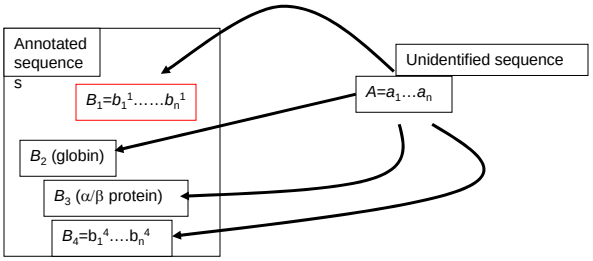
Sequence alignment



Annotation

Typical sequences

- Table 1: A sample of protein sequences. Typical lengths of protein sequences vary from a few tens to a few hundreds. The single letter notation (one character for an amino acid) is used.
- Serine Protease inhibitor

EDICSLPPEV GPCRAGFLKF AYYSELNKCK LFTYGGCQGN ENNFETLQAC XQA

- Amicyanin alpha
- MRALAFAAALAAFSATAALAAGALEAVQEAPAGSTEVKIAKMKFQTPEVR IKAGSAVTWTNTEALPHNVHFKSGPGVEKDVEGPMLRSNQTYSVKFNAPG TYDYICTPHPFMKGKVVVE
- Major Histocompatability Complex (class I)
- MAVMAPRTLV LLLSGALALT QTWAGSHSMR YFSTSVSRPG RGEPRFIAVG YVDDTQFVRF
- DSDAASQRME PRAPWIEQEG PEYWDRNTRN VKAHSQTDRV DLGTLRGYYN QSEDGSHTIQ
- RMYGCDVGSD GRFLRGYQQD AYDGKDYIAL NEDLRSWTAA DMAAEITKRK WEAAHFAEQL
- RAYLEGTCVE WLRRHLENGK ETLQRTDAPK THMTHHAVSD HEAILRCWAL SFYPAEITLT
- WQRDGEDQTQ DTELVETRPA GDGTFQKWAA VVVPSGQEQR YTCHVQHEGL PEPLTLRWEPSSQPTIPIVG IIAGLVLFGA VIAGAVVAAV RWRRKSSDRK GGSYSQAASS DSAQGSDVSLTACKV

Score of amino acid substitutions

$$S(a,b) = \lambda \log\left[P(a,b) / P(a) P(b) \right]$$

Blosum matrix

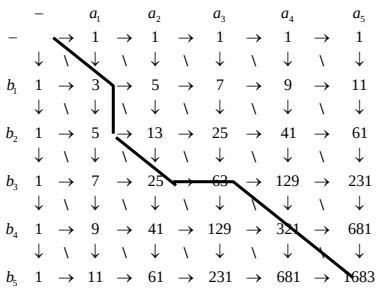
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	
:A	5	-2	-1	-2	-1	-1	-1	0	-2	-1	-2	-1	-1	-3	-1	1	0	-3	-2	0	:A
:R	-2	7	-1	-2	-4	1	0	-3	0	-4	-3	3	-2	-3	-3	-1	-1	-3	-1	-3	:R
:N	-1	-1	7	2	-2	0	0	0	1	-3	-4	0	-2	-4	-2	1	0	-4	-2	-3	:N
:D	-2	-2	2	8	-4	0	2	-1	-1	-4	-4	-1	-4	-5	-1	0	-1	-5	-3	-4	:D
:C	-1	-4	-2	-4	13	-3	-3	-3	-3	-2	-2	-3	-2	-2	-4	-1	-1	-5	-3	-1	:C
:Q	-1	1	0	0	-3	7	2	-2	1	-3	-2	2	0	-4	-1	0	-1	-1	-1	-3	:Q
:E	-1	0	0	2	-3	2	6	-3	0	-4	-3	1	-2	-3	-1	-1	-1	-3	-2	-3	:E
:G	0	-3	0	-1	-3	-2	-3	8	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4	:G
:H	-2	0	1	-1	-3	1	0	-2	10	-4	-3	0	-1	-1	-2	-1	-2	-3	2	-4	:H
:I	-1	-4	-3	-4	-2	-3	-4	-4	-4	5	2	-3	2	0	-3	-3	-1	-3	-1	4	:I
:L	-2	-3	-4	-4	-2	-2	-3	-4	-3	2	5	-3	3	1	-4	-3	-1	-2	-1	1	:L
:K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	6	-2	-4	-1	0	-1	-3	-2	-3	:K
:M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	7	0	-3	-2	-1	-1	0	1	:M
:F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	8	-4	-3	-2	1	4	-1	:F
:P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	10	-1	-1	-4	-3	-3	:P
:S	1	-1	1	0	-1	0	-1	0	-1	-3	-3	0	-2	-3	-1	5	2	-4	-2	-2	:S
:T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	2	5	-3	-2	0	:T
:W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	1	-4	-4	-3	15	2	-3	:W
:Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	2	8	-1	:Y
:V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	-3	-1	5	:V
*	* A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	*

Typical alignment

```
EKLGKLQYSLDYDFQNNQLLVGIIQ-AAEL-PALDMGGTSDPYVKVFLLPD-K-KKKFE
ERRGRIYIQAHID-R--EVLIVVVRDAKNLVP-MDPNGLSDPYVKLKLIPDPKSESKQK

TKVHRKTLNPVFNEQFTFKVPYSELGGKTLVMAVYDFDRFSKHDIIGEFKVPMNTVD-F
TKTIKCSLNPEWNETFRFQLKESDKD-RRLSVEIWDWDLTSRNDFMGSLSFGISELQKA
```

Counting alignments (how to obtain the best alignment??)



Counting non-degenerate alignments

Sequences

$$A = a_1 a_2 \ldots a_m$$

$$B = b_1 b_2 \ldots b_n$$

alignment

$$\{a_1, -\}\{-, b_1\}\{a_2, b_2\}\ldots\{a_m, b_k\}\ldots\{-, b_n\}$$

Counting alignments (continue)

$$alignment \rightarrow label$$

$$\{(a_1 b_1)(a_2 b_2)\} \rightarrow \{b_1 b_2\}$$

$$\{(a_1 -)(a_2 -)(-b_1)(-b_2)\} \rightarrow \{a_1 a_2\}$$

$$\{(-b_1)(-b_2)(a_1 -)(a_2 -)\} \rightarrow \{a_1 a_2\}$$

$$\{(a_1 -)(a_2 -)(a_3 b_1)(-b_2)(-b_3)\} \rightarrow \{a_1 a_2 b_1\}$$

The new labeling can be viewed as the selection of elements (**m** is the length of the sequence **A** and is always the length of the label) from **n+m** pool of objects. The order of the selections does not matter (see second example) since the objects are already labeled with their location in the alignment. The number of alignment is therefore

$$N(n,m) = \binom{n+m}{m}$$

Dynamics programming

$T(n,m)$ The optimal score for aligning a sequence length n against a sequence length m

Assuming that a very kind fellow gave us the optimal scores for the following alignment:

$$T(n-1,m-1))$$
$$T(n-1,m)$$
$$T(n,m-1)$$

can we construct the score ? $T(n,m)$

Yes…

Dynamic programming: Continue

We consider three possibilities to obtain an alignment of n against m amino acids.

Option A: align n-1 against m-1 amino acids score $T(n-1,m-1)$ extend the alignment by a(n)/b(m) with a score S(an,bm)

$$T(n-1,m-1)+S(a_n,b_n)$$

Option B: align n amino acids against m-1 amino acids with a score $T(n,m-1)$ extend the alignment by -/(b(m) with a score g for a gap

$$T(n,m-1)+g$$

Option C: align n-1 amino acids against m amino acids with a score $T(n-1,m)$ Extend the alignment by a(n)/- with a corresponding score of g

$$T(n-1,m)+g$$

To decide which of the three options is optimal we need to compare the score of the three options A, B, C

Dynamic programming: Decision

$$T(n,m) = \max \begin{bmatrix} T(n-1,m-1) + S(a_n, b_m) \\ T(n,m-1) + g \\ T(n-1,m) + g \end{bmatrix}$$

How to start??

$$T(1,-) = T(-,1) = g$$

And continue (for example…) by

$$T(a_1,b_1) = \max \begin{bmatrix} T(a_1,-) + g \\ T(-,b_1) + g \\ T(0,0) + S(a_1,b_1) \end{bmatrix} = \max \begin{bmatrix} 2g \\ 2g \\ S(a_1,b_1) \end{bmatrix}$$

Here we start…

	–		a_1		a_2		a_3		a_4		a_5
–	0	→	g	→	$2g$	→	$3g$	→	$4g$	→	$5g$
		\	↓	\	↓	\	↓	\	↓	\	↓
b_1	g	→		→		→		→		→	
		\	↓	\	↓	\	↓	\	↓	\	↓
b_2	$2g$	→		→		→		→		→	
	↓	\	↓	\	↓	\	↓	\	↓	\	↓
b_3	$3g$	→		→		→		→		→	
	↓	\	↓	\	↓	\	↓	\	↓	\	↓
b_4	$4g$	→		→		→		→		→	
	↓	\	↓	\	↓	\	↓	\	↓	\	↓
b_5	$5g$	→		→		→		→		→	

Pseudo code

- A simple pseudo code to create the dynamic matrix is given below
- /* fill the first (zero) column and the first (zero) row */
- T(0,0) = 0
- Do I=1:n
- T(I,0) = I*g
- End do
- Do I=1:m
- T(0,I) = I*g
- End do
- /* Now fill the rest of the matrix picking the maximum value
- from the three possibilities */
- Do I = 1:n
- Do J = 1:m
- T(I,J) = max[T(I-1,J)+g,
- T(I,J-1)+g,
- T(I-1,J-1)+S(a(i),b(j))]
- End do
- End do

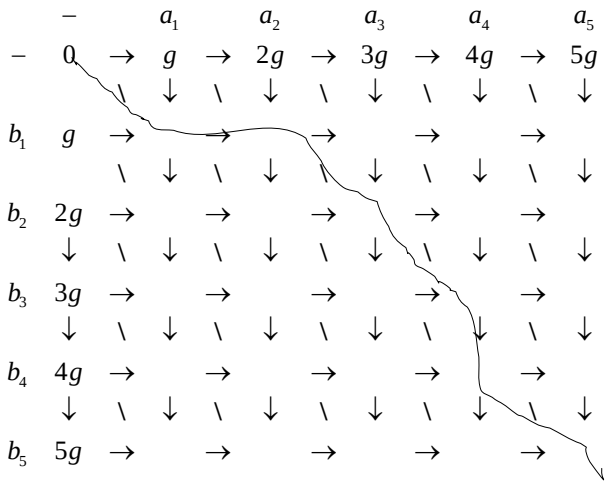
Tracing back an alignment

$$T(n,m) \overset{?}{=} T(n-1,m-1) + S(a_n, b_m)$$

$$T(n,m) \overset{?}{=} T(n-1,m) + g$$

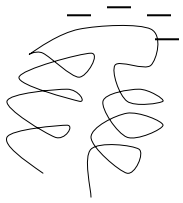
$$T(n,m) \overset{?}{=} T(n,m-1) + g$$

Tracing back an alignment



Can be made more efficient by adding a backward pointers to the table

Gap opening and extension



Influencing the score…
And creating correlations

$$\begin{matrix} a_n & - \\ b_{m-2} & b_{m-1} \end{matrix} \rightarrow \begin{matrix} a_n & - & - \\ b_{m-2} & b_{m-1} & b_m \end{matrix} \qquad g_o$$

$$\begin{matrix} a_{n-1} & a_n \\ b_{m-2} & b_{m-1} \end{matrix} \rightarrow \begin{matrix} a_{n-1} & a_n & - \\ b_{m-2} & b_{m-1} & b_m \end{matrix} \qquad g_e$$

Gap opening and extension Options

$$(1)\begin{matrix} \dots & a_{n-1} \\ \dots & b_{m-1} \end{matrix} \qquad (2)\begin{matrix} \dots & a_{n-1} \\ \dots & - \end{matrix} \qquad (3)\begin{matrix} \dots & - \\ \dots & b_{m-1} \end{matrix}$$

$$(1)T(n,m) \qquad (2)T_{-}(n,m) \qquad (3)T^{-}(n,m)$$

Recursive relations

$$T(n+1,m+1) = \max\begin{Bmatrix} T(n,m) + S(a_n, b_n) \\ T_{\_}(n,m) + S(a_n, b_n) \\ T^{-}(n,m) + S(a_n, b_n) \end{Bmatrix}$$

$$T_{\_}(n+1,m+1) = \max\begin{Bmatrix} T(n,m+1) + g_o \\ T_{\_}(n,m+1) + g_e \\ T^{-}(n,m+1) + g_o \end{Bmatrix}$$

$$T_{\_}(n+1,m+1) = \max\begin{Bmatrix} T(n+1,m) + g_o \\ T_{\_}(n+1,m) + g_o \\ T^{-}(n+1,m) + g_e \end{Bmatrix}$$