Examination Bioinformatics

Linköpings Universitet, IDA, Statistik

Course:	732A51 Bioinformatics
Date:	2021/02/08, 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)

A) **(5p)** The overall base composition of a new (in 2006) bacterial strain of the *Chlorobium* genus is G=C= 46.7% and A=T= 53.3%. In a random (all positions independent and identically distributed) sequence of 2,572,079 bp nucleotides with these proportions, what is the expected number of occurrences of the sequences CTAG, TCAG and GCTA?
B) **(2p)** The number 2,572,079 bp is A) is not taken randomly. Find how it is related to the problem. You will have to do an internet search (you are permitted for this question to search anywhere online during the exam). A starting point for the online search can be the article describing the new strain, `VoglGlaeserPfannesWannerOvermann_Chlorobium_2006.pdf`. Do **not** read it in detail during the exam, think what can be the key piece of information for finding the origin of this number. Write (very briefly) how this number is related to the bacterial strain in question.

Problem 2 (8p)

For each matrix A) PAM250 (file `PAM250.png`) and B) BLOSUM62 (file `BLOSUM62.png`) answer the following

1. What are/is the most probable substitution? If there are too many to write out, then provide the value in the table and three examples.
2. What are/is the least probable substitution? If there are too many to write out, then provide the value in the table and three examples.
3. Which substitution is more probable W→F or H→R?
4. Which substitution is more probable M→Q or K→T?

TIP: The numbers in the matrices are some transformations of the original observed data. Exactly how this transformation was done and what do the actual numbers mean is not that important for this question.

Problem 3 (5p)

A) **(3p)** You are given, in Newick format, the following two rooted trees with tip labels

$$T_1 = (A, (((E, D), C), ((G, F), (B, H))))$$

and

$$T_2 = ((((H, B), (F, G)), (C, (D, E))), A).$$

Are they identical in topology?
B) **(2p)** Draw all possible rooted, without tip labels, trees with 3, 4 and 5 tips.