Exercises – Lecture 2
Stochastic Processes and Markov Chains, Part 2

Question 1
Question 1a (without `R`*)*
The transition matrix of Markov chain is:

$$\begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}.$$

Find the stationary distribution of this Markov chain in terms of a and b, and interpret your results.

Question 1b (without `R`*)*
For which a and b is the Markov chain reversible?

Question 1c (without `R`*)*
For which a and b is the Markov chain periodic?

Question 2
Consider the transition matrix:

$$\mathbf{P} = \begin{pmatrix} 0.1500 & 0.3500 & 0.3500 & 0.1500 \\ 0.1660 & 0.3340 & 0.3340 & 0.1660 \\ 0.1875 & 0.3125 & 0.3125 & 0.1875 \\ 0.2000 & 0.3000 & 0.3000 & 0.2000 \end{pmatrix}.$$

Question 2a (with `R`*)*
Calculate the stationary distribution of the transition matrix $\mathbf{P}$ analytically and through matrix multiplication.

Question 2b (with `R`*)*
The GC content of a DNA sequence is defined as the percentage of `C`'s and `G`'s on the total number of bases of the sequence: $(\#\texttt{C} + \#\texttt{G})/(\#\texttt{A} + \#\texttt{C} + \#\texttt{G} + \#\texttt{T}) * 100\%$. Calculate the `GC` content for an infinitely long DNA sequence generated in accordance with the sampling model above (the first order Markov dependent stochastic process described by transition matrix above). Compare this to your answer to Question 4a of the previous set of exercises (corresponding to Lecture 1).

Question 3

Question 3a (without `R`*)*
Proof that irreducible, aperiodic first-order Markov chains, defined by a transition matrix with a nonsymmetric off-diagonal zero entry, are irreversible.

Question 3b (without `R`*)*
Proof that a symmetric transition matrix has a uniform stationary distribution.

Question 3c (without `R`*)*
Use the answer to question 3b to show that all first order Markov chains with an associated symmetric transition matrix are reversible.

Question 4

Consider the transition matrix:

$$\mathbf{P} = \begin{pmatrix} 0.40 & 0.10 & 0.30 & 0.20 \\ 0.00 & 0.20 & 0.80 & 0.00 \\ 0.00 & 0.60 & 0.40 & 0.00 \\ 0.30 & 0.30 & 0.10 & 0.30 \end{pmatrix}.$$

Question 4a (without `R`*)*
Draw the state diagram that corresponds to the transition matrix $\mathbf{P}$ above.

Question 4b (without `R`*)*
Is this Markov chain associated with the transition matrix above irreducible?

Question 4c (without `R`*)*
On the basis of Question 3a and 3b, which limiting behavior do you expect?

Question 4d (with `R`*)*
Perform a spectral decomposition of $\mathbf{P}$, using `R`'s `eigen`-function, and store the results in an object called `Pdecomp`. Verify that all eigenvalues are smaller or equal than 1 (in an absolute sense).

Question 4e (with `R`*)*
Having performed the spectral decomposition, then `Pdecomp$vectors` contains the right eigenvectors. Verify that the left eigenvector are given by the inverse of `Pdecomp$vectors`. Reconstruct $\mathbf{P}$ from the matrices of left and right eigenvectors and the diagonal matrix with eigenvalues on the diagonal.

Question 4f (with `R`*)*
Using the spectral decomposition, investigate how fast the influence of the initial value washes out (i.e., how fast do you reach the stationary distribution).

Question 5

Question 5a (without `R`*)*
Give the definition of stationary distribution of a Markov chain.

On the basis of their chemical properties the nucleotides are divided into two groups, the so-called purines (`A` en `G`) and pyrimidines (`C` en `T`). Assume the DNA may be modeled by a first order Markov process. This Markov process a multinomial initial distribution with probability 0.4 for a purine and 0.1 for a pyrimidine. With respect to the elements of the transition matrix:

- the probability of a purine at position $j+1$ in the DNA, if position j is occupied by a *different* purine, is equal to 0.15.
- the probability of a pyrimidine at position $j+1$ in the DNA, if position j is occupied by a purine, is equal to 0.25.
- the probability of a pyrimidine at position $j+1$ in the DNA, if position j is occupied by a *different* pyrimidine, is equal to 0.35.
- the probability of a purine at position $j+1$ in the DNA, if position j is occupied by a pyrimidine, is equal to 0.20.

Question 5b (beginning without, but end with `R`*)*
Give the transition matrix and its corresponding stationary distribution.

The DNA is not one long, uniform sequence of nucleotides, but can (very) crudely be divided in non-coding and coding pieces of DNA. Coding DNA contains the information for the formation of a protein, non-coding DNA does not. Assume the non-coding and coding sequences alternate every tenth nucleotide. Given that a particular piece of the DNA encodes for a protein or not, the sequence of nucleotides may be described by a first order Markov process. The transition probabilities of coding DNA is given by the transition matrix of question 5b, while in non-coding DNA every nucleotide has a probability of 0.6 to be succeeded by a *different* nucleotide, with equal probability for the different nucleotides to be drawn.

Question 5c (beginning without, but end with `R`*)*
Give the transition matrix and its corresponding stationary distribution of non-coding DNA.

Question 5d (with `R`*, but also without* `R` *when using the answers to Question 5b and 5c)*
Give the stationary distribution (which is to be interpreted as the proportion of time spent in a particular state) of DNA regularly composed of the coding and non-coding sequences (as described above).

It is nonsense to presume that non-coding and coding DNA alternate so regularly. Replace this assumption. Let the length of both non-coding and coding DNA be Poisson distributed with $\lambda = 10$ and $\lambda = 5$, respectively.

Question 5e (with `R`*, but also without* `R` *when using the answers to Question 5b and 5c)*
Keep in mind the definition given under Question 4a, and give the stationary distribution of DNA composed of the coding and non-coding sequences with random lengths (as described above). Hint: calculate the expected length of both non-coding and coding DNA sequences, and combine this with the answers to Questions 5b and 5c.

Question 6
Reconsider the DNA sequence analyzed in Question 5b of the previous set of exercises (correspond-

ing to Lecture 1). The nucleotide frequencies of this sequences are:

$$\begin{pmatrix} \texttt{\#A} & \texttt{\#C} & \texttt{\#G} & \texttt{\#T} \end{pmatrix} = \begin{pmatrix} 186 & 242 & 265 & 307 \end{pmatrix},$$

and the di-nucleotide frequencies:

$$\begin{pmatrix} \texttt{\#AA} & \texttt{\#AC} & \texttt{\#AG} & \texttt{\#AT} \\ \texttt{\#CA} & \texttt{\#CC} & \texttt{\#CG} & \texttt{\#CT} \\ \texttt{\#GA} & \texttt{\#GC} & \texttt{\#GG} & \texttt{\#GT} \\ \texttt{\#TA} & \texttt{\#TC} & \texttt{\#TG} & \texttt{\#TT} \end{pmatrix} = \begin{pmatrix} 58 & 60 & 0 & 68 \\ 39 & 77 & 83 & 43 \\ 43 & 0 & 87 & 134 \\ 46 & 105 & 95 & 61 \end{pmatrix}.$$

Question 6a (without `R`*)*
Give the period set of the motif `GAGA`.

Question 6b (without `R`*)*
How many times do you expect to observe the motif `GAGA` in the sequence? Hereto obtain the ML estimates of the transition matrix from the mono- and di-nucleotide frequencies given above.

Question 6c (with `R`*)*
The motif `TGTC` appears 17 times in the sequence. How exceptional is this?