Examination Bioinformatics

Linköpings Universitet, IDA, Statistik

Course:	732A51 Bioinformatics
Date:	2022/01/12, 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 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.

Problem 1 (7p)

A) **(3p)** Describe the difference between the Hamming and Levenshtein (edit) distances. Illustrate this with one example
B) **(4p)** In the lecture we described various dynamic programming algorithms to perform sequence alignments. Explain why there is a need for so many algorithms for this problem in sequence analysis. What is the key mathematical property differentiating these algorithms?

Problem 2 (9p)

You are given the following matrix with distances between four sequences A, B, C and D

$\mathbf{D} =$

	A	B	C	D
A	0	0.3	0.5	0.6
B		0	0.6	0.5
C			0	0.9
D				0

A) **(4p)** Using the Neighbour Joining algorithm manually create a tree joining these sequences. Present all your calculations, plot the resulting tree. Do not forget about presenting branch lengths.
B) **(2p)** Propose how you would root this tree. Could a fifth sequence be useful in doing this? If yes, what property could make it very helpful?
C) **(3p)** Use the `ape::nj()` function to create a tree from **D**. Plot this tree and present the plot. Report on the branch lengths inside this tree. Does it agree with your tree from A)? Do the branch lengths agree?

Problem 3 (4p)

Read in the file `MA.RData`. In it you will find a matrix `X` that contains points for the **MA** plot from a microarray experiment. The first column of `X` (called `x`) are the "A" values ($(\log_2\text{CY5}+\log_2\text{CY3})/2$) and the second column of `X` (called `y`) are the "M" values ($\log_2(\text{CY5}/\text{CY3})$). Make a plot of `X`. Is this how you would expect a MA plot to look like? If not propose and implement a method to correct it. Plot the corrected data points. Which, if any, of the data points would you consider to be candidates for being differentially expressed? What method do you use to decide, describe and implement it. Plot these candidate points in a different colour on your plot.
TIP: No specialized microarray software is required for this task.