Examination Bioinformatics

Linköpings Universitet, IDA, Statistik

Course:	732A51 Bioinformatics
Date:	2021/01/13, 8–12
Teacher:	Krzysztof Bartoszek
Provided 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 id]_[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. If you also need to provide some hand–written derivations please number each page according to the pattern: Question number . page in question number i.e. Q1.1, Q1.2, Q1.3,. . ., Q2.1, Q2.2, . . ., Q3.1, ... Scan/take photos of such derivations preferably into a single pdf file but if this is not possible multiple pdf or .bmp/.jpg/.png files are fine. Please do not use other formats for scanned/photographed solutions. Please submit all your solutions via LISAM or e–mail. If emailing, please email them to **BOTH** krzysztof.bartoszek@liu.se and KB_LiU_exam@protonmail.ch . During the exam you may ask the examiner questions by emailing them to KB_LiU_exam@protonmail.ch **ONLY**. Other exam procedures in LISAM.

Problem 1 (7p)

There are three models described below for a signal of length five: i.i.d., weight matrix and first–order Markov chain. For each of the sequences $AGTCTGCC$ and $CGCGTATA$ find the probability of the sequence given the model, for each of the three models (so your answer should consist of six probabilities).
(i) i.i.d. The probabilities of the four nucleotides are $P(A) = 0.2$, $P(C) = 0.4$, $P(G) = 0.3$ and $P(T) = 0.1$.
(ii) Weight Matrix. The weight matrix (for the nucleotide ordering: A, C, G, T) is

$$\begin{bmatrix} 0.7 & 0.4 & 0.25 & 0.2 & 0.25 & 0.25 & 0.8 & 0 \\ 0.1 & 0.3 & 0.05 & 0.1 & 0.25 & 0.25 & 0.05 & 0.4 \\ 0.1 & 0.1 & 0.65 & 0.6 & 0.15 & 0.25 & 0.05 & 0.4 \\ 0.1 & 0.2 & 0.05 & 0.1 & 0.35 & 0.25 & 0.1 & 0.2 \end{bmatrix}.$$

(iii) First–Order Markov chain. The initial distribution is $P(A) = 0.25$, $P(C) = 0.25$, $P(G) = 0.25$ and $P(T) = 0.25$. The transition matrix (for the nucleotide ordering: A, C, G, T) is

$$\begin{bmatrix} 0.2 & 0.6 & 0.1 & 0.1 \\ 0.5 & 0.25 & 0.1 & 0.15 \\ 0.05 & 0.95 & 0 & 0 \\ 0.4 & 0.2 & 0.25 & 0.15 \end{bmatrix}.$$

Problem 2 (7p)

Manually, without using any software, find the best fit of the DNA sequence $CCGC$ inside the sequence $ATCCTGCA$, scoring $+2$ for a match, -2 for a mismatch and -1 for a gap.

Problem 3 (6p)

In Fig. 1 you have a phylogeny provided. Encode this tree *manually* in the `R` **ape** package's `phylo` format. Afterwords write code that plots your `phylo` object exactly (with the exception that the branch lengths need not be written on the plot) as it is in Fig. 1 (you will need to read the documentation of `?plot.phylo`).

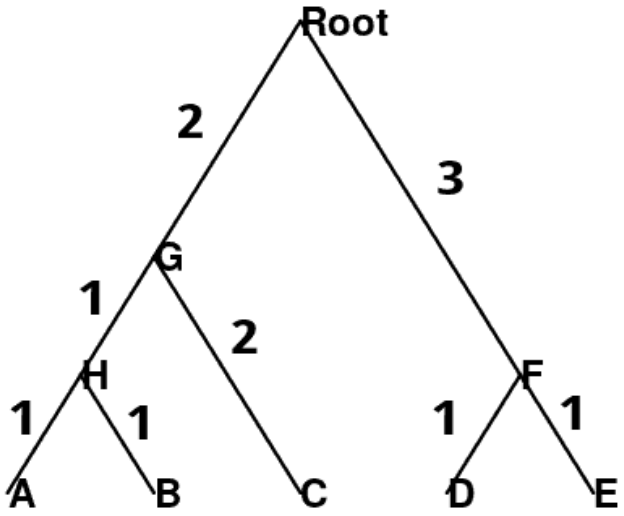

Figure 1: Phylogeny for Problem 3. The numbers next to the branches are the branch lengths. Next to each node and tip, the label of the node/tip is provided.