Examination Bioinformatics

Linköpings Universitet, IDA, Statistik

Course:	732A51 Bioinformatics
Date:	2024/02/12, 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 exam account id]_[own file description]** 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, ...

Problem 1 (6p)

In the file `S_protein.fasta` you will find part of the DNA of the SARS–CoV–2 isolate Wuhan–Hu–1 genome, related to the S surface glycoprotein (so–called spike protein) deposed under Gene ID: 43740568. The goal of this exercise is to find the protein code of the S surface glycoprotein, i.e., correctly translate the nucleotide sequence into the protein sequence. Your final provided sequence (in a fasta file) should be only the sequence of the protein without any extras. Explain what you did, why and motivate why you think your translation is correct.

TIP: You might want to look at **ape**'s functions: `ape::complement()`, `ape::read.FASTA()`, `ape::trans()` and the options they have.

Problem 2 (7p)

You are given two DNA sequences GGCGT and GGTCCT. Choose some distance function (for which these two sequences are valid input) and calculate the distance between these two sequences under the chosen function. Perform (manually) a global alignment between the two sequences. Choose all necessary parameters yourself (however they should be non–trivial, i.e., **NOT** 0, and furthermore meaningful). Do not forget to report the dynamic programming matrix and how one obtains the optimal solution from it. Explain what an alignment is from a biological point of view.

Problem 3 (7p)

In the figure below you can find three candidate, labelled, without branch lengths phylogenies for the clade of six lizard species $\{S_1, S_2, S_3, S_4, S_5, S_6\}$. It was observed that some of these species possess legs while others are limbless:

S_1	S_2	S_3	S_4	S_5	S_6
legs	legs	legless	legless	legs	legs .

For each tree provide the *most parsimonious* assignment(s) of limb possession 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.