Examination Bioinformatics

Linköpings Universitet, IDA, Statistik

Course code and name:	732A51 Bioinformatics
Date:	2019/02/13, 8–12
Assisting teacher:	Krzysztof Bartoszek
Allowed aids:	The help material is included in the zip file **exam_help_material.zip**. In the zip **exam_material.zip** file you will also find data files and help scripts for your exam.

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]** 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 (6p)

Define an HMM with the following parameters:
Three hidden states, S_1, S_2, S_3,
alphabet $A = \{1, 2\}$,

transition matrix for hidden layer $P = \begin{bmatrix} 1/2 & 1/2 \\ 1 & 0 \end{bmatrix}$, initial distribution $\pi = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$,

emission probabilities by hidden states ($b_i(j)$ is the probability that a hidden state that equals i will emit observed state j):
$b_1(1) = 1/2$, $b_1(2) = 1/2$,
$b_2(1) = 1$, $b_2(2) = 0$.

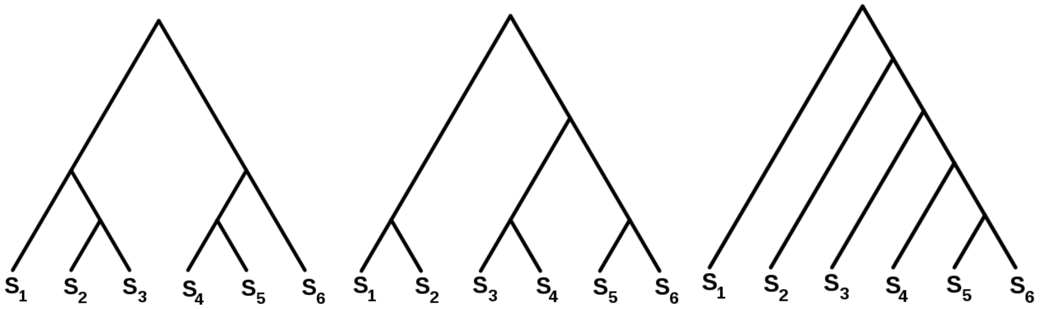
What are all possible state sequences for the following observed sequences $\mathcal{O}$, and what is the probability of each observed sequence?
(a) $\mathcal{O} = 2, 1, 1$
(b) $\mathcal{O} = 1, 2, 2$

Problem 2 (7p)

In the figure below you can find three candidate, labelled, without branch lengths phylogenies for the clade of six bird species $\{S_1, S_2, S_3, S_4, S_5, S_6\}$. It was observed that some of these species lay their eggs in nests on the ground, while others in nests on trees:

S_1	S_2	S_3	S_4	S_5	S_6
ground	ground	ground	tree	tree	tree .

For each tree provide the *most parsimonious* assignment(s) of nesting strategy to the internal nodes of the tree and also provide the parsimony score. Provide a justification why your provided assignment(s) are the most parsimonious.



Problem 3 (7p)

In the file `BirdSpeicesDNA.fasta` you are provided with the DNA sequences of the six birds described in Problem 2. Read the sequences into `R` using the code in `ReadBirdSpeicesDNA.R`. Construct a phylogenetic tree of the species. Interpret what the estimated phylogeny is telling you about the relationships between the six species. Do you see any (there might not be any) connection with the nesting strategies from the table in Problem 2? Justify.
TIP: You could look at the functions: `ape::dist.dna(tree,model="raw")` and `ape::nj()`.