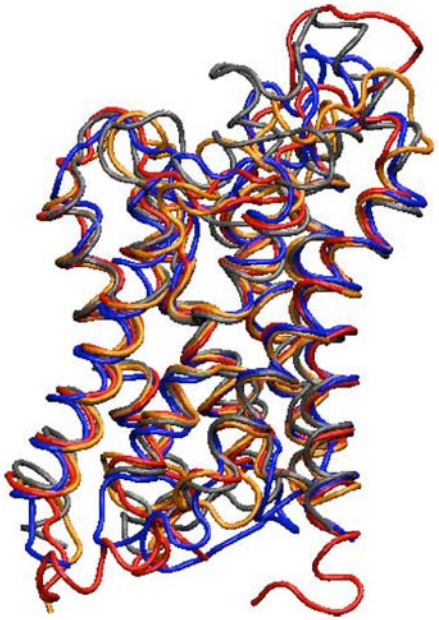
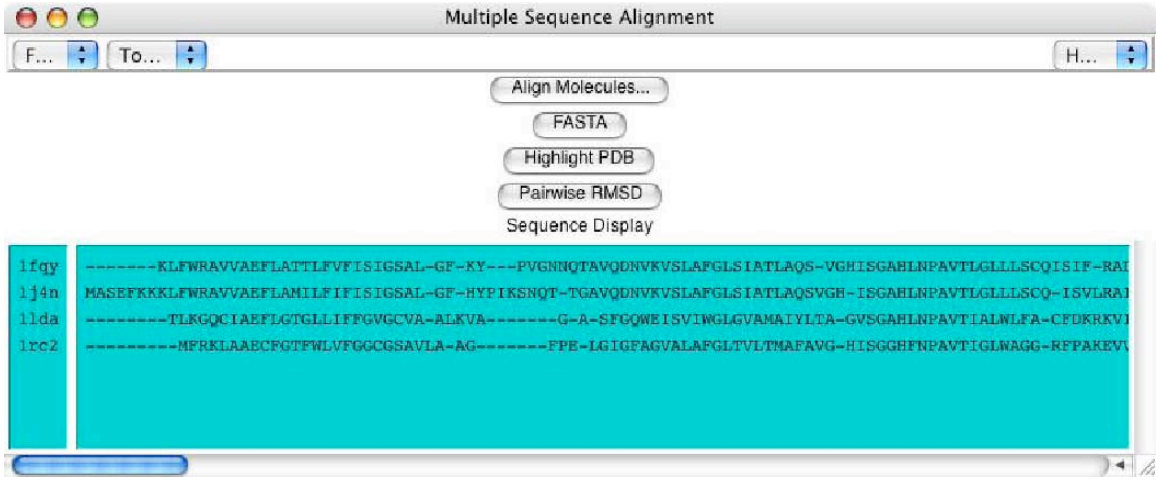
Sequence and Structure Alignment

Z. Luthey-Schulten, UIUC

Chicago, 2005

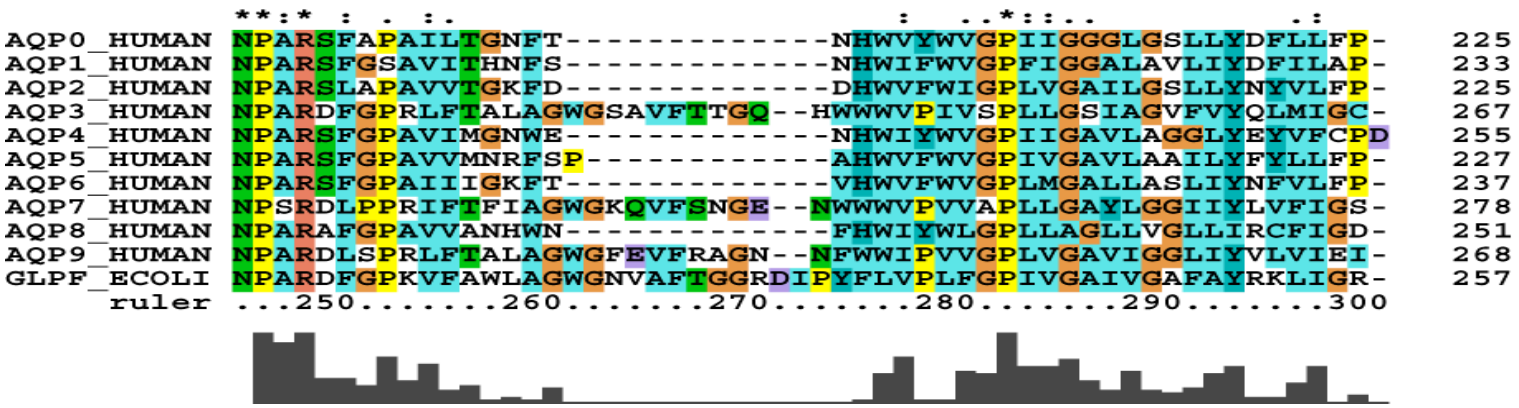


VMD 1.83



Why Look at More Than One Sequence?

1. Multiple Sequence Alignment shows patterns of conservation



2. What and how many sequences should be included?

3. Where do I find the sequences and structures for MS alignment?

4. How to generate pairwise and multiple sequence alignments?

Sequence-Sequence Alignment

- Smith-Watermann
- Needleman-Wunsch

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

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

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
- Needleman-Wunsch

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

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

Sequence-Structure Alignment

- Threading
- Hidden Markov, Clustal

Structure-Structure Alignment

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

SCOP, Astral

PDB

Sequence Database Searches

- Blast and Psi-Blast NCBI Swiss Prot

Swiss-Prot

Protein knowledgebase

TrEMBL

Computer-annotated supplement to Swiss-Prot

UniProt
the universal protein resource

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...)

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

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)

NiceProt View of Swiss-Prot: P47865

Printer-friendly view | Submit update | Quick BlastP search

[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.

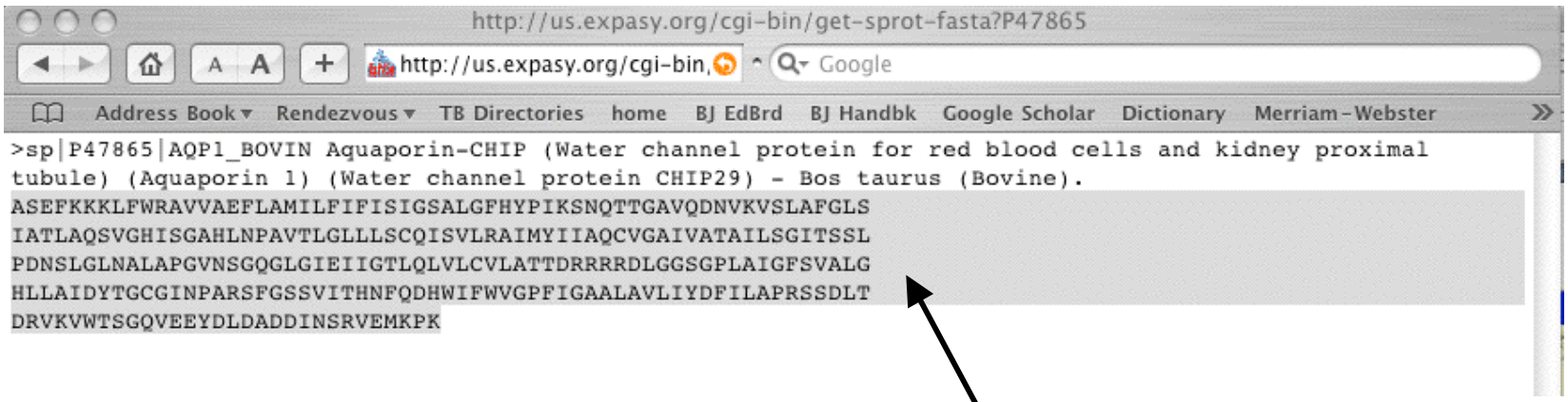
Entry information	
Entry name	**AQP1_BOVIN**
Primary accession number	**P47865**
Secondary accession numbers	None
Entered in Swiss-Prot in	Release 33, February 1996
Sequence was last modified in	Release 44, July 2004
Annotations were last modified in	Release 45, October 2004
Name and origin of the protein	
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

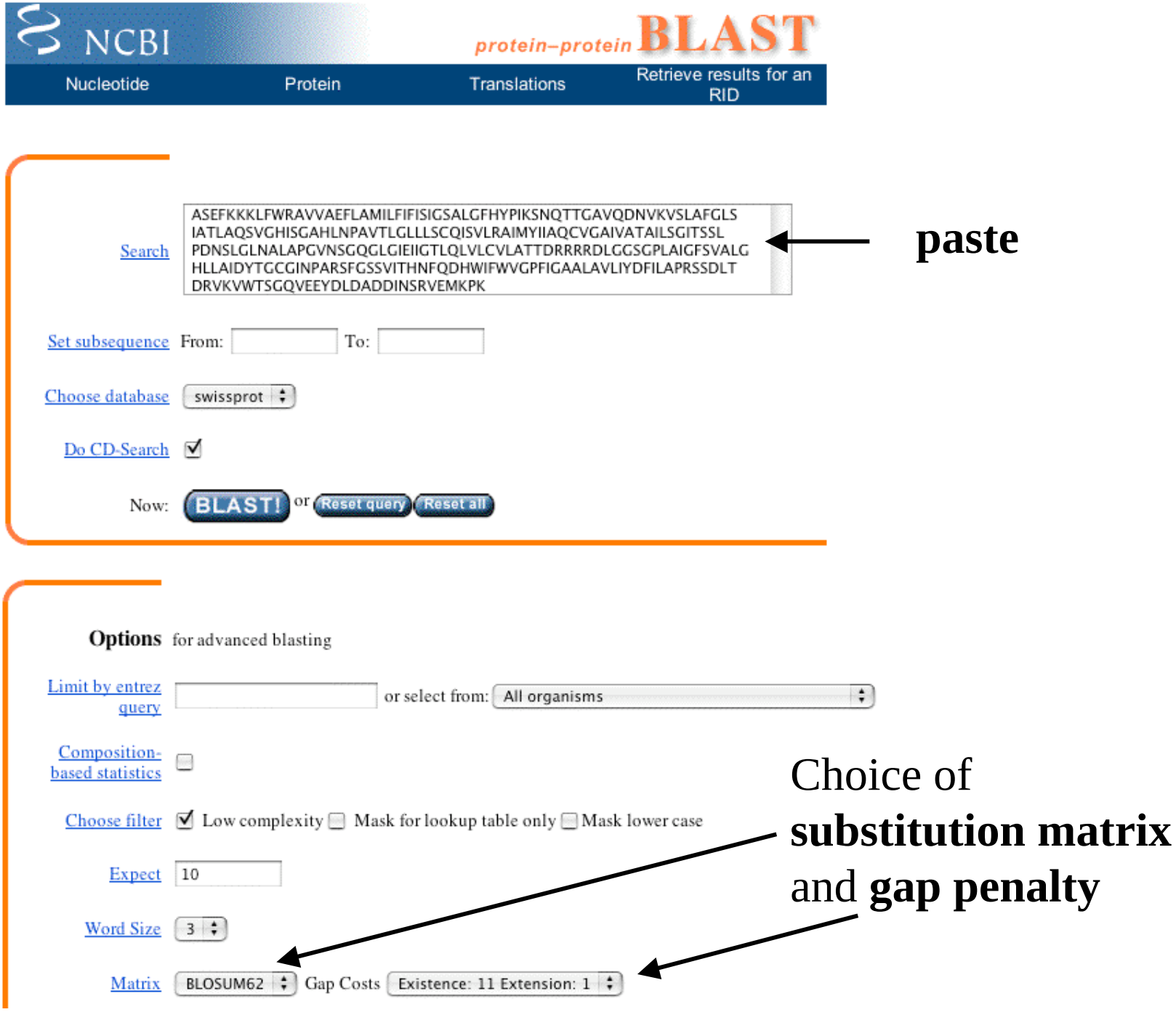
Length: **270 AA** | Molecular weight: **28669 Da** | CRC64: **F3ECAD45DCCDB309** [This is a checksum on the sequence]

```
        10          20          30          40          50          60
ASEFKKKLFW RAVVAEFLAM ILFIFISIGS ALGFHYPIKS NQTTGAVQDN VKVSLAFGLS
        70          80          90         100         110         120
IATLAQSVGH ISGAHLNPAV TLGLLLSCQI SVLRAIMYII AQCVGAIVAT AILSGITSSL
       130         140         150         160         170         180
PDNSLGLNAL APGVNSGQGL GIEIIGTLQL VLCVLATTDR RRRDLGGSGP LAIGFSVALG
       190         200         210         220         230         240
HLLAIDYTGC GINPARSFGS SVITHNFQDH WIFWVGPFIG AALAVLIYDF ILAPRSSDLT
       250         260         270
DRVKVWTSGQ VEEYDLDADD INSRVEMKPK
```

P47865 in FASTA format



cut


NCBI
protein–protein BLAST
Nucleotide
Protein
Translations
Retrieve results for an RID
Search
ASEFKKKLFWRAVVAEFLAMILFIFISIGSALGFHYPIKSNQTTGAVQDNVKVSLAFGLS
IATLAQSVGHISGAHLNPAVTLGLLLSCQISVLRAIMYIIAQCVGAIVATAILSGITSSL
PDNSLGLNALAPGVNSGQGLGIEIIGTLQLVLCVLATTDRRRRDLGGSGPLAIGFSVALG
HLLAIDYTGCGINPARSFGSSVITHNFQDHWIFWVGPFIGAALAVLIYDFILAPRSSDLT
DRVKVWTSGQVEEYDLDADDINSRVEMKPK
paste
Set subsequence
From:
To:
Choose database
swissprot
Do CD-Search
Now:
BLAST!
or
Reset query
Reset all
Options for advanced blasting
Limit by entrez query
or select from:
All organisms
Composition-based statistics
Choose filter
Low complexity
Mask for lookup table only
Mask lower case
Expect
10
Word Size
3
Matrix
BLOSUM62
Gap Costs
Existence: 11 Extension: 1
Choice of
substitution matrix
and gap penalty

NCBI formatting BLAST

Nucleotide | Protein | Translations | 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 1102423974-1792-209874599488.BLASTQ4

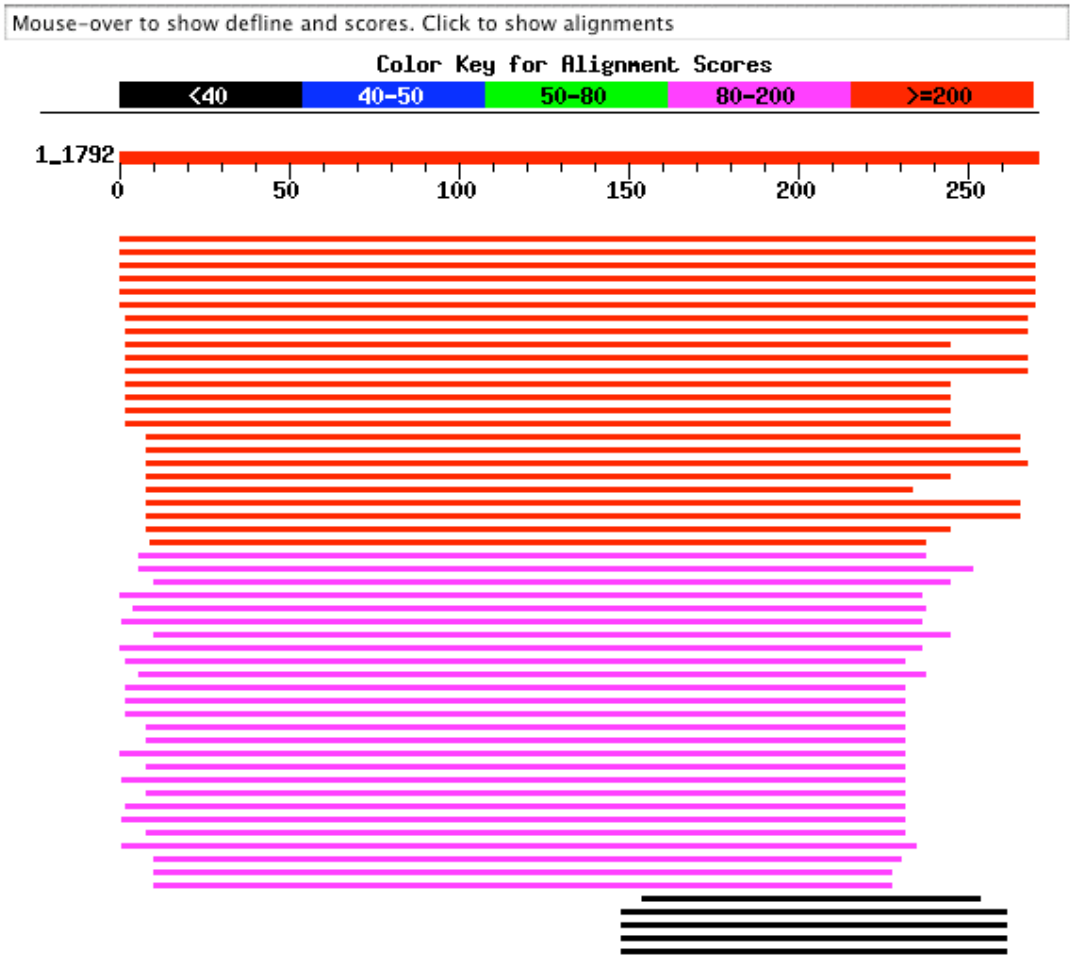
Format! or Reset all

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.

Taxonomy reports

Distribution of 164 Blast Hits on the Query Sequence



Sequences producing significant alignments:		Score (bits)	E 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|Q02013|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|AQPA_RANES	Aquaporin FA-CHIP	431	e-121	
gi|730026|sp|Q06019|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|Q923J4|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|Q9WTY4|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|Q9NHW7|AQP_AEDAE	Aquaporin AQPAe.a	201	1e-51	
gi|32469580|sp|Q25074|AQP_HAEIE	Aquaporin (Water channel 1)...	192	5e-49	
gi|2497939|sp|Q13520|AQP6_HUMAN	Aquaporin 6 (Aquaporin-2 li...	192	9e-49	G
gi|47115531|sp|Q9WTY0|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|Q9V5Z7|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|Q8C4A0|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|Q9ZV07|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

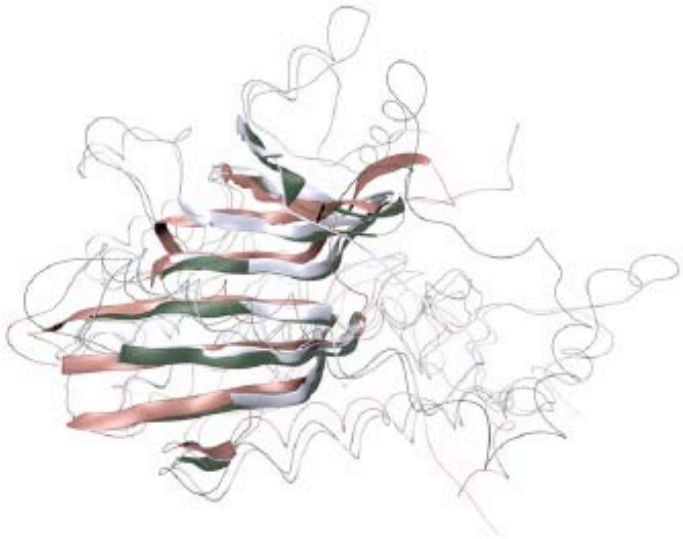
Score = 119 bits (299), Expect = 6e-27
Identities = 70/186 (37%), Positives = 105/186 (56%), Gaps = 12/186 (6%)

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

Query: 113 LSGITSSLP--DNSLGL--NALAP----GVNSGQGLGIEIIGTLQLVLCVLATTDRRRRD 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

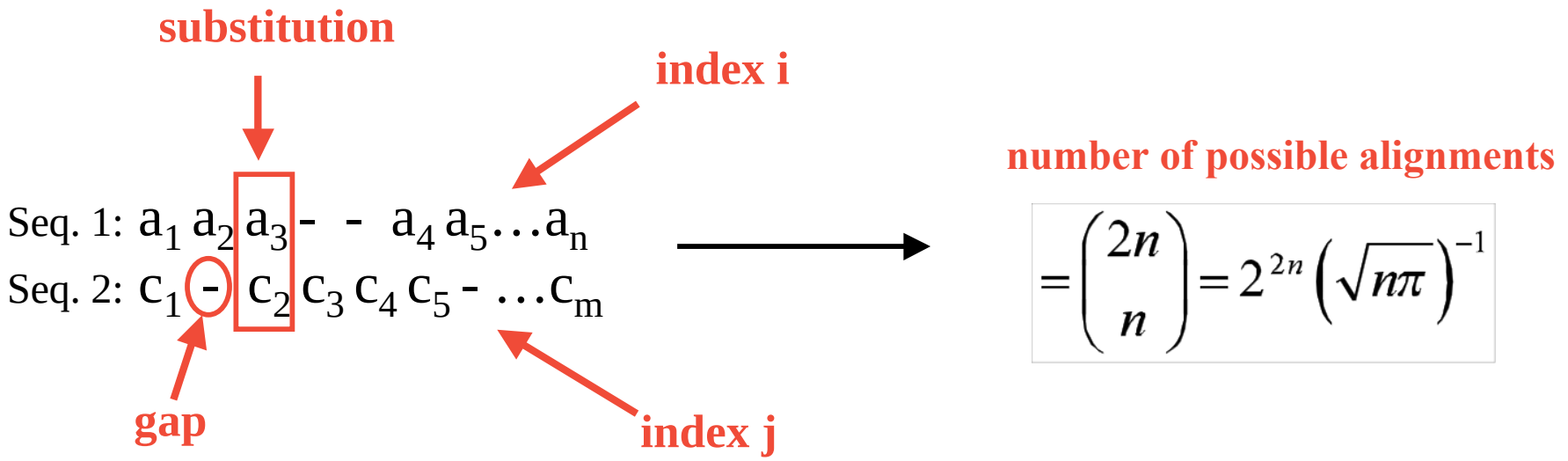
Tutorial for the material of this lecture available



Rommie Amaro
Felix Autenrieth
Brijeet Dhaliwal
Barry Isralewitz
Zaida Luthey-Schulten
Anurag Sethi
Taras Pogorelov

June 2004

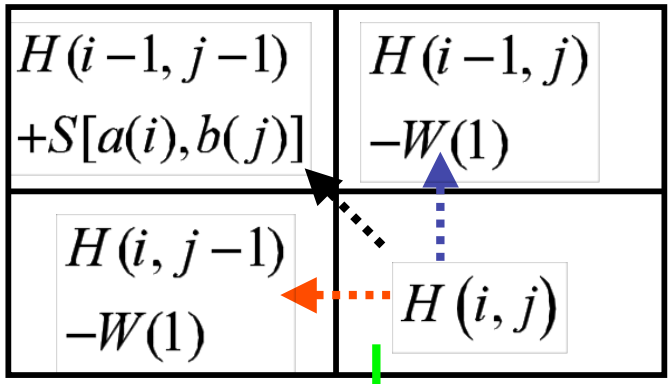
Sequence Alignment & Dynamic Programming



Smith-Waterman alignment algorithm

objective function

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



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: $a_1\ a_2\ a_3$ - - $a_4\ a_5 \ldots a_n$

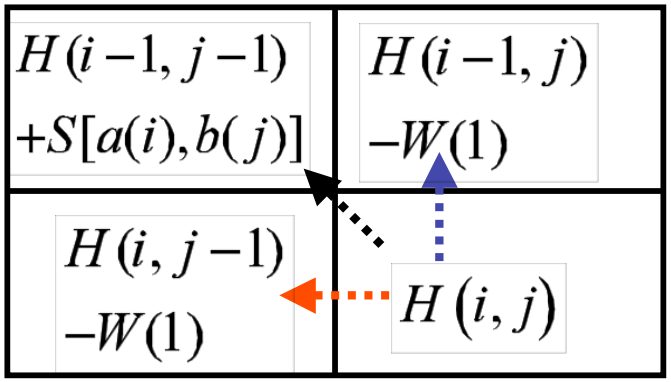
Seq. 2: c_1 - $c_2\ c_3\ c_4\ c_5$ - $\ldots c_m$

number of possible alignments:

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

Needleman-Wunsch alignment algorithm

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



Score Matrix H: Traceback

S : 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	-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

gap penalty W = - 6

Needleman-Wunsch Global Alignment

Similarity Values

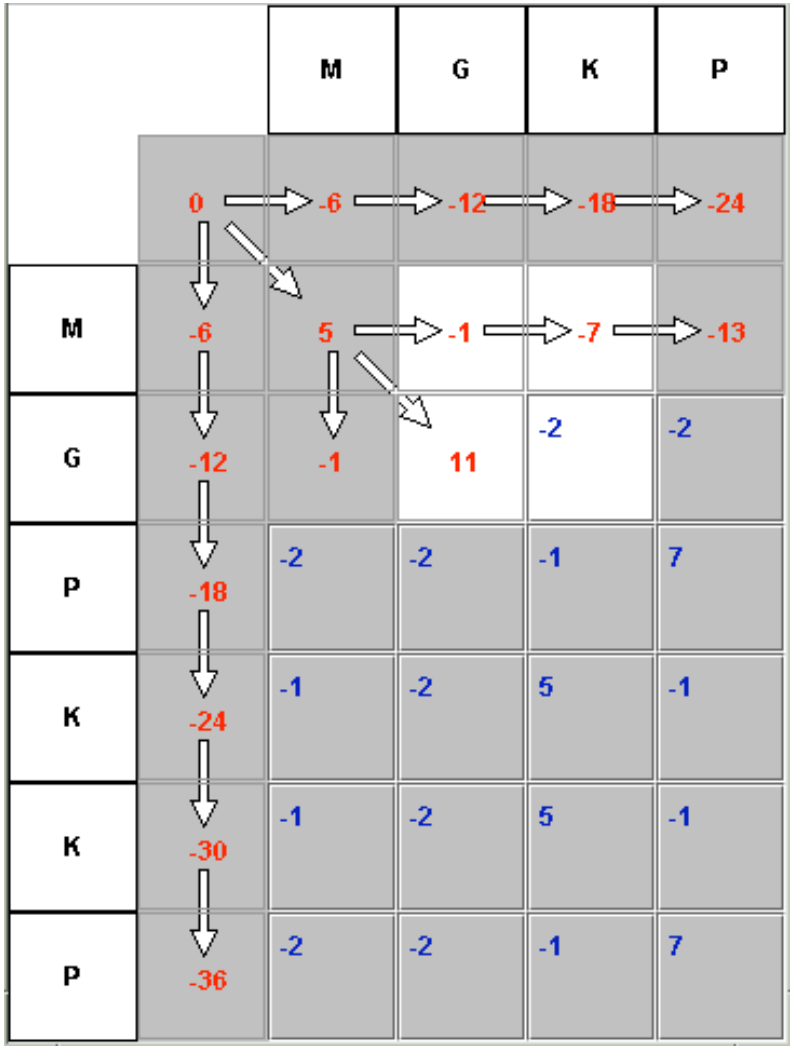
		M	G	K	P
M		5	-3	-1	-2
G		-3	6	-2	-2
P		-2	-2	-1	7
K		-1	-2	5	-1
K		-1	-2	5	-1
P		-2	-2	-1	7

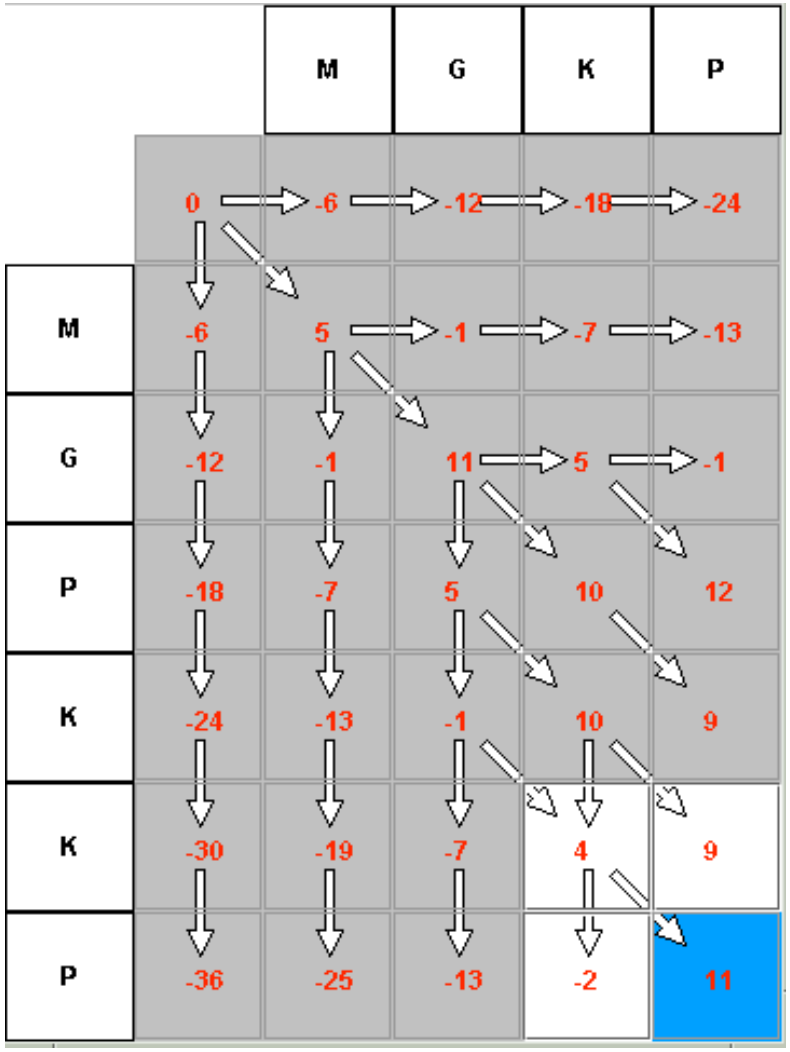
Initialization of Gap Penalties

		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

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

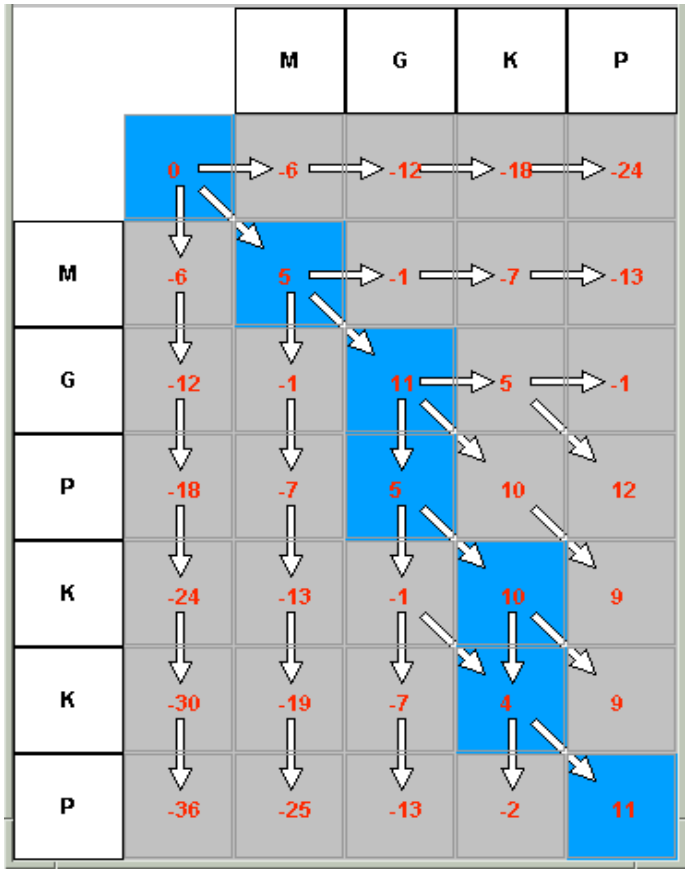
Filling out the Score Matrix H





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

Traceback and Alignment



The Alignment

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

Traceback (blue) from optimal score

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

Protein Structure Prediction

1-D protein sequence

SISSIRVKSKRIQLG…

3-D protein structure

Homology Modeling/ FR

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

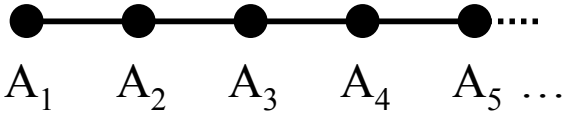
Target Sequence

SISSRVKSKRIQLGLNQAELAQKV------GTTQ…
QFANEFKVRRIKLGYTQ----TNVGEALAAVHGS…

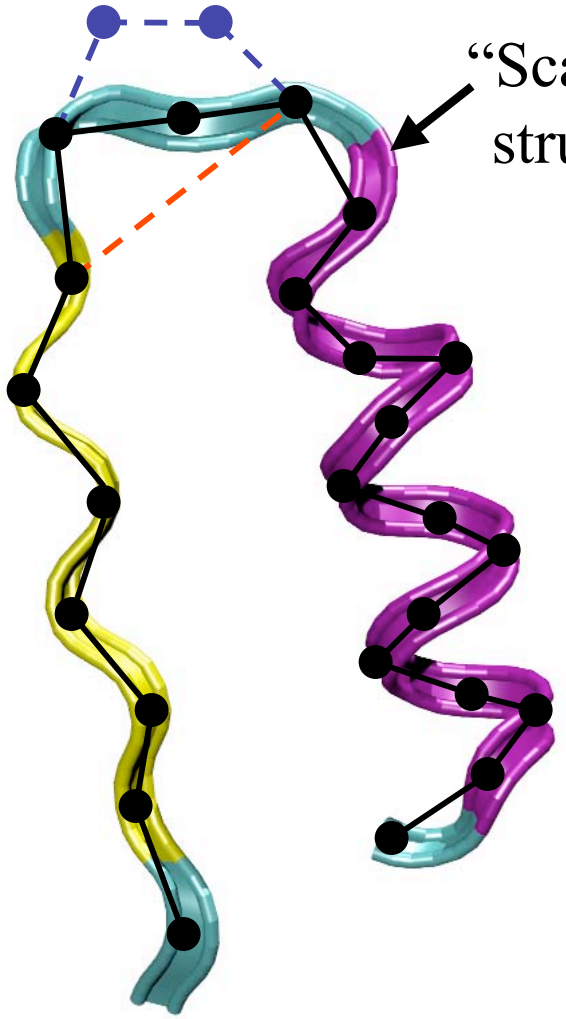
Known structure(s)

Sequence-Structure Alignment

Target sequence



Alignment between target(s) and scaffold(s)



1. Energy Based Threading*

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

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

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

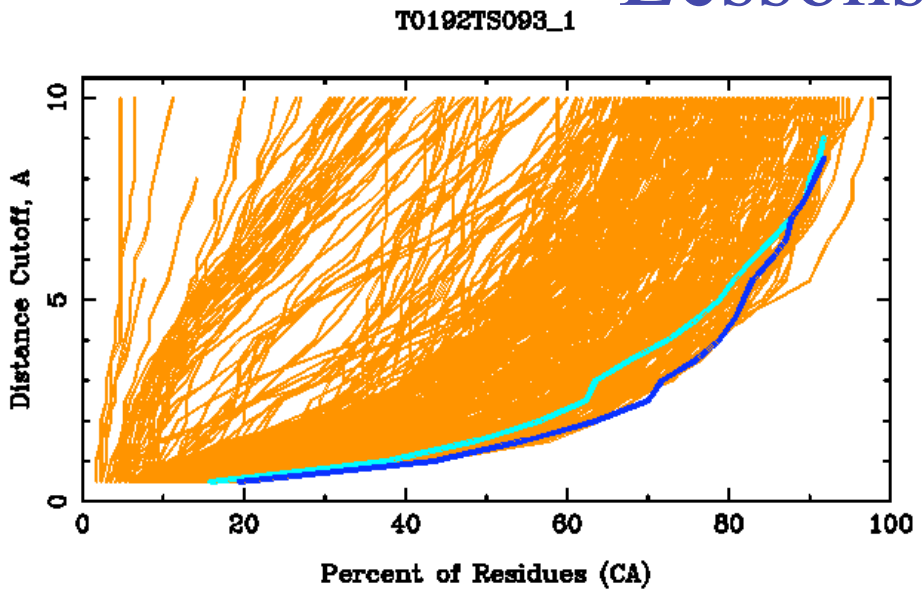
2. Sequence – Structure Profile Alignments

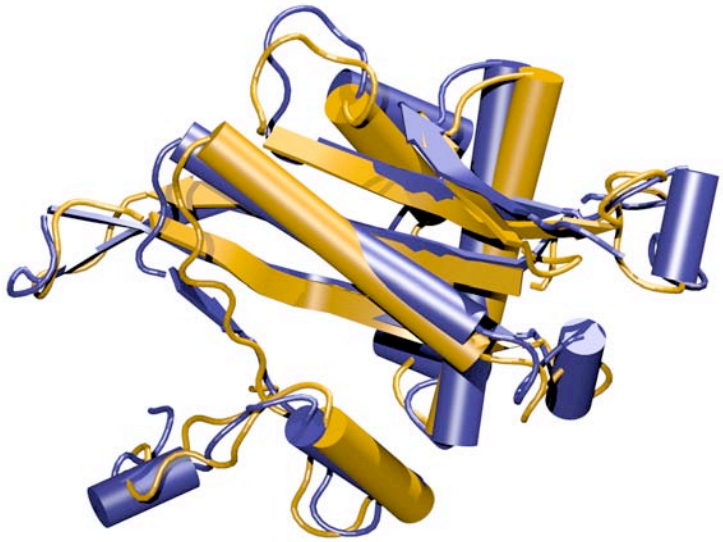
Clustal, Hidden Markov (HMMER, PSSM)
with position dependent gap penalties

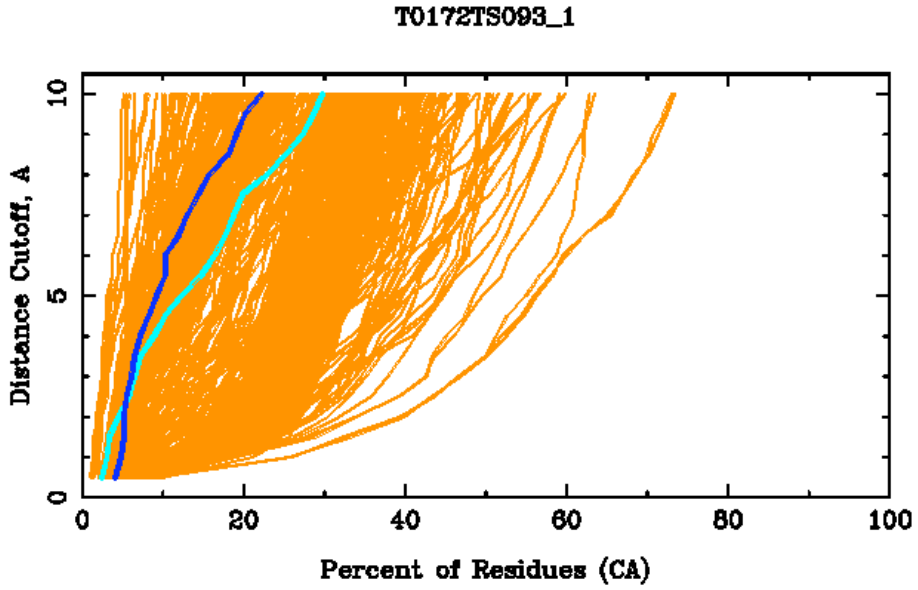
*R. Goldstein, Z. Luthey-Schulten, P. Wolynes (1992, PNAS), K. Koretke et.al. (1996, Proteins)

CM/Fold Recognition Results from CASP5

Lessons Learned







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 redunancy

Structural Domains

Structural Classification of Proteins

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

Lineage:

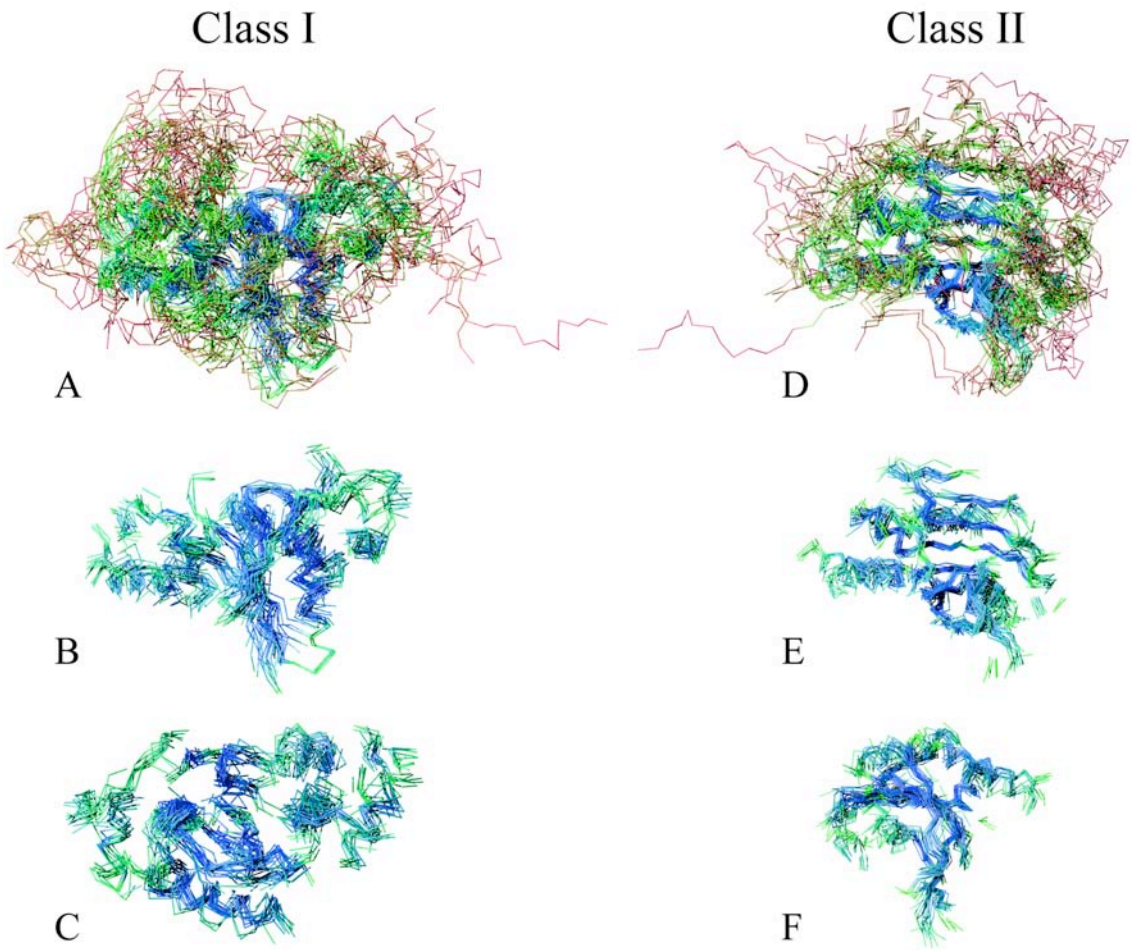
1. Root: scop
2. Class: All beta proteins
3. Fold: OB-fold
   barrel, closed or partly opened n=5, S=10 or S=8; greek-key
4. Superfamily: Nucleic acid-binding proteins
5. Family: Anticodon-binding domain
   barrel, closed; n=5, S=10
6. Protein: Aspartyl-tRNA synthetase (AspRS)
   this is N-terminal domain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

•Dynamic Programming (Smith-Waterman) through P matrix gives optimal set of equivalent residues.
•This set is used to re-superpose the two chains. Then iterate until alignment score is unchanged.
•This procedure is performed for all pairs.

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

Multiple Structural Alignments

STAMP – cont'd

2. Refine Initial Alignment & Produce Multiple Structural Alignment

Alignment score:

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

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

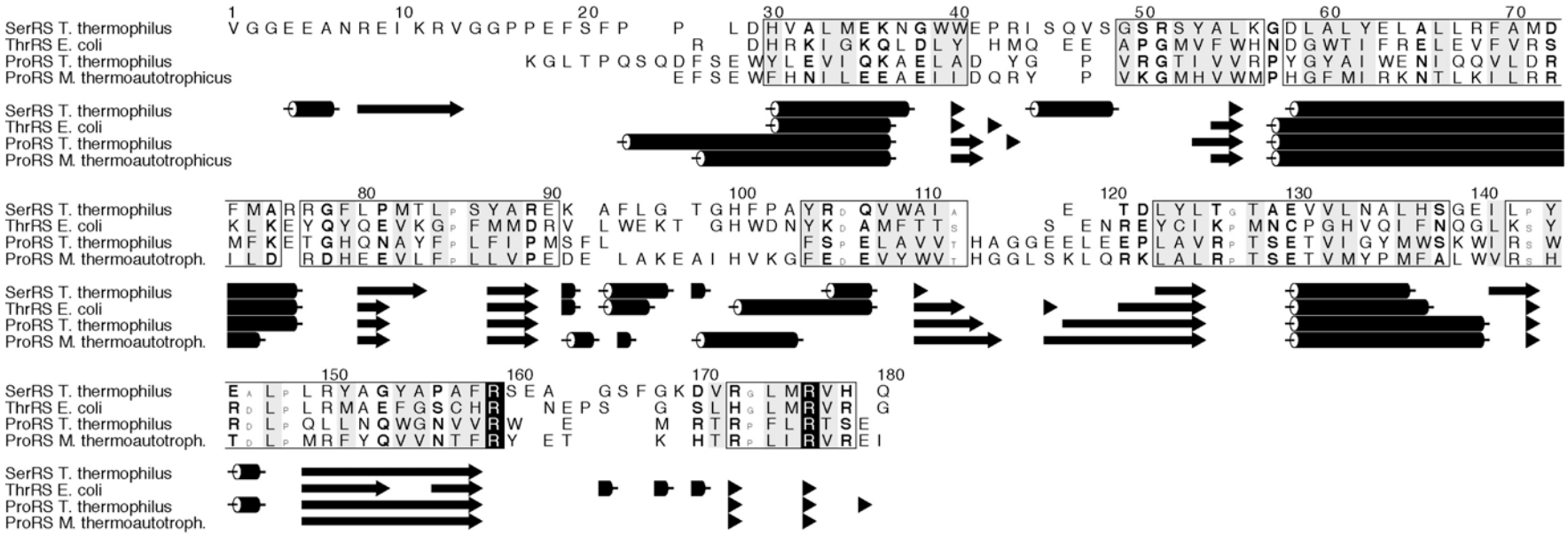
L_p, L_A, L_B -- length of alignment, sequence A, sequence B

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

Multiple Alignment:

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

Variation in Secondary Structure STAMP Output



Eargle, O'Donoghue, Luthey-Schulten, UIUC 2004

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

Stamp Output/Clustal Format

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

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

SerRS-T_thermophilus          TDLYLTGTAEVVLNALHSGEILPYEALPLRYAGYAPAFRSEA--GSFGKDVRGLMRVH-Q
ThrRS-E_coli                  REYCIKPMNCPGHVQIFNQGLKSYRDLPLRMAEFGSCHR--NEPS--G-SLHGLMRVR-G
ProRS-T_thermophilus          EPLAVRPTSETVIGYMWSKWIRSWRDLPQLLNQWGNVVRW--E----M-RTRPFLRTSE-
ProRS-M_thermoautotrophicus   RKLALRPTSETVMYPMFALWVRSHTDLPMRFYQVVNTFRY-ET----K-HTRPLIRVREI
space                         ------------------------------------------------------------
SerRS-T_thermophilus          SSEEE-S-THHHHHHHTTT-EEEGGG-SEEEEEEEEE------S--SSTTTTTTS-S-E
ThrRS-E_coli                  EEEEE-S-SHHHHHHHHTSS--BTTT-SEEEEE--EEE------G--G-G-BTTTB-S-E
ProRS-T_thermophilus          EEEEE-S-SHHHHHHHHHHH--BGGG--EEEEEEEEEE---------S-S-BTTTB-SE-
ProRS-M_thermoautotrophicus   EEEEE-SSSHHHHHHHHHHH--BTTT--EEEEEEEEEE---------S---BTTTB-SEE
```

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

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

II Database search

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

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

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

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

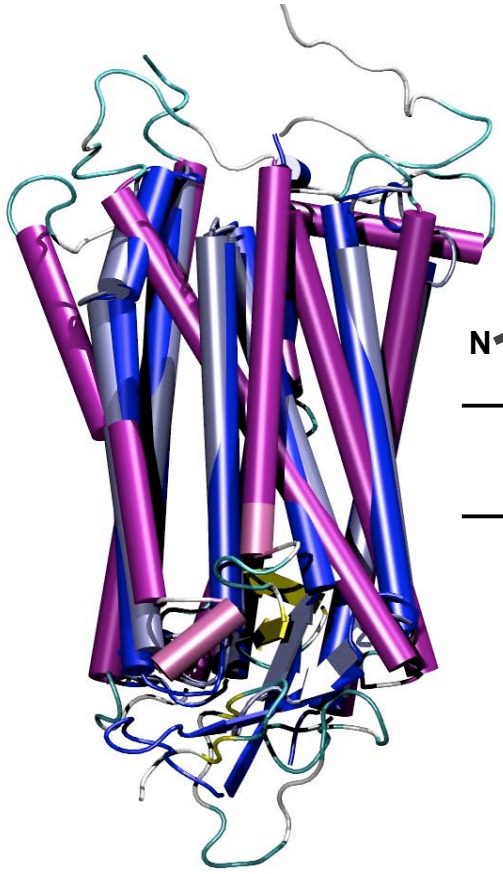
$$H(i,j) = max \begin{cases} H(i-1,j-1) + Sc(S_i^j) \\ H(i-1,j) + \delta \\ H(i,j-1) + \delta \end{cases}$$

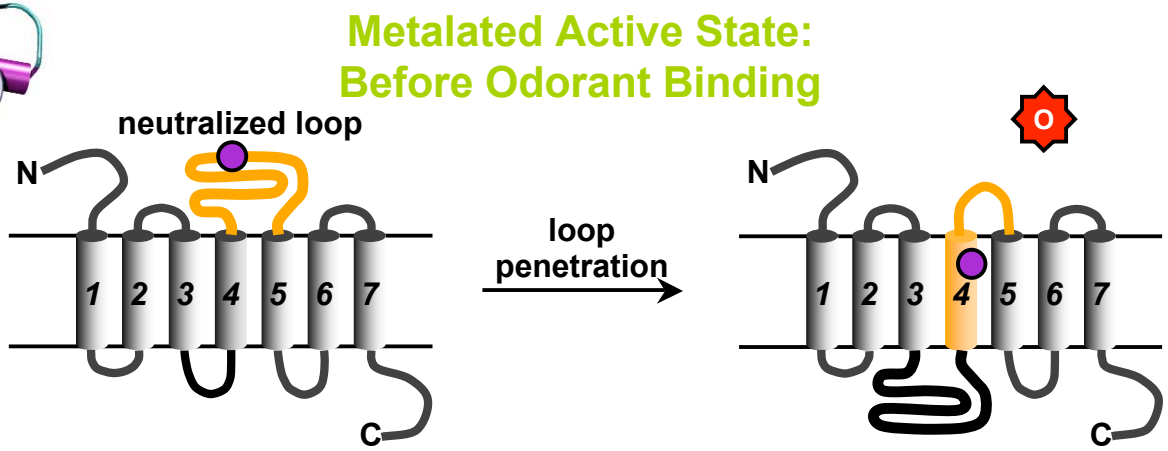
for $j=1,2,\ldots,m_{aln}$ and $i=1,2,\ldots,N_{res}$

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

Altschul, et. al., *Nucleic Acids Research*, **25:**3389-3402. Also useful: http://apps.bioneq.qc.ca/twiki/bin/view/Knowledgebase/PSSM

Hidden Markov Models of Transmembrane Proteins



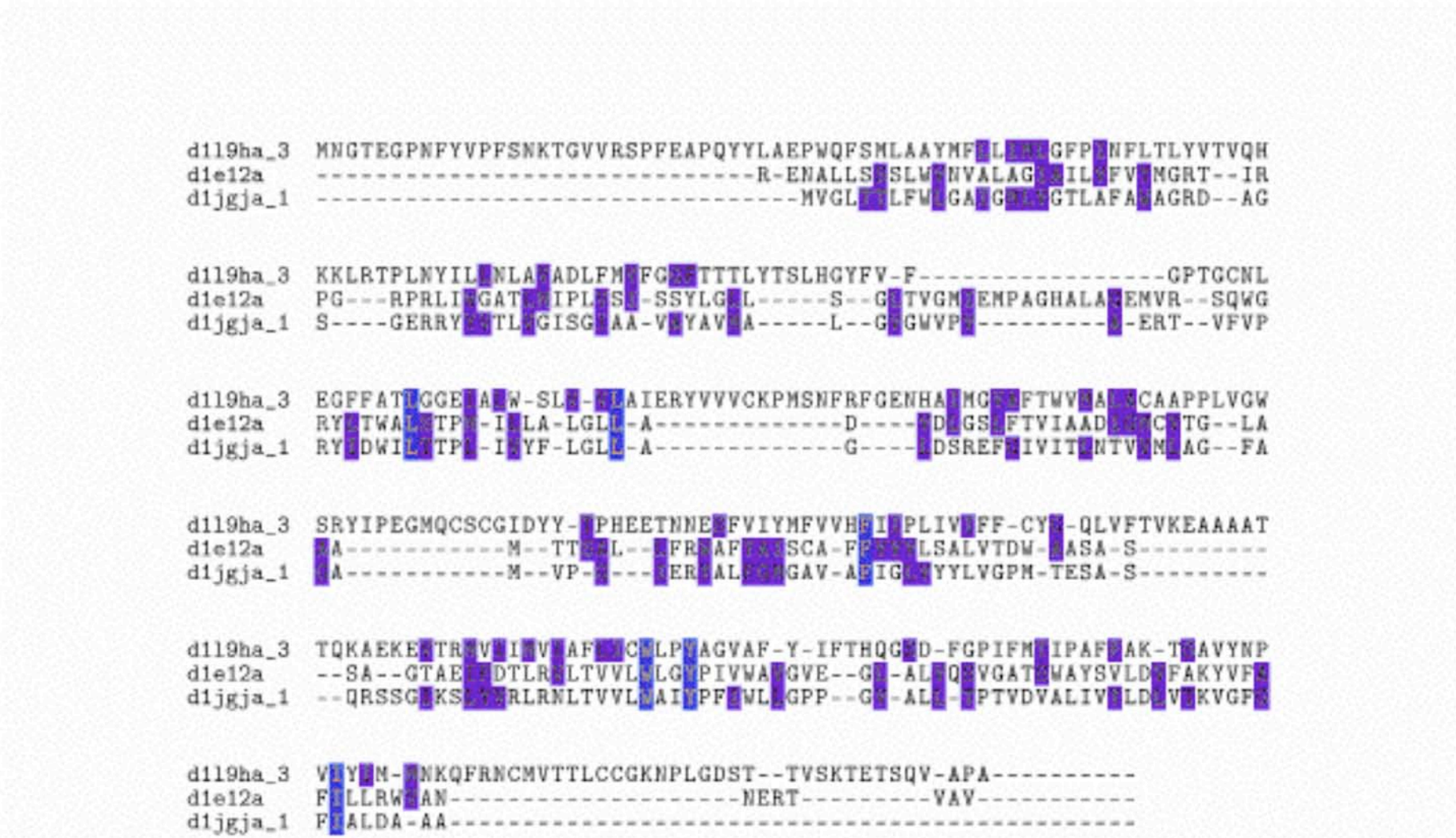


Bacteriorhodpsin/Rhodopsins

Olfactory Receptor/Bovine Rhodopsin

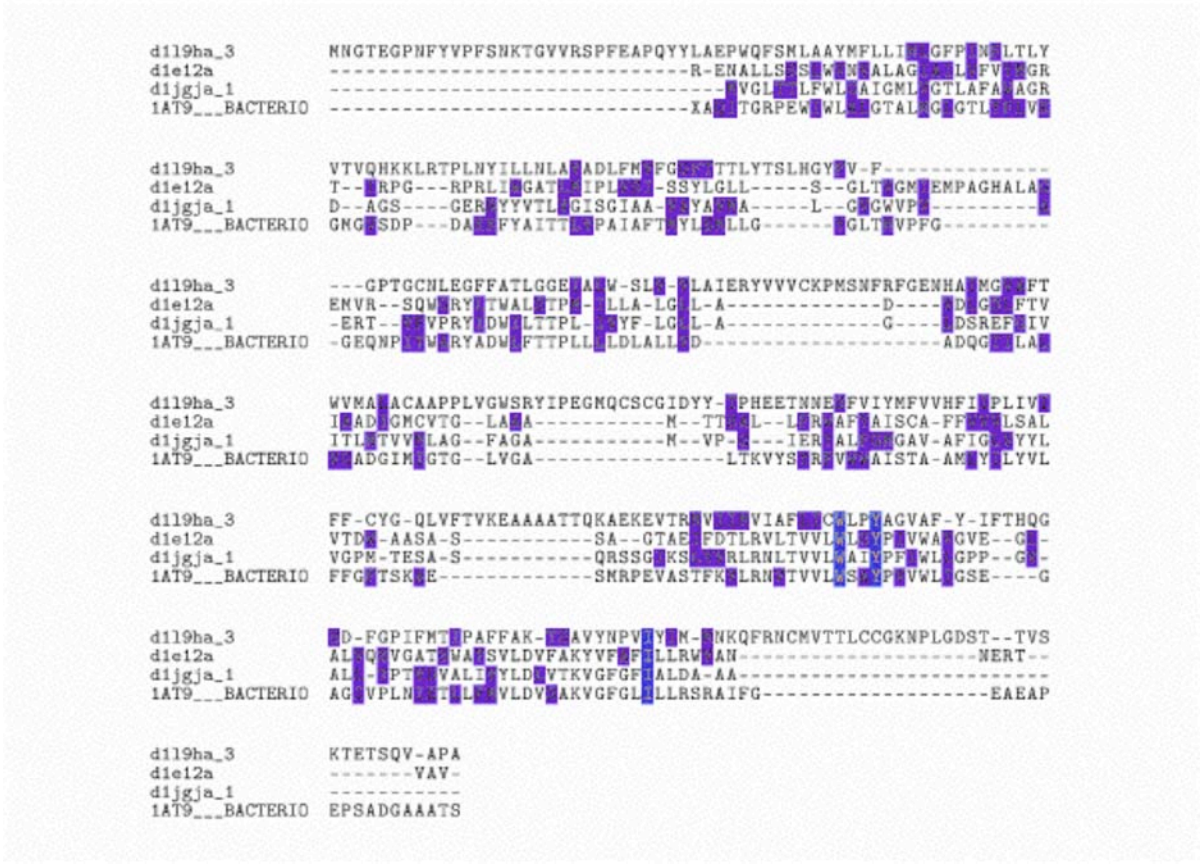
J. Wang, Z. Luthey-Schulten, K. Suslick (2003) *PNAS* **100**(6):3035-9

Stamp Profile



Sethi and Luthey-Schulten, UIUC 2003

Clustal Profile-Profile Alignment



Sethi and Luthey-Schulten, 2003

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

• **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**
	• **Proteomics and sequence analysis tools**
	  ◦ 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, Asia, Australia, Europe
PROF	B. Rost : PROFile 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 (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

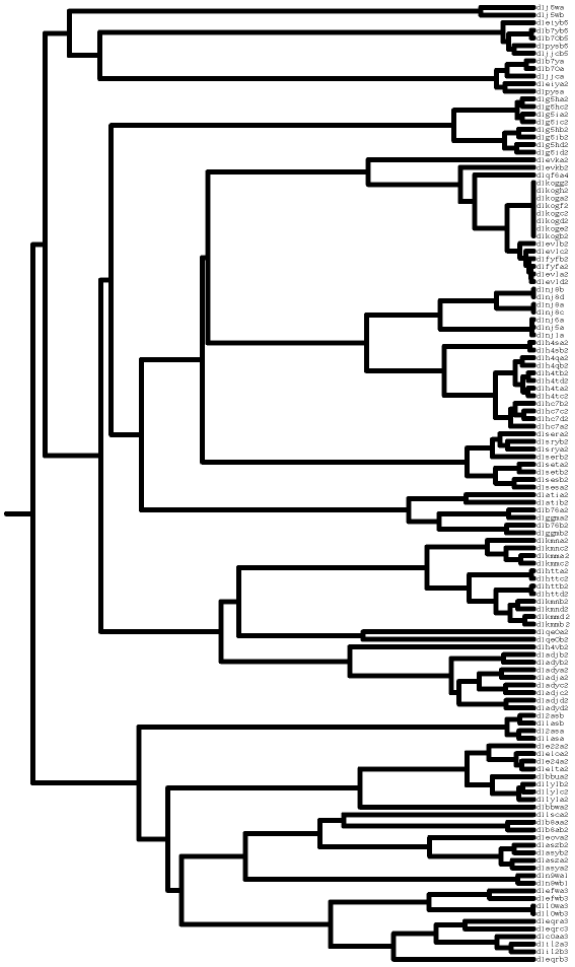
Cont'd - Web Resources

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_a 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 gragments to do a Monte Carlo Fragment Insertion conformational search for peices 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.

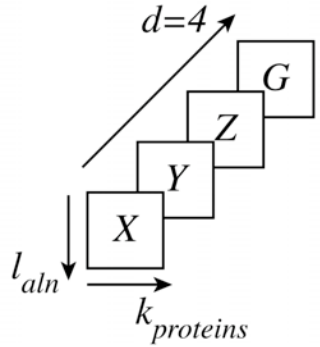
Non-redundant Representative Sets of Structures

Too much information
129 Structures

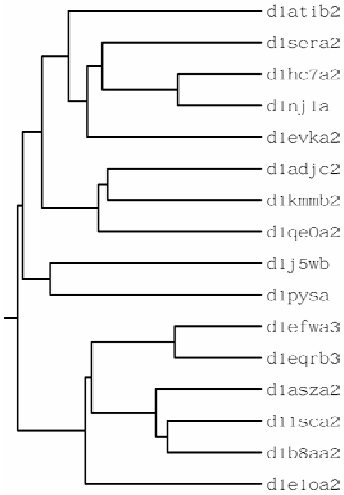


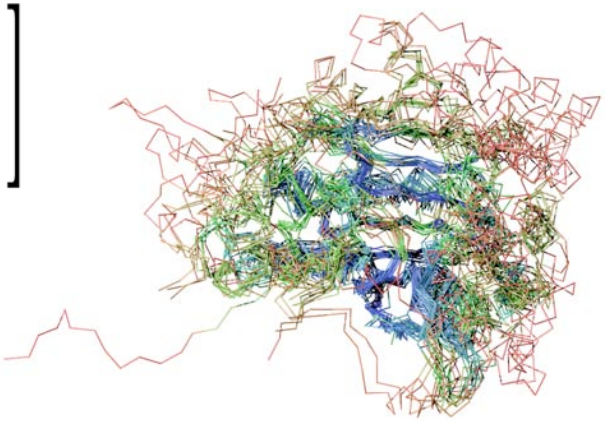
Multidimensional QR factorization of alignment matrix, A.

$$A = \left[\quad \right]$$



Economy of information
16 representatives





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

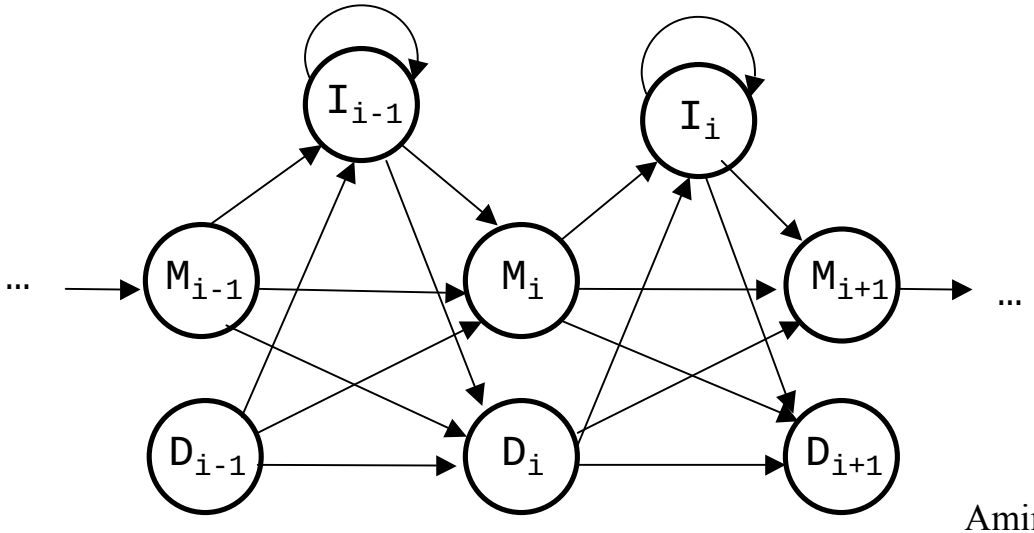
P. O'Donoghue and Z. Luthey-Schulten (2004) *J. Mol. Biol.*, in press.

HMM-based approach

	1	2	3	4	5
Sequence 1	-	B	B	-	C
Sequence 2	C	C	-	-	C
Sequence 3	C	B	C	C	B
State	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

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

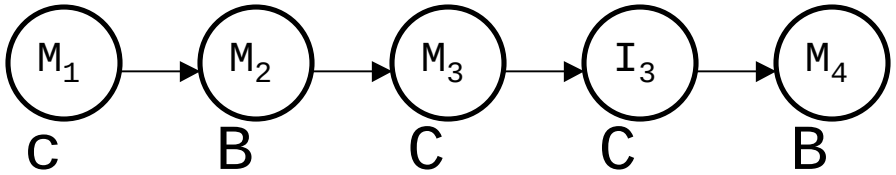
$P(C|I) = 7/11 = 0.64$

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

Leads to affine gap penalty.

$P(-|D) = 1$.

State Path of Sequence 3



Durbin, Eddy, Krogh and Mitchison, Biological Sequence Analysis, Cambridge, 1998

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,…,A_N) is:

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

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

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

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

$$P(MSA, \pi|M) = \prod_{k=1}^{N_{seq}} P(A^k, \pi^k|M)$$

where A^k represents the k^{th} aligned sequence.

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

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

$$logP_j = \max_{0\leq i\leq j-1} logP_{ij} = \max_{0\leq i\leq j-1} (logP_i + ST_{ij} + AM_j + AI_{ij})$$

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 jth column.

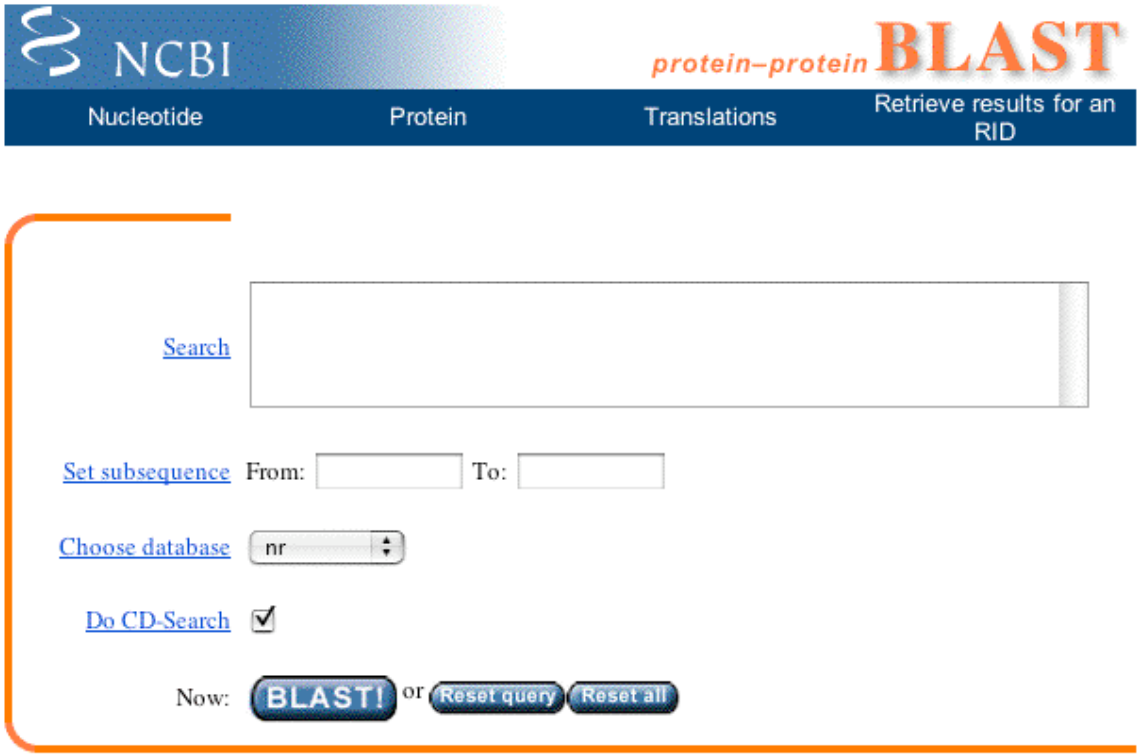
HMM-based approach

New sequence aligned using dynamic programming.

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

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

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

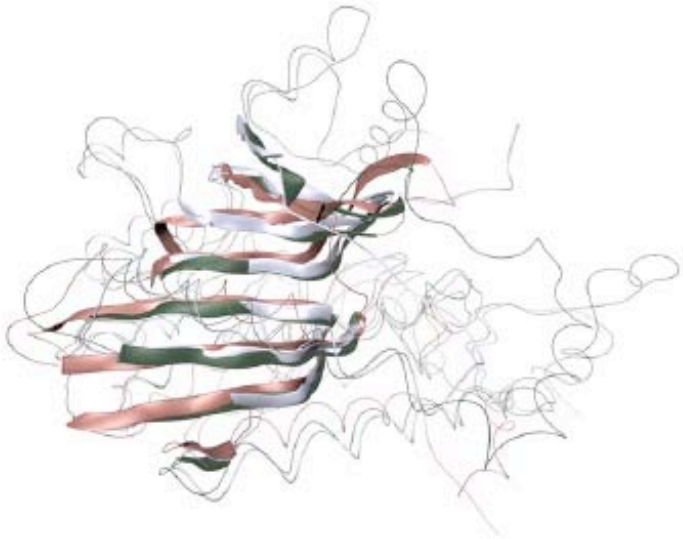

NCBI
protein–protein BLAST
Nucleotide
Protein
Translations
Retrieve results for an RID
Search
Set subsequence
From:
To:
Choose database
nr
Do CD-Search
Now:
BLAST!
or
Reset query
Reset all

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