

MS2a, Week 1, Model Solutions

Rune Lyngsø

October 19, 2011

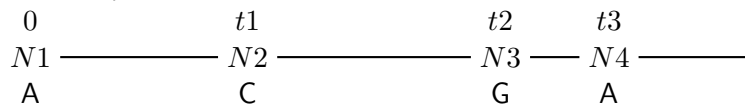
A From rates to probabilities

Describing evolution in terms of rates that describes what happens in a very short time interval is easy, but what is needed is a description of what happens during longer time interval. We only observe sequences at the leaves of a phylogeny, and what happens at the internal branches is hidden. Several, or even numerous events, can have taken place, while still only being observable as the net result of one nucleotide changing into another. It is thus necessary to be able to go from descriptions of instantaneous events to the accumulated result over a time interval.

We model nucleotide evolution as a continuous time discrete Markov process. Let us assume that all nucleotides evolve at the same rate and that one jumps to the alternative nucleotides with the same probability, known as the Jukes-Cantor model. More realistic models exist, but though the mathematics become more involved, conceptually they are equivalent to the simple model we assume here. The rate matrix of our model is parameterised by α , the rate of change from one particular nucleotide to another, and has the following form:

		To			
		A	C	G	T
From	A	-3 α	α	α	α
	C	α	-3 α	α	α
	G	α	α	-3 α	α
	T	α	α	α	-3 α

The trajectory of a process determined by Q starting at time t in state N_1 (for instance A) could look like this



The probability that we have to wait more than time T from the change to $N(i-1)$ until it again changes to N_i is $P\{t_i - t_{i-1} > T\} = e^{-3\alpha T}$, i.e. exponentially

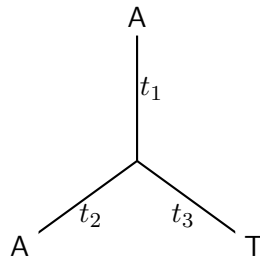
distributed with intensity 3α . The expected number of events (substitutions) in a time interval of length t is $3\alpha t$. Let $P(t)$ denote the matrix of probabilities of change over time t , i.e. the (i, j) entry in $P(t)$, $P_{i,j}(t)$, is the probability that nucleotide i has changed into nucleotide j after time t . We know that $P(t) = e^{tQ} = I + tQ + t^2Q^2/2 + \dots + t^kQ^k/k! + \dots$ (where $I = Q^0$ is identity matrix). Note that if t is small then $I + tQ$ is a good approximation to $P(t)$, which corresponds to all substitutions being observable because there is at most one event in the evolutionary trajectory. For this simple model of substitution, one can realise that $Q^i = (-4)^{i-1}$ for $i \geq 1$. With a bit of rearrangements of the exponential expansion, this yields

$$P_{i,j}(t) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-4\alpha t} & \text{if } i = j \\ \frac{1}{4} - \frac{1}{4}e^{-4\alpha t} & \text{if } i \neq j \end{cases}$$

by converting the exponential expansion on matrix powers to an exponential expansion on a real number.

- a. What is the probability of observing (A, A, T) in an alignment column

With only three observed sequences there is only one possible tree topology relating them, namely



as time reversibility allows us to place the root anywhere we want to without changing the data probability. If we denote the time from the node at the centre of the tree to each of these sequences by t_i for $i \in \{1, 2, 3\}$ as shown in the illustration, we need to sum over all possible nucleotides we could have at the centre the probability that

this nucleotide evolved into the three observed nucleotides. This is

$$\begin{aligned}
& \sum_{c \in \{A, C, G, T\}} \pi_c P(t_1)_{c, A} P(t_2)_{c, A} P(t_3)_{c, T} \\
&= \frac{1}{4} \left(\left(\frac{1}{4} + \frac{3}{4} e^{-4\alpha t_1} \right) \left(\frac{1}{4} + \frac{3}{4} e^{-4\alpha t_2} \right) \left(\frac{1}{4} - \frac{1}{4} e^{-4\alpha t_3} \right) \right. \\
&\quad + \left(\frac{1}{4} - \frac{1}{4} e^{-4\alpha t_1} \right) \left(\frac{1}{4} - \frac{1}{4} e^{-4\alpha t_2} \right) \left(\frac{1}{4} - \frac{1}{4} e^{-4\alpha t_3} \right) \\
&\quad + \left(\frac{1}{4} - \frac{1}{4} e^{-4\alpha t_1} \right) \left(\frac{1}{4} - \frac{1}{4} e^{-4\alpha t_2} \right) \left(\frac{1}{4} - \frac{1}{4} e^{-4\alpha t_3} \right) \\
&\quad \left. + \left(\frac{1}{4} - \frac{1}{4} e^{-4\alpha t_1} \right) \left(\frac{1}{4} - \frac{1}{4} e^{-4\alpha t_2} \right) \left(\frac{1}{4} + \frac{3}{4} e^{-4\alpha t_3} \right) \right) \\
&= \frac{1}{256} \left((1 + 3e^{-4\alpha t_1}) (1 + 3e^{-4\alpha t_2}) (1 - e^{-4\alpha t_3}) \right. \\
&\quad + 2(1 - e^{-4\alpha t_1}) (1 - e^{-4\alpha t_2}) (1 - e^{-4\alpha t_3}) \\
&\quad \left. + (1 - e^{-4\alpha t_1}) (1 - e^{-4\alpha t_2}) (1 + 3e^{-4\alpha t_3}) \right) \\
&= \frac{1}{64} \left(1 + 3e^{-4\alpha(t_1+t_2)} - e^{-4\alpha(t_1+t_3)} - e^{-4\alpha(t_2+t_3)} - 2e^{-4\alpha(t_1+t_2+t_3)} \right)
\end{aligned}$$

- b. If there on average is 10^{-8} substitutions/(position \times year), how many events would you expect in 1000 base pair long sequences that had a common ancestor 5 million years ago?

Observe that the total time separating two sequences that had a common ancestor 5 million years ago is 10 million years. We now get 10^3 positions $\times 10^7$ years $\times 10^{-8}$ substitutions/(position \times year) = 100 substitutions.

How many differences would you expect to observe between the two sequences (assuming they are only affected by the substitution process, and not for example insertions and deletions)?

The probability that we observe different nucleotides in the same position is $\frac{3}{4} (1 - e^{-4\alpha t})$. The number of expected events in time t is $3\alpha t$, so in this case $\alpha t = \frac{1}{3} 10^7$ years $\times 10^{-8}$ substitutions/(position \times year) = $10^{-1}/3$ substitutions/position so we get 93.62 expected positions with observed differences.

What would your answers be if they had a common ancestor 50 million years ago? How would the approximation $I + Qt$ be in the two cases?

With ten times as large a t we would expect to see ten times as many events, *i.e.* 1000 substitutions. However, we would only expect to observe differences in 552.30 of the 1000 positions. Clearly the approximation of assuming all events are observable starts to break down at this time scale, as we expect to observe only slightly more than half the events. The probability that no event occurs in a position in 100 million years is 0.368, so of the expected 447.70 seemingly unchanged positions we would expect 79.82 to have experienced substitutions before reverting back to the original nucleotide.

- c. Kimura's 2 parameter substitution model is a slightly more realistic model than Jukes–Cantor, allowing different substitution rates between transitions (when an A changes to a G or vice versa, or when a C changes to a T or vice versa) and transversions (all other changes). The rate matrix can be written as

		To			
		A	G	C	T
From	A	$-\alpha - 2$	α	1	1
	G	α	$-\alpha - 2$	1	1
	C	1	1	$-\alpha - 2$	α
	T	1	1	α	$-\alpha - 2$

where the four nucleotides have been sorted so we first have the purines and then the pyrimidines. It may look like this only has a single parameter, α , but as we cannot separate rates and time we have chosen a unit of time corresponding to one transversion to simplify the matrix – if transversion rate was β , we can divide all matrix entries by β if we just use βt as time instead of t and let α denote the relative difference between transition and transversion rate.

This matrix is slightly harder to exponentiate by finding regularities in the powers of Q , compared to the Jukes–Cantor model. However, it can still be done. Calculate Q^2 , Q^3 , Q^k (Hint: try to write the entries depending on α as a power of $(\alpha + 1)$ and a term not depending on α) and then $P(t)$.

$$\begin{aligned}
Q^2 &= \begin{bmatrix} (-2-\alpha)^2 + \alpha^2 + 2 & 2(-2-\alpha)\alpha + 2 & 2(-2-\alpha+\alpha) & 2(-2-\alpha+\alpha) \\ 2(-2-\alpha)\alpha + 2 & (-2-\alpha)^2 + \alpha^2 + 2 & 2(-2-\alpha+\alpha) & 2(-2-\alpha+\alpha) \\ 2(-2-\alpha+\alpha) & 2(-2-\alpha+\alpha) & (-2-\alpha)^2 + \alpha^2 + 2 & 2(-2-\alpha)\alpha + 2 \\ 2(-2-\alpha+\alpha) & 2(-2-\alpha+\alpha) & 2(-2-\alpha)\alpha + 2 & (-2-\alpha)^2 + \alpha^2 + 2 \end{bmatrix} \\
&= \begin{bmatrix} 2\alpha^2 + 4\alpha + 6 & -2\alpha^2 - 4\alpha + 2 & -4 & -4 \\ -2\alpha^2 - 4\alpha + 2 & 2\alpha^2 + 4\alpha + 6 & -4 & -4 \\ -4 & -4 & 2\alpha^2 + 4\alpha + 6 & -2\alpha^2 - 4\alpha + 2 \\ -4 & -4 & -2\alpha^2 - 4\alpha + 2 & 2\alpha^2 + 4\alpha + 6 \end{bmatrix} \\
&= \begin{bmatrix} -(-4) - (-2)(\alpha+1)^2 & -(-4) + (-2)(\alpha+1)^2 & -4 & -4 \\ -(-4) + (-2)(\alpha+1)^2 & -(-4) - (-2)(\alpha+1)^2 & -4 & -4 \\ -4 & -4 & -(-4) - (-2)(\alpha+1)^2 & -(-4) + (-2)(\alpha+1)^2 \\ -4 & -4 & -(-4) + (-2)(\alpha+1)^2 & -(-4) - (-2)(\alpha+1)^2 \end{bmatrix}
\end{aligned}$$

$$\begin{aligned}
Q^3 &= \begin{bmatrix} -4\alpha^3 - 12\alpha^2 - 12\alpha - 20 & 4\alpha^3 + 12\alpha^2 + 12\alpha - 12 & 16 & 16 \\ 4\alpha^3 + 12\alpha^2 + 12\alpha - 12 & -4\alpha^3 - 12\alpha^2 - 12\alpha - 20 & 16 & 16 \\ 16 & 16 & -4\alpha^3 - 12\alpha^2 - 12\alpha - 20 & 4\alpha^3 + 12\alpha^2 + 12\alpha - 12 \\ 16 & 16 & 4\alpha^3 + 12\alpha^2 + 12\alpha - 12 & -4\alpha^3 - 12\alpha^2 - 12\alpha - 20 \end{bmatrix} \\
&= \begin{bmatrix} -(-4)^2 - (-2)^2(\alpha+1)^3 & -(-4)^2 + (-2)^2(\alpha+1)^3 & (-4)^2 & (-4)^2 \\ -(-4)^2 + (-2)^2(\alpha+1)^3 & -(-4)^2 - (-2)^2(\alpha+1)^3 & (-4)^2 & (-4)^2 \\ (-4)^2 & (-4)^2 & -(-4)^2 - (-2)^2(\alpha+1)^3 & -(-4)^2 + (-2)^2(\alpha+1)^3 \\ (-4)^2 & (-4)^2 & -(-4)^2 + (-2)^2(\alpha+1)^3 & -(-4)^2 - (-2)^2(\alpha+1)^3 \end{bmatrix}
\end{aligned}$$

$$Q^k = \begin{bmatrix} -(-4)^{k-1} - (-2)^{k-1}(\alpha+1)^k & -(-4)^{k-1} + (-2)^{k-1}(\alpha+1)^k & (-4)^{k-1} & (-4)^{k-1} \\ -(-4)^{k-1} + (-2)^{k-1}(\alpha+1)^k & -(-4)^{k-1} - (-2)^{k-1}(\alpha+1)^k & (-4)^{k-1} & (-4)^{k-1} \\ (-4)^{k-1} & (-4)^{k-1} & -(-4)^{k-1} - (-2)^{k-1}(\alpha+1)^k & -(-4)^{k-1} + (-2)^{k-1}(\alpha+1)^k \\ (-4)^{k-1} & (-4)^{k-1} & -(-4)^{k-1} + (-2)^{k-1}(\alpha+1)^k & -(-4)^{k-1} - (-2)^{k-1}(\alpha+1)^k \end{bmatrix}$$

That the expression for Q^k is indeed correct can be easily, but tediously, proved by induction. It holds for all $k > 1$, but it has to be remembered that $Q^0 = \mathbf{I}$. Hence, for the exponential expansion $P(t) = e^{Qt} = \sum_{i=0}^{\infty} \frac{(Qt)^i}{i!} = \mathbf{I} + \sum_{i=1}^{\infty} \frac{(Qt)^i}{i!}$ we need to treat the first term specially. For the remaining terms, however, we can deal with the sum entry by entry, with

$$\begin{aligned}
\sum_{i=1}^{\infty} \frac{(-4)^{i-1} t^i}{i!} &= \frac{-1}{4} \left(\sum_{i=0}^{\infty} \frac{(-4t)^i}{i!} - 1 \right) = \frac{1}{4} - \frac{1}{4} e^{-4t} \\
\sum_{i=1}^{\infty} \frac{(-2)^{i-1} ((\alpha+1)t)^i}{i!} &= \frac{-1}{2} \left(\sum_{i=0}^{\infty} \frac{(-2(\alpha+1)t)^i}{i!} - 1 \right) = \frac{1}{2} - \frac{1}{2} e^{-2(\alpha+1)t}
\end{aligned}$$

Bringing it all together, we get the transition probability matrix

$$P(t) = \begin{bmatrix} \frac{1}{4} + \frac{1}{4}e^{-4t} + \frac{1}{2}e^{-2(\alpha+1)t} & \frac{1}{4} + \frac{1}{4}e^{-4t} - \frac{1}{2}e^{-2(\alpha+1)t} & \frac{1}{4} - \frac{1}{4}e^{-4t} & \frac{1}{4} - \frac{1}{4}e^{-4t} \\ \frac{1}{4} + \frac{1}{4}e^{-4t} - \frac{1}{2}e^{-2(\alpha+1)t} & \frac{1}{4} + \frac{1}{4}e^{-4t} + \frac{1}{2}e^{-2(\alpha+1)t} & \frac{1}{4} - \frac{1}{4}e^{-4t} & \frac{1}{4} - \frac{1}{4}e^{-4t} \\ \frac{1}{4} - \frac{1}{4}e^{-4t} & \frac{1}{4} - \frac{1}{4}e^{-4t} & \frac{1}{4} + \frac{1}{4}e^{-4t} + \frac{1}{2}e^{-2(\alpha+1)t} & \frac{1}{4} + \frac{1}{4}e^{-4t} - \frac{1}{2}e^{-2(\alpha+1)t} \\ \frac{1}{4} - \frac{1}{4}e^{-4t} & \frac{1}{4} - \frac{1}{4}e^{-4t} & \frac{1}{4} + \frac{1}{4}e^{-4t} - \frac{1}{2}e^{-2(\alpha+1)t} & \frac{1}{4} + \frac{1}{4}e^{-4t} + \frac{1}{2}e^{-2(\alpha+1)t} \end{bmatrix}$$

It is an easy check that $P(t) = \mathbf{I}$ and $P'(0) = Q$, as they should be for a correct solution, for this matrix.

B From probabilities to rates

Assume we have observed two sequences that have evolved from a common ancestor under just the Jukes-Cantor model of nucleotide substitution, i.e. we know they are correctly aligned as

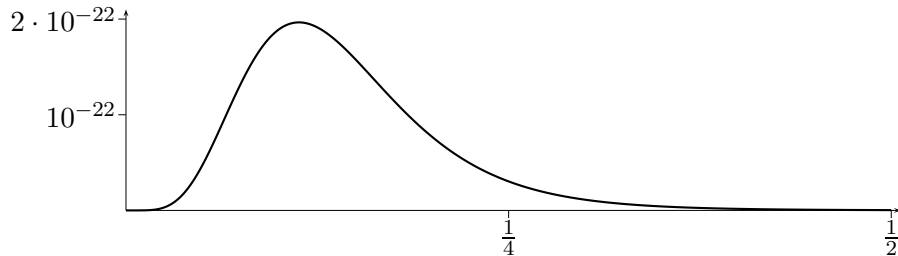
```
ACGTTGACCTCAAATTTGCTCT
ACGGTGACGTCACAAATGCACT
```

- d. Write (plot if you can) likelihood function of this alignment as a function of αt , assuming that different positions evolve independently.

We don't need to worry about the actual content of each alignment column, only whether they are the same or different. This follows as the probability of observing a column with identical nucleotides is $\frac{1}{4} + \frac{3}{4}e^{-4\alpha t}$ and the probability of observing a column with different nucleotides is $\frac{1}{4} - \frac{1}{4}e^{-4\alpha t}$ under the Jukes-Cantor model, regardless of the actual nucleotides. There are sixteen columns with identical nucleotides and six columns with different nucleotides, so the probability of the alignment is

$$L(\alpha t) = \left(\frac{1}{4} \left(\frac{1}{4} + \frac{3}{4}e^{-4\alpha t} \right) \right)^{16} \left(\frac{1}{4} \left(\frac{1}{4} - \frac{1}{4}e^{-4\alpha t} \right) \right)^6,$$

when we remember to also include the probability under the stationary distribution of the ancestral sequence. Plotting this for αt in the range from 0 to $1/2$ we get



e. What is the αt that makes the alignment most likely?

$$\begin{aligned}\frac{d}{d(\alpha t)}L(\alpha t) &= -48e^{-4\alpha t} \left(\frac{1}{4} + \frac{3}{4}e^{-4\alpha t}\right)^{15} \left(\frac{1}{4} - \frac{1}{4}e^{-4\alpha t}\right)^6 \\ &\quad + 6e^{-4\alpha t} \left(\frac{1}{4} + \frac{3}{4}e^{-4\alpha t}\right)^{16} \left(\frac{1}{4} - \frac{1}{4}e^{-4\alpha t}\right)^5 \\ &= 6e^{-4\alpha t} \left(\frac{1}{4} + \frac{3}{4}e^{-4\alpha t}\right)^{15} \left(\frac{1}{4} - \frac{1}{4}e^{-4\alpha t}\right)^5 \left(\frac{1}{4} + \frac{3}{4}e^{-4\alpha t} - 2 + 2e^{-4\alpha t}\right)\end{aligned}$$

From this it is relatively easy to deduce that the maximum likelihood estimate of αt is $\frac{\ln(11/7)}{4}$.