Computational Biology HS 19 - Exercise 3
Answers to Theory Questions

Toby Law

5 November 2019

Question 1

What happens if you try to compute the K80 distance between two very dis-similar sequences? Take for instance the sequences AACTCA and TTAGTG.

For the example AACTCA and TTAGTG, all the positions in the alignment do not match, which makes the proportion of transitions (S) and transversions (V) add up to 1 (i.e, S+V = 1). When we try to compute the K80 distance ($\hat{d}$), log(1-2S-V) and log(1-2V) will give log(-S), and the log of any number $\leq$ 0 cannot be evaluated. How K80 distance is defined assumes that between two sequences, the proportion of transversions cannot be greater or equal to 0.5 and the proportion of transitions cannot be greater or equal to 0.25, so K80 distance cannot be used on too dissimilar sequences.

Question 2

Assume you are using the UPGMA algorithm, and your initial distance matrix has multiple equal minimal entries. Does your implementation of the algorithm influence the output tree?

In the event of a polytomy, where the distance between two nodes and another are the minimal entries (min(D) = d[S1,S2] = d[S1,S3]), because UPGMA does not allow cherries with more than two tips, the implementation of the UPGMA algorithm could yield two possible trees. But if the minimum entries are distances between 2 non-overlapping pairs of nodes (min(D) = d[S1,S2] = d[S3,S4]), no matter which is evaluated first, we would still get two equivalent trees, as the two pairs of nodes will be clustered at the same node height.

Question 3

Say we know that several sampled sequences have evolved under a JC69 model. We calculate a JC69 distance matrix for these sequences, apply the UPGMA algorithm to construct a tree, and calculate distances from the tree. However, our tree distances do not match the JC69 distances. Name two features of the sampling scheme or the true evolutionary process that might cause such a discrepancy. (2 points)

The UPGMA algorithm assumes that all sequences are sampled at the same point in time. This is reflected by ultrametric branches for the UPGMA tree, where branches of each clustered to each tip of a cherry is equidistant. It also assumes that sequences evolve according to a strict molecular clock, which is imposed by the fixed substitution rate α specified for the JC69 evolution model. This assumption is reflected by the length of the branches being proportional to genetic distances between species and calendar time. However, in reality evolution does not occur at a constant rate across the genome and between species, since mutation rates might differ across sites, and there might be different selective pressures on the phenotypic level. We also cannot tell if the sequence we are sampling are at the same time point or same evolutionary level. This would cause discrepancy between the tree distances and the JC69 distances.

Question 4

Can any of the potential problems you raised in your answer to question 3 be taken into account using an alternative algorithm mentioned in the lectures? If yes, which problem(s) and with which algorithm? If no, please give a short justification.

Assuming a fixed evolution rate across all sites in the genome leads to an underestimation of sequence distances. The JC69+γ model replaces the constant substitution rate λ with λR, where R is a γ-distributed random variable. This accounts for substitution rate variation across different sites in the genome, while the average substitution rate remains the same overall. To account for sequences being sampled at different time points in evolutionary time, we could use the neighbour-joining algorithm, which clusters pairs of nodes together like the UPGMA algorithm, but infer unrooted trees with branch lengths that are defined in number of substitutions, and does not have any proportionality to calendar time.