**Bioinformatics Research Center**
**Aarhus University**

**Statistical Methods in Bioinformatics**
**Asger Hobolth**
**January 14, 2015**

# Exam

## General information

The report should be printed and handed in before noon (12.00 am) Friday January 16, 2015, to me (office 324, Building 1110) or Ellen Noer (office 319, Building 1110). You should also e-mail the report to me (asger@birc.au.dk).

## Writing and organization of the report

The report should be written in danish or english. Part of the exam is to write clear and documented R code. In the report you should include the R programs that you have written in an Appendix.

## Project description

The project falls in three parts:

1. In the first part of the project we consider the strand-symmetric model. The model is compared to other models for DNA sequence evolution.
2. In the second part of the project we analyse a pair of homologous sequences. We find an appropriate description of how the sequences have evolved, and determine the evolutionary distance (in terms of expected number of substitutions) between the sequences.
3. In the third part we consider a hidden Markov model (HMM) along a sequence alignment. We estimate the parameters of the HMM and discuss methods for decoding the hidden states conditional on the data.

### Part I: The strand-symmetric model

In the double helix structure of DNA the two strands coil around each other. The nucleotides in the two strands are joined to each other according to the base pair rule: $a$ with $t$ and $g$ with $c$ (see Figure 1).

The strand-symmetric model (SSM) is symmetric with respect to the two strands. This means, for example, that the rate $Q_{ag} = Q_{tc}$ since the substitution $a \to g$ in one strand corresponds to the substitution $t \to c$ in the other strand.

(a) Let $Q$ be the $4 \times 4$ rate matrix for the SSM corresponding to the states $\{a, g, c, t\}$. Argue that the strand-symmetric model has the constraints

$$Q_{ag} = Q_{tc}, \;\; Q_{ac} = Q_{tg}, \;\; Q_{at} = Q_{ta}, \;\; Q_{ga} = Q_{ct}, \;\; Q_{gc} = Q_{cg}, \;\; Q_{gt} = Q_{ca},$$

and therefore can be written

$$Q = \begin{bmatrix} \cdot & q & s & u \\ r & \cdot & v & t \\ t & v & \cdot & r \\ u & s & q & \cdot \end{bmatrix},$$

where states are in the order $\{a, g, c, t\}$. Here diagonal entries are such that each row sums to zero. Note that the model has 6 parameters.

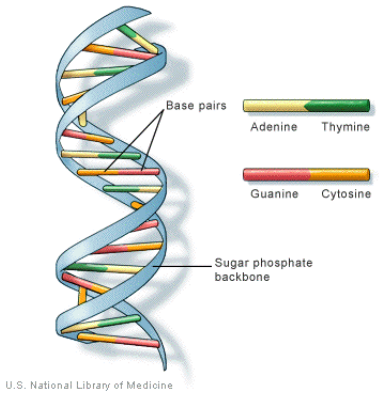



Figure 1: Double helix structure of DNA. The nucleotide $a$ (Adenine) pairs with $t$ (Thymine) and $g$ (Guanine) pairs with $c$ (Cytosine). The strand-symmetric model is symmetric with respect to the two strands.

(b) Recall that the general time-reversible model (GTR) has 9 parameters and can be parametrized by

$$Q = \begin{bmatrix} \cdot & \alpha\varphi_g & \beta\varphi_c & \gamma\varphi_t \\ \alpha\varphi_a & \cdot & \delta\varphi_c & \epsilon\varphi_t \\ \beta\varphi_a & \delta\varphi_g & \cdot & \eta\varphi_t \\ \gamma\varphi_a & \epsilon\varphi_g & \eta\varphi_c & \cdot \end{bmatrix},$$

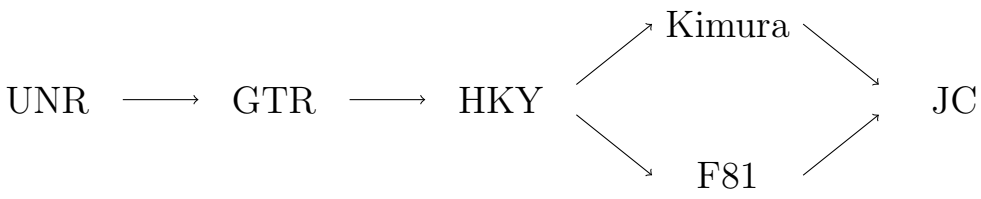
where $(\varphi_a, \varphi_g, \varphi_c, \varphi_t)$ is the stationary distribution and $(\alpha, \beta, \gamma, \delta, \epsilon, \eta)$ are free parameters.

We now want to define a reversible *and* strand-symmetric model. Show that such a model can be parametrized by

$$Q = \begin{bmatrix} \cdot & \alpha\varphi_2 & \beta\varphi_2 & \gamma\varphi_1 \\ \alpha\varphi_1 & \cdot & \delta\varphi_2 & \beta\varphi_1 \\ \beta\varphi_1 & \delta\varphi_2 & \cdot & \alpha\varphi_1 \\ \gamma\varphi_1 & \beta\varphi_2 & \alpha\varphi_2 & \cdot \end{bmatrix},$$

where $(\varphi_1, \varphi_2, \varphi_2, \varphi_1)$ is the stationary distribution and $(\alpha, \beta, \gamma, \delta)$ are free parameters. Note that the model has 5 free parameters. We call the model the strand-symmetric reversible model (abbreviated SSRM).

(c) We have previously considered the relation between various DNA evolutionary models:




Here UNR is the unrestricted model where all off-diagonal entries in the rate matrix are non-negative (12 parameters in total), and $A \to B$ means that $B$ is nested in $A$.

Extend this hierarchy of models to include the SSM and SSRM.

## Part II: Pairwise sequence analysis

Consider a pairwise alignment of two homologous sequences $S_1$ and $S_2$. Ignoring gaps and assuming sites are independent and identically distributed, the alignment can be summarized in the following table $\{n_{ij}\}$, $(i,j) \in \{a,g,c,t\}$, where e.g. $n_{ag} = 2380$ means that in 2380 sites we observed the pair $(a,g)$ in the sequence alignment:

| | $a$ | $g$ | $c$ | $t$ | $\sum$ |
|---|---|---|---|---|---|
| $a$ | 2380 | 126 | 43 | 22 | 2571 |
| $g$ | 112 | 2297 | 31 | 46 | 2486 |
| $c$ | 40 | 26 | 2277 | 105 | 2448 |
| $t$ | 27 | 56 | 97 | 2315 | 2495 |
| $\sum$ | 2559 | 2505 | 2448 | 2488 | 10000 |

We describe the relation between the two sequences using a continuous time Markov chain on the state space $\{a,g,c,t\}$. We only consider reversible models and therefore we can move the root to the first sequence $S_1$. Furthermore the likelihood only depends on the parameters through the rate matrix $Qt$ where $t$ is the time between the two sequences and $Q$ is the rate matrix. We choose the scaling determined by setting $t = 1$.

(d) Consider the 2-parameter Kimura model with rate matrix

$$Q = \begin{pmatrix} \cdot & \alpha & \beta & \beta \\ \alpha & \cdot & \beta & \beta \\ \beta & \beta & \cdot & \alpha \\ \beta & \beta & \alpha & \cdot \end{pmatrix},$$

with diagonal entries $-\alpha - 2\beta$. Determine the maximum likelihood estimates for $\alpha$ and $\beta$ and the corresponding maximum log-likelihood value. What is the expected number of substitutions between the two sequences?

(e) We now consider the strand-symmetric reversible model (SSRM) from Part I of the project with uniform stationary distribution $\varphi = (1,1,1,1)/4$.
Argue that the model has four free parameters and can be parametrized by

$$Q = \begin{bmatrix} \cdot & \alpha & \beta & \gamma \\ \alpha & \cdot & \delta & \beta \\ \beta & \delta & \cdot & \alpha \\ \gamma & \beta & \alpha & \cdot \end{bmatrix}.$$

Use the EM-algorithm for discretely observed continuous-time Markov chains to determine the maximum likelihood estimates for the four parameters $(\alpha,\beta,\gamma,\delta)$ in the strand-symmetric reversible model with uniform stationary distribution. Explain the M-step in the EM-algorithm in detail. What is the expected number of substitutions between the two sequences in the SSRM with uniform stationary distribution?

(f) Test the hypothesis $H : \beta = \gamma = \delta$ in the SSRM with uniform stationary distribution, and briefly discuss your finding.

## Part III: Hidden Markov Models

The data and a few commands for reading the data into R can be downloaded from the homepage (in the Exam directory; the files are called HMMExam.dat and HMMExam.R).

The purpose of this part of the project is to extend the emission probabilities in the hidden Markov model (HMM) from Exercise 2 in Project 1 from the Jukes-Cantor model to the Kimura model. As in

Exercise 2 in Project 1 we consider an alignment of two homologous sequences `S1` and `S2`, and want to divide the sequences into functional regions (hidden state 1) and neutral regions (hidden state 2). The sequence alignment is of length $L = 250$ (no gaps are present).

For each of the two hidden states the emission probabilities are determined by the Kimura model with rate matrix

$$Q = \begin{pmatrix} \cdot & \alpha & \beta & \beta \\ \alpha & \cdot & \beta & \beta \\ \beta & \beta & \cdot & \alpha \\ \beta & \beta & \alpha & \cdot \end{pmatrix}.$$

where entries are in the order $(a, g, c, t)$ and the diagonal entries are $-\alpha - 2\beta$. EG Section 14.3.2 shows that

$$\exp(Q) = \begin{pmatrix} p_{sa} & p_{ts} & p_{tv} & p_{tv} \\ p_{ts} & p_{sa} & p_{tv} & p_{tv} \\ p_{tv} & p_{tv} & p_{sa} & p_{ts} \\ p_{tv} & p_{tv} & p_{ts} & p_{sa} \end{pmatrix},$$

where

$$\begin{aligned} p_{ts} &= p_{ts}(\alpha, \beta) = \frac{1}{4} + \frac{1}{4}e^{-4\beta} - \frac{1}{2}e^{-2(\alpha+\beta)} \\ p_{tv} &= p_{tv}(\beta) = \frac{1}{4} - \frac{1}{4}e^{-4\beta} \\ p_{sa} &= 1 - p_{ts} - 2p_{tv}. \end{aligned}$$

Let $(\alpha_1, \beta_1)$ be the parameters from the Kimura model that determines the emission probabilities from hidden state 1, and similarly let $(\alpha_2, \beta_2)$ determine the emission probabilities from hidden state 2. Neutrally evolving positions have on average more substitutions than functional positions and therefore $\alpha_2 + 2\beta_2 > \alpha_1 + 2\beta_1$.

Define $\gamma_i = p_{ts}(\alpha_i, \beta_i)$ to be the probability for a transition, and similarly define $\delta_i = p_{tv}(\beta_i)$ to be the probability for a transversion and $\epsilon_i = 1 - \gamma_i - 2\delta_i$ to be the probability for observing the same two nucleotides in hidden state $i$ $(i = 1, 2)$.

(g) Argue that the emission probabilities are determined by the matrix $b_i(a)$ in Table 1, where $i = 1, 2$ corresponds to the two hidden states and $a = 1, \ldots, 16$ corresponds to the 16 possible alignment columns. Explain why the emission probabilities sum to 1 within each row.

| alignment column | `S1` / `S2` | `A` `A` | `A` `G` | `A` `C` | `A` `T` | `G` `A` | `G` `G` | `G` `C` | `G` `T` | `C` `A` | `C` `G` | `C` `C` | `C` `T` | `T` `A` | `T` `G` | `T` `C` | `T` `T` |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| hidden state | 1 | $\frac{\epsilon_1}{4}$ | $\frac{\gamma_1}{4}$ | $\frac{\delta_1}{4}$ | $\frac{\delta_1}{4}$ | $\frac{\gamma_1}{4}$ | $\frac{\epsilon_1}{4}$ | $\frac{\delta_1}{4}$ | $\frac{\delta_1}{4}$ | $\frac{\delta_1}{4}$ | $\frac{\delta_1}{4}$ | $\frac{\epsilon_1}{4}$ | $\frac{\gamma_1}{4}$ | $\frac{\delta_1}{4}$ | $\frac{\delta_1}{4}$ | $\frac{\gamma_1}{4}$ | $\frac{\epsilon_1}{4}$ |
| | 2 | $\frac{\epsilon_2}{4}$ | $\frac{\gamma_2}{4}$ | $\frac{\delta_2}{4}$ | $\frac{\delta_2}{4}$ | $\frac{\gamma_2}{4}$ | $\frac{\epsilon_2}{4}$ | $\frac{\delta_2}{4}$ | $\frac{\delta_2}{4}$ | $\frac{\delta_2}{4}$ | $\frac{\delta_2}{4}$ | $\frac{\epsilon_2}{4}$ | $\frac{\gamma_2}{4}$ | $\frac{\delta_2}{4}$ | $\frac{\delta_2}{4}$ | $\frac{\gamma_2}{4}$ | $\frac{\epsilon_2}{4}$ |

Table 1: Emission probabilities from the two hidden states. We have $\epsilon_1 = 1 - \gamma_1 - 2\delta_1$ and $\epsilon_2 = 1 - \gamma_2 - 2\delta_2$.

(h) The transition matrix of the HMM is determined by the $2 \times 2$ matrix

$$P = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix},$$

and the initial state is given by $\pi = (1/2, 1/2)$.

The emission probabilies for the neutrally evolving regions are known and determined by $(\alpha_2, \beta_2) = (0.5, 0.2)$. The parameters $\alpha_1$ and $\beta_1$ for the functional regions are unknown.

We want to estimate the free parameters $\theta = (\alpha_1, \beta_2, p_{11}, p_{22})$ using the EM-algorithm. Let $\mathcal{O} = (\mathcal{O}_1, \ldots, \mathcal{O}_L)$ be the observed sequence alignment and suppose we know the hidden state sequence $Q = (q_1, \ldots, q_L)$.

Show that the full likelihood

$$L(\theta; Q, \mathcal{O}) \quad = \quad \pi_{q_1} b_{q_1}(\mathcal{O}_1) \prod_{l=2}^{L} \Big( p_{q_{l-1}, q_l} b_{q_l}(\mathcal{O}_l) \Big)$$

is proportional to

$$L(\theta; Q, \mathcal{O}) \propto p_{11}^{N_{11}} p_{12}^{N_{12}} p_{21}^{N_{21}} p_{22}^{N_{22}} \gamma_1^{N_1^\gamma} \delta_1^{N_1^\delta} \epsilon_1^{N_1^\epsilon}, \tag{1}$$

and describe in words the meaning of $N_{ij}$, $N_1^\gamma$, $N_1^\delta$ and $N_1^\epsilon$.

(i) Describe the M-step in the EM-algorithm (explain the maximization of (1)).

(j) Describe what type of calculations are needed in the E-step of the EM-algorithm.

(k) Implement the EM-algorithm with initial values

$$(\alpha_1, \beta_1, p_{11}, p_{22}) = (0.1, 0.1, 0.9, 0.95),$$

and provide the maximum likelihood estimates for the four free parameters.

(l) Decode the sequence alignment using first the Viterbi algorithm and second posterior decoding. Make a plot that summarizes the data and corresponding decoding. Do the regions with few substitutions correspond to predicted functional regions?