Exercise sheet 1 in DNA sequence analysis

Krzysztof Bartoszek
732A51 Bioinformatics
Department of Computer and Information Science, Linköping University

28 XI 2024 (R35)

The exercises here are taken from the below textbooks

BE M. Borodovsky, S. Ekisheva., 2006, Problems and Solutions in Biological Sequence Analysis, Cambridge University Press.

EG W. J. Ewens, G. R. Grant., 2005, Statistical Methods in Bioinformatics, 2nd ed. Springer.

MM J. Momand, A. McCurdy., 2017, Concepts in Bioinformatics and Genomics, Oxford University Press.

Notes
codon: DNA triplet coding for an amino acid
nt: nucleotides, 3nt means a sequence of 3 nucleotides
oligonucleotide: a short DNA/RNA molecule (sequence from our perspective)
polymerase chain reaction: a procedure used to amplify (make multiple copies of) a specific DNA fragment
Prokaryota: single cell organisms, that lack membrane–bound organelle. They are divided into Archaea and Bacteria.
restriction enzyme: a macromolecule (large molecule/protein) that cleaves (cuts) DNA into fragments at *restriction sites* (specific places in DNA sequence)
Other biological terms are explained in the text of the exercises.

Exercise 1 (BE: Problem 1.6)
In the herpesvirus genome, nucleotides C, G, A, and T occur with frequencies 35/100, 35/100, 15/100, and 15/100, respectively. Assuming the independence model for the genome, what is the probability that a randomly selected 15nt long DNA fragment contains eight C's or G's and seven A's or T's?

Exercise 2 (BE: Problem 1.7)
A *DNA primer* used in the polymerase chain reaction is a one–strand DNA fragment designed to bind (to hybridize) to one of the strands of a target DNA molecule. It was observed that primers can hybridize not only to their perfect complements, but also to DNA fragments of the same length having one or two mismatching nucleotides. If the genomic DNA is "sufficiently

long", how many different DNA sequences may bind to an eight nucleotide long primer? The notion of "sufficient length" implies that all possible oligonucleotides of length 8 are present in the target genomic DNA.

Exercise 3 (EG: Problem 5.5)
There are three models described below for a signal of length five: i.i.d., weight matrix and first–order Markov. For each of the sequences $CCGAT$ and $CATAT$ find the probability of the sequence given the model, for each of the three models (so your answer should consist of six probabilities).
(i) i.i.d. The probabilities of the four nucleotides are $P(A) = 0.2$, $P(C) = 0.1$, $P(G) = 0.1$ and $P(T) = 0.6$.
(ii) Weight Matrix. The weight matrix (for the nucleotide ordering: A, C, G, T) is

$$\begin{bmatrix} 0.2 & 0.3 & 0.2 & 0.1 & 0.1 \\ 0.1 & 0.2 & 0.15 & 0.6 & 0.6 \\ 0.3 & 0.4 & 0.6 & 0.1 & 0.15 \\ 0.4 & 0.1 & 0.05 & 0.2 & 0.15 \end{bmatrix}.$$

(iii) First–Order Markov chain. The initial distribution is $P(A) = 0.2$, $P(C) = 0.1$, $P(G) = 0.1$ and $P(T) = 0.6$. The transition matrix (for the nucleotide ordering: A, C, G, T) is

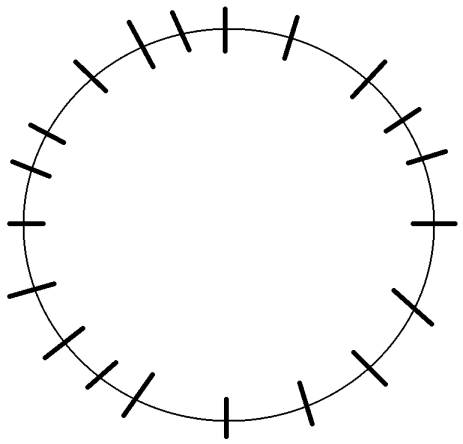
$$\begin{bmatrix} 0.1 & 0.8 & 0.05 & 0.05 \\ 0.35 & 0.1 & 0.1 & 0.45 \\ 0.3 & 0.2 & 0.2 & 0.3 \\ 0.6 & 0.1 & 0.25 & 0.05 \end{bmatrix}.$$

Exercise 4 (BE: Problem 1.8)
A DNA sequencing reaction is performed with an error rate of 10%, thus a given nucleotide is wrongly identified with probability 0.1. To minimize the error rate, DNA is sequenced by $n = 3$ independent reactions, the newly sequenced fragments are aligned, and the nucleotides are identified by the following majority rule. The type of nucleotide at a particular position is identified as α, $\alpha \in \{T, C, A, G\}$, if more nucleotides of type α are aligned in this position than all other types combined. If at an alignment position no nucleotide type appears more than $n/2$ times, the type of nucleotide is not identified (type N). What is the expected percentage of
(a) correctly and
(b) incorrectly identified nucleotides?
(c) What is the probability that at a particular site identification is impossible?
(d) How does the result of (a) change if $n = 5$; what about for $n = 7$? Assume that there are only substitution type errors (no insertions or deletions) with no bias to a particular nucleotide type.

Exercise 5 (BE: Problem 1.9)
Due to redundancy of genetic code, a sequence of amino acids could be encoded by several DNA sequences. For a given ten amino acid long protein fragment, what are the lower and upper bounds for the number of possible DNA sequences that could carry code for this protein fragment?

Exercise 6 (BE: Problem 1.10)
Life forms from planet XYZ were discovered to have a DNA and protein basis with proteins consisting of twenty amino acids. By analysis of the protein composition, it was determined that the average frequencies of all amino acids excluding Met and Trp were equal to 1/19, while the frequencies of Met and Trp were equal to 1/38. Given the high temperature on the XYZ surface, it was speculated that the DNA has an extremely high $G + C$ content. What could be the highest average $G+C$ content of protein–coding regions (given the average amino acid composition as stated above) if the standard (the same as on planet Earth) genetic code is used to encode XYZ proteins?

Exercise 7 (BE: Problem 1.11)
A restriction enzyme is cutting DNA at a palindromic site 6nt long. Determine the probability that a circular chromosome, a double–stranded DNA molecule of length $L = 84000$nt, will be cut by the restriction enzyme into exactly twenty fragments. It is assumed that the DNA sequence is described by the independence model with equal probabilities of nucleotides T, C, A, and G. Hint: use the Poisson distribution.
Assume that there are no edge effects. Furthermore, ignore any mutual dependencies between the restriction sites (where the DNA enzyme cuts). By this we mean that if a restriction site begins at position i, this does not make any difference for an overlapping restriction site to begin at position j ($|i - j| \leq 6$), e.g. the sequence $AAAAAA$ satisfies this. Since the genome is circular we need 20 cuts (twenty restriction sites to cut the genome into twenty pieces). See graphic below, the restriction sites and cuts are marked with bold lines on the circular genome.



Exercise 8 (BE: Problem 1.12)
Determine the average length of the restriction fragments produced by the six–cutter (length of recognition site is 6nt) restriction enzyme $SmaI$ with the restriction site $CCCGGG$. Consider
(a) a genome with a $G + C$ content of 70% and
(b) a genome with a $G + C$ content of 30%.

It is assumed in both (a) and (b) that the genomic sequence can be represented by the independence model with probabilities of nucleotides such that $P(G) = P(C)$, $P(A) = P(T)$. Note that enzyme $SmaI$ cuts the double strand of DNA in the middle of site $CCCGGG$.

Just as in Exercise 7 assume that there are no edge effects. Furthermore, ignore any mutual dependencies between the restriction sites (where the DNA enzyme cuts). By this we mean that if a restriction site begins at position i, this does not make any difference for an overlapping restriction site to begin at position j ($|i-j| \leq 6$), e.g. the sequence $AAAAAA$ satisfies this.

Exercise 9 (BE: Problem 1.14)
A prokaryotic protein–coding gene normally consists of an uninterrupted sequence of nucleotide triplets, codons. This sequence starts with a specific start codon (ATG is most frequent) and ends with one of the three stop codons: TAA, TAG, TGA. A sequence with such a structure is called an "open reading frame" (ORF). However, not every ORF found in prokaryotic genomic DNA is a functional gene. Assuming that ATG is the only possible start codon, what is the length distribution of ORFs that occur by chance? Consider an independence model with equal probabilities of four nucleotide types. Remember that there are three possible stop codons.

Exercise 10 (BE: Problem 1.19)
It is known that CpG–islands in high eukaryotes are relatively rich with CpG dinucleotides, while these dinucleotides are discriminated in the rest of a chromosome. It is assumed that the frequency of occurrences of CpG dinucleotides in a CpG–island can be approximated by the Poisson distribution with twenty–five CpG dinucleotides per 250nt long fragment on average, while in the rest of the DNA this average is ten CpG per 250nt. Suggest the Bayesian type algorithm for CpG–island identification. Assume that the probabilities of a 250nt fragment being or not being a CpG island are a priori equal (i.e. both are 0.5). How will this algorithm characterize a 250nt long DNA fragment containing nineteen CpG dinucleotides?

Exercise 11 (MM: Problem 6, p. 330)
Assume that five cells are arranged in a line. A molecule will remain in the cell it currently occupies with probability 0.9, and will move to one of the two adjacent cells with equal probability. If it is at the end of the chain of cells, it moves with probability 0.1 to the single adjacent cell. State the transition matrix and draw the associated state diagram.

Exercise 12 (MM: Problem 7, p. 330)
An $n \times n$ matrix with non–negative entries is called *stochastic* if the entries in each row sum to 1. The matrix is called *doubly stochastic* if each row and each column sum to 1.
(a) For each of the following matrices determine whether it is stochastic or doubly stochastic.

$$P_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad P_2 = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 2/3 & 0 \\ 1/3 & 0 & 2/3 \end{bmatrix}$$

$$P_3 = \begin{bmatrix} 1/3 & 1/3 & 0 & 1/3 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 0 & 1 & 0 \\ 1/2 & 0 & 0 & 1/2 \end{bmatrix} \quad P_4 = \begin{bmatrix} 1/2 & 0 & 0 & 0 & 1/2 \\ 1/3 & 0 & 1/3 & 1/3 & 0 \\ 0 & 1/3 & 0 & 1/3 & 1/3 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

(b) Draw the state diagrams of the Markov chains corresponding to these transition matrices. Indicate absorbing states.

(c) A matrix is called *regular* is the associated Markov chain is such that every state can be reached from every other state (not necessarily in one step), which means that there is at least one power of the matrix P which has a non–zero entry corresponding to the desired transition. Determine which of the matrices in part (a) is regular.

Exercise 13 (BE: Problem 3.10)
A prokaryotic gene is a continuous sequence of nucleotide triplets, codons. A gene starts with a start codon ATG and ends with one of three stop codons: TAA, TAG, TGA. Calculate the number of free parameters in such a codon model. The data set contains on the order of 300000 codons. Would it be feasible to estimate a second order Markov chain from this data set?

Exercise 14 (EG: Problem 12.1)
Define an HMM λ with the following parameters: Three states, S_1, S_2, S_3,
alphabet $A = \{1, 2, 3\}$,

$$P = \begin{bmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \qquad \pi = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix},$$

$b_1(1) = 0.5$, $b_1(2) = 0.5$, $b_1(3) = 0$,
$b_2(1) = 0.5$, $b_2(2) = 0$, $b_2(3) = 0.5$,
$b_3(1) = 0$, $b_3(2) = 0.5$, $b_3(3) = 0.5$.
What are all possible state sequences for the following observed sequences $\mathcal{O}$, and what is $p(\mathcal{O}|\lambda)$?

(a) $\mathcal{O} = 1, 2, 3$.
(b) $\mathcal{O} = 1, 3, 1$.