PART 1:

**Q1. Which model had the HIGHEST log likelihood value? Briefly describe this model, its assumptions in terms of base frequencies and equal or unequal substitution rates, and any other parameters it may include.**

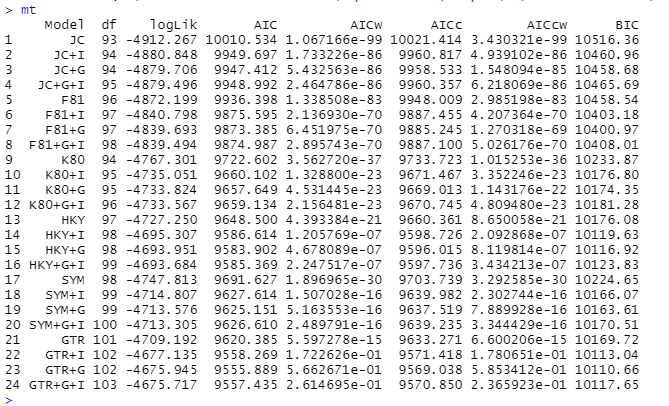


Fig: 1

As shown in the Fig:1, the model named GTR+G+I has the highest value of log likelihood (as -4675.717)

GTR = General Time-Reversible, G = Rate of Variation and I = Invariable Sites. This model is a time reversal model which means that a base can change to another base and then change back again to the original one, to exemplify: G -> T and from T -> G. It considers this substitution as 1 and makes all other 5 substitutions rates relative to the G-T transversion and 4 relative nucleotide frequencies. It assumes separate (unequal) rate parameters for each type of nucleotide substitution and assumes different frequency parameters for each nucleotide type. Here, the importance of I is that the nucleotides at some sites remain unchanged over time in all sequences.

**Q2. 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? Is it different from the model with the highest likelihood? Describe this model in terms of its assumptions.**

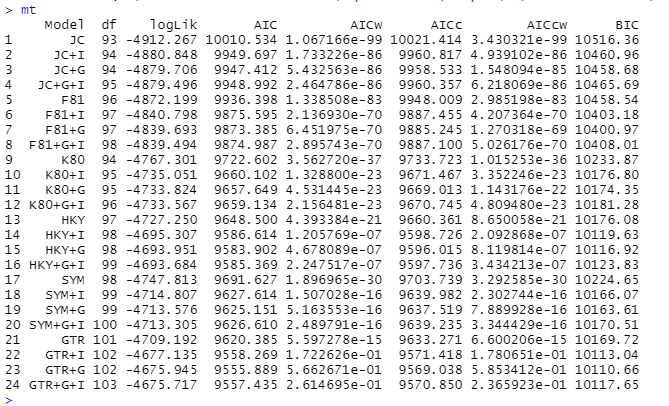


Fig: 1 (referring again)

A) For this, let us again refer to Fig:1. So, as we can see, the lowest BIC score is of the model named GTR + G, which means General Time-Reversible + Rate of Variation. So, here just like the answer 1, we refer to the model having variation in the rates with a reversibility feature but it has the variation in the sites unlike the previous model (GTR+G+I). So here the proportions and/or compositions of nucleotides differ while using this model.

PART 2:

**Q3. Describe any differences in clades that you observe between the UPGMA and NJ trees (be sure to include a plot of each tree in your answer!). In particular, are viruses from the same year always in the same clade in each tree? Do viruses collected in later years branch off from the viruses of previous years (like we would expect)?**

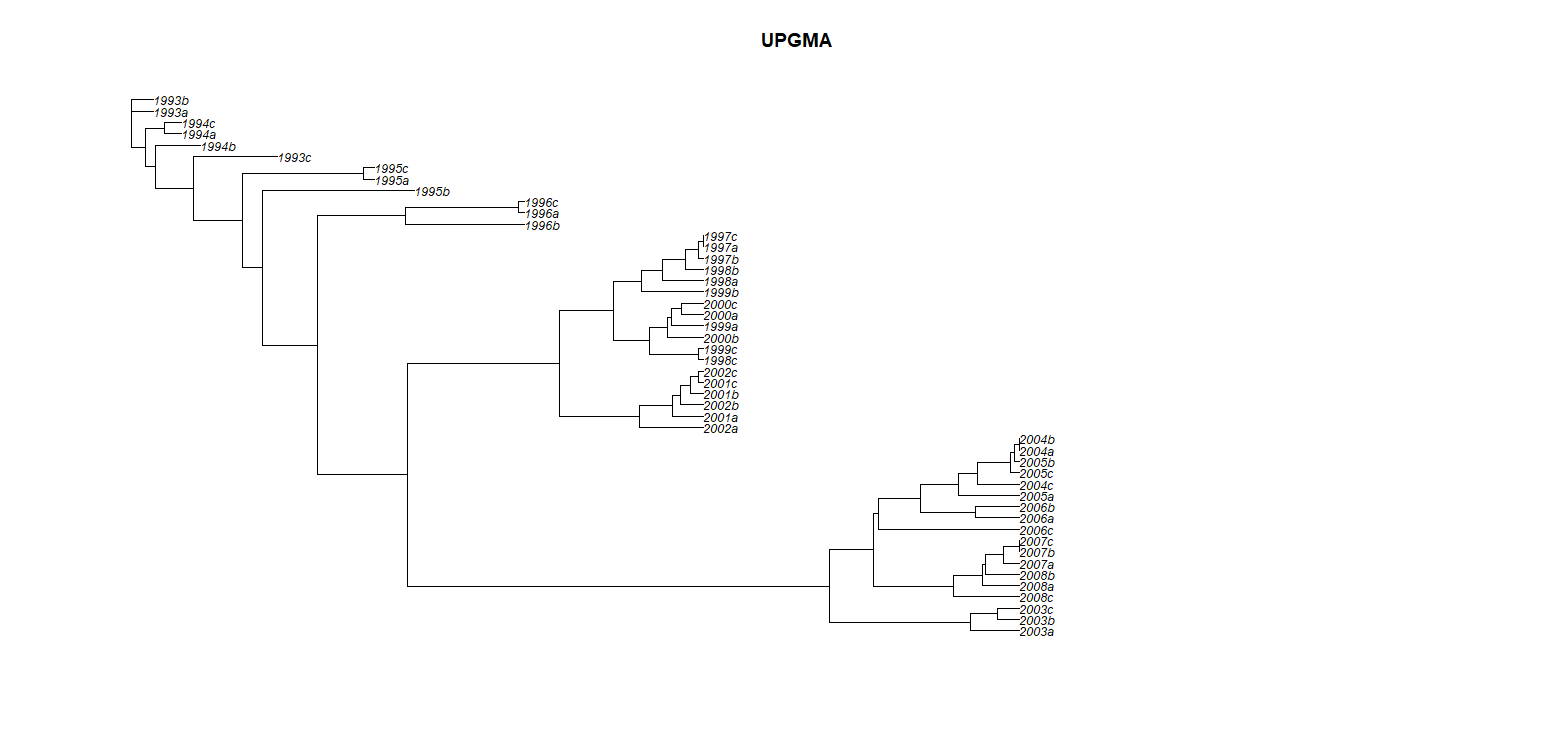
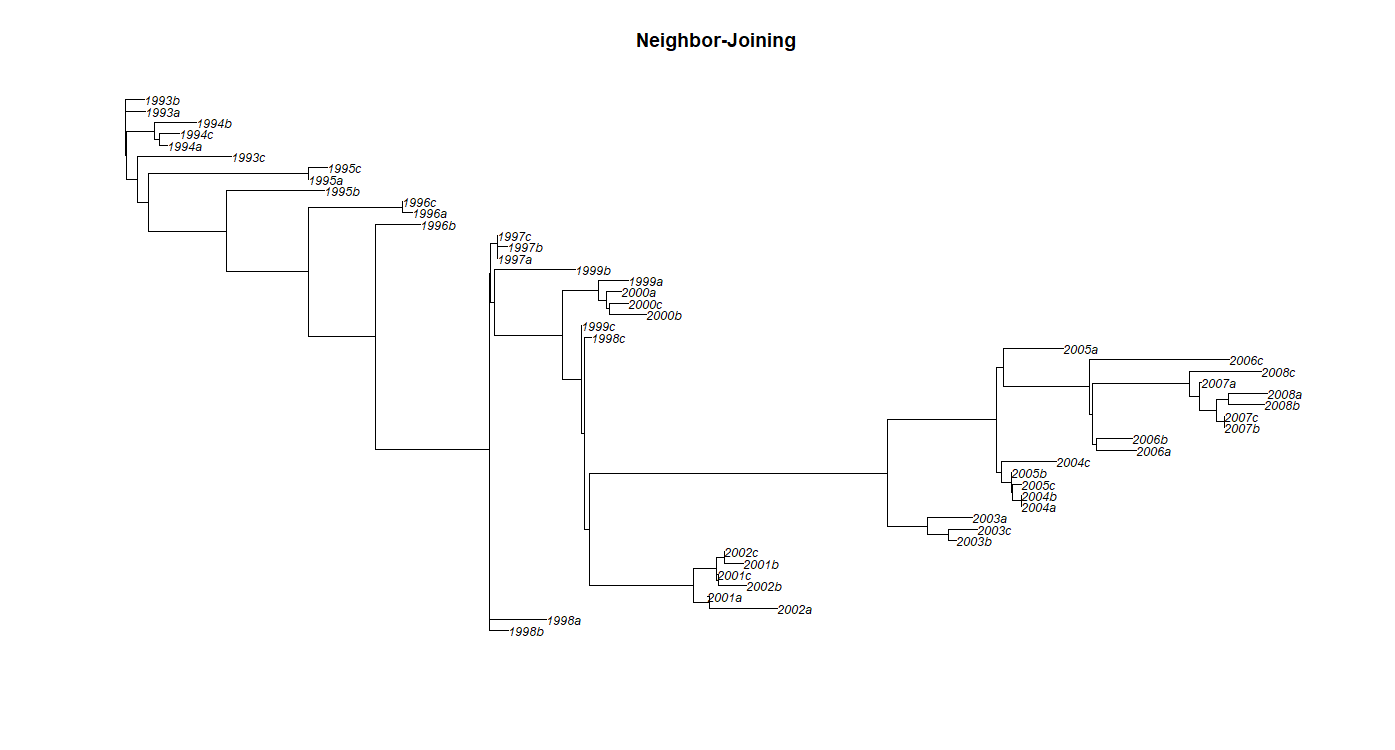


Fig: 2 A, B NJ and UPGMA trees respectively

A) We can see the plots for neighbor joining and UPGMA in the images above respectively. Focusing on the beginning of the trees arising from both methods, taking an example of 1994, in the case of NJ the branches a and c fall under one clade and b under another without creating a different node/splitting (except the one showing difference). Whereas for the UPGMA method, a and c fall under one clade and b falls under another by creating another node. Also, considering 1997, the tree in case of NJ method shows a polytomy for a, b and c whereas in case of UPGMA, a and c are under 1 clade and b falls under another creating a separate node. Again, in the case of NJ, proper evolutionary path can be understood, whereas in UPGMA, it shows a really brief detail compared to NJ.

PART 3:

**Q4. 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?**

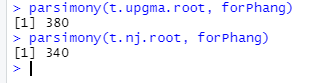


Fig: 3 Parsimony scores

A) As shown in the figure, the parsimony scores for UPGMA and NJ trees are 380 and 340 respectively, showing that the score for NJ trees is lower making it better than UPGMA. Here NJ is most parsimonious.

**Q5. 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?**

A) Based on the definition of parsimony, we can define a tree to be more parsimonious when it goes through lesser changes in terms of evolution, when a tree uses less changes to explain the evolution from the ancestor to the current generation. Also, from the diagram we can see that the NJ tree gives explanation in detail of ancestor history whereas UPGMA is briefing it out. The nodes show the perfect splitting of the trees and which clade falls where, taking this in mind, though NJ has lesser nodes making it less evolutionary changed, it explains with every splitting about the ancestral history.

PART 4:

**Q6. 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.**

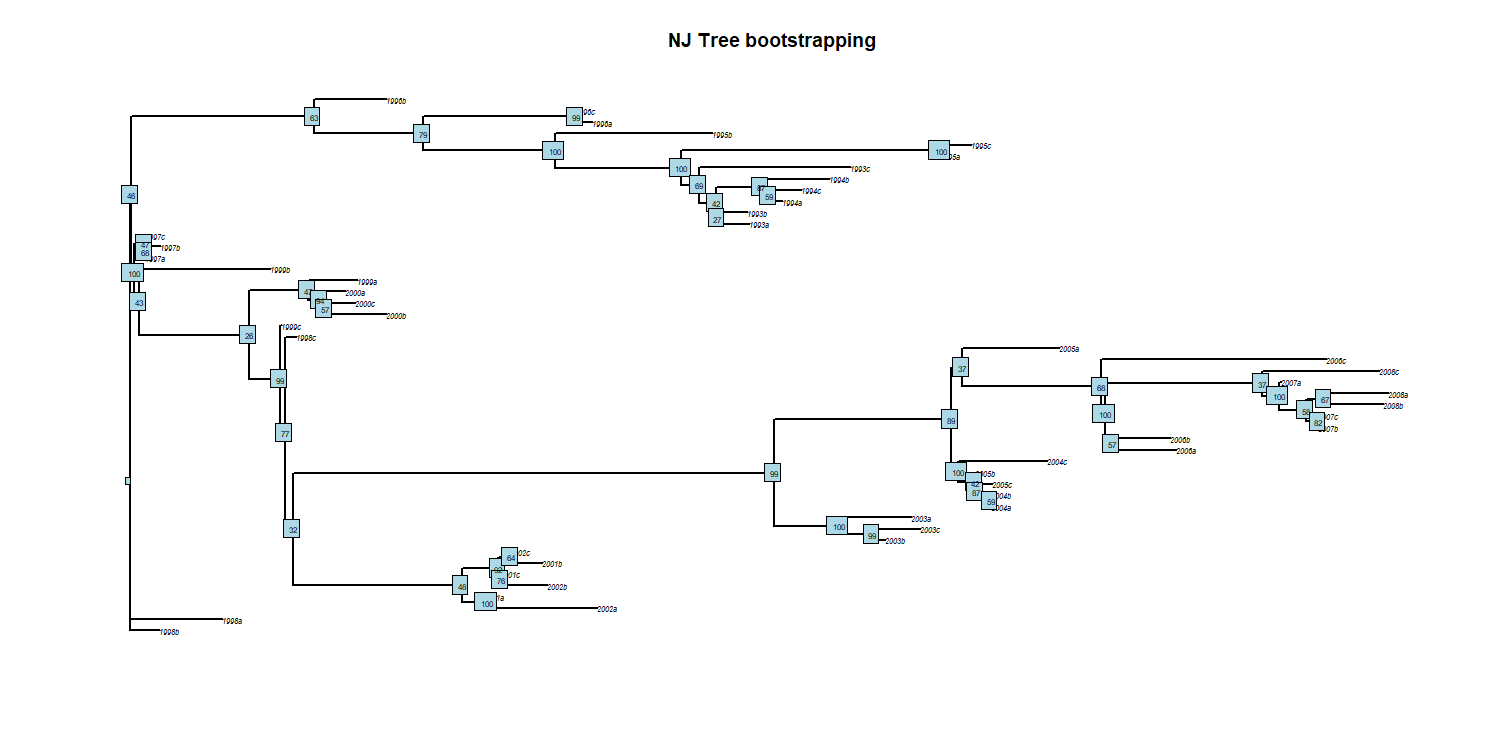


Fig: 4 Bootstrapping for NJ tree

A) Yes, there are nodes having poor bootstrapping support. There are 12 of these nodes. 1993a has the weakest support (: 27 score) and it is just in the beginning of the tree in the clade with 1993b.

PART 5:

**Q7. Show 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.**

A) The final tree is shown below. Overall, branching is occurring more in between years 1999 to 2005. When we look at the tree, we can see that initially the new species arose from the previous virus form until around 1999 - 2000. From this year, trends show branching of the trees, like for example, 2000a and 2000c arise from 2000b in a branched form. Also for 1997, it shows a polytomy. Huge clusters can be seen for the year 2001 - 2002. At the later part of the tree, branching decreases and new species again arise from the previous one without branching.

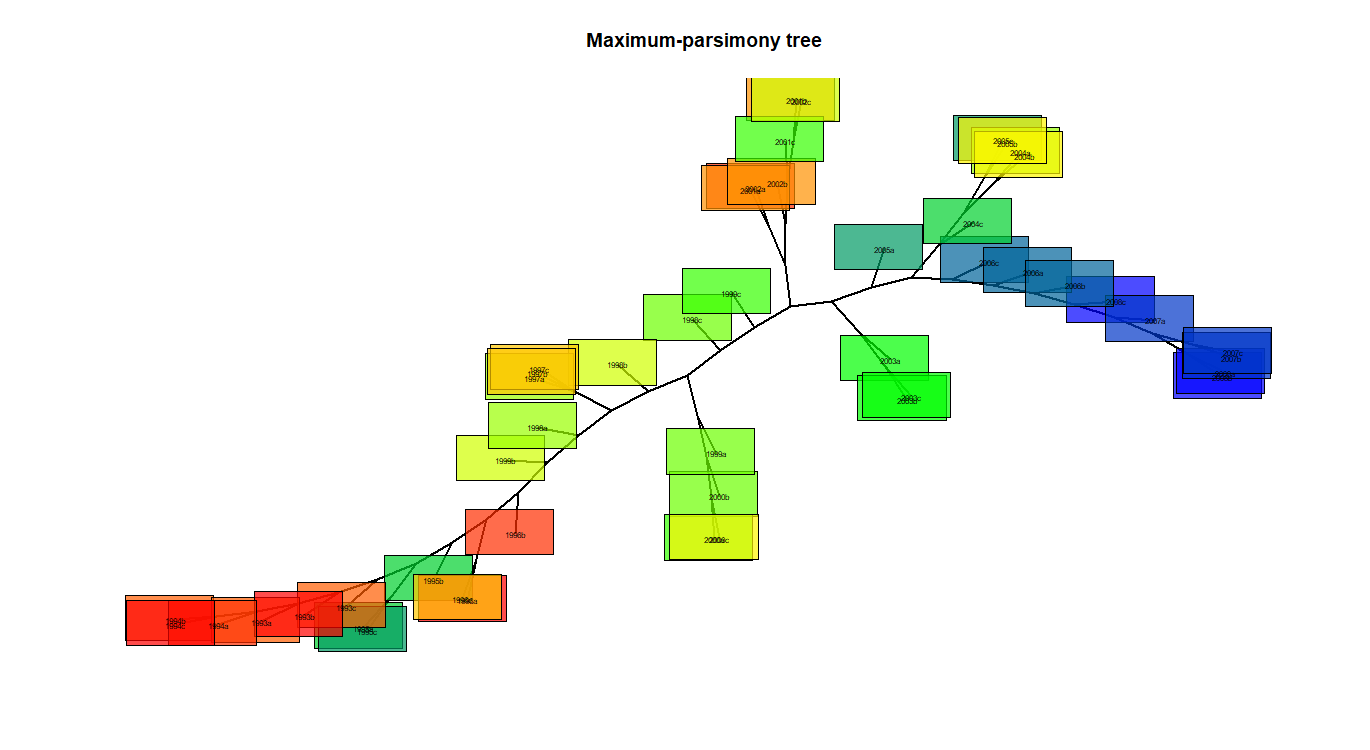
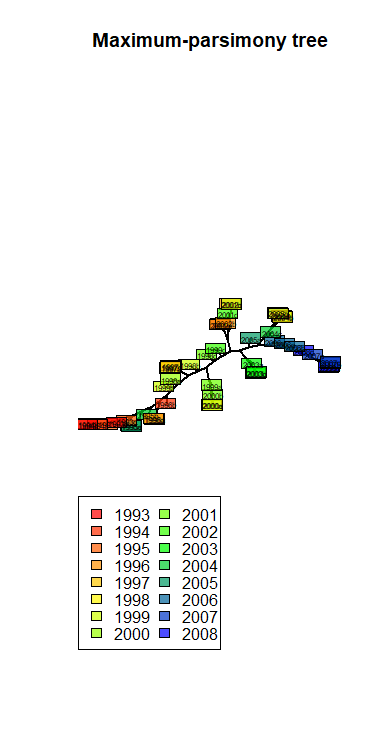


Fig: 5 A, B Maximum - parsimony tree for NJ.