Assignments III - Markovian modeling and Bayesian learning

11. Choose a tree topology with 5 leaf nodes and simulate continuous-time Markov chains (CTMCs) according to the Felsenstein F81 and the so called Hasekawa-Kishino-Yano (HKY) time-reversible models. A particularly useful program for such simulations is Seq-Gen (http://tree.bio.ed.ac.uk/software/seqgen/), but there are several alternatives available as well (e.g. you can use code for R or Matlab). Use the same stationary nucleotide distributions under both generating models and choose the generator matrix parameters such that the models are not identical. Choose the time frame for the simulation such that a reasonable amount of sequence differences is present in the data. Investigate and compare the patterns of molecular variation that emerge under these models. Investigate also the effect of simulated sequence lengths on the variation visible in data by considering both short (say 100 bases) and long sequences (say 10,000 bases).

12. Repeat the simulation and investigations in assignment #11 by including Gamma distributed transition rate heterogeneity in the models. Set the parameter of the Gamma distribution equal to 1/2. Compare the amount of variation present in this and the previous data, in particular pay attention to the presence of sequence sites with inflated levels of variation compared to the average.

13. For data sets in assignments #11 and #12 you know the time since most recent common ancestor (TMRCA) of the sequences (since you simulated them). Describe the so called *molecular clock* hypothesis (see e.g. the book on phylogenetic models by Timo Koski) and investigate the effect of the Gamma distributed transition rate heterogeneity on a rough quantification of TMRCA by comparison with the case where no rate heterogeneity was present.

14. Simulate noisy Markov chain under the HMM specified on p. 217 in Koski's book on HMMs. This model is defined by considering the hidden process as a Markov chain $\{X_n\}_{n=0}^{\infty}$ with $S = \{0, 1\}$ as the state space and the transition probability matrix $A = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$. The observations are then represented by the process $\{Y_n\}_{n=0}^{\infty}$ with $\mathcal{O} = \{0, 1\}$ as the state space, such that

$$\begin{aligned} P(Y_0 &= o_0, ..., Y_n = o_n | X_0 = s_0, ..., X_n = s_n; B) \\ &= \prod_{l=0}^{n} \epsilon^{|o_l - j_l|} \cdot (1-\epsilon)^{1-|o_l - j_l|}, \end{aligned}$$

where the emission probability matrix equals $B = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix}$. This model is a noisy Markov chain where the probability of having an error in an observation equals ϵ. Consider how ML learning of transition parameters behaves on

these data when ordinary Markov(1) chain model is used instead of the correct model. Run the simulation with different error probabilities ϵ and investigate whether the errors in ML learning under ordinary Markov(1) model behave differently with different values of ϵ.

15. Consider a non-standard HMM introducing a segmental structure on a DNA sequence. Assume a binary Markov chain as the state process $\{X_n\}_{n=0}^{\infty}$, with $S = \{0, 1\}$ where the two states correspond to 'CpG rich' and CpG poor' areas within an observed DNA sequence. Assume that the transition probability matrix for the chain equals $A = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$, where p is fairly small, say 0.01. Define then a first order Markov chain $\{Y_n\}_{n=0}^{\infty}$ with $\mathcal{X} = \{A, C, G, T\}$, such that the invariant distribution of the emitted symbols reflects the 'CpG rich' and CpG poor' states, that is states $\{C, G\}$ have high probabilities when $X_l = 0$ and low probabilities when $X_l = 1$, respectively. For details on a similar Markov(0) model see the article by Braun & Muller. Investigate how ML estimation of the transition probabilities behaves if the data from the HMM are used in an ordinary stationary Markov(1) chain model. Visualize the output generated by the HMM.