BIOINFORMATICS TUTORIAL – DYNAMIC PROGRAMMING

AIM: obtaining the best alignment between two protein sequences by using cumulative matrices

a) Building the alignment between the sequences **APQE** & **LAPD**, by using the scores assigned by the **PAM250** scoring matrix and a penalty value for INDELs (insertions/deletions or GAPs) of **-8**, by considering 2 cases:

1) we do not penalize INDELs at the alignment start, in other words it doesn't matter whether the amino acids at the first position of the two sequences will correspond to each other (i.e. will be aligned) or not. This implies values of the first row and first column of the cumulative matrix all equal to "0";
2) we penalize INDELs at the alignment start, that means we force the condition that the amino acids at the first position of the two sequences will correspond to each other (i.e. will be aligned). This implies penalty scores for the first row and first column of the cumulative matrix that we will set up as multiple of "-8".

b) By using the cumulative matrices built at point a), deriving the best alignment of sequence **APQ** with sequence **LAPD**.

c) Let's suppose that the **APQE** sequence is our query sequence for an homology search in a database that only contains the **LAPD** sequence.

Let's assume that we are using an homology search software such as FASTA, which will shuffle the query sequence and will compare it with the sequences in the database, in order to calculate the random score distribution (from scores obtained for the shuffled sequences, therefore with no evolutionary significance).

In the hypothesis that the software has randomly shuffled the query sequence **APQE** three times, obtaining for instance the sequences PEQA, QPAE, EAPQ and scores for the corresponding alignments with the **LAPD** sequence in the database of –1, 1 e 3, respectively,

1) let's calculate the average and standard deviation for the random score distribution, and
2) let's say whether the alignments we obtained above between APQE and LAPD are meaningful (in other words, whether the obtained alignments scores are high enough for us to hypothisize that the two sequences are evolutionarily related).

Standard deviation formula to be used is:

$$s_{N-1} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}.$$

where N is the number or random alignements, x_i the score obtained for each random alignment and $\overline{x}$ the avergae of scores obtained for the random alignments.

WARNING. The **c)** point of the tutorial is only for educational purposes. Homology searches on such short sequences is generally pointless from a biological point of view.

	G	A	V	L	I	P	S	T	D	E	N	Q	K	R	H	F	Y	W	M	C
G	5																			
A	1	2																		
V	-1	0	4																	
L	-4	-2	2	6																
I	-3	-1	4	2	5															
P	0	1	-1	-3	-2	6														
S	1	1	-1	-3	-1	1	2													
T	0	1	0	-2	0	0	1	3												
D	1	0	-2	-4	-2	-1	0	0	4											
E	0	0	-2	-3	-2	-1	0	0	3	4										
N	0	0	-2	-3	-2	0	1	0	2	1	2									
Q	-1	0	-2	-2	-2	0	-1	-1	2	2	1	4								
K	-2	-1	-2	-3	-2	-1	0	0	0	0	1	1	5							
R	-3	-2	-2	-3	-2	0	0	-1	-1	-1	0	1	3	6						
H	-2	-1	-2	-2	-2	0	-1	-1	1	1	2	3	0	2	6					
F	-5	-3	-1	2	1	-5	-3	-3	-6	-5	-3	-5	-5	-4	-2	9				
Y	-5	-3	-2	-1	-1	-5	-3	-3	-4	-4	-2	-4	-4	-4	0	7	10			
W	-7	-6	-6	-2	-5	-6	-2	-5	-7	-7	-4	-5	-3	-2	-3	0	0	17		
M	-3	-1	2	4	2	-2	-2	-1	-3	-2	-2	-1	0	0	-2	0	-2	-4	6	
C	-3	-2	-2	-6	-2	-3	0	-2	-5	-5	-4	-5	-5	-4	-3	-4	0	-8	-5	12

PAM 250