + \log(ST_{D_jI_j}) \end{cases}$$

$$H_j^D(i) = \max \begin{cases} H_{j-1}^M(i) + \log(ST_{M_{j-1}D_j}) \\ H_{j-1}^I(i) + \log(ST_{I_{j-1}D_j}) \\ H_{j-1}^D(i) + \log(ST_{D_{j-1}D_j}) \end{cases}$$



NCBI

Nucleotide

Protein

*protein-protein* **BLAST**

Translations

Retrieve results for an  
RID

[Search](#)

[Set subsequence](#) From:  To:

[Choose database](#)

[Do CD-Search](#)

Now:  or

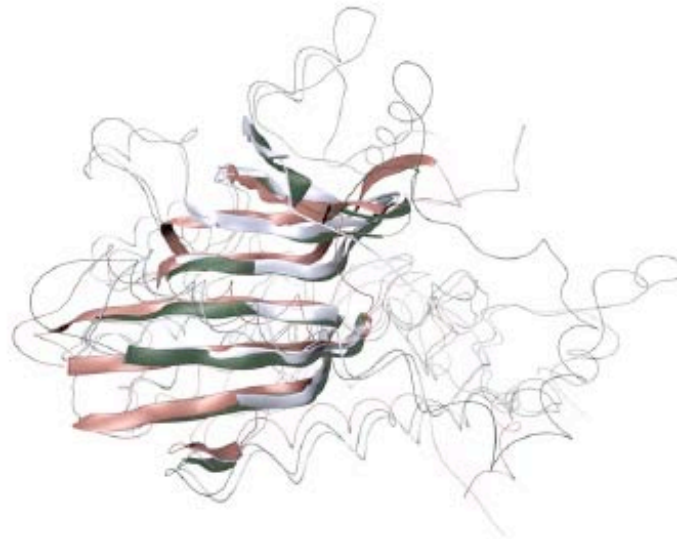
University of Illinois at Urbana-Champaign  
Luthey-Schulten Group  
Theoretical and Computational Biophysics Group  
Summer School 2004 - University of Western Australia, Perth

# Sequence Alignment Algorithms

---

*Tutorial for the  
material of this  
lecture available at*

<http://www.ks.uiuc.edu/Training/Tutorials/>



Rommie Amaro  
Felix Autenrieth  
Brijeet Dhaliwal  
Barry Isralewitz

Zaida Luthey-Schulten  
Anurag Sethi  
Taras Pogorelov

June 2004

```

f1l9ha_  MNGTEGPNFYVPFSNKTGVVRSPEAPQYYLAEPWQFSMLAAYMFL IMLGFPINFLTLY
f1e12a_  -----R--ENALLSSSLWVNVALAGIAILVFVYMGR
f1jgja_  -----MVGLTTLFMLGAIGMLVGTLAFAWAGR
                                     ::      :.

f1l9ha_  VTVQHKKLRTP LNYILLNLAVADLFMVFGGFTTTL YTS LHGYFVF--G-----
f1e12a_  TI--RPG---RPRLIWGATLMIPLVSI S-SYLGLLS-----G--LTVGMIEMPAGHALAG
f1jgja_  DA--GSG----ERRYYVTLVGISGIAAV-AYAVMAL-----G--VGWVPVA-----
                                     .      .      .:      .

f1l9ha_  ----PTGCNLEGGFATLGGEIALWSLVVLAIERYVVVCKPMSNFRFGENHAIMGVAFTWV
f1e12a_  EMVRS--QWGRYL TWALSTPMILLALGLLA-D-----V----DLGSLFTVIA
f1jgja_  E-RTV--FVPRYIDWILTTP LIVYFLGLLA-G-----L----DSREFGIVIT
                                     . : * : : * :**      .      . . .

f1l9ha_  MALACAAPPLVGWSRYIPEGMQCSCGIDYYT-PHEETNNE SFVIYMFVWHFIIPLIVIFF
f1e12a_  ADIGMCVTGL-A-A-A-----MT-TSAL-L--FRWAFYAISCAF-FVVVLSALVT
f1jgja_  LNTVVMLAGF-A-G-A-----MV-PG-I----ERYALFGMGAVA-FIGLVYYLVG
                                     . : . .      :      . : :      . * : : : .

f1l9ha_  C-YGQ-LVFTVKEAAAATTQKAEKEVTRMVIIMVIAFLICWLPYAGVAFY-IFTHQGSDF
f1e12a_  DWA-ASA-SS-----AG--TAEIFDTRLRVLTVVLWLGYPVWAVGVEGL-A-LV
f1jgja_  PMT-ESA-SQ-----RSSGIKSLYVRLRNLTVVLWAIYPIWLLGPPGV-A-LL
                                     :      :      : : * * . :      . .

f1l9ha_  -GPIFMTIPAFFAKT-SAVYNPVIYIMMN-KQFRNCMVTTLCCGKNPLGDSTT-VSKTET
f1e12a_  QSVGATSWAYSVLDVFAKYVFAFILLRWVANNE-----R-TV-----
f1jgja_  -TPTVDVALIVYLDLVTKVGFIALDAA-A-----
                                     . :      . * :

f1l9ha_  SQVA-PA
f1e12a_  ----AV-
f1jgja_  -----

```

d1l9ha\_ MNGTEGPNFYVPPFSNKTGVVRSPPFEAPQYYLAEPWQFSMLAAYMFLLI MLGFPINFLTLY  
d1e12a\_ -----R--ENALLSSSLWVNVVALAGIAILVFVYMGR  
d1jgja\_ -----MVGLTTLFWLGAIGMLVGTLAFAWAGR  
d1at9\_ -----AQITGRPEWIWLALGTALMGLGTYFLVK

. :.

d1l9ha\_ VTVQHKKLRTPNLYILLNLAVADLFMVFGGFTTTLTSLHGYFVF--G-----  
d1e12a\_ TI--RPG---RPRLIWGATLMIPLVSIS--SYLGLLS-----G--LTVGMIEMPAGHALAG  
d1jgja\_ DA--GSG---ERRYVVTLVGISGIAAV--AYAVMAL-----G--VGWVPVA-----  
d1at9\_ GMGVSDP---DAKKFYAITTLVPAIAFT--MYLSMLL-----GYGLTMVPPFG-----

. . :

d1l9ha\_ ----PTGCNLEGGFATLGGELALWSLVVLAIERVYVVKPMSNFRFGENHAIMGVAFTWV  
d1e12a\_ EMVRS--QWGRYLTWALSTPMILLALGLLA-D-----V----DLGSLFTVIA  
d1jgja\_ E--RTV--FVPRYIDWILTTPLIVYFLGLLA-G-----L----DSREFGIVIT  
d1at9\_ --EQNPIYWARYADWLFTTLLLLLDALLVDA-----DQGTILALVG

. : : \* :\*

d1l9ha\_ MALACAAPPLVGWSRYIPEGMCSCGIDYYT-PHEETNNEFVIYMFVVHFI IPLIVIFF  
d1e12a\_ ADIGMCVTGL-A-A-A-----MT-TSAL-L--FRWAFYAISCAF-FVVVLSALVT  
d1jgja\_ LNTVVMLAGF-A-G-A-----MV-PG-I----ERYALFGMGAVA-FIGLVYYLVG  
d1at9\_ ADGIMIGTGLVG-----ALTKVYSYRFVWVAISTAAMLYILYVLFVG

. : . . : :.

d1l9ha\_ C-YGQ-LVFTVKEAAAATTQKAEKEVTRMVIIMVIAFLICWLPYAGVAFY-IFTHQGSDF  
d1e12a\_ DWA-ASA-SS-----AG--TAEIFDTRLRVLTVVLWLGYPVWAVGVEGL-A-LV  
d1jgja\_ PMT-ESA-SQ-----RSSGIKSLYVRLRNLTVVLWAIYPIWLLGPPGV-A-LL  
d1at9\_ FTSKAES-----MRPEVASTFKVLRNVTVVLWSAYPVVWLVIGSEGA---GI

: : \* \* . :

d1l9ha\_ -GPIFMTIPAFFAKT-SAVYNPVIYIMMN-KQFRNCMVTTLCCGKNPLGDSTT-VSKTET  
d1e12a\_ QSVGATSWAYSVLDVFAKYVFAFILLRWVANNE-----R-TV-----  
d1jgja\_ -TPTVDVALIVYLDLVTKVGFGLDAA-A-----  
d1at9\_ VPLNIETLLFMVLDVSAKVGFLILLRSRAIFG-----

. : .\* :

d1l9ha\_ SQVA-PA  
d1e12a\_ ----AV-  
d1jgja\_ -----  
d1at9\_ -----