Examination Bioinformatics

Linköpings Universitet, IDA, Statistik

Course code and name:	732A51 Bioinformatics
Date:	2019/08/30, 8–12
Assisting teacher:	Krzysztof Bartoszek
Allowed aids:	The help material is included in the zip file **exam_help_material.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]_[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)

Visualize the similarity between the sequences DOROTHYHODGKIN and DOROTHYMARYCROWFOOTHODGKIN by means of a dotplot. Consider the package `seqinr`. You should notice that there are a lot of matches between fragments of the two sequences that do not correspond to each other. Where are they located on the dotplot? What do you need to change in your visualization code to remove these? Draw the dotplot again but now without the additional, wrong matches. You might find the function `strsplit()` helpful.

Dorothy Hodgkin (1910–1994) was a British chemist who developed protein crystallography, for which she won the Nobel Prize in Chemistry in 1964 (`https://en.wikipedia.org/wiki/Dorothy_Hodgkin`).

Problem 2 (6p)

Comparisons of DNA sequences of homologous chromosomes in different people show that, on average, one of every 700 base pairs of noncoding DNA is different. About 95% of the human genome is noncoding. Estimate the number of polymorphisms (i.e. positions in the DNA that differ between people) in the human genome to give some idea of the number of potential DNA markers (i.e. positions in the DNA that differentiate between people).

Problem 3 (7p)

In the overall yeast transcriptional regulatory network the number of incoming connections to target genes (i.e. the number of other genes or RNA fragments influencing that gene's expression levels) follows a geometric distribution. That is the probability that a gene is controlled by k transcriptional regulators (other genes or RNA fragments) equals $(1 - e^{-\alpha})e^{-\alpha k}$ for $k = 0, 1, 2, \ldots$.
a) (1p) What is the ratio of the expected number of target genes with four input connections to the expected number with two input connections when $\alpha = 0.8$?
b) (3p) Write a short program (or do analytically, but coding is probably easier) to find the maximum value of k for which at least 1% of the genes are controlled by at least k transcriptional factors for $\alpha = 0.8$.
c) (3p) Determine the expected number of transcriptional factors in terms of α. Provide a plot of this relationship. You can do this part through simulations and/or analytically and/or numerically, your choice. In all cases you still need to provide a plot.