

# MS2a, Exercises Week 6

Rune Lyngsø

November 12, 2009

## A Probabilistic Alignment

- 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  $s_1 = AG$  and  $s_2 = G$ ?
- What is the probability of the most probable alignment of these two sequences?
- What is the most probable alignment?
- What is the probability of observing  $s_1$  and  $s_2$  as non-homologous sequences, *i.e.* assuming they are not descendents from the same ancestral sequence?
- 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

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

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. What are the probabilities as  $t \rightarrow \infty$ ? What would you expect for the two alignments in a time reversible model? Can you explain this phenomenon?

## 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

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