Introduction to Bioinformatics

JTMS-19

Marc-Thorsten Hütt mhuett@constructor.university
Felix Jonas fjonas@constructor.university

Algorithms for pairwise sequence alignments

What is this session about?

Pairwise sequence alignment is discussed in detail (scoring functions, Needleman-Wunsch algorithm, Smith-Waterman algorithms). Some background on heuristic sequence alignment methods (FastA and BLAST) is given.

How can you revise the material after the session?

Read Baxevanis/Oullette chapter 11

Read Durbin et al. chapters 2.1– 2.5

Read Frédérique Galisson, The Fasta and Blast programs, 2000

alternative reading: Hütt/Dehnert chapters 3.1.1-3.1.3, 3.2.1

DNA sequences: the case of global sequence alignment

	0	1 T	2 G	3 C	4 A	5 T	6 A
0	0	0	0	0	0	0	0
1 A	0	0	0	0	1	1	1
2 T	0	1	1	1	1	2	2
3 C	0	1	1	2	2	2	2
4 T	0	1	1	2	2	3	3
5 G	0	1	2	2	2	3	3
6 A	0	1	2	2	3	3	4
7 T	0	1	2	2	3	4	4

```
-tgcat-a-
at-c-tgat
```

	0	1 T	2 G	3 C	4 A	5 T	6 A
0	0	0	0	0	0	0	0
1 A	0	0	0	0	1	1	1
2 T	0	1	1	1	1	2	2
3 C	0	1	1	2	2	2	2
4 T	0	1	1	2	2	3	3
5 G	0	1	2	2	2	3	3
6 A	0	1	2	2	3	3	4
7 T	0	1	2	2	3	4	4

```
---tgcata
atctg-at-
```

$$x = C, G, A, T, C, C, T, G, T$$
$$y = C, A, T, C, G, C, C, T, T$$

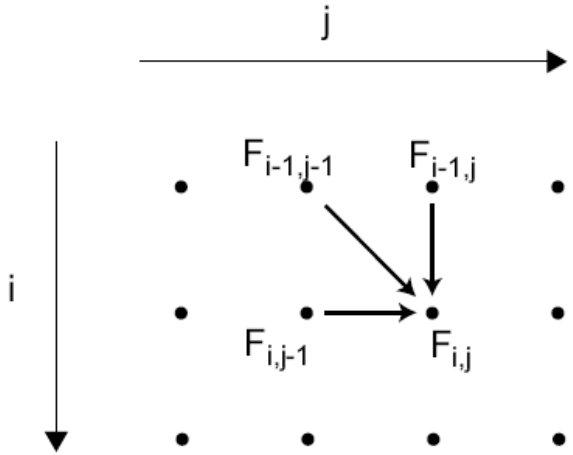
$$x_i = y_j : \quad F_{i,j} = \max \begin{cases} F_{i-1,j} \\ F_{i,j-1} \\ F_{i-1,j-1} + 1 \end{cases}$$

$$x_i \neq y_j : \quad F_{i,j} = \max \begin{cases} F_{i-1,j} \\ F_{i,j-1} \\ F_{i-1,j-1} \end{cases}$$

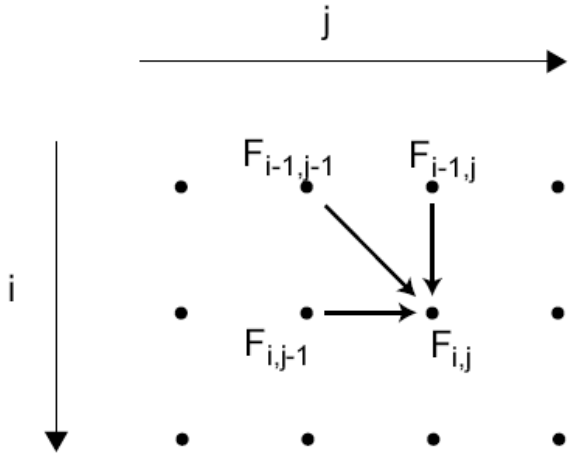
		C	A	T	C	G	C	C	T	T
	0	0	0	0	0	0	0	0	0	0
C	0	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	2	2	2	2	2
A	0	1	2	2	2	2	2	2	2	2
T	0	1	2	3	3	3	3	3	3	3
C	0	1	2	3	4	4	4	4	4	4
C	0	1	2	3	4	4	5	5	5	5
T	0	1	2	3	4	4	5	5	6	6
G	0	1	2	3	4	5	5	5	6	6
T	0	1	2	3	4	5	5	5	6	7

```
CGAT--CCTGT
|  ||   ||||  |
C-ATCGCCT-T
```

moving around in the dotplot plane



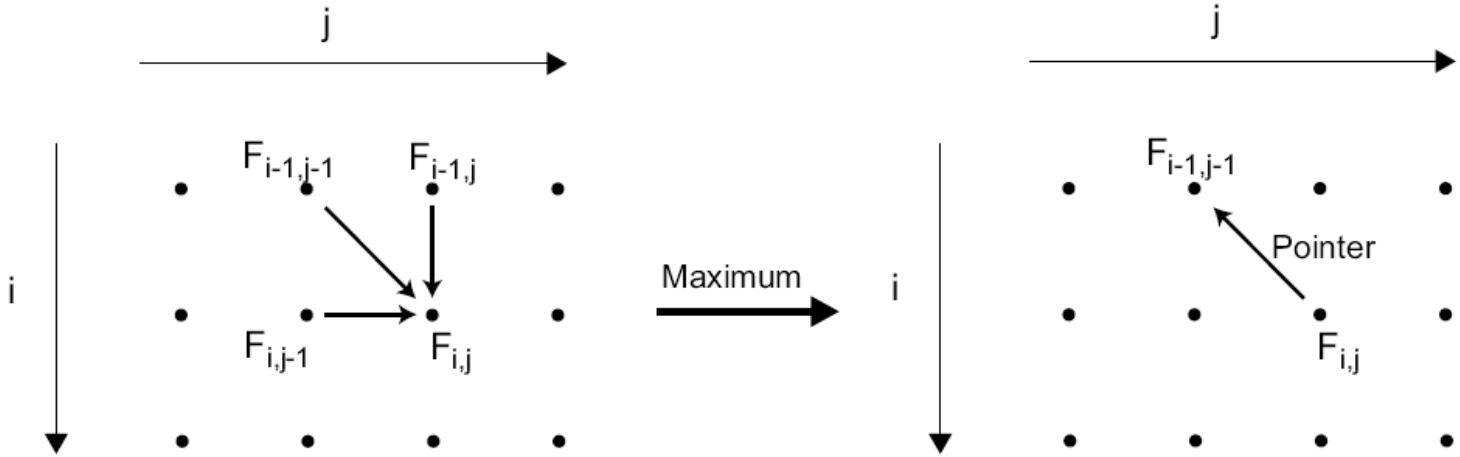
moving around in the dotplot plane



translation of a path into an alignment

- **horizontal segment: gap in x + sympol in y**
- **vertical segment: symbol in x + gap in y**
- **diagonal segment: symbol in x + symbol in y**

moving around in the dotplot plane



EXAMPLE x = CACTG
y = ATG

- initialize the matrix
- fill out the entries (including the arrows)

  heuristic for the arrows
  - in case of ties, prefer diagonals
  - highlight the remaining ambiguous cases and see, if they matter

- extract the optimal alignment

EXAMPLE x = CACTG
y = ATG

- initialize the matrix
- fill out the entries (including the arrows)

  heuristic for the arrows
  - in case of ties, prefer diagonals
  - highlight the remaining ambiguous cases and see, if they matter

- extract the optimal alignment

		C	A	C	T	G
	0	0	0	0	0	0
A	0	0	1	1	1	1
T	0	0	1	1	2	2
G	0	0	1	1	2	3

```
C  A  C  T  G
   |     |  |
-  A  -  T  G
```

Global sequence alignment for protein sequences: the Needleman-Wunsch algorithm

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

BLOSUM 62

1. x_i is matched with y_j:

$$\left.\begin{array}{c} \ldots\ x_i \\ | \\ \ldots\ y_j \end{array}\right\} \rightarrow F_{i,j} = F_{i-1,j-1} + s(x_i, y_j)$$

2. x_i is matched with a gap:

$$\left.\begin{array}{c} \ldots\ x_i \\ | \\ \ldots\ - \end{array}\right\} \rightarrow F_{i,j} = F_{i-1,j} - d$$

3. y_j is matched with a gap:

$$\left.\begin{array}{c} \ldots\ - \\ | \\ \ldots\ y_j \end{array}\right\} \rightarrow F_{i,j} = F_{i,j-1} - d$$

1. x_i is matched with y_j:

$$\left.\begin{array}{c} \ldots\ x_i \\ | \\ \ldots\ y_j \end{array}\right\} \rightarrow F_{i,j} = F_{i-1,j-1} + s(x_i, y_j)$$

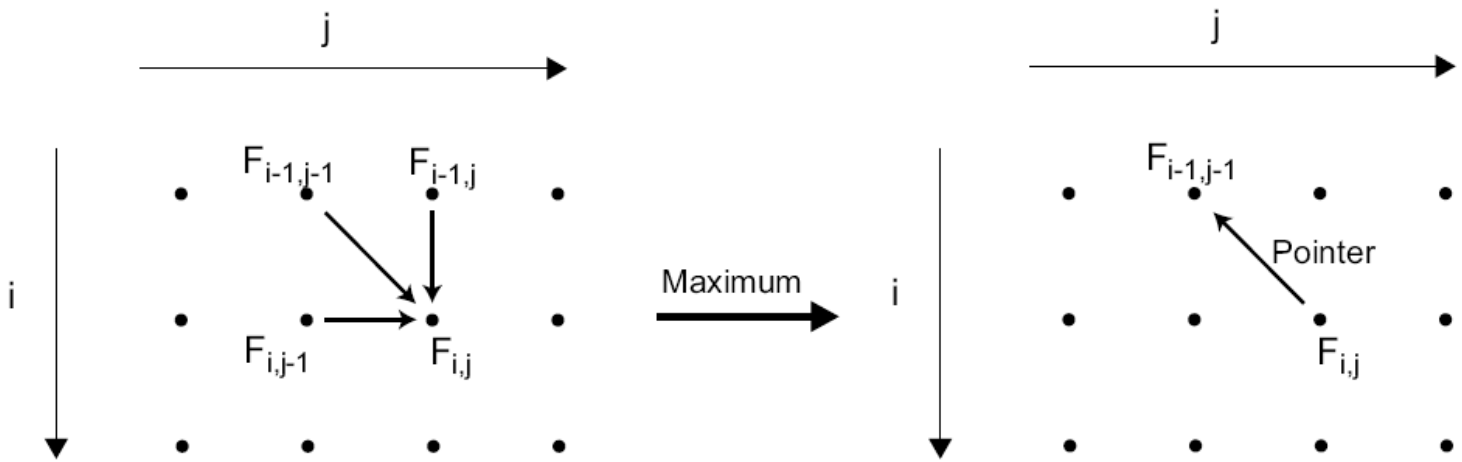
2. x_i is matched with a gap:

$$\left.\begin{array}{c} \ldots\ x_i \\ | \\ \ldots\ - \end{array}\right\} \rightarrow F_{i,j} = F_{i-1,j} - d$$

3. y_j is matched with a gap:

$$\left.\begin{array}{c} \ldots\ - \\ | \\ \ldots\ y_j \end{array}\right\} \rightarrow F_{i,j} = F_{i,j-1} - d$$

$$F_{i,j} = \max \begin{cases} F_{i-1,j-1} + s(x_i, y_j) \\ F_{i-1,j} - d \\ F_{i,j-1} - d \end{cases}$$


j
i
$F_{i-1,j-1}$
$F_{i-1,j}$
$F_{i,j-1}$
$F_{i,j}$
Maximum
j
i
$F_{i-1,j-1}$
Pointer
$F_{i,j}$

	-	H	E	A	G	A	W	G	H	E	E
-	0	-8	-16	-24							
P	-8										
A	-16										
W	-24										
H											
E											
A											
E											

[from Durbin et al. (1998)]

	-	H	E	A	G	A	W	G	H	E	E
-	0	-8	-16	-24							
P	-8	-2									
A	-16										
W	-24										
H											
E											
A											
E											

$$F_{i-1,j-1} + s(P,H) = 0 - 2$$

$$F_{i-1,j} - d = -8-8$$

$$F_{i,j-1} - d = -8-8$$

[from Durbin et al. (1998)]

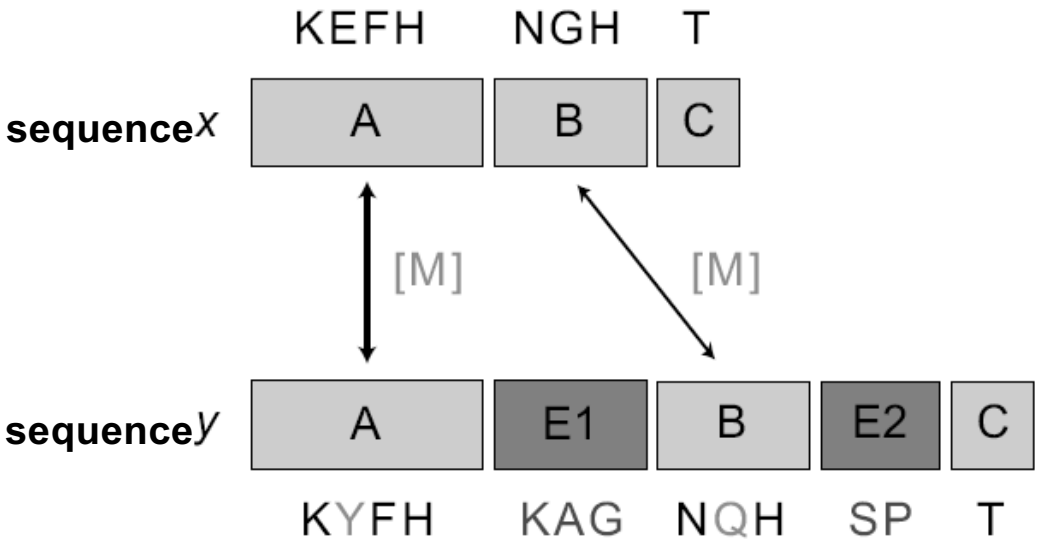
		H	E	A	G	A	W	G	H	E	E
	0	**–8**	**–16**	–24	–32	–40	–48	–56	–64	–72	–80
P	–8	–2	–9	**–17**	**–25**	–33	–42	–49	–57	–65	–73
A	–16	–10	–3	–4	–12	**–20**	–28	–36	–44	–52	–60
W	–24	–18	–11	–6	–7	–15	**–5**	**–13**	–21	–29	–37
H	–32	–14	–18	–13	–8	–9	–13	–7	**–3**	–11	–19
E	–40	–22	–8	–16	–16	–9	–12	–15	–7	**3**	–5
A	–48	–30	–16	–3	–11	–11	–12	–12	–15	**–5**	2
E	–56	–38	–24	–11	–6	–12	–14	–15	–12	–9	**1**

HEAGAWGHE-E

--P-AW-HEAE

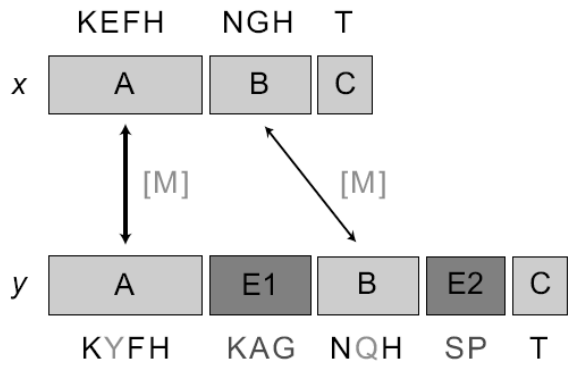
[from Durbin et al. (1998)]

$$x = K, E, F, H, N, G, H, T$$
$$y = K, Y, F, H, K, A, G, N, Q, H, S, P, T$$



		K	Y	F	H	K	A	G	N	Q	H	S	P	T
	0	-5	-10	-15	-20	-25	-30	-35	-40	-45	-50	-55	-60	-65
K	-5	5	0	-5	-10	-15	-20	-25	-30	-35	-40	-45	-50	-55
E	-10	0	3	-2	-5	-9	-14	-19	-24	-28	-33	-38	-43	-48
F	-15	-5	3	9	4	-1	-6	-11	-16	-21	-26	-31	-36	-41
H	-20	-10	-2	4	17	12	7	2	-3	-8	-13	-18	-23	-28
N	-25	-15	-7	-1	12	17	12	7	8	3	-2	-7	-12	-17
G	-30	-20	-12	-6	7	12	17	18	13	8	3	-2	-7	-12
H	-35	-25	-17	-11	2	7	12	15	19	14	16	11	6	1
T	-40	-30	-22	-16	-3	2	7	10	15	18	13	17	12	11

```
KEFHN-G--H--T
|  ||   |   |  |
KYFHKAGNQHSPT
```



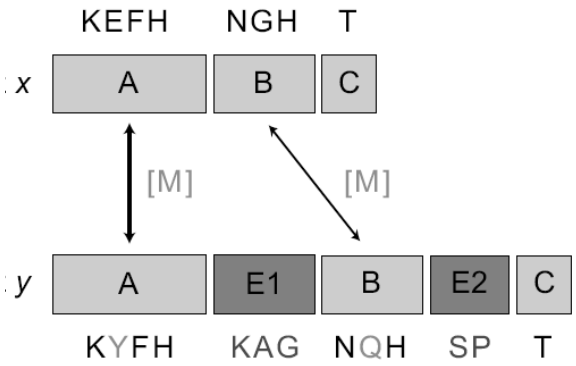

```
KEFHN-G--H--T
|  ||   |   |  |
KYFHKAGNQHSPT
```

Local sequence alignment for protein sequences: the Smith-Waterman algorithm

$$F_{i,j} = \max \begin{cases} 0 \\ F_{i-1,j-1} + s(x_i, y_j) \\ F_{i-1,j} - d \\ F_{i,j-1} - d \,. \end{cases}$$

```
KEFHN-GH
|  || |
KYFHKAGN
```

		K	Y	F	H	K	A	G	N	Q	H	S	P	T
	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	0	5	0	0	0	5	0	0	0	1	0	0	0	0
E	0	1	3	0	0	1	4	0	0	2	1	0	0	0
F	0	0	4	9	4	0	0	1	0	0	1	0	0	0
H	0	0	2	4	17	12	7	2	2	0	8	3	0	0
N	0	0	0	0	12	17	12	7	8	3	3	9	4	0
G	0	0	0	0	7	12	17	18	13	8	3	4	7	2
H	0	0	2	0	8	7	12	15	19	14	16	11	6	5
T	0	0	0	0	3	7	7	10	15	18	13	17	12	11




```
KEFHN-GH
|  ||   |
KYFHKAGN
```

How probable is a particular sequence similarity?

```
KEFHN-G--H--T
|  ||   |   |   |     S=11
KYFHKAGNQHSPT
```

```
-H-HN-GFE-T-K
   |   |          S=-10
KYFHKAGNQHSPT
```

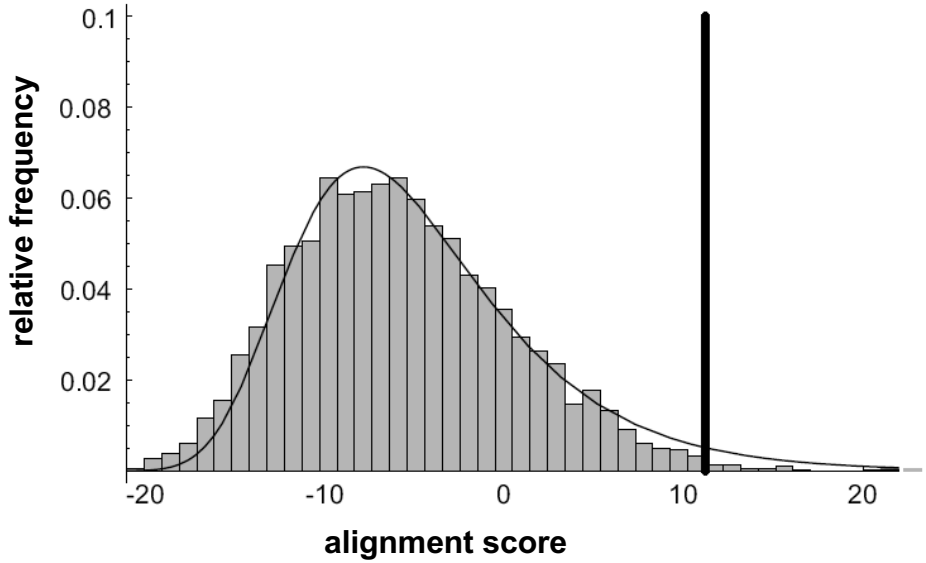
```
E-F-K--HTHG-N
  | |     |        S=-5
KYFHKAGNQHSPT
```

```
N--HKTGFEH---
   || |   |        S=1
KYFHKAGNQHSPT
```

```
GENFTHK----H---
   | ||     |       S=-19
-KYF-HKAGNQHSPT
```

```
GEF----NKHT-H
  |     | |         S=-9
KYFHKAGNQHSPT
```

How probable is a particular sequence similarity?



How probable is a particular sequence similarity?

z-score of an alignment score S:

- alignments of all permutations of x and y
- set of alignment scores $\{S_i\}$
- average value A and standard deviation w for $\{S_i\}$

$$\text{z-score}(S) = \frac{S - A}{w}$$

z-score	
0	alignment score S coincides with that of the surrogate data
> 5	S is significant

How probable is a particular sequence similarity?

other quantitative assessments of an alignment score S:

P-value
probability to obtain the score S with a random alignment

$P \leq 10^{-100}$ exact match,
$10^{-100} \leq P \leq 10^{-50}$ nearly identical (SNPs)
$10^{-50} \leq P \leq 10^{-10}$ homology certain
$10^{-5} \leq P \leq 10^{-1}$ usually distant relative
$P > 10^{-1}$ probably insignificant

E-value
expected frequency of S in a random alignment with a database

$E \leq 0.02$	significant result
$0.02 < E < 1$	unclear; homology uncertain
$E \geq 1$	score corresponds to a random alignment

P = E * {size of the database}

How probable is a particular sequence similarity?

heuristic methods of sequence alignment

FastA = fast Alignment

D.J. Lipman and W.W.R. Pearson. Rapid and sensitive protein similarity searches. *Science*, 227:1435–1441, 1985.

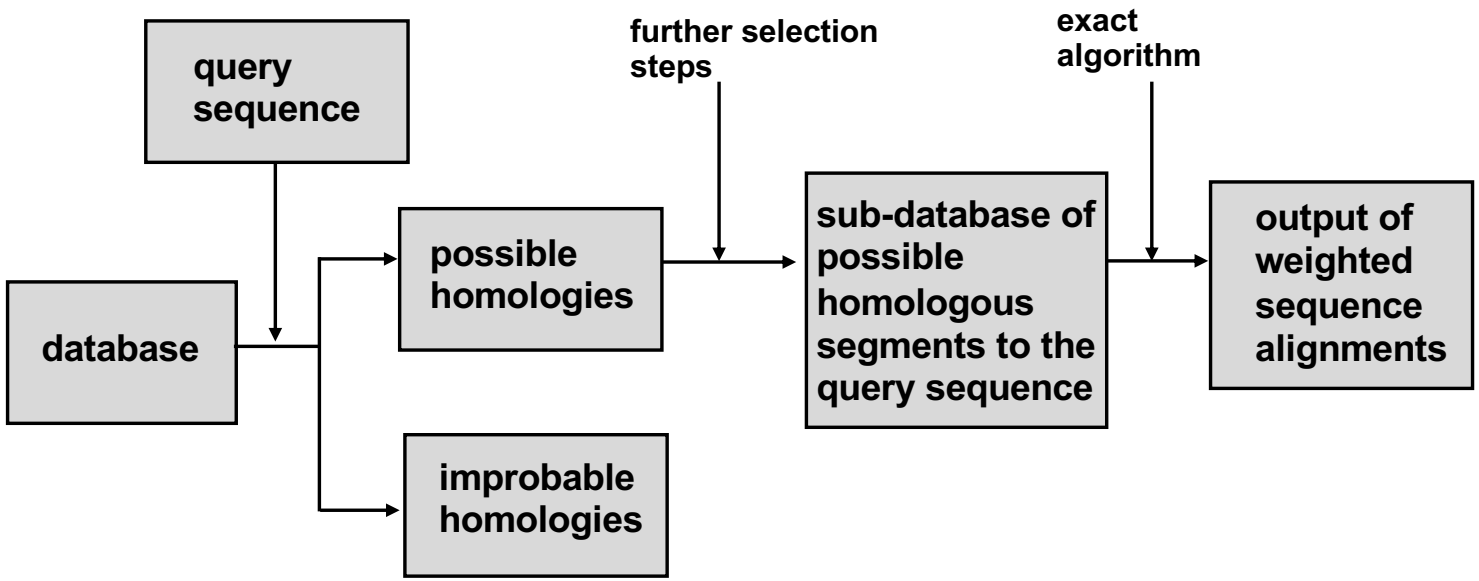
W.R. Pearson and D.J. Lipman. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, 85:2444–2448, 1988.

BLAST = Basic Local Alignment Search Tool

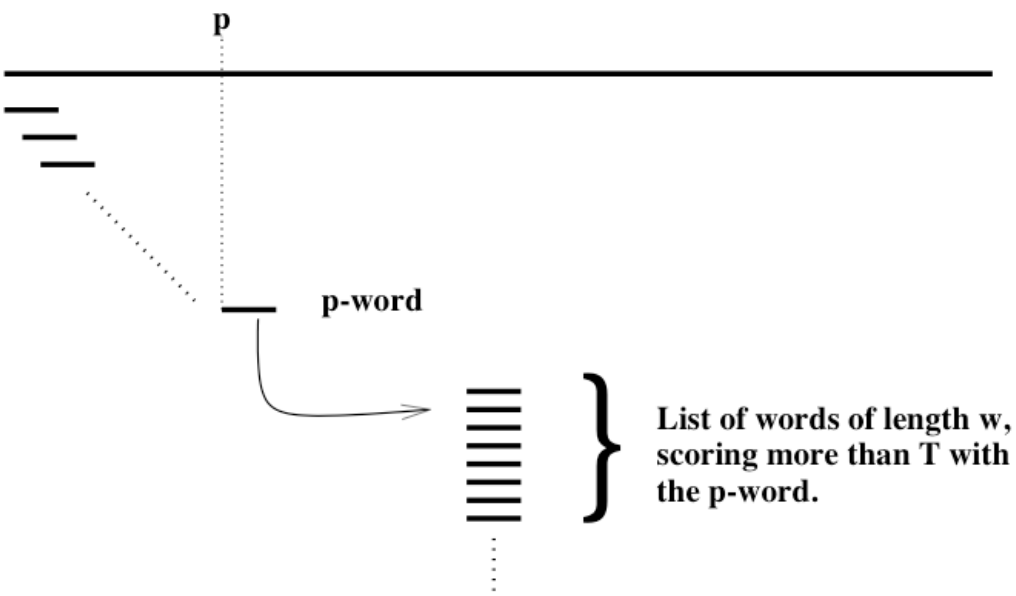
S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, 1990.

[a good introduction to these methods:
Frédérique Galisson, The fasta and blast programs, 2002]

Looking inside BLAST



A: For each position p of the query, find the list or words of length w scoring more than T when paired with the word starting at p:



Looking inside BLAST

query sequence: QLNFSAGW

(1) parameters

word length w = 2
score threshold T = 8

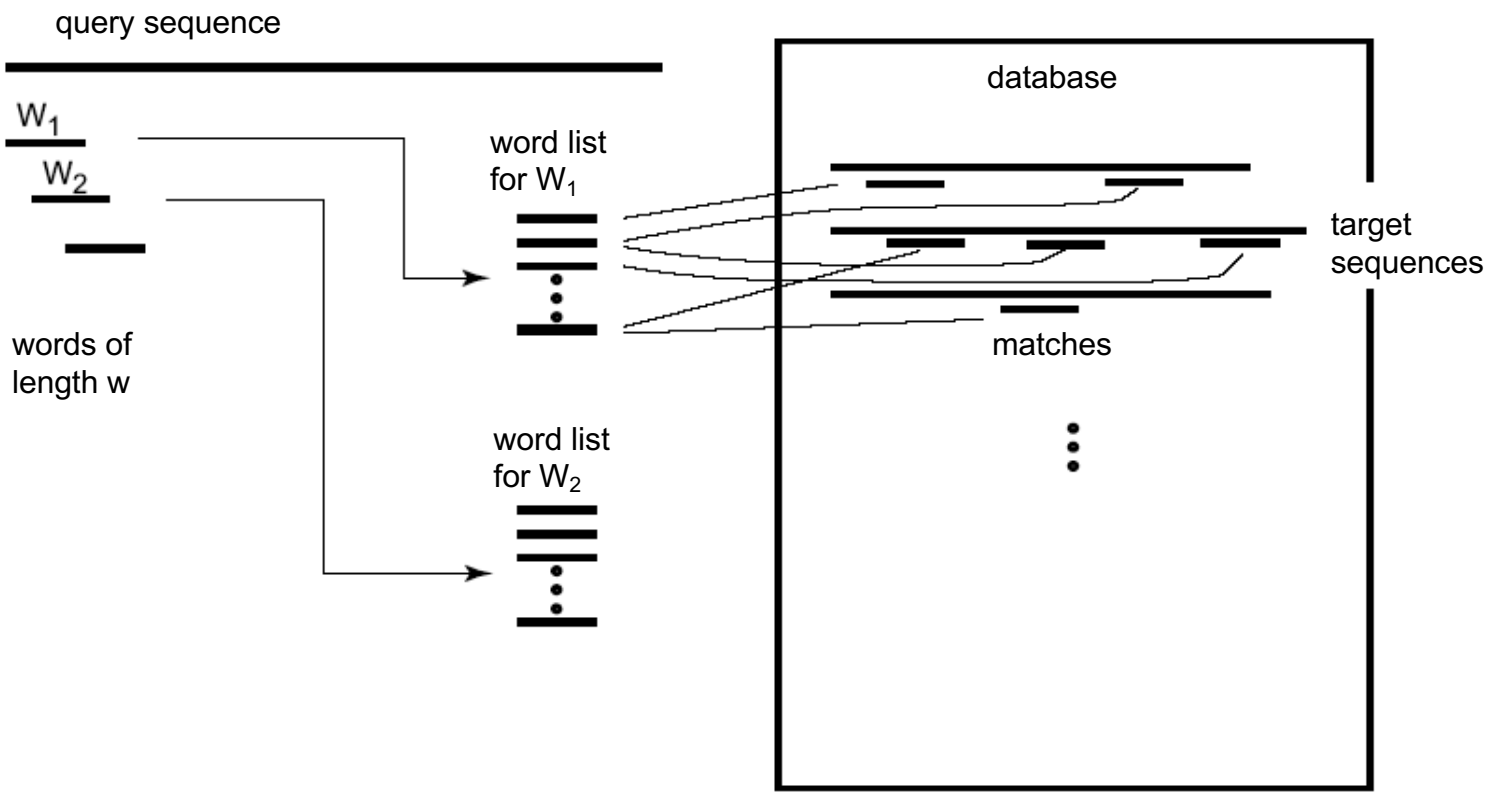
(2) determine all words of length w in the query sequence:

QL LN NF FS SA AG GW

(3) for each word determine a word list with an alignment score larger than (or equal to) the threshold T:

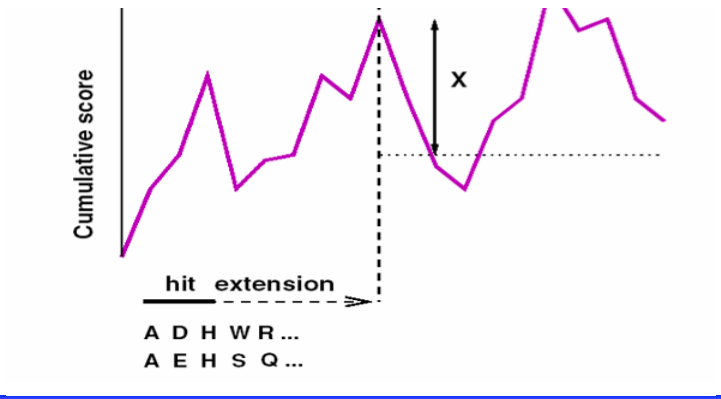
QL:	**QL=11, QM=9, HL=8, ZL=9**
LN:	**LN=9, LB=8**
NF:	**NF=12, AF=8, NY=8, DF=10, ...**
...	

Looking inside BLAST

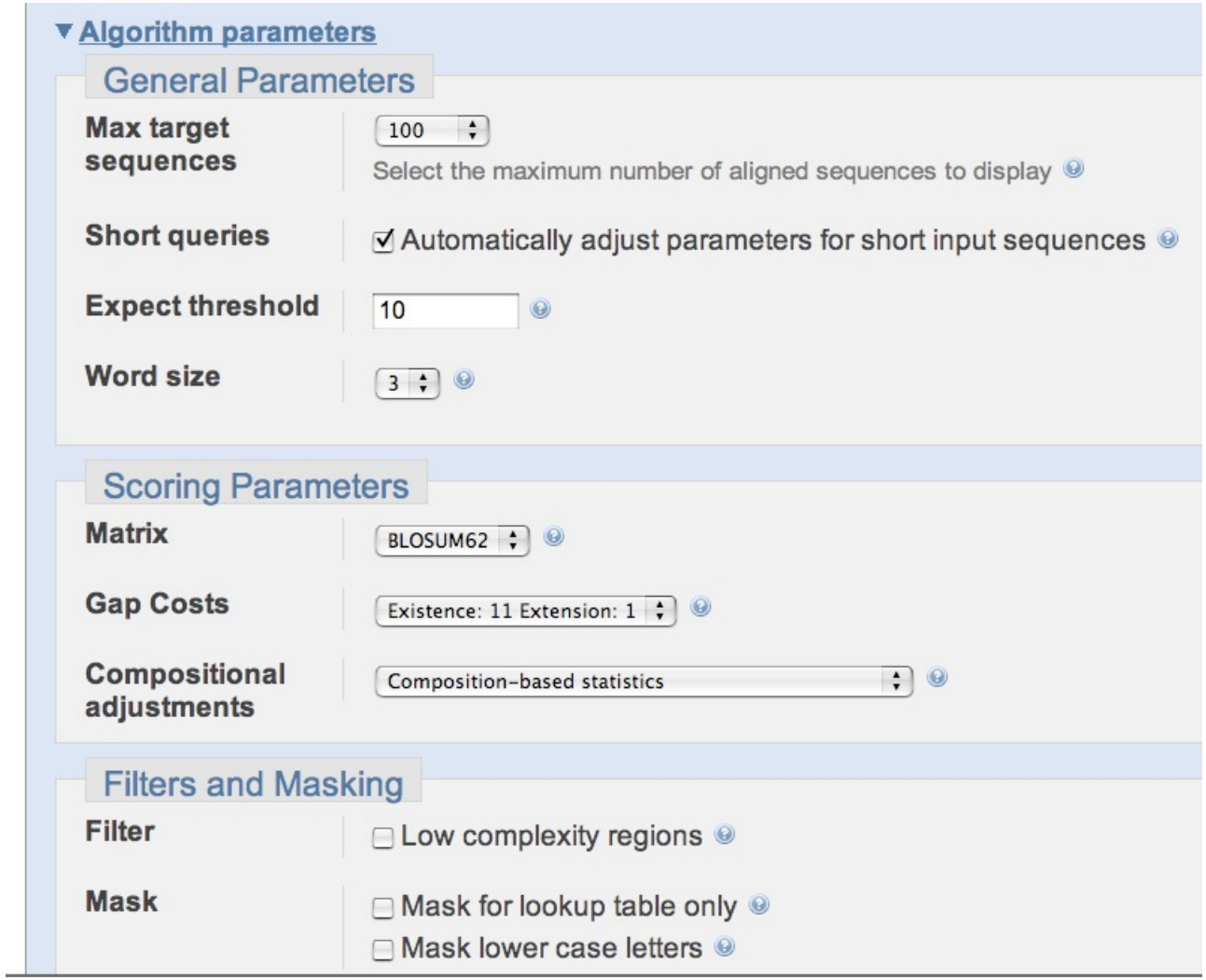


C: For each word match («hit»), extend ungapped alignment in both directions. Stop when S decreases by more than X from the highest value reached by S.

C: For each word match («hit»), extend ungapped alignment in both directions. Stop when S decreases by more than X from the highest value reached by S.



Looking inside BLAST



Looking inside BLAST

heuristic methods of sequence alignment

FastA = fast Alignment

D.J. Lipman and W.W.R. Pearson. Rapid and sensitive protein similarity searches. *Science*, 227:1435–1441, 1985.

W.R. Pearson and D.J. Lipman. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, 85:2444–2448, 1988.

BLAST = Basic Local Alignment Search Tool

S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, 1990.

[a good introduction to these methods:
Frédérique Galisson, The fasta and blast programs, 2002]

Looking inside FastA

EMBL-EBI | EB-eye Search | All Databases | Enter Text Here | Go | Reset | Advanced Search | Give us feedback

Databases | Tools | EBI Groups | Training | Industry | About Us | Help | Site Index

- Help Index
- General Help
- Formats
- Gaps
- Matrix
- References
- FASTA Help
- MView Help
- VisualFASTA Help

- View all FASTA's at EBI
- FASTA Programmatic Access

- Database Information
  - UniProt
  - UniParc

- Similar Applications
  - FASTA
  - BLAST
  - MPsrch
  - ScanPS
  - SSEARCH

FASTA Related Literature

Search for FASTA related literature in Medline... more

EBI > Tools > Similarity & Homology

FASTA and SSEARCH - Protein Similarity Search

Provides sequence similarity searching against protein databases using the FASTA and SSEARCH programs. **SSEARCH** does a rigorous Smith-Waterman search for similarity between a query sequence and a database. **FASTA** can be very specific when identifying long regions of low similarity especially for highly diverged sequences. You can also conduct sequence similarity searching against nucleotide databases or complete proteome/genome databases using the FASTA programs.

Download Software

PROGRAM	DATABASES	RESULTS	SEARCH TITLE	YOUR EMAIL
FASTA	Protein	interactive	Sequence	

UniProt Knowledgebase
UniProtKB/Swiss-Prot
UniProt Clusters 100%
UniProt Clusters 100% (SEG filter

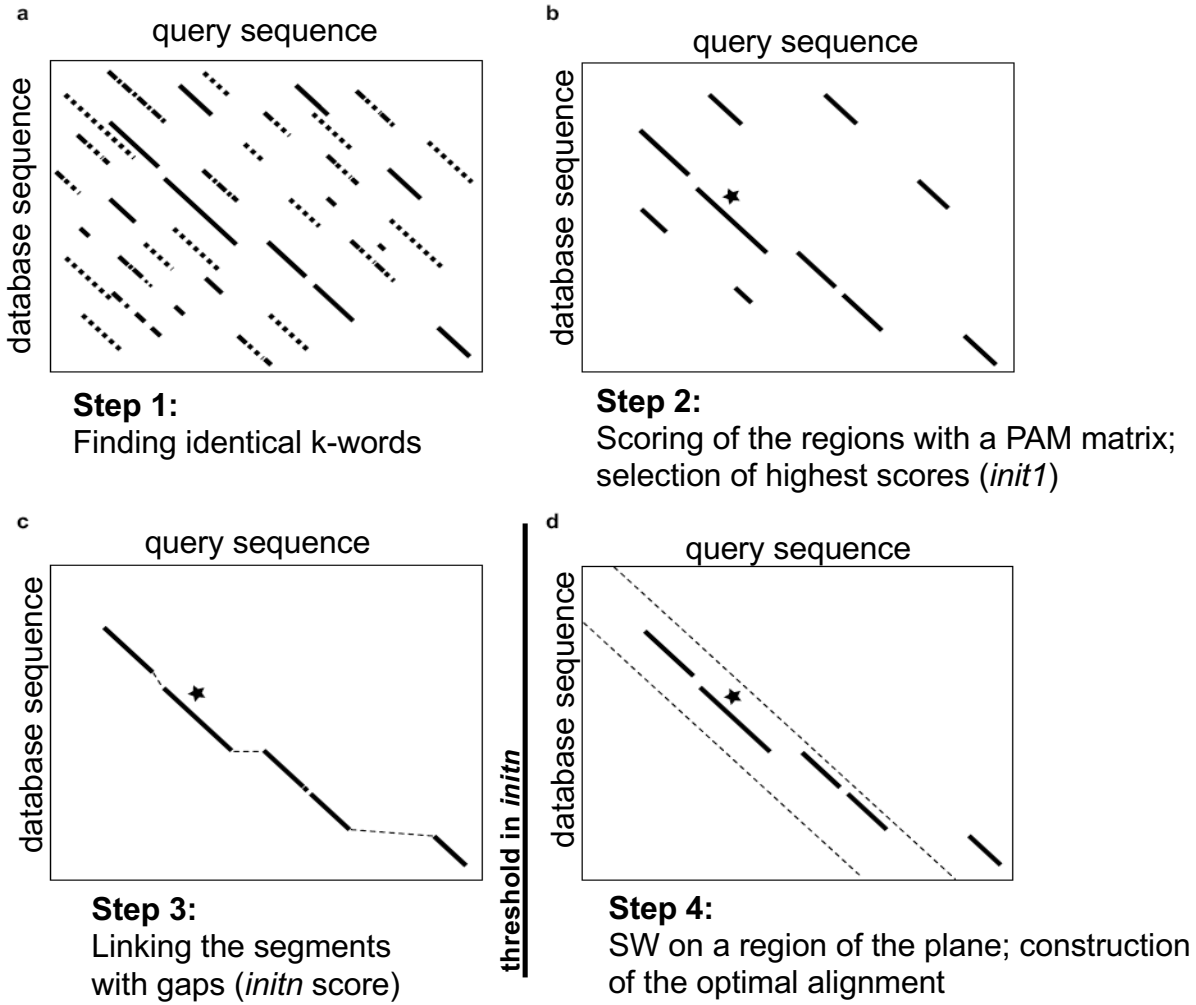
MATRIX	GAP OPEN	GAP EXTEND	KTUP	EXPECTATION UPPER VALUE	EXPECTATION LOWER VALUE
BLOSUM50	-10	-2	2	10.0	default

DNA STRAND	HISTOGRAM	MOLECULE TYPE
none	no	Protein

SCORES	ALIGNMENTS	SEQUENCE RANGE	DATABASE RANGE	FILTER	STATISTICAL ESTIMATES
50	50	START-END	START-END	none	Regress

Enter or Paste a PROTEIN Sequence in any format: Help

Looking inside FastA



Step 1:
Finding identical k-words

Step 2:
Scoring of the regions with a PAM matrix; selection of highest scores (*init1*)

Step 3:
Linking the segments with gaps (*initn* score)

Step 4:
SW on a region of the plane; construction of the optimal alignment

$$
\begin{array}{lll}
(1) & x_i \longleftrightarrow & y_j \\
(2) & x_i \longleftrightarrow & - \\
(3) & - \longleftrightarrow & y_j
\end{array}
$$

$M_{i,j}$: **best score up to (i, j) in the case (1)**

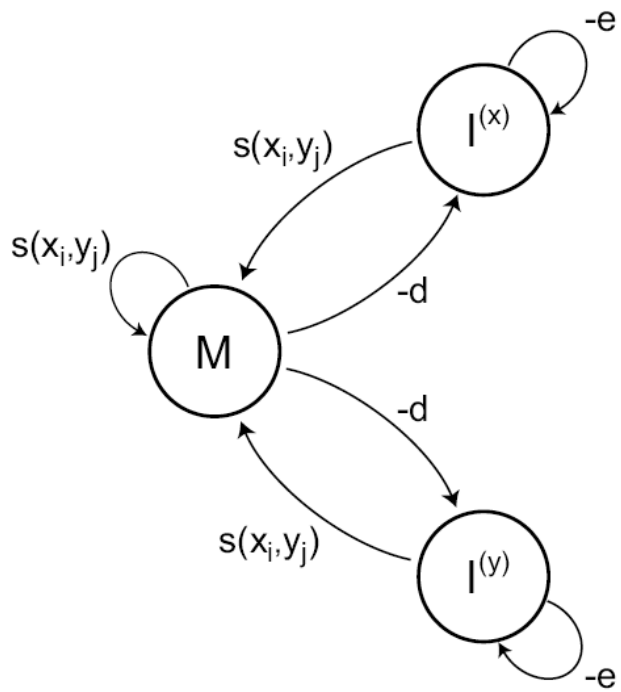
$I^{(x)}_{i,j}$: **best score up to (i, j) in the case (2)**

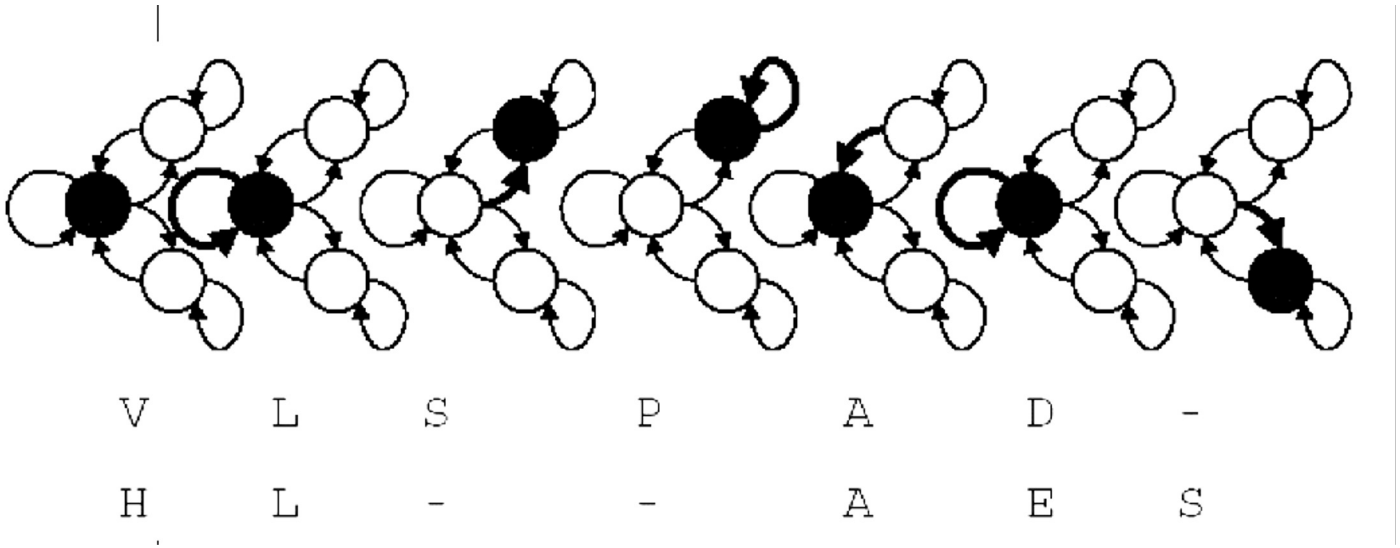
$I^{(y)}_{i,j}$: **best score up to (i, j) in the case (3)**

$$
M_{i,j} = \max \begin{cases} M_{i-1,j-1} + s(x_i, y_j) \\ I^{(x)}_{i-1,j-1} + s(x_i, y_j) \\ I^{(y)}_{i-1,j-1} + s(x_i, y_j) \end{cases}
$$

$$
I^{(x)}_{i,j} = \max \begin{cases} M_{i-1,j} - d \\ I^{(x)}_{i-1,j} - e \end{cases}
$$

$$
I^{(y)}_{i,j} = \max \begin{cases} M_{i,j-1} - d \\ I^{(y)}_{i,j-1} - e \end{cases}
$$





[from Durbin et al. (1998)]