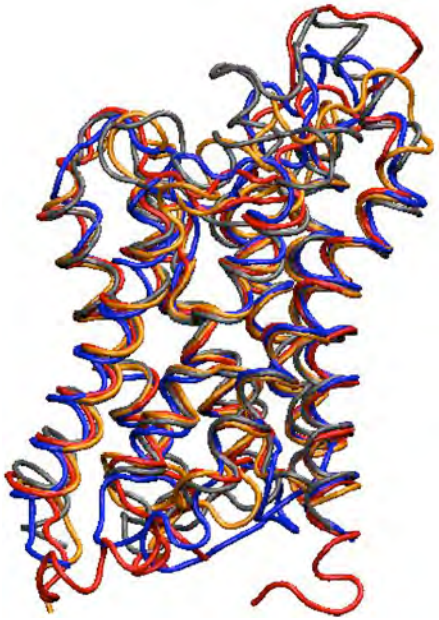
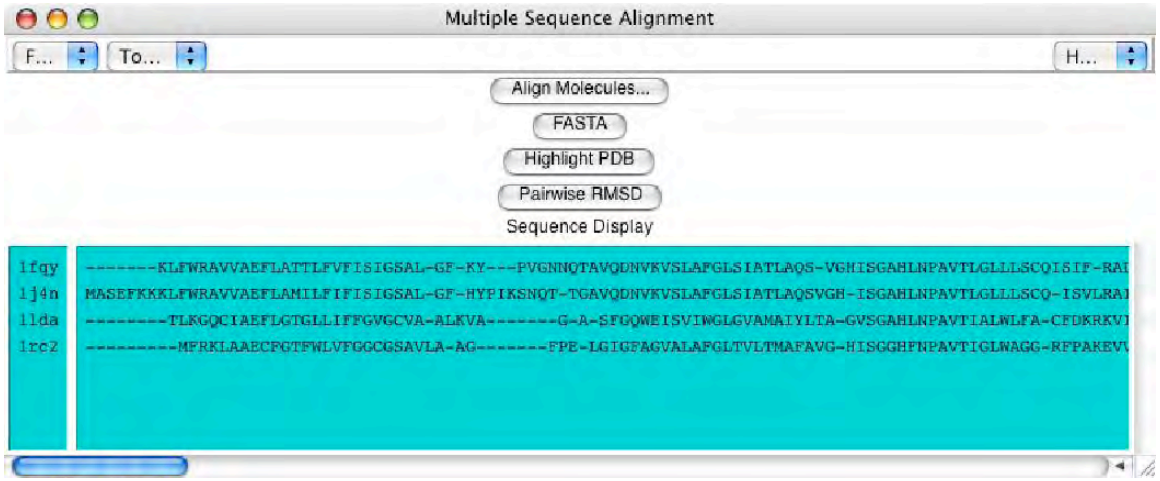
Sequence and Structure Alignment

Z. Luthey-Schulten, UIUC

Boston, December 2004





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

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

Structure-Structure Alignment

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

Sequence Database Searches

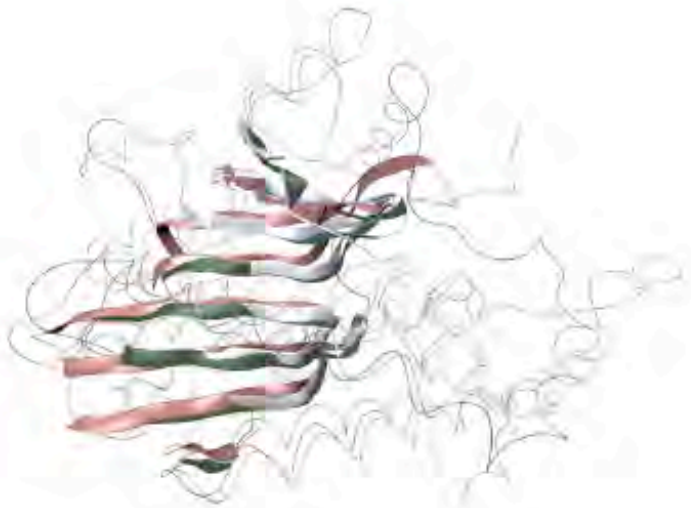
- Blast and Psi-Blast

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

Sequence Alignment Algorithms

Tutorial for the material of this lecture available at

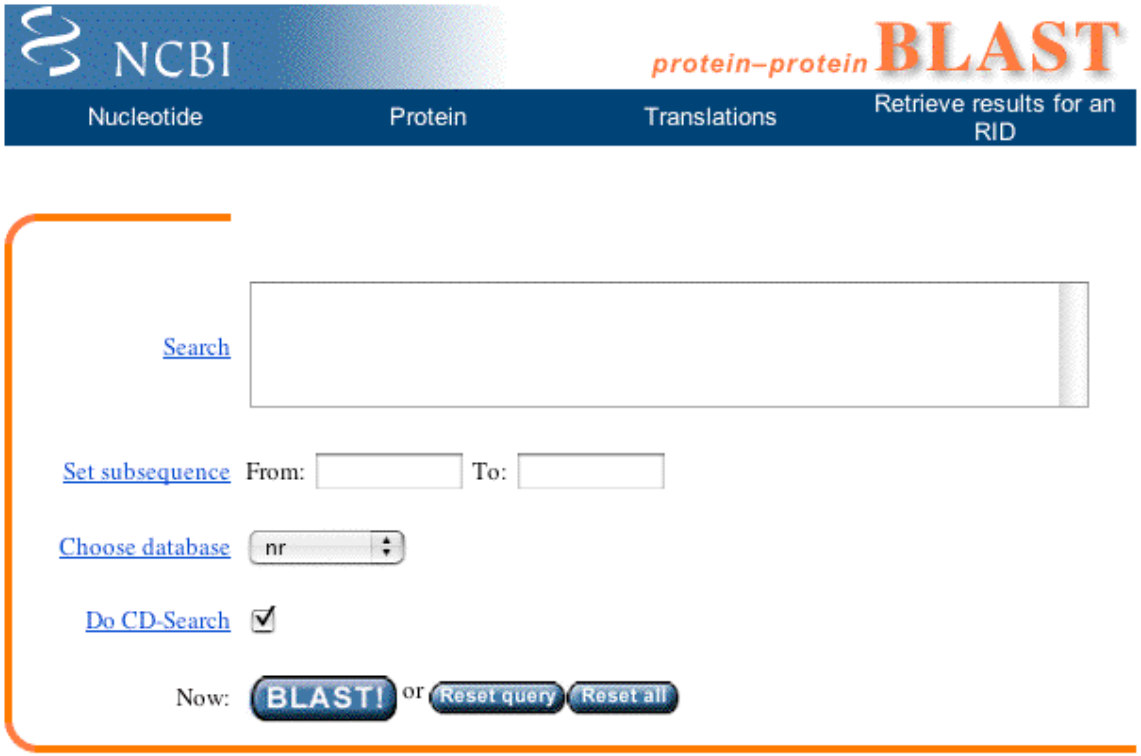
http://www.ks.uiuc.edu/Training/Tutorials/



Rommie Amaro
Felix Autenrieth
Brijeet Dhaliwal
Barry Isralewitz

Zaida Luthey-Schulten
Anurag Sethi
Taras Pogorelov

June 2004


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

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.

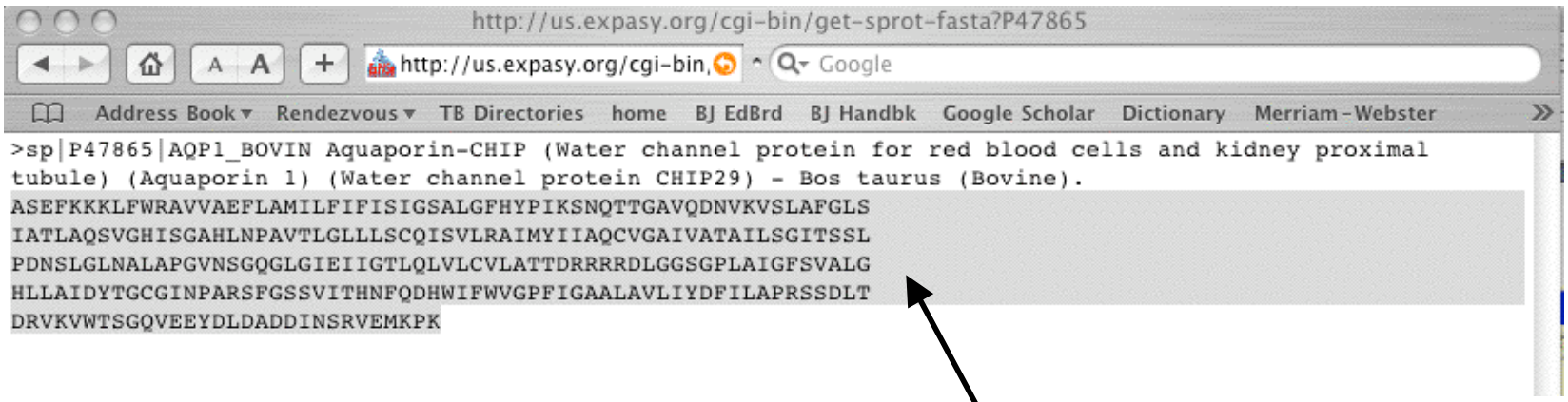
Entry information	
Entry name	**AQP1_BOVIN**
Primary accession number	**P47865**
Secondary accession numbers	None
Entered in Swiss-Prot in	Release 33, February 1996
Sequence was last modified in	Release 44, July 2004
Annotations were last modified in	Release 45, October 2004
Name and origin of the protein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

paste

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

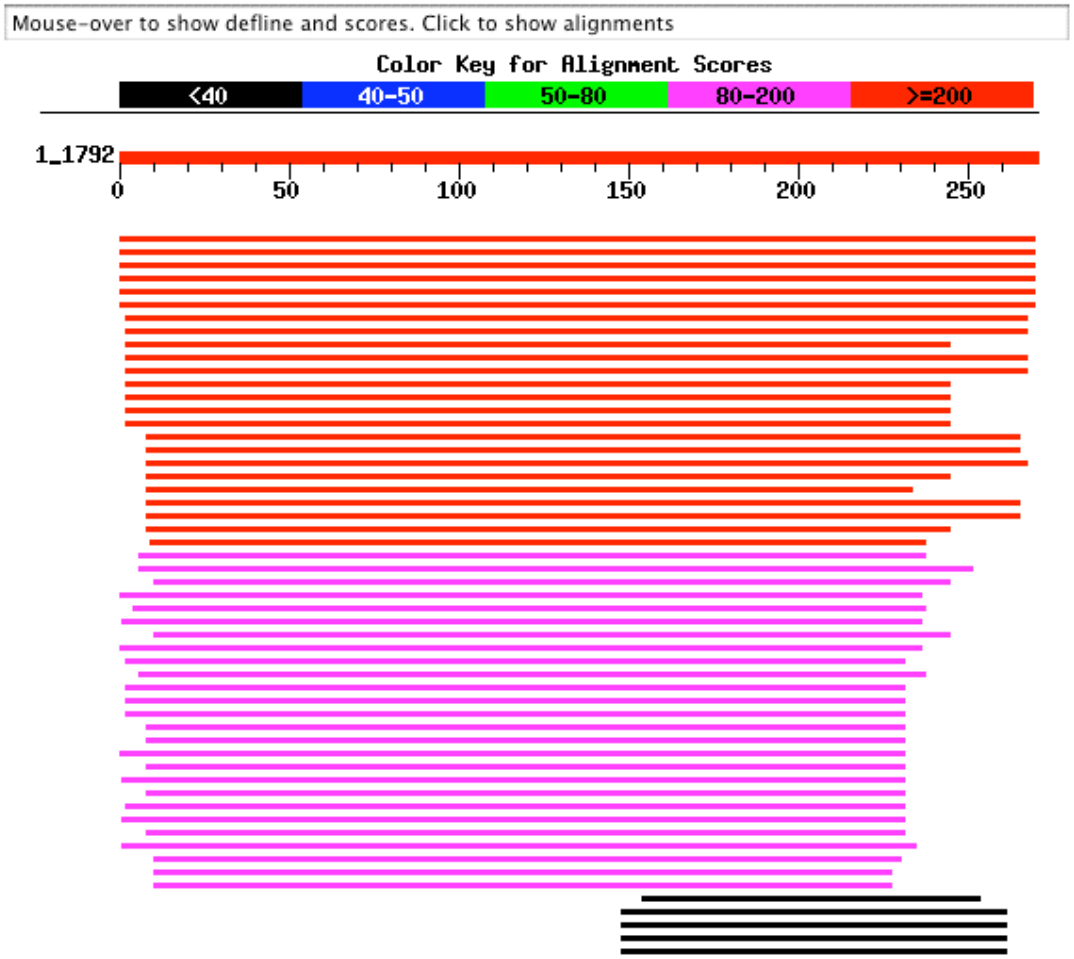
Format! or Reset all

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

>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%)

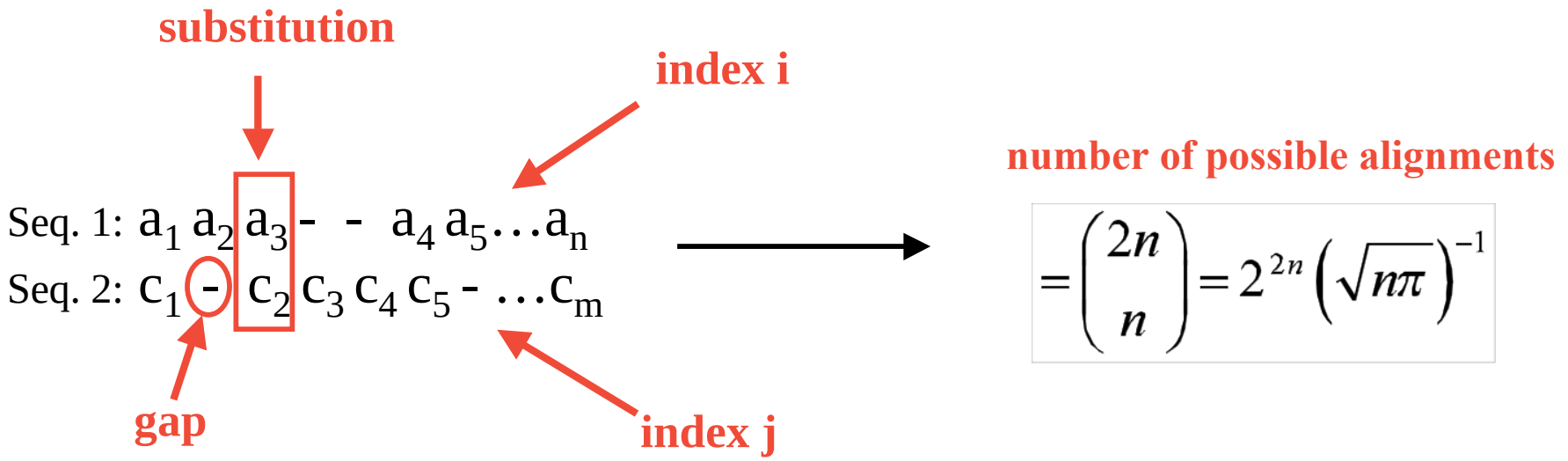
```
Query: 53  VSLAFGLSIATLAQSVGHISGAHLNPAVTLGLLLSCQISVLRAIMYIIAQCVGAIVATAI 112
           V+ AFGL++ T+A ++GHISG HLNPAV+ GL++  +      + Y+IAQ +GAI+A  +
Sbjct: 40  VAFAFGLTVLTMAFAIGHISGCHLNPAVSFGLVVGGRFPAKELLPYVIAQVIGAILAAGV 99

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
```

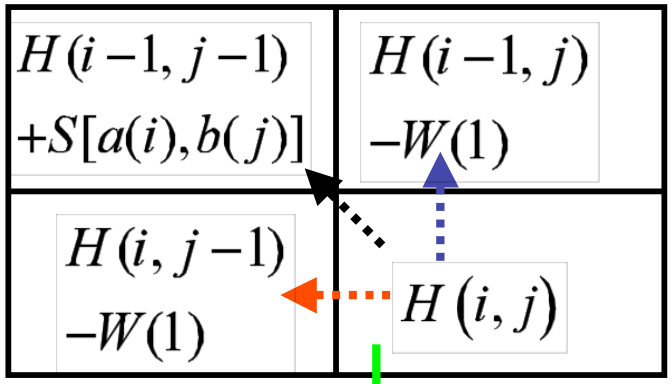
Sequence Alignment & Dynamic Programming



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

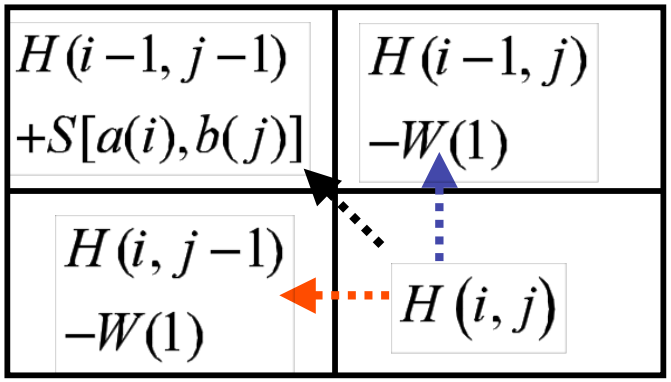
Seq. 2: c_1 - $c_2\ c_3\ c_4\ c_5$ - $\ldots c_m$

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

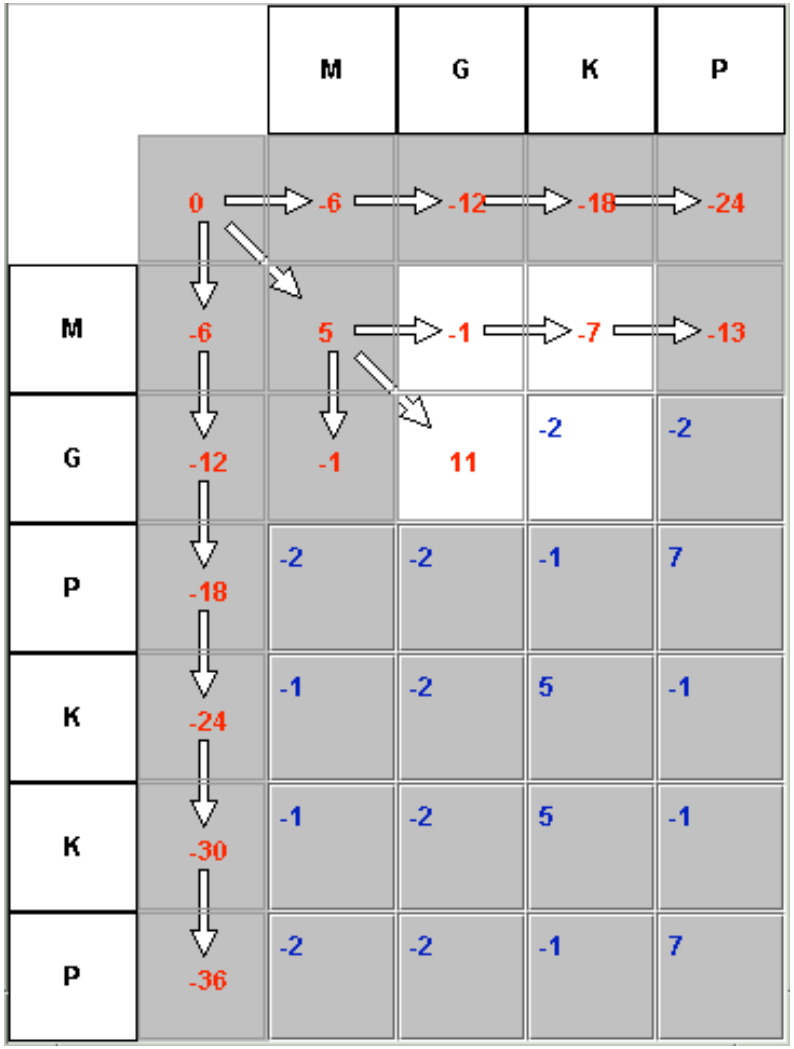
		M	G	K	P
M		5	-3	-1	-2
G		-3	6	-2	-2
P		-2	-2	-1	7
K		-1	-2	5	-1
K		-1	-2	5	-1
P		-2	-2	-1	7

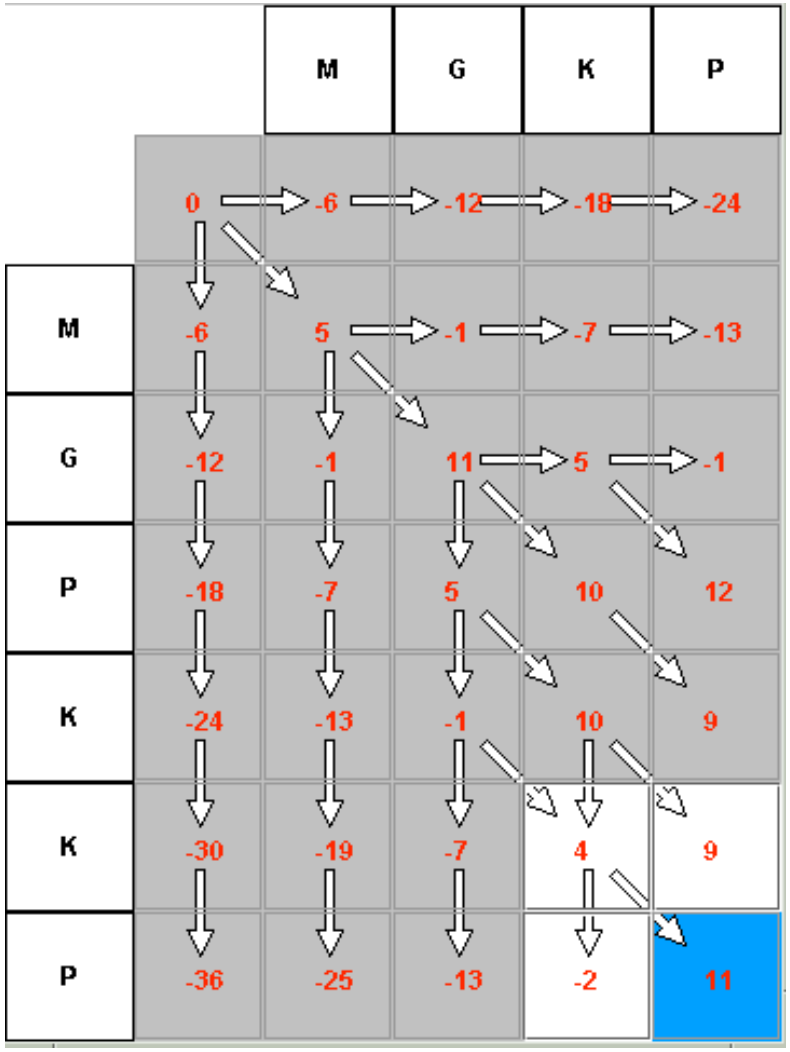
Initialization of Gap Penalties

		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

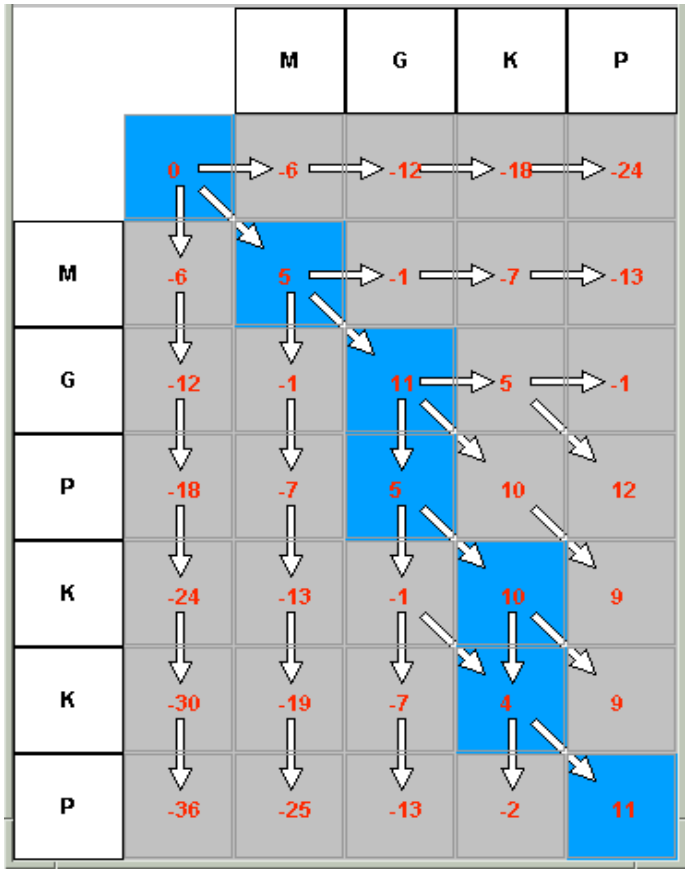
Filling out the Score Matrix H





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

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

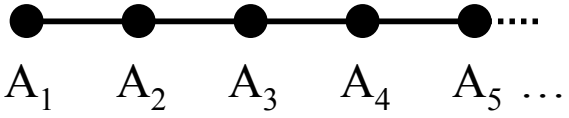
Target Sequence

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

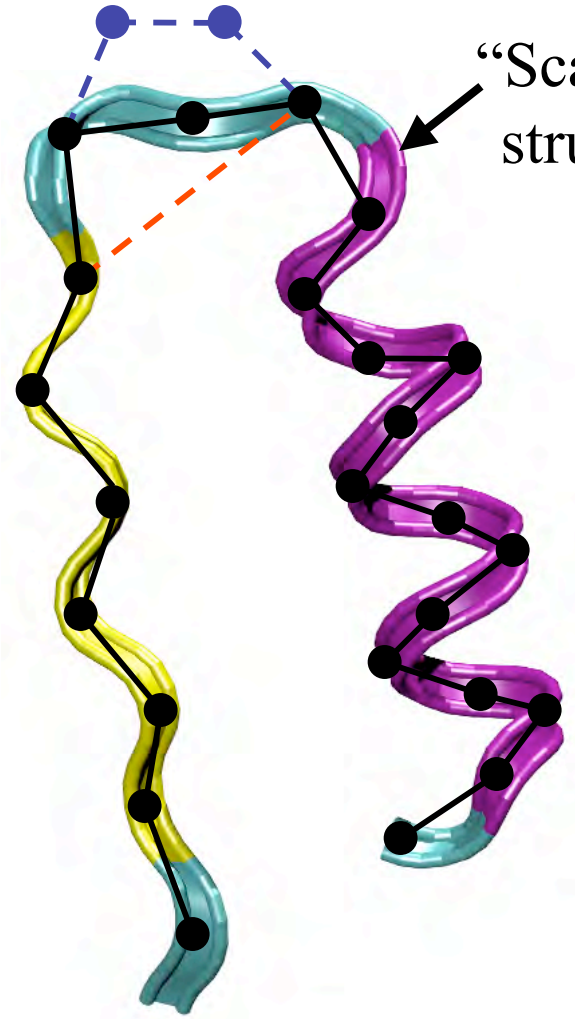
Known structure(s)

Sequence-Structure Alignment

Target sequence



Alignment between target(s) and scaffold(s)



1. Energy Based Threading*

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

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

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

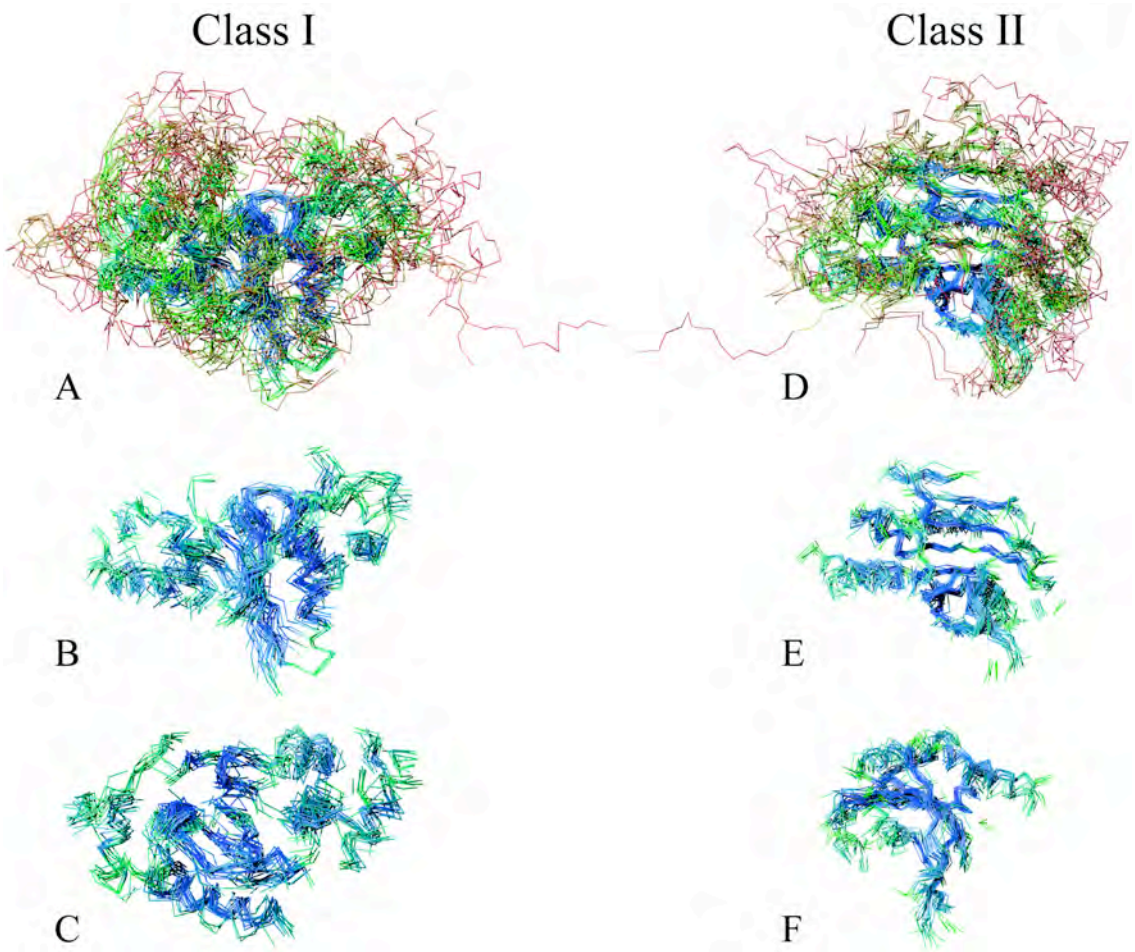
2. Sequence – Structure Profile Alignments

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)

Profile - Multiple Structural Alignments

Representative Profile of AARS Family



Structural Profiles

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

Add more structural information!

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.

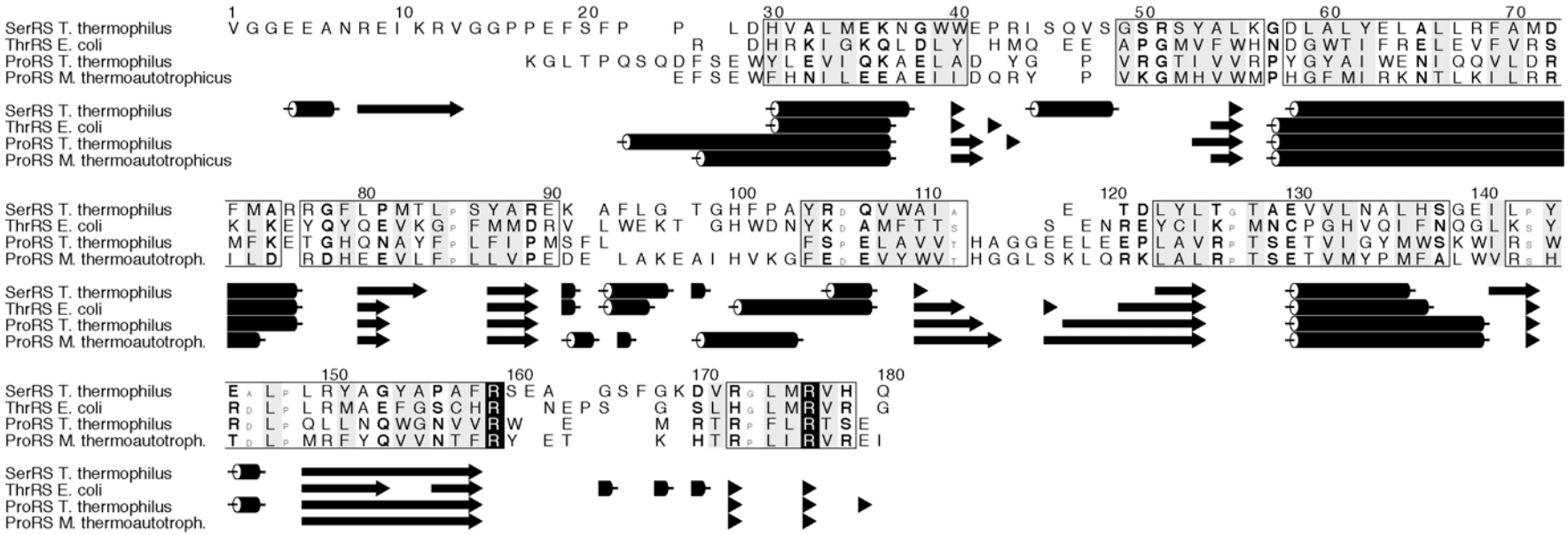
d_{ij} -- distance between i & j

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.

Variation in Secondary Structure STAMP Output



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

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