

Bioinformatics I & II

Wayne Delpont & Sergei Kosakovsky Pond

Practical 2: Saturation

In Practical 1 you would have written some code to generate a random sequence of length N and then a mutated version of that sequence, in which we set the probability of mutation to 0.05 . In that example, most of us assumed that there was a probability of a mutation occurring at each site along the sequence and implemented a mutation model which considered each site in turn and decided whether that site was mutated, or not. If mutated, we changed the nucleotide, if not we left the nucleotide as is. This is analogous to the mutation process, but ignores one important property. Mutations can occur at random across the sequence, such that one site may exhibit multiple mutations (*e.g.* from A to C to G). In this case, we know the state A in the first sequence, and the state G, in the mutated sequence. However, we don't observe the intermediate state. We will, therefore, underestimate the number of mutations.

a) In this practical you will explore the degree to which these mutations are underestimated given the mutation rate. Modify your code from the first practical to the more realistic mutation process described above and determine the relationship between the observed and expected divergence. In your code you will need to generate a random sequence, generate a mutated sequence, keep track of how many mutations occurred (expected divergence), and estimate the number of mutations between the two sequences (observed divergence).

b) Jukes and Cantor proposed a correction for the genetic distance, d , between two sequences.

$$d = -3/4 \ln(1 - 4/3p) \quad (1)$$

where p is the observed genetic distance.

Does this correction resolve the problem above?