# Genome Sequence Analysis Assignment

**1**. Write a program which reads a $K \times K$ transition matrix and a $1 \times K$ initial distribution from a file and outputs N elements of a Markov chain.

**2**. Write a program which reads a sequence of state indices (i.e. elements of $\{1, \ldots, K\}$) and infers a maximum likelihood Markov chain transition matrix and initial distribution.

**3**. Implement a HMM by modifying your program from (1), adding an emitted variable at each point in the chain. Use it to simulate 115 emitted values from the following model:

$$S = \{0, 1\}, V = \{1, 2, 3, 4, 5\}$$

$$\mathbf{A} = \begin{pmatrix} 0.8 & 0.2 \\ 0.1 & 0.9 \end{pmatrix}, \mu^0 = (0.5, 0.5), \mathbf{B} = \begin{pmatrix} 0.2 & 0.5 & 0.2 & 0.1 & 0 \\ 0 & 0.1 & 0.4 & 0.4 & 0.1 \end{pmatrix}$$

where $S$ is the hidden state space, $V$ the emission space, $\mathbf{A}$ the transition matrix, $\mu^0$ the initial distribution and $\mathbf{B}$ the emission matrix.

Plot the resulting sequences of hidden and emitted states on the same graph.

**4**. Implement the forward algorithm with scaling to calculate the likelihood of an emitted sequence given the model, reading the emitted sequence from a file.

**5**. Download chromosome III of the *Saccharomyces cerevisiae* (yeast) genome sequence from Ensembl or NCBI. Calculate its `GC` content (the fraction of bases which are `G` or `C`) in 100 bp windows. Choose an appropriate binning scheme to represent the `GC` content of each window such that the resulting sequence corresponds to an emission sequence for the model in (3).

Calculate the log likelihood of this `GC` sequence under the model in (3).

**6**. Extend your programs in (3) and (4) to implement Baum-Welch estimation of model parameters $\mathbf{A}$, $\mathbf{B}$ and $\mu^0$, given an emitted sequence.

Using the encoded `GC` sequence for the yeast genomic sequence produced in (5), estimate new parameters for the HMM specified in (3). Calculate the log likelihood of the data under this new model.

**7**. Infer the most likely sequence of hidden states under the model with your new parameters. Plot some of the output along with the corresponding `GC` content, and demonstrate the significance of your inference in relation to annotation in this and other genome sequences. Are there any extensions or adaptations you could make to your model which might make it more useful for this purpose?