Research Assignment 1 - Tree Accuracy on 1000L1 dataset

Dylan Irlbeck

March 2019

1 Dataset source

This simulated dataset consisted of 1000 taxa with long gap lengths, and 20 different replicates, generated under the GTR model. This data is originally a part of the SATe-I open source datasets. To perform my computations, I iterated over all 20 replicates, and ran FastTree2 and PAUP* NJ on the true alignment (*rose.aln.true.fasta*) with the commands shown below.

2 NJ Commands

PAUP* MacOSX (Mountain Lion 10.8 or later) was run as

```
echo "ToNEXUS format=FASTA fromFile=[input alignment fasta file]
toFile=[alignment nexus file];
  exe [alignment nexus file]; NJ distance=logDet showtree=No;
  savetrees file=[output species tree file] format=newick;" | ./paup4a164_osx -n
```

3 FastTree 2 Commands

To enforce double precision, I complied FastTree with

```
gcc -DUSE_DOUBLE -O3 -finline-functions -funroll-loops -Wall -o FastTree FastTree.c -lm
```

FastTree Version 2.1.10 Double Precision was then run as

```
./FastTree -nt -gtr [input true alignment] > [output tree file]
```

The-`nt` and `-gtr` flags were used to infer a tree from an input nucleotide alignment (`-nt`) under the GTR+CAT model of evolution (`-gtr`)

4 Computing Error Commands

False positives and false negatives were computed using Dendropy Version 4.4.0 as

```
[fp, fn] = false_positives_and_negatives(tr1, tr2)
```

where `tr1` and `tr2` are Dendropy tree objects, and `fp` and `fn` are the number of false positives and false negatives, respectively.

5 Checking if trees are binary

To check if the estimated trees were binary or not (necessary for computing FP and FN rates), the number of internal edges were computing using Dendropy Version 4.4.0 as

```
ei1 = len(tr1.internal_edges(exclude_seed_edge=True))
ei2 = len(tr2.internal_edges(exclude_seed_edge=True))
```

where `tr1` and `tr2` are Dendropy tree objects, and `ei1` and `ei2` are the corresponding internal edge counts. The output of running this command with the inferred tree (both FastTree and NJ gave the same result) as the `tr1` argument and the true true as the `tr2` argument gave me the following output:

```
Number of edges in t1: 997, Number of edges in t2: 995
```

Since the number of edges in the estimated tree is $n-3$, where n is the number of taxa, we can conclude the estimated tree is binary, and calculate our FP and FN rates as the FP/FN count divided by 997. Furthermore, we can conclude the reference tree is not binary, since the number of internal edges is 995.

6 Results

The graphs of FP and FN rates for each method ran on all the 1000L1 replicates are shown below:

Average Results:

```
Average FP rate for NJ is  0.43104312938816447
Average FN rate for NJ is  0.43450351053159475
Average FP rate for FastTree is  0.10516549648946841
Average FN rate for FastTree is  0.1086258776328987
```

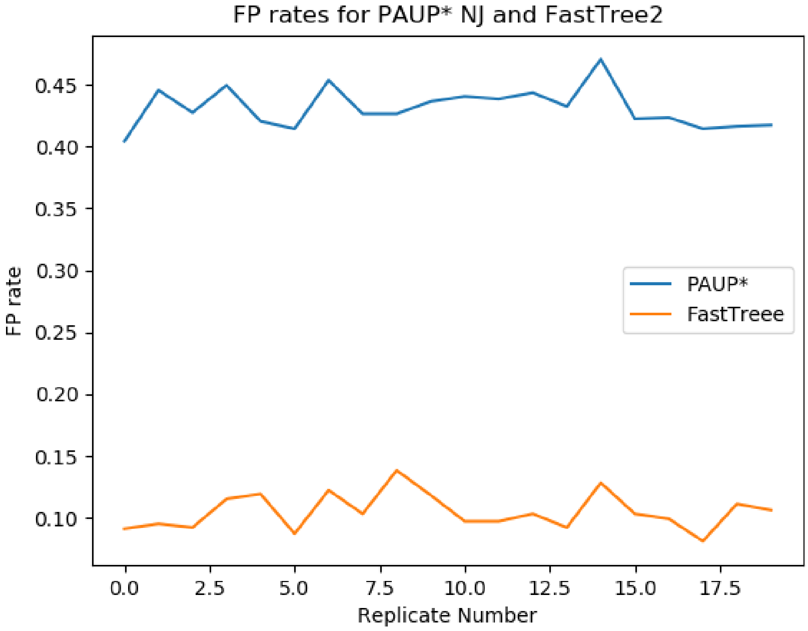

Figure 1: FP rates

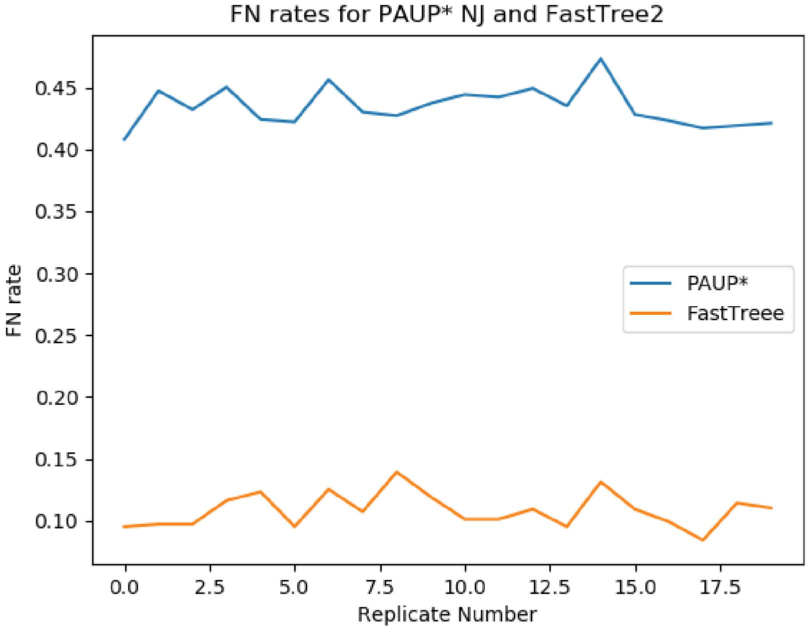

Figure 2: FN rates

7 Questions

Difference between *rose.tt* and *rose.mt*?

How the distances are calculated when running NJ?

In my command used for PAUP* Neighbor Joining, I use the flag `NJ distance=logDet` to indicate that the distance calculation was done using the logDet pairwise distance method. Since the logdet distance correction can be used to compute distances for any of the sub-models of the GM Model, and the 1000L1 replicates were generated under the GTR model, it is an appropriate method for computing distances.

8 Github Repo

All the data and code for my analysis can be found here.