

# MS2a, Exercises Week 6, Model Solution

Rune Lyngsø

November 19, 2009

## A Probabilistic Alignment

- a. Assume the TKF91 model of sequence evolution with nucleotide substitution described by the Jukes-Cantor single parameter model. Let parameters be  $st = 0.2$ ,  $\mu t = 0.1$ , and  $\lambda t = 0.09$ . What is the likelihood of observing homologous sequences  $s1 = AG$  and  $s2 = G$ ?

At first sight it may appear that we are missing information if we are to apply the equations in [1] (observe there is a typo in the expression for  $\beta(t)$  as the  $t$  in the denominator should be part of the exponent – this is correct in the lecture overheads), as we do not have values of  $\lambda$  and  $\mu$  separated from  $t$ . And *e.g.*  $\beta(t)$  will vary with a reciprocal rescaling of birth/death parameters vs. time as  $\lambda$  and  $\mu$  appear in the denominator without being multiplied by  $t$ . So simultaneously increasing  $\lambda$  and  $\mu$  by a factor of  $x$  while decreasing  $t$  with the same factor will not change  $\lambda t$  and  $\mu t$ , but will decrease  $\beta(t)$  by a factor of  $x$ . However, whenever we use  $\beta(t)$  in the equations it is multiplied by either  $\lambda$  or  $\mu$ , cancelling out this effect. Moreover, the only other places where  $\lambda$  and  $\mu$  occur unmultiplied by  $t$  are as multiplicands of  $\beta(t)$ . Hence, the end result is independent of a reciprocal rescaling. This shouldn't surprise as you should be used to inseparability of time and rates by now.

There are five possible alignments postulating  $s2$  as a descendant of  $s1$ :

A G	A G	– A G	A – G	A G –
G –	– G	G – –	– G –	– – G

The first two have an immortal link with zero descendants, one surviving ancestral character with no further descendants, and one ancestral character that died with no surviving descendants. The first alignment has the surviving ancestral character change, while the second alignment has no observed change to the surviving ancestral character. This

gives probabilities

$$p_1 = p_0^I(t)p_1^H(t)p_0^N(t) \left( \frac{1}{4} - \frac{1}{4}e^{-4st} \right) \approx 9.58 \cdot 10^{-3}$$

$$p_2 = p_0^I(t)p_1^H(t)p_0^N(t) \left( \frac{1}{4} + \frac{3}{4}e^{-4st} \right) \approx 4.09 \cdot 10^{-2}$$

The last two alignments have an immortal link with zero descendants and two ancestral characters that dies, one with no descendants and one with one descendant. This gives probabilities

$$p_4 = p_5 = p_0^I(t)p_1^N(t)p_0^N(t)\pi_G \approx 7.38 \cdot 10^{-5}$$

Finally, the middle alignment has one descendant from the immortal link and two ancestral characters that died with no descendants, which gives probability

$$p_3 = p_1^I(t)\pi_G p_0^N(t)^2 \approx 1.57 \cdot 10^{-4}$$

All these probabilities should be multiplied with the probability of observing ancestral sequence AG, which is

$$p_{AG} = q_2\pi_A\pi_G = \frac{81}{16,000} \approx 5.06 \cdot 10^{-3}$$

The total probability of observing these two sequences as homologous sequences is thus

$$p_{AG}(p_1 + p_2 + p_3 + p_4 + p_5) \approx 2.57 \cdot 10^{-4}$$

- b. What is the probability of the most probable alignment of these two sequences?
- c. What is the most probable alignment?

The highest of the probabilities is  $p_2$ , so the most probable alignment is

A G  
- G

which has probability

$$p_2 p_{AG} \approx 2.07 \cdot 10^{-4}$$

- d. What is the probability of observing  $s_1$  and  $s_2$  as non-homologous sequences, *i.e.* assuming they are not descendants from the same ancestral sequence?

The probability of observing the sequences as non-homologous is

$$p_{AG}p_G = q_2\pi_A\pi_Gq_1\pi_G = \frac{729}{6,400,000} \approx 1.14 \cdot 10^{-4}$$

- e. The TKF91 model can be viewed as a composition of two models, an insertion/deletion process that defines a distribution over alignment structures, and a substitution process that defines a distribution over the sequences observed in the alignment. Ignoring the sequence content and just focusing on the alignment structure, write up the probability expressions for the two alignment structures

$$\begin{array}{cc} \# - & - \# \\ - \# & \# - \end{array}$$

assuming that the top sequence is the ancestor and the bottom sequence the descendant. The  $\#$  character is known as a Felsenstein wildcard and indicates a marginalisation over all possible characters, as in Felsenstein's tree peeling algorithm.

In the first alignment we have no descendants from the immortal link, and one ancestral character that did not survive but left one descendant. Hence, we get probability

$$p_1 = p_0^I(t)p_1^N(t) = I(t)N(t) = (1 - e^{-\mu t} - \mu\beta(t))(1 - \lambda\beta(t))^2$$

for this alignment. In the second alignment, we have one descendant from the immortal link, and one ancestral character that died with no descendants. This gives probability

$$p_2 = p_1^I(t)p_0^N(t) = I(t)B(t)E(t) = \mu\lambda\beta(t)^2(1 - \lambda\beta(t))$$

If the ancestral sequence length is considered part of the alignment structure, each of these probabilities should be multiplied by  $q_2 = (1 - \frac{\lambda}{\mu})\frac{\lambda}{\mu}$ .

What are the probabilities as  $t \rightarrow \infty$ ?

It is a prerequisite for the model to be meaningful that  $\lambda < \mu \Leftrightarrow \lambda - \mu < 0$ . We now get

$$\lim_{t \rightarrow \infty} \beta(t) = \lim_{t \rightarrow \infty} \frac{1 - e^{(\lambda - \mu)t}}{\mu - \lambda e^{(\lambda - \mu)t}} = \frac{1}{\mu}.$$

Using this, we now get

$$\lim_{t \rightarrow \infty} p_1 = \lim_{t \rightarrow \infty} (1 - e^{-\mu t} - \mu\beta(t))(1 - \lambda\beta(t))^2 = (1 - 0 - 1) \left(1 - \frac{\lambda}{\mu}\right)^2 = 0$$
$$\lim_{t \rightarrow \infty} p_2 = \lim_{t \rightarrow \infty} \mu\lambda\beta(t)^2(1 - \lambda\beta(t)) = \frac{\lambda}{\mu} \left(1 - \frac{\lambda}{\mu}\right).$$

What would you expect for the two alignments in a time reversible model?  
Can you explain this phenomenon?

We are using the time reversibility to place the root of the tree relating the two sequences at an arbitrary point, in particular allowing us to choose one of the observed sequences as the ancestral sequence. The two alignment structures corresponds to the same alignment of two sequences, just with the choice of which is the ancestor and which is the descendant toggled. So it would seem fair to expect the two alignment structures to have the same probability. However, the probability of the first tends to 0 as  $t \rightarrow \infty$  while the other tends to a non-zero value (assuming  $0 < \lambda < \mu$ ). Hence, the two alignment structures do not have the same probability, and our choice of ancestor may affect which alignment is the most probable.

This shouldn't really surprise us. As time progresses, there is an increasing probability, in the limit of  $t \rightarrow \infty$  tending to 1, that any ancestral character has died. The immortal link, by nature, never dies. So as time increases, there will be a stronger and stronger bias towards piling up insertions at the beginning of the alignment. The model does not have detailed time reversibility, as individual evolutionary trajectories may have different probabilities depending on the direction of time. However, it does have overall time reversibility, as the total probability of the observation, summing over all evolutionary trajectories, does not depend on the direction of time. For reasonable values of  $t$  and sizable observations, the bias towards having insertions at the beginning of the alignment will moreover be defeated by the preference for matching identical characters in the substitution model. For large  $t$ , it may make sense to check how much the alignment depends on the choice of ancestral sequences, though.

## B RNA Secondary Structure

- a. A string is a palindrome if it reads the same from left to right as from right to left. So the name of the Swedish pop group ABBA is a palindrome,

but ABAB is not. Any single character word is a palindrome. What is the longest palindrome you can find in the sequence

ACGAGTGCGCATTCTCAAAACACCGGCCACTATCACCGGCCACCACCGGCCACTATGACTCCATTACTC

The longest palindrome is indicated above, and is of length 26.

- b. What is the fewest number of palindromes you can split the above sequence into? The palindromes, when joined together should spell out exactly the above sequence, e.g. ABA, C, A would be a valid split of ABACA, but ABA, ACA would not.

One cover using 18 palindromes is

A C GAG T GCG C A TT CTC AAAA CACCGGCCAC TATCACCGGCCACCACCGGCCACTAT G A CTC CATTAC T C

- c. How could you systematically determine the fewest palindromes a string can be split into?

One approach would be to for a position to try all possible of the last palindrome in the cover. The minimum number of palindromes needed to cover the sequence up to a position would be one plus the minimum number of palindromes needed to cover the remaining part of the sequence for all choices. We can write this as a recursion in  $C(i)$ , the minimum number of palindromes needed to cover the sequence up to position  $i$ :

$$C(i) = \min\{1 + C(j - 1) \mid 1 \leq j \leq i \wedge s[j..i] \text{ a palindrome}\}$$

Initial condition is  $C(0) = 0$ .

What if the two halves of even length palindromes were allowed to occur non-contiguously, e.g. AB...BA, A, C would be a valid split of the string ABACAB.

To be able to do this in a systematic way, we need to be able to check the minimum number of palindromes needed to cover the sequence between the two non-contiguous parts of an even length palindrome. Hence, we need a recursion in  $C(i, j)$ , the minimum number of palindromes needed to cover the sequence from position  $i$  to position  $j$ :

$$C(i, j) = \min\{\min\{1 + C(i, k - 1) \mid i \leq k \leq j \wedge s[k..j] \text{ a palindrome}\}, \min\{1 + C[i+k+1, j-k-1] \mid 0 \leq k \leq \frac{j-i-1}{2} \wedge s[i..i+k] \cdot s[j-k..j] \text{ a palindrome}\}\}$$

Initial condition is  $C(i, i - 1) = 0$  for  $1 \leq i \leq |s|$  and  $s \cdot t$  denotes concatenation of strings  $s$  and  $t$ .

## References

- [1] I. Miklós, Ádám Novák, R. Satija, R. Lyngsø, and J. Hein. Stochastic models of sequence evolution including insertion-deletion events. *Statistical Methods in Medical Research*, 2009. Accepted.