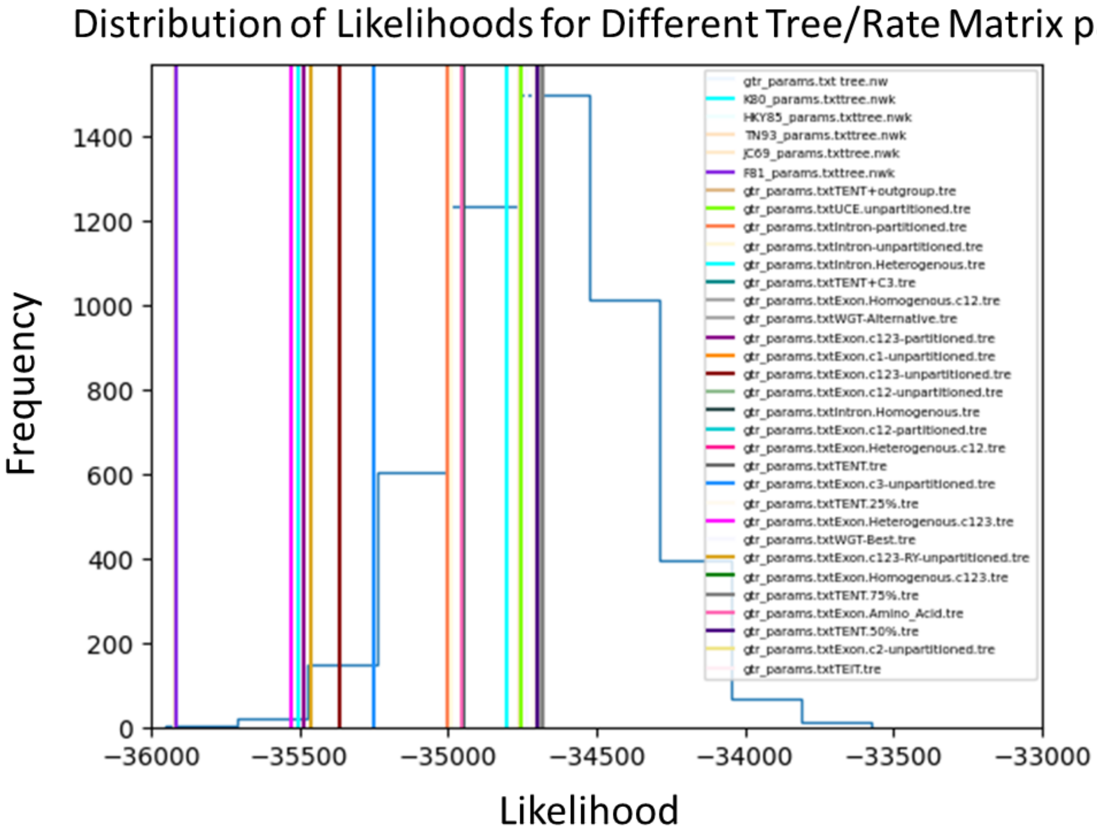
Nathan Newbury **HW9 Writeup**

The main tree/rate matrix pair was used to create my null distribution. For each alternate tree/rate matrix pair, I compute an alternate likelihood. Then I find the percentage of my null distribution below my alternate likelihood. This probability is called the p-value. A p-value of zero means the alternate likelihood was smaller than the entire null distribution. A p-value of 1 means the alternate likelihood was larger than the entire null distribution. In this case, a smaller p-value means we can reject the alternative and a larger p-value means we cannot reject the alternative. A p-value closer to 1 may be statistically significant in indicating which alternate tree/rate matrix pair was used to generate the data. **To summarize, a larger p-value means more confidence in the alternative choice** (this aligns with how the assignment is worded, but may be (1- p-value) in the traditional sense). Below we can see the null distribution and various alternate likelihood cutoffs for sequence 10:



We can reject trees with p-value < .05, we are uncertain about trees with .95> p-value > .05 and trees with p-value > .95 are statistically more likely than the main tree/ rate parameters pair. Any tree/ rate matrix pair with p-value ~>= .5 could be a better choice than the null choice. Here is a list of trees that I can reject, and a list of trees than I cannot reject:

List of Alternate Trees:

Reject Trees	Uncertain Trees

Intron.Heterogenous	TENT+outgroup
Exon.Homogenous.c12	UCE.unpartitioned
Exon.c123-partitioned	Intron-partitioned
Exon.c1-unpartitioned	Intron-unpartitioned
Exon.c123-unpartitioned	TENT+C3
Exon.c12-unpartitioned	WGT-Alternative
Exon.c12-partitioned	Intron.Homogenous
Exon.Heterogenous.c12	TENT
Exon.c3-unpartitioned	TENT.25%
Exon.Heterogenous.c123	WGT-Best
Exon.c123-RY-unpartitioned	TENT.75%
Exon.Homogenous.c123	Exon.Amino_Acid
Exon.c2-unpartitioned	TENT.50%
TEIT	

Of the trees that we cannot reject, all TENT trees and WGT trees performed the best with p-values near .5. We cannot statistically note any difference between these trees and the main tree.

Here are the parameters that I can reject, and the parameters that I cannot reject.

List of Alternate Parameters:

Reject Parameters	Uncertain Parameters
JC69	K80
F81	HKY85
	TN93

Based on the p-values for alternate rate parameters, we can reject JC69 and F81. While no rate parameters reached a p-value greater than the p-value of the null pair (~=.5), we cannot make a statistically significant claim for rate parameters K80, HKY85 and TN93. Each of these parameters are statistically no worse than the gtr parameters.

If I could not choose the null pair as my model, I would choose the “TENT+outgroup” tree (or any “TENT” tree). This tree consistently recorded high p-values. Because no parameters reached a p-value of .5, I would use the original “gtr”parameters. That being said, the K80 parameters could not statistically be rejected. We can represent “K80” parameters with only a single parameter. “K80” may be a good choice if we want to reduce the rate matrix parameter count.

The bootstrapping was performed using the resampling method seen below:

```
def resample(seqs):
    seqsnew = dict()
    keyslist = list(seqs.keys())

    filterindices =
np.random.choice(len(seqs[keyslist[0]]),len(seqs[keyslist[0]]))
    for key in seqs:

        alnstring = seqs[key]
        seqsnew[key] = ''.join(itemgetter(*filterindices)(alnstring))

    return seqsnew
```

With this method, for the main tree and rate matrix, I first create a list of sample indices. I make a random index choice for the entire length of 1 sequence. I then use these sample indices to create a new sequence for each leaf of the tree. It is important that every leaf of the tree uses the same sample indexes. Once I have resampled all of the sequences n times, I can find n different likelihoods and create the “null distribution”.

This assignment showed the difficulty in converging to the “correct” tree or “correct” parameters. Many trees can have similarly high likelihoods. This assignment also showed the power of bootstrapping. Without bootstrapping, it would take almost unreasonable amounts of data to find any statistical significance. With a cutoff p-value of 5% and bootstrapping for 5000 iterations, it is still hard to find statistically significant models. I would alter the procedure to use bootstrapping to find the alternate distributions. With two distributions it is easier to make significant inferences. Additionally, tree likelihood could not be the only metric for model efficacy. Maybe some applications value the number of parameters, or a certain computation time. In this way the exercise is narrow.