hw2_geneticDistanceMethods_substitutionModels

Daisy Fry-Brumit

2022-09-19

Background

For this exercise, you will be using a dataset that consists of 48 DNA sequences from the seasonal influenza (flu) virus.[1] These 48 samples were collected in the U.S. over 16 years (from 1993 - 2008), with 3 samples from each year. The samples are labeled by the year they were collected. The goals for the this exercise are to determine which substitution models best fit these data, and to use some classic phylogenetic methods in order to see if we can capture how the virus evolved over time.

Part 1: Align Sequences and Compare Models

Question 1 (5 pts)

The Bayesian Information Criterion (BIC) is a method for comparing models that includes penalties for more complex models. With this method, the best one is the model with the LOWEST BIC score. In your test, which model is this? Describe this model in terms of its assumptions. The lowest scoring model here is one of the Tamura-Nei models: TrN+G(4). The TrN models operate under the assumptions that

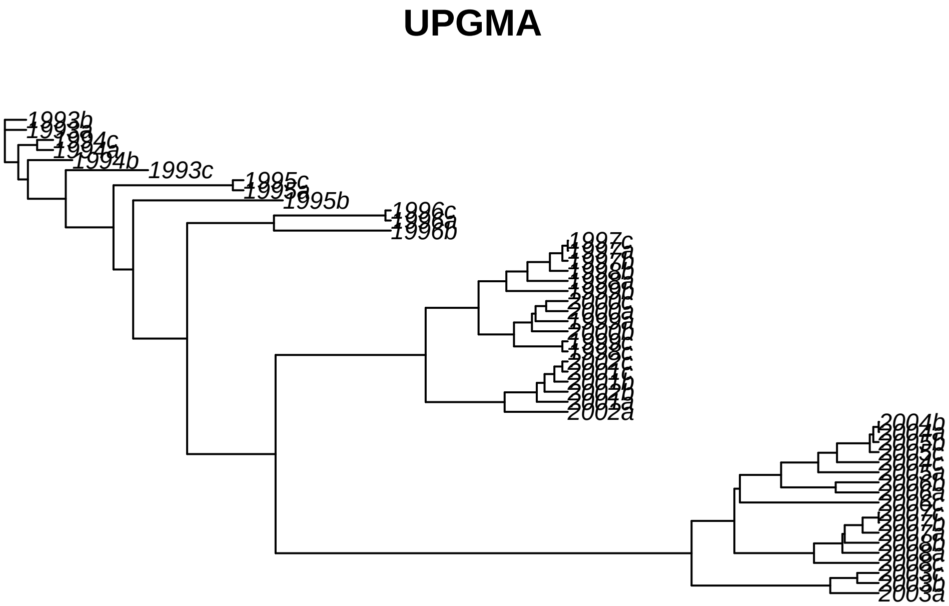
- base frequencies are unequal
- rates of subsittution are unequal between transitions and transversions
  - further, rates are unqual between the different types of transitions

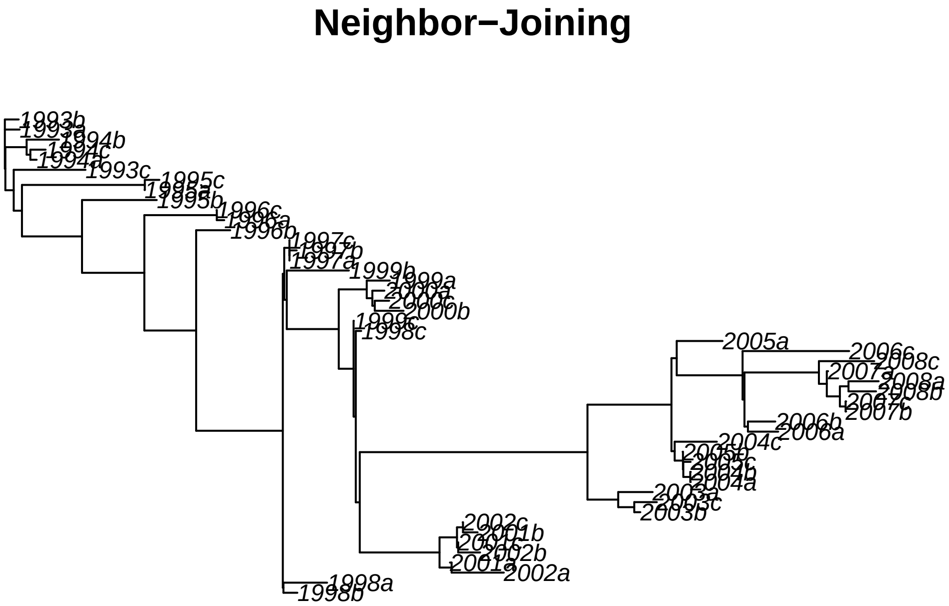
As such, it is a more complex model and includes 3 parameters (transition1, transition2, and transversion) to describe rates of substitution. The TrN + G(4) model specifically assumes that the rates of substitution follow a gamma(4) distribution through the length of the sequence.

Question 2 (1 pt)

What model did you use in your command, and which option did you use for gamma? I used TN93 (T92 does not include gamma correction) and applied gamma correction.

Part 2: Create and Compare Trees





Question 3 ()

Question 4 ()

Question 5 ()

Question 6 ()

Part 3: Find the Most Parsimonious Tree

Question 7 ()

Question 8 ()

Question 9 ()

References

[1] flu_seqs.fasta