TU Dortmund
Department of Statistics
Prof. Dr. Jörg Rahnenführer
Marieke Stolte

Summer term 2023
02.05.2023

Statistics in Genetics (Bioinformatics)
Exercise Sheet 5
Graded Repetition Sheet

Review task 1 (5P): Markov Chains I

Consider a Markov chain with 3 states and the following transition matrix:

$$\mathcal{P}_{n,n+1} = \begin{pmatrix} 0.6 & 0.2 & p_{13} \\ 0.3 & 0.4 & p_{23} \\ 0.1 & 0.2 & p_{33} \end{pmatrix}$$

(a) Calculate the missing entries in the transition matrix.

(b) Calculate the stationary distribution by hand. All steps need to be explicitly written down.

(c) Assume that one starts in state 3. What is the probability that one is in state 1 after two steps? For $m \to \infty$, what value does the probability of being in state 1 after m steps converge to?

Review task 2 (3P): Markov Chains II

Let the transition matrix P of a Markov chain be a doubly stochastic matrix, i.e. all entries are non-negative and all row sums as well as all column sums are equal to 1:

- $P_{i,j} \geq 0 \, \forall i, j \in 1, \ldots, k$,
- $\sum_{j=1}^{k} P_{ij} = 1 \, \forall i \in 1, \ldots, k$,
- $\sum_{i=1}^{k} P_{ij} = 1 \, \forall j \in 1, \ldots, k$.

(a) Show that the uniform distribution $\pi = (\frac{1}{k}, \ldots, \frac{1}{k})$ is a stationary distribution of P.

(b) Under what condition on P is the Markov chain also time-reversible?

Review task 3 (5P): Markov Chains III

Let $X(t)$, $t \in \mathbb{N}_0$, be a homogeneous Markov chain with state space $\mathcal{X} = \{1, \ldots, 6\}$ and transition matrix

$$P = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 1/2 & 1/4 \\ 0 & 0 & 0 & 1/2 & 1/4 & 1/4 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

(a) Which states can be reached from state 1?

(b) Is the Markov chain irreducible or reducible?

(c) Which states are periodic and which states are aperiodic? Specify the period for the periodic states.

(d) Which states are recurrent and which states are transient?

Justify each of your answers.

Review task 4 (10P): Phylogenetics: Maximum-Likelihood

Consider the following 4 aligned sequences of length 7.

Sequence	1	2	3	4	5	6	7
s_1	A	C	G	T	A	G	A
s_2	A	C	A	T	G	T	A
s_3	C	C	G	A	C	G	A
s_4	C	C	T	A	T	T	C

The likelihood of the tree topology $((s_1, s_2), (s_3, s_4))$ is sought for sequence positions 1 and 6. Use the notation for maximum likelihood estimation from the lecture when calculating the likelihood of the tree topology $((s_1, s_2), (s_3, s_4))$ for both sequence positions. For simplified calculation, assume all edge lengths to be of equal length, i.e. use the following transition matrix for bases A, G, C, and T that is independent of edge lengths (in that order, K2P model):

$$P = \begin{pmatrix} 0.7 & 0.2 & 0.05 & 0.05 \\ 0.2 & 0.7 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.7 & 0.2 \\ 0.05 & 0.05 & 0.2 & 0.7 \end{pmatrix}.$$

Review task 5 (7P): Sequence Alignments

Given are the following two sequences:

Sequence 1	$D\,A\,T\,A$
Sequence 2	$D\,E\,L\,T\,A$

To calculate scores in the sequence alignment, use the BLOSUM62 matrix and set the GAP cost constant to the value 5, i.e., the alignment of each amino acid with a GAP gives a score of -5, respectively.

(a) Using the *Needleman-Wunsch* algorithm, create the alignment (path) matrix and calculate the optimal *global* alignment.

(b) Find the optimal *local* alignment using the *Smith-Waterman* algorithm.

Upload the processed exercise sheets in the Moodle until Tuesday, 09.05.2023, 10:00. Please note the information listed in the Moodle on the submission formalities for this course.