Bioinformatics Research Center
Aarhus University

Advanced Statistical Methods in Bioinformatics
Asger Hobolth
November 22, 2018

Handout 13

Lectures in Week 48

EG Section 14.3 (Evolutionary Models: Continuous Time Markov Chains). The section describes a number of continuous time Markov chains and their application in molecular evolution. Instead of lecturing about this section I have prepared the exercises below. I expect you to finish the exercises during this week; I will be available on Monday and Wednesday, and Noor is available on Thursday during the usual hours of lectures and exercise class.

Exercises in Week 48

1. **EG Section 11.7.3: The Stationary Distribution**
   Show that EG equation (11.34) is equivalent to $\varphi Q = 0$ where $\varphi = (\varphi_1, \ldots, \varphi_s)$ and $Q = (q_{ij})$ is the $s \times s$ rate matrix with diagonal entries $q_{jj} = -q_j$, $j = 1, \ldots, s$.

2. **EG Section 11.7.4: Detailed Balance and Reversibility**
   Consider a CTMC with rate matrix Q. The CTMC is reversible if for all i, j and t

   $$\varphi_i P_{ij}(t) = \varphi_j P_{ji}(t). \tag{1}$$

   The Detailed Balance criterion is

   $$\varphi_i q_{ij} = \varphi_j q_{ji} \quad \text{for all } i, j. \tag{2}$$

   EG claims that these two conditions are the same; the purpose of this exercise is to verify their claim.

   (i) Show that (1) implies (2).
   Hint: Divide both sides in (1) by t and consider the limit $t \to 0$.

   (ii) Show that (2) implies (1).
   Hint: First show (using induction) that (2) implies that $\varphi_i (Q^n)_{ij} = \varphi_j (Q^n)_{ji}$ for all $n \geq 1$. Second use the matrix series expansion of $P_{ij}(t) = (e^{Qt})_{ij}$ to verify (1).

   (iii) An alternative proof that (2) implies (1) is based on matrix manipulation and is as follows: Write (2) as $D_\varphi Q = Q^T D_\varphi$ and (1) as $D_\varphi e^{Qt} = (e^{Qt})^T D_\varphi$, where D_φ is the diagonal matrix with φ along its diagonal and T denotes matrix transpose. Now use the matrix series expansion of e^{Qt} to show that (2) implies (1).

3. **EG Section 14.3.4: HKY model**
   Let $\varphi = (\varphi_a, \varphi_g, \varphi_c, \varphi_t)$ be a probability vector (i.e. entries are non-negative and sum to one) and consider the HKY model with rate matrix

   $$Q = \begin{pmatrix} \cdot & \alpha\varphi_g & \beta\varphi_c & \beta\varphi_t \\ \alpha\varphi_a & \cdot & \beta\varphi_c & \beta\varphi_t \\ \beta\varphi_a & \beta\varphi_g & \cdot & \alpha\varphi_t \\ \beta\varphi_a & \beta\varphi_g & \alpha\varphi_c & \cdot \end{pmatrix}.$$

   where entries are in the order (a, g, c, t) and the diagonal entries are such that rows sum to zero.
   (i) Show that φ is the stationary distribution for the HKY model.
   (ii) Argue that the HKY model has 5 free parameters.

An alternative way of parameterising the HKY model is as follows: The HKY model is determined by a rate matrix Q given by

$$q_{ij} = \begin{cases} \kappa\varphi_j & \text{transition} \\ \varphi_j & \text{transversion,} \end{cases}$$

for $i \neq j$, with q_{ii} such that each row in Q sums to 0.
(iii) Argue that $\kappa = \alpha/\beta$ and why κ is called the transition-to-transversion rate ratio.

4. **Sliding the Root (or Felsensteins Pulley Principle)**
Consider a reversible continuous-time Markov chain with stationary distribution φ. Show that

$$\sum_i \varphi_i P_{ix_1}(t_1) P_{ix_2}(t_2) = \varphi_{x_1} P_{x_1x_2}(t_1 + t_2) = \varphi_{x_2} P_{x_2x_1}(t_1 + t_2).$$

Note that the consequence of this equation is very important: We can reverse the time on a branch! The equation is the basis for considering unrooted trees instead of rooted trees (see EG bottom of page 514 for more information).

5. **EG Section 14.3.1: The Jukes-Cantor model**
Consider two DNA sequences an evolutionary distance $2t$ apart. Suppose p is the probability of observing different nucleotides at a given site.

(i) Consider the Jukes-Cantor model with rate α for a change. Argue for the estimating equation

$$p = \frac{3}{4} - \frac{3}{4} e^{-8\alpha t}.$$

(ii) Show that the solution to the equation is given by

$$\alpha t = -\frac{1}{8} \log\Big(1 - \frac{4}{3} p\Big),$$

and argue that we require $0 \leq p < 3/4$.

(iii) The expected number of substitutions is given by $\nu = 2t \cdot 3\alpha \cdot n = 6n\alpha t$, where n is the length of the two DNA sequences. Show that

$$\nu = -\frac{3n}{4} \log\Big(1 - \frac{4}{3} p\Big).$$

6. **EG Section 14.3.1: The Jukes-Cantor model**
Suppose two DNA sequences are compared using the Jukes-Cantor model. Their lengths are $n = 3000$, the number of constant sites is 2700, and the number of differences is 300. Show that $\hat{p} = 0.10$, and use the previous exercise to show that the expected number of substitutions is $\nu = 320$.

Argue that the variance of $\hat{p}$ is $p(1-p)/n$. Provide an approximate 95% confidence interval for p and transform the interval to an approximate 95% condidence interval for ν.

7. **EG Section 14.3.2: The Kimura Model**
Consider 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 $-\alpha - 2\beta$.

(i) Use the symmetry of the rate matrix to argue that

$$\begin{aligned} P_{aa}(t) = P_{gg}(t) = P_{cc}(t) = P_{tt}(t) \\ P_{ag}(t) = P_{ga}(t) = P_{ct}(t) = P_{tc}(t) \\ P_{ac}(t) = P_{at}(t) = \cdots = P_{tg}(t). \end{aligned}$$

(ii) Solve the forward Kolmogorov equations, i.e. show that

$$\begin{aligned} P_{aa}(t) &= \frac{1}{4} + \frac{1}{4}e^{-4\beta t} + \frac{1}{2}e^{-2(\alpha+\beta)t} \\ P_{ag}(t) &= \frac{1}{4} + \frac{1}{4}e^{-4\beta t} - \frac{1}{2}e^{-2(\alpha+\beta)t} \\ P_{ac}(t) &= \frac{1}{4} - \frac{1}{4}e^{-4\beta t}. \end{aligned}$$

Hint: Determine (and solve) seperate differential equations for $P_{ac}(t), P_{aa}(t) + P_{ag}(t)$ and $P_{aa}(t) - P_{ag}(t)$.

8. **EG Section 14.3.2: The Kimura model**
Consider two DNA sequences an evolutionary distance $2t$ apart. Suppose p_1 and p_2 are the probabilities of observing a transition and a transversion at a given site, respectively.

(i) Use the previous exercise to argue for the two estimating equations

$$\begin{aligned} p_1 &= \frac{1}{4} + \frac{1}{4}e^{-4\beta(2t)} - \frac{1}{2}e^{-2(\alpha+\beta)(2t)} \\ p_2 &= \frac{1}{2} - \frac{1}{2}e^{-4\beta(2t)}. \end{aligned}$$

(ii) Show that the solutions to the two equations are given by

$$\begin{aligned} 4(\alpha+\beta)t &= -\log(1-2p_1-p_2) \\ 8\beta t &= -\log(1-2p_2), \end{aligned}$$

and argue that we require $0 < p_2 < 1/2$ and $0 < 2p_1 + p_2 < 1$.

(iii) The expected number of substitutions is given by $\nu = 2t \cdot (\alpha + 2\beta) \cdot n = 2n(\alpha + 2\beta)t$, where n is the length of the two DNA sequences. Show that

$$\nu = \frac{n}{2}\Big(-\log(1-2p_1-p_2)\Big) + \frac{n}{4}\Big(-\log(1-2p_2)\Big).$$

9. **EG Section 14.3.2: The Kimura model**
Suppose two DNA sequences are compared using the Kimura model. Their lengths are $n = 3000$, the number of constant sites is 2700, the number of sites with a transition is 210, and the number of sites with a transversion is 90. Show that $p_1 = 0.07$, $p_2 = 0.03$, and use the previous exercise to show that the expected number of substitutions is $\nu = 326$.

10. Problem 14.10 in EG page 496. Regarding (ii):

(a) Show that for the JC model

$$\hat{\nu} = N\Big(\hat{p} + \frac{2}{3}\hat{p}^2\Big),$$

to second order, and for the Kimura model

$$\hat{\nu} = N\Big(\hat{p}_1 + \hat{p}_2 + \hat{p}_1^2 + \hat{p}_1\hat{p}_2 + \frac{3}{4}\hat{p}_2^2\Big),$$

to second order.

(b) Show that if $\hat{p}_1 = \hat{p}/3$ and $\hat{p}_2 = 2\hat{p}/3$, then the two expressions are the same.

(c) Show that EG expression (14.47) reduces to EG (14.35) if $\hat{p}_1 = \hat{p}/3$ and $\hat{p}_2 = 2\hat{p}/3$.

11. **GTR model**

The GTR model is the most general time-reversible model for DNA on the nucleotide level. Let $\varphi = (\varphi_a, \varphi_g, \varphi_c, \varphi_t)$ be a probability vector (i.e. entries are non-negative and sum to one) and consider the GTR model with rate matrix

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

where entries are in the order (a, g, c, t) and the diagonal entries are such that rows sum to zero.
(i) Show that the GTR model is reversible with stationary distribution φ.
(ii) Argue that the GTR model has 9 free parameters.