T-61.5120 Computational Genomics (5 credit points), Exam 25th October 2011

Solve **All 5** problems. Use of calculator is allowed.

1. Definitions and concepts

   (a) Explain shortly the following terms

   - Orthologs (1.5p)
   - Paralogs (1.5p)
   - Analogs (1.5p)
   - Xenologs (1.5p)
   - PSI-BLAST (3p)
   - Base covariation analysis (3p)

2. Dot Matrix

   (a) Using the Dot matrix method, produce examples of the following four cases. Use sequences of length $\sim$ 10 nucleotides

   - Similar subsequences, (1p)
   - Insertion/deletion (1p)
   - Repeat of subsequence (1p)
   - Inverted subsequence. (1p)

   (b) Describe filtering in the Dot Matrix concept (4p)

   (c) Describe how the Dot Matrix concept can be applied to RNA secondary structure analysis (4p)

3. Use the Needleman-Wunsch dynamic programming algorithm to construct a global alignment of the amino acid sequences "STETES" and "TEST". Use the BLOSUM62 substitution matrix (on the last page) and gap penalty -10 (both for opening a gap and for expanding a gap).

   (a) Fill the initializied score and trace-back matrix. (3p +3p)

Score Matrix

		S	T	E	T	E	S
	0	-10	-20	-30	-40	-50	-60
T	-10	1					
E	-20						
S	-30						
T	-40						

Trace-back Matrix

		S	T	E	T	E	S
		←	←	←	←	←	←
T	↑	↖					
E	↑						
S	↑						
T	↑						

(b) Write down the optimal global alignment, when gaps in the end **are assigned** gap penalties. (3p)

(c) Write down the optimal global alignment, when gaps in the end **are NOT assigned** penalties. (3p)

4. Assume a transition probability matrix $P1$, where $p_{i,j}$ (the element on the ith row and $j$the coloumn) is the probability that nucleotide i is substituted with nucleotide j during 1 million years.

(a) How can we calculate the probabilities over a longer time period, say k million years. (Note: this method is used in calculating PAMk matrices) (3p)

(b) What assumption do we have to make. (3p)

(c) Given the following transition probability matrix $P1$, what is the probability that A is substituted with T in 3 million years. (6p)

$P1 =$

	A	C	G	T
A	0.97	0.01	0.01	0.01
C	0.01	0.97	0.01	0.01
G	0.01	0.01	0.97	0.01
T	0.01	0.01	0.01	0.97

5. The score S of a local alignment follows the extreme value distribution. Thus the probability that the score S exceeds x is

$$P(S \geq x) = 1 - \exp\left(-Kmn\, e^{(-\lambda x)}\right)$$

where m and n are the length of the sequences and K and λ are constants.

Assume $m = n = 250$, substitution matrix BLOSUM62 (on the last page) and $K = 0.1$ and $\lambda = 0.3$. Note: The Blosum62 matrix values are in half bits and the parameter values (K and λ) have been given assuming the score is in half bits.

(a) Calculate the alignment scores for the local alignments (3p)

i)

$$FWLEVEGF$$
$$FWLDVQGF$$

ii)

$$FWLEVEGFE$$
$$FWLDVQGFQ$$

iii)

$$FWLEVEGFVQ$$
$$FWLDVQGFVE$$

(b) For each alignment what is the probability to get by chance a score as high as this? Are the alignment scores significant? (3p)

(c) Assume the above alignments have resulted from a database search, including 10 sequences of length 250. Are the alignments significant? (6p)

The BLOSUM62 substitution matrix

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