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

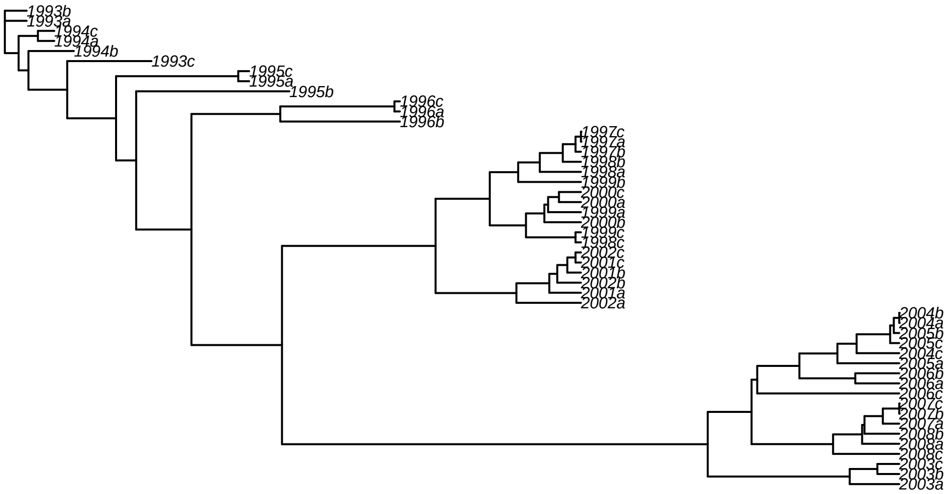
As such, it is a more complex model and includes 3 parameters (transition1, transition2, and transversion) to describe rates of substitution. The TrN + G(4) model specifically assumes that the rates of substitution follow a gamma(4) distribution through the length of the sequence.

Question 2 (1 pt)

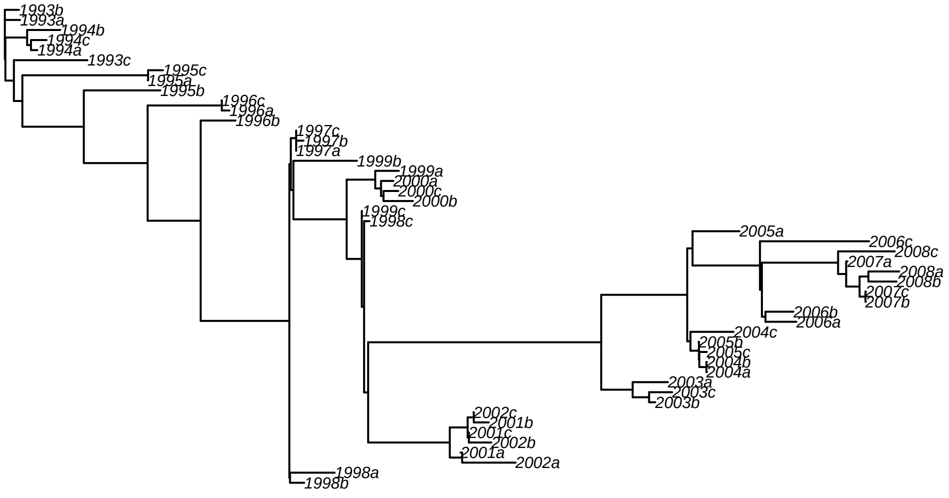
What model did you use in your command, and which option did you use for gamma? I used TN93 (T92 does not include gamma correction) and applied gamma correction.

Part 2: Create and Compare Trees

UPGMA



Neighbor−Joining



Question 3 (5pts)

Provide your plots for the rooted UPGMA and rooted Neighbor-Joining trees (be sure they are labeled so I can tell which is which!). Describe any differences you see in the overall appearance of the two trees. Do the same clades appear in each tree? Do viruses collected in later years branch off from the viruses of previous years (like we would expect)? UPGMA and NJ do mix the chronology/grouping between the viruses collected in 1994 and 1993, and 1998 with the early 2000's. Other than that, the clades appear in ways that do make chronological sense with regards to their collection times, especially with UPGMA output.

UPGMA outputs in a cleaner way that is easier for me to read, while the NJ method is harder to decipher at first glance. It appears that we have 2 separate instances of early 2000's clades resulting from 1999. NJ also jumps straight from 2003 to 2005 in one location. All in all, it strikes me as a messier tree that is harder to make sense of.

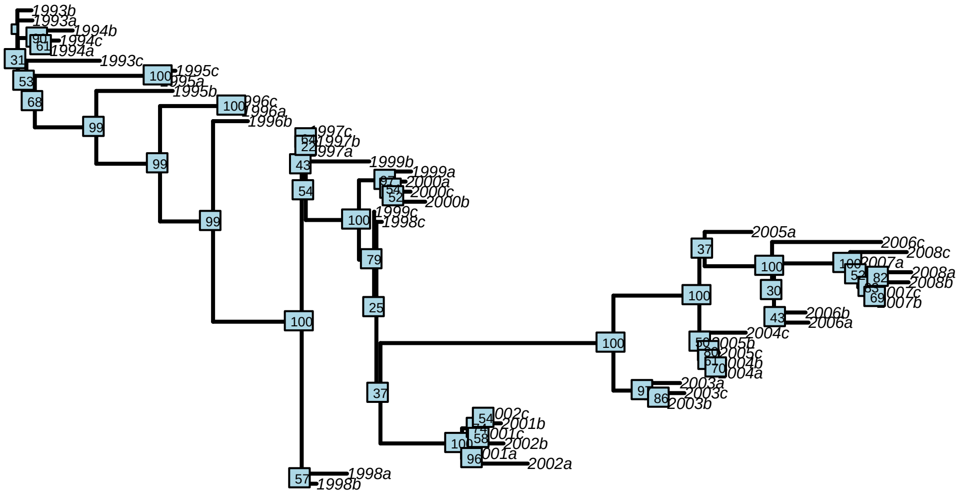
Question 4 (3pts)

What is the symmetric difference, branch score difference, and path difference between your two trees? Which number implies the biggest difference in the trees? Based on what you know about the definition of these metrics, why would one score be much higher than the others? Does your highest score make sense given what you see when you plot the two trees? See below for difference score outputs. As you can see, the path difference is much, much higher than our others metrics, implying a difference in not just the clades, but the paths taken to generate the clades in the trees. This makes sense as the inner nodes differ from the UPGMA and NJ generated trees (while Robinson-Foulds/symmetric difference is more concerned with final clade output) according to our plots.

Table 1: Tree Dist Output

	x
symmetric.difference	42.0000000
branch.score.difference	0.0205406
path.difference	135.0814569
quadratic.path.difference	0.3760019

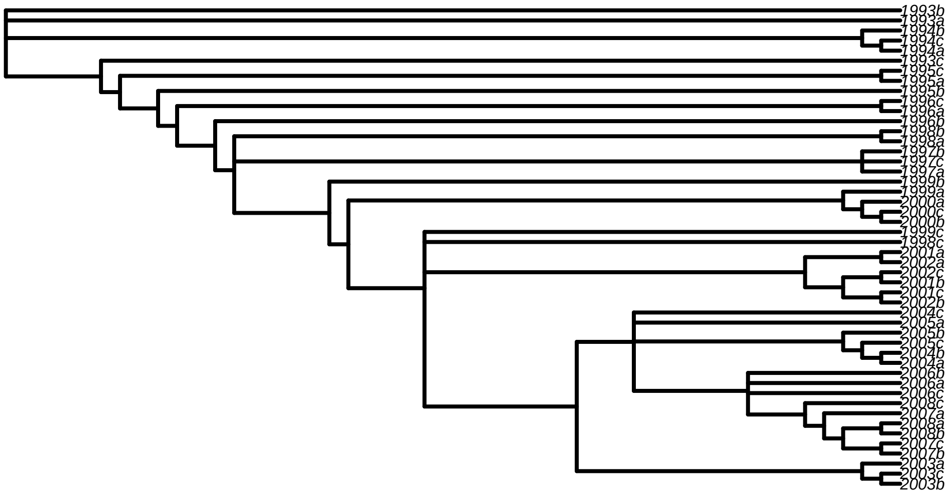
NJ W/ Bootstrapping Values



Question 5 (3pts)

Provide a plot of your bootstrapped tree. Are there any nodes in your tree that seem to have very weak bootstrap support (i.e. less than 50%)? How many of these weakly supported nodes do you see? Which node has the weakest support, and what clade is this node at the base of? If it is hard to see exactly where the nodes are with the bootstrap labels on them, you might also want to refer back to your original plot without the labels to help you There is one node early on in the tree (around 1944a and 1993c) with 41% bootstrapping support, two more in the 1977 clade with 25 and 37%, 40% that stands between 1998 and 2001+ groups, and 36% support that reults in a terminal branch leading to 2005a.

Majority Rule Consensus Tree



Question 6 (1pt)

Provide a plot of your majority rule consensus tree.

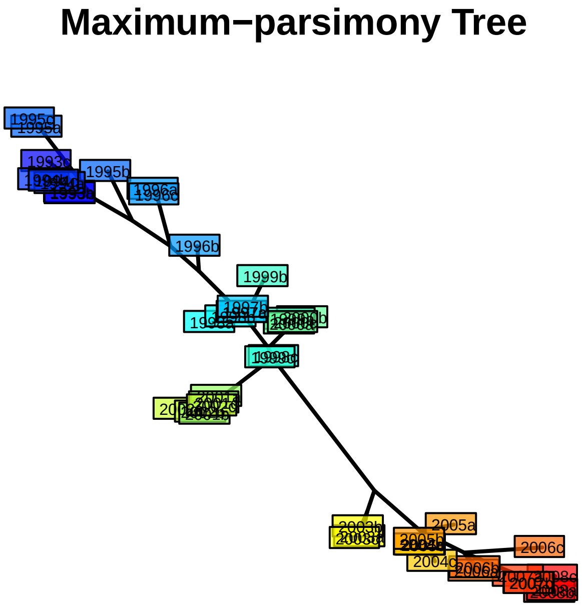
Part 3: Find the Most Parsimonious Tree

Question 7 (3pts)

What are your parsimony scores for your UPGMA and Neighbor-Joining trees? Given that a lower score is better, which of your trees is the most parsimonious? Based on what you know about the definition of parsimony, what does it mean when we say that one tree is more parsimonious than another? i.e. How should we interpret this result? UPGMA tree yields a score of 380, and nj yields 340, so they aren't too far apart but NJ performs better here. Basically this means that our neighbor joining generated tree can be explained in fewer evolutionary/mutant steps than our UPGMA generated tree. By some schools of thought, this might make the neighbor joining generated tree a better tree to accept, though there is some debate about that in the relevant communities.

Question 8 (1pt)

Were you able to improve on the parsimony score you got with the neighbor-joining tree? No!



Question 9 (3pts)

Provide the plot for your final tree. Even though there are still some uncertain relationships, can you now see any kind of trend with respect to the years the viruses were sampled and where they appear on the tree? Describe your new tree and the trends that you see. This tree does show a pretty cool and consistent trend that clades tend to form around general chronological points (based on when they were samples), with earlier years seen in cool colors at the top left of this figure (above), a slightly jumbled collection of points from 1997 to 2001 in greens and yellows in the center, and later collection years from 2003 and beyond in warm colors on the bottom right of the figure.

References

[1] flu_seqs.fasta