Examination Bioinformatics

Linköpings Universitet, IDA, Statistik

Course code and name:	732A51 Bioinformatics
Date:	2020/01/13, 8–12
Assisting teacher:	Krzysztof Bartoszek
Allowed aids:	The help material is included in the zip file **exam_help_material_732A51.zip**.
Grades:	A= $[18-20]$ points B= $[16-18)$ points C= $[14-16)$ points D= $[12-14)$ points E= $[10-12)$ points F= $[0-10)$ points
Instructions:	Provide a detailed report that includes plots, conclusions and interpretations. Give motivated answers to the questions. If an answer is not motivated, the points are reduced. Provide all necessary codes in an appendix. If you are asked to do plots, then make sure that they are informative, have correctly labelled axes, informative axes limits and are correctly described. Points may be deducted for poorly done graphs. Name your digital part solution files as: **[your exam account]_[own file description].[format]** If you have problems with creating a pdf you may submit your solutions in text files with unambiguous references to graphics and code that are saved in separate files There are **THREE** assignments (with sub–questions) to solve. Include all code that was used to obtain your answers in your solution files. Make sure it is clear which code section corresponds to which question. Your code should be complete and readable, possible to run by copying directly into a script. Comment directly in the code whenever something needs to be explained or discussed.

Problem 1 (7p)

Define an HMM with the following parameters:
Three hidden states, S_1, S_2, S_3,
alphabet $A = \{C, G\}$,

transition matrix for hidden layer $P = \begin{bmatrix} 3/4 & 1/4 \\ 0 & 1 \end{bmatrix}$, initial distribution $\pi = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$,

emission probabilities by hidden states ($b_i(j)$ is the probability that a hidden state that equals i will emit observed state j):
$b_1(C) = 1/4$, $b_1(G) = 3/4$,
$b_2(C) = 1/2$, $b_2(G) = 1/2$.

What are all possible state sequences for the following observed sequences $\mathcal{O}$, and what is the probability of each observed sequence?
(a) $\mathcal{O} = C, C, G$
(b) $\mathcal{O} = G, G, C$

Problem 2 (6p)

You are given two DNA sequences ACCCAT and ATCGTA. Consider the Hamming distance and calculate the distance between the two sequences.

It is known that the genome can tolerate better a mutation purine to purine or pyrimidine to pyrimidine (transitions) than between purines and pyrimidines (transversions). Modify the Hamming distance to take this into account and recalculate the distance between the two sequences.

Problem 3 (7p)

You have observed a bivariate trait (X, Y) amongst five species, $s1$, $s2$, $s3$, $s4$ and $s5$. The measurements are provided in the file `tm.csv`. Based on this data propose a phylogeny relating these species. The phylogeny should also contain some proposals for the branch lengths. Justify and explain your proposed tree.