Examination Bioinformatics

Linköpings Universitet, IDA, Statistik

Course:	732A51 Bioinformatics
Date:	2023/01/10, 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 (7p)

We consider a HMM with hidden states, S_1, S_2, on the alphabet $A = \{1, 2\}$,

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

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

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 (6p)

Calculate the dynamic programming matrix for aligning the DNA sequences $CGTTC$ and $CGATC$, scoring $+1$ for a match, 1 for a mismatch, and with a linear gap penalty of $d = 1$. Provide all optimal global alignments. What are their scores?

Problem 3 (7p)

Three simple models to describe a five nucleotide DNA sequence are an i.i.d. model, a position weight matrix and a first–order Markov chain. We consider the following models.

1. i.i.d. model with probabilities of the four nucleotides equalling $P(A) = 0.1$, $P(C) = 0.4$, $P(G) = 0.4$ and $P(T) = 0.1$,

2. position weight matrix model (for the nucleotide ordering: A, C, G, T)

$$
\begin{bmatrix}
0.4 & 0.2 & 0.2 & 0.1 & 0.2 \\
0.1 & 0.3 & 0.4 & 0.2 & 0.5 \\
0.1 & 0.3 & 0.3 & 0.5 & 0.1 \\
0.4 & 0.2 & 0.1 & 0.2 & 0.2
\end{bmatrix},
$$

3. first–order Markov Chain model with initial distribution $P(A) = 0.4$, $P(C) = 0.1$, $P(G) = 0.1$ and $P(T) = 0.4$ and transition matrix (for the nucleotide ordering: A, C, G, T) is

$$
\begin{bmatrix}
0.1 & 0.4 & 0.4 & 0.1 \\
0.1 & 0.2 & 0.6 & 0.1 \\
0.1 & 0.4 & 0.3 & 0.2 \\
0.2 & 0.4 & 0.3 & 0.2
\end{bmatrix}.
$$

Consider the sequence $TCGCG$. What is its probability under models 1—3? Can you say anything about its possible location in the genome?