

# MS2a, Exercises Week 4

Rune Lyngsø

November 3, 2011

## A Score Based Alignment

Define a similarity score  $w$  on the four nucleotides such that

$$w(X, Y) = \begin{cases} 10 & \text{if } X = Y \\ 2 & \text{if } X \neq Y \text{ but } X \text{ can be changed to } Y \text{ by a transition} \\ 0 & \text{otherwise} \end{cases}$$

Furthermore, let an indel have a *dissimilarity* of  $g = 10$ .

To find the maximum 'similarity' between two sequences,  $s_1 = \text{CTAGGA}$  and  $s_2 = \text{TTGTG}$ , (taken over all possible alignments) you should use the recursion

$$S_{i,j} = \max \{S_{i-1,j-1} + w(s_1[i], s_2[j]), S_{i,j-1} - g, S_{i-1,j} - g\}$$

With initial conditions

$$S_{i,j} = \begin{cases} 0 & \text{if } i = j = 0 \\ -\infty & \text{if } i < 0 \text{ or } j < 0 \end{cases}$$

- a. Fill out the following table according to the recursion

G						
T						
G						
T						
T						
0	C	T	A	G	G	A



- e. Argue from the basic model (i.e., not by differentiating the expression above) that the following differential equation holds:

$$\begin{aligned} \frac{d}{dt} p \left( \begin{array}{c} \# \\ - \\ \# \\ \# \\ \dots \\ \# \\ k \end{array} \right) (t) = & (k-1) \lambda p \left( \begin{array}{c} \# \\ - \\ \# \\ \# \\ \dots \\ \# \\ k-1 \end{array} \right) (t) \\ & + (k+1) \mu p \left( \begin{array}{c} \# \\ - \\ \# \\ \# \\ \dots \\ \# \\ k+1 \end{array} \right) (t) \\ & + \mu p \left( \begin{array}{c} \# \\ \# \\ \# \\ \dots \\ \# \\ k+1 \end{array} \right) (t) \\ & - k(\mu + \lambda) p \left( \begin{array}{c} \# \\ - \\ \# \\ \# \\ \dots \\ \# \\ k \end{array} \right) (t) \end{aligned}$$

- f. Complete the following expression for the initial condition, i.e. what should go instead of the two question marks, and justify your answer.

$$p \left( \begin{array}{c} \# \\ - \\ \# \\ \# \\ \dots \\ \# \\ k \end{array} \right) (?) = ?$$

- g. 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  $s2 = G$ ?
- h. What is the probability of the most probable alignment of these two sequences?
- i. What is the most probable alignment?
- j. What is the probability of observing  $s1$  and  $s2$  as non-homologous sequences, i.e. assuming they are not descendents from the same ancestral sequence?
- k. 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. Ignore the sequence content and just focus on the alignment structure. Prove that the length equilibrium distribution is stationary (i.e., the distribution does not change over time).
- l. Still ignoring the sequence content, write up the probability expressions for the two alignment structures

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

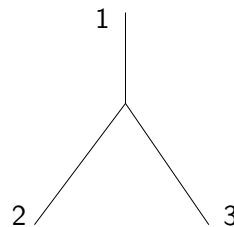
assuming that the top sequence is the ancestor and the bottom sequence the descendant. Here the # character denotes a Felsenstein wildcard, rather than an immortal link, and indicates a marginalisation over all possible characters, as in Felsenstein's tree peeling algorithm.

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?

- m. For two or more sequences, recursions exist that allow the calculation of the probability of the complete sequences as a function of the evolutionary parameters. For  $k$  sequences and a given phylogenetic relationship,  $T$ , these recursions can be viewed as hidden Markov models with the following set of hidden states: a start state,  $S$ , an end state,  $E$ , and the set of possible assignments of # (unspecified nucleotide) and - (absence of nucleotide) to the nodes of  $T$ , such that the nodes assigned # form a connected subset of the nodes of  $T$  (i.e., you can go from any of the #s to any other of the #s without stepping on a -).

The following diagram shows a small phylogeny relating 3 sequences.



In how many ways can we assign #s and -s to the nodes of this phylogeny (including the internal node) such that the #s form a (non-empty) connected set?

## References

- [1] I. Miklós, Ādám Novák, R. Satija, R. Lyngsø, and J. Hein. Stochastic models of sequence evolution including insertion-deletion events. *Statistical Methods in Medical Research*, 18(5):453–85, 2009.