

MS2a, Exercises Week 5

Rune Lyngsø

November 6, 2009

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

- b. What is the maximum similarity score between the two sequences s_1 and s_2 ?
- c. Find an alignment with this similarity score.
- d. Is the alignment you found unique, or are there more than one alignment achieving the maximum similarity score?

B Recombination

- a. Can we find a tree for the data set

Pan	TTATCC
Gorilla	TTGTTC
Pongo	CCACCC
Hylobates	CCGTTC

such that only one substitution is required in each position? If yes, provide such a tree. If no, why not?

- b. Compute the minimum number of substitutions required for the above data set for each of the three possible unrooted tree topologies, e.g. by using Fitch's algorithm (or just eye-balling it if you feel confident about doing this).
- c. Assume that apart from substitutions, you are also allowed events arbitrarily changing the tree topology between consecutive sites (this is a simplification of recombination events – recombination events only allow certain changes to tree topology). What is the minimum number of events you need to explain the above data set. Give an ancestral recombination graph explaining the data set with this number of events.
- d. How many recombination nodes are there in the ancestral recombination graph (ARG) you constructed in c? Can you construct a data set by permuting the columns in the above data set that requires more recombination nodes for any ARG explaining it? If yes, give an example.

How many different marginal trees does the ARG you constructed in c have (a marginal tree is the tree relating the species at a particular position)? Can you construct a data set by permuting the columns in the above data set that has more different marginal trees in any ARG explaining it? If yes, give an example.