Exercises – Lecture 1
Stochastic Processes and Markov Chains, Part1

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

$$\begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}.$$

If the initial probability distribution (at time $t = 0$) is $(0.8, 0.2)^T$, what is the probability that at time $t = 3$ the state occupied is E_1?

Question 2 *(without* `R`*)*
There are two models described below for a signal of length five: idd, 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 two models (so your answer should consist of four probabilities).

(i) iid: The probabilities of the four nucleotides are $\{p_A = 0.2, p_C = 0.1, p_G = 0.1, p_T = 0.6\}$.

(ii) First order Markov: The initial distribution is $\{p_A = 0.2, p_C = 0.1, p_G = 0.1, p_T = 0.6\}$, and the transition matrix (for the nucleotide ordering `A`, `C`, `G`, `T`) is:

$$\begin{pmatrix} 0.10 & 0.80 & 0.05 & 0.05 \\ 0.35 & 0.10 & 0.10 & 0.45 \\ 0.30 & 0.20 & 0.20 & 0.30 \\ 0.60 & 0.10 & 0.25 & 0.05 \end{pmatrix}.$$

Question 3
The `sample` function can be used to generate random sequences in R. For example, the syntax:

```
> nucleotides <- c("A", "C", "G", "T")
> p0 <- c(0.2, 0.3, 0.3, 0.2)
> sample(nucleotides, 10, replace=TRUE, prob=p0)
[1] "A" "C" "A" "T" "A" "G" "C" "G" "C" "A"
```

generates a DNA sequence from alphabet=`nucleotides` with length=10 and multinomial probabilities equal to `p0`.

This can also be done using a `for`-loop:

```
> DNAseq <- sample(nucleotides, 1, replace=TRUE, prob=p0)
> for (i in 2:10){
```

```
>    DNAseq <- c(DNAseq, sample(nucleotides, 1, replace=TRUE, prob=p0))
> }
```
Although this is less efficient.

Question 3a (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: $(\#\mathtt{C}+\#\mathtt{G})/(\#\mathtt{A}+\#\mathtt{C}+\#\mathtt{G}+\#\mathtt{T})*100\%$. Calculate the GC content for an infinitely long DNA sequence generated in accordance with the sampling model above. Confirm this by simulation: generate a DNA sequence of (say) length 100000 and calculate its GC content.

Question 3b (without R*)*
Write down the transition matrix of Markov chain from which the DNA sequence above has been generated.

Question 4
Consider the transition matrix:

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

Now use `sample()` to generate DNA sequences according to a first order Markov chain. Hereto use the transition matrix given above. We can only simulate one symbol at a time, because we need to keep track of the current state.

```
> DNAseq <- sample(nucleotides, 1, replace=TRUE, prob=p0)
> DNAseq <- c(DNAseq, sample(nucleotides, 1, replace=TRUE,
                  P[nucleotides==DNAseq[1], ]))
> DNAseq
[1] "C" "A"
```

The `P[nucleotides==DNAseq[1], ]` statement returns the row in the transition matrix that corresponds to the matching character in the alphabet `nucleotides`. Try `P[nucleotides=="A", ]`, `P[nucleotides=="C", ]`, `P[nucleotides=="G",]` and `P[nucleotides=="T",]` to see what happens.

Question 4a (with R*)*
Use the transition matrix given in Question 4 and simulate a DNA sequence of (say) length 100000. Hint: use a `for`-loop. Calculate the GC content for this sequence and compare it to the answer of Question 3a.

Question 5

Question 5a (with R*)*
Given the following sequence (available at the lecture website):

```
TCCATCGTCCAACCTCGTCATTACGTTTAATATGGTACCAATGTGTGGGTCGATTGTTC
CGGTGACTACCGGTGTGTTCAACGGTTTACGTGGGTGTGTTCTAATCTCACCCCGTGAA
CTCTCGTAAAATATGTTGTCCCACTCTTGTCTCCTACGATCGGTCGTGTTCCCGTCCTC
```

```
GTTCGTTCATACGTGATCACGGTATGTGGGATTCGTCGTTACCCCCGGTGACGGATGGA
TGTCAATGTGTCGATTTCGTGGTCAATGTGGGGGTGTCGTGGGGTGGTATCTTCGGTGG
TTCCGACGGGGGGTCCCTCTCGTGGGGTCTATATGGTTCTACGTCTACCCTCAATGAAC
GTCGTTGTTCACCCATGATTTCGGATCAAATGGTGGTTATGTTTCGTGAATCGACTGAA
CCCCGGGGTGGTGGTCATCAAATCTTGTCGAAACCGGTGGGTGTTGAACAATAATGACA
CTCCGTGATCTTGTTCATTAACGTGACCGAACGTCCGGTCGAATGTCGGTCCTCAAATC
GGAATTGTTAAACTGGTCAATCGACACCGTTCGTTCCTGTGTGGTATTATGTACCTCCC
TTCTCCGTGTCGTACTCGTCGTTTCGTGTCATATGGATACTCAACGGGGGTTCCACGTC
CCCTAATGAATGTCGGTGATGTCCCCCCGTATTCCTCACCGGTGTAACCCACGAAATTG
TGGGGACGGTACGTCAACCTTGTCGATACGAAACGTTCGTGTATGTGATGTGTGGGTGT
TTCCAATAACGAATACACGTCCTGTGGGATTCCCTATAACGTGGTCCCGGTCGGGGGTG
AAACCACGGTCCTACCGGTTCATACGTGTGTCGTAACCGTGGGTTAAATGTCCGGATCT
CGGAACAACCACTTGACCTTGGTTGTGTATGTCCCGGACTCGGGGGTCCATAATCCTGT
TGATCGTGTTTCCCTAATCTGTCCACCCGTATCAAATTTGTCCAATTTCGGGGATG
```

Maximize the likelihood with respect to the parameters, give the estimated transition matrix, and draw the state diagram.

To read in a string in `R`:

```
> DNAseq <- c("AAGTCAGT")
```

To select a letter, say the 5th, from this string:

```
> substr(DNAseq, 5, 5)
```

To obtain the number of characters of a string:

```
> nchar(DNAseq)
```

Question 5b (in principle without `R`*, although a calculator is practical)*
Using the estimated transition matrix in Question 5b, estimate the likelihood of the following motifs: `AAAAAA`, `CTGCAG` and `ACCGGT`. You may wish to assume that the first nucleotide is given, hence $P(Y_0 = y_0) = 1$.

Question 5c (in principle without `R`*, although a calculator is practical)*
For the sequence given in Question 5a, test using the χ^2-test whether the independence model could have been assumed.

Question 5d (with and without `R`*)*
Repeat Question 5a, but now assume that the transition matrix is symmetric. Hint: use the derivation of the likelihood and the corresponding estimators presented in Lecture 1, and simplify as much as possible using the symmetry assumption.

Question 6 *(in principle without* `R`*, although a calculator is practical)*
Regions on the genome that have many more `CG` dimers (and in fact more C and G nucleotides) than elsewhere on the DNA are called CpG islands. From a set of human DNA sequences we have learned that the nucleotide sequence of a CpG island and of remainder of the DNA are modeled

by first order Markov models with the transition matrices:

$$\mathbf{P}_1 = \begin{pmatrix} 0.180 & 0.274 & 0.426 & 0.120 \\ 0.171 & 0.368 & 0.274 & 0.188 \\ 0.161 & 0.339 & 0.375 & 0.125 \\ 0.079 & 0.355 & 0.384 & 0.182 \end{pmatrix} \quad \text{and} \quad \mathbf{P}_2 = \begin{pmatrix} 0.300 & 0.205 & 0.285 & 0.210 \\ 0.322 & 0.298 & 0.078 & 0.302 \\ 0.248 & 0.246 & 0.298 & 0.208 \\ 0.177 & 0.239 & 0.292 & 0.292 \end{pmatrix},$$

respectively.

Given the following stretches of genomic sequence, decide whether they come from a CpG island or not. Hint: calculate the likelihood of a stretch under both first order Markov models. You may wish to assume that the first nucleotide is given, hence $P(Y_0 = y_0) = 1$.
- Stretch 1: `GGTGGTCATCAAATCTTGTCGA`
- Stretch 2: `ACGTTTAATATGGTACCAATGT`
- Stretch 3: `ACGGGGGGTCCCTCTCGTGGGG`

Question 7
Assume a first order Markov model with state space $\mathcal{S} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, initial distribution $\boldsymbol{\pi}$ and transition matrix $\mathbf{P}$. Let $\ldots, X_{t-2}, X_{t-1}, X_t, X_{t+1}, X_{t+2}, \ldots$ be the sequence generated by this model. Furthermore, write $P(X_t = \mathtt{A}) = \psi_{\mathtt{A}}$, $P(X_t = \mathtt{C}) = \psi_{\mathtt{C}}$, $P(X_t = \mathtt{G}) = \psi_{\mathtt{G}}$, and $P(X_t = \mathtt{T}) = \psi_{\mathtt{T}}$ for all t.

Question 7a (without `R`*)*
Write $P(X_{t+1} = \mathtt{A}, X_t = \mathtt{T})$ in terms of the transition probabilities of $\mathbf{P}$ and the ψ's.

Question 7b (without `R`*)*
Write $P(X_{t+1} = \mathtt{A}, X_{t-1} = \mathtt{T})$ in terms of the transition probabilities of $\mathbf{P}$ and the ψ's.

Question 7c (without `R`*)*
Write $P(X_{t+1} = \mathtt{A}, X_t = \mathtt{T}, X_{t-1} = \mathtt{T})$ in terms of the transition probabilities of $\mathbf{P}$ and the ψ's.

Question 7d (without `R`*)*
Write $P(X_{t+1} = \mathtt{A}, X_t = \mathtt{T} \,|\, X_{t-1} = \mathtt{T})$ in terms of the transition probabilities of $\mathbf{P}$ and the ψ's.

Question 7e (without `R`*)*
Write $P(X_{t+1} = \mathtt{A}, X_{t-1} = \mathtt{T} \,|\, X_t = \mathtt{T})$ in terms of the transition probabilities of $\mathbf{P}$ and the ψ's.

Question 7f (without `R`*)*
Write $P(X_{t+1} = \mathtt{A}, X_{t-2} = \mathtt{C} \,|\, X_t = \mathtt{T})$ in terms of the transition probabilities of $\mathbf{P}$ and the ψ's.