HW3

Vishal Rana

September 20, 2020

Results

The following are the results for false positives and false negatives for estimated trees using FastTree and Neighbour-joining using FastME. RF is calculated as $\frac{FN+FR}{2n-6}$.

	1000M1		1000M4	
Method and Model	False Negative	False Positive	False Negative	False Positive
FastTree JC69	124.65	121.05	80.30	56.55
FastTree GTR	107.25	103.65	73.85	50.10
FastME (NJ) p-distance	231.15	227.55	190.30	166.45
FastME (NJ) JC69	211.20	207.60	158.50	134.65
FastME (NJ) LogDet	248.05	244.45	167.40	143.55

	1000M1			1000M4		
Method and Model	FN Rate	FP Rate	RF	FN Rate	FP Rate	RF
FastTree JC69	0.125	0.121	0.123	0.081	0.057	0.069
FastTree GTR	0.108	0.104	0.106	0.074	0.050	0.062
FastME (NJ) p-distance	0.232	0.228	0.230	0.191	0.167	0.179
FastME (NJ) JC69	0.212	0.208	0.210	0.159	0.135	0.147
FastME (NJ) LogDet	0.249	0.245	0.247	0.168	0.144	0.156

It was expected that the algorithms would work better on 1000M4 than 1000M1 because of less number of gaps in the alignment, which is what we observed across all experiments. Also, it was expected that FastTree with GTR model would perform better than FastTree with JC69 model for both datasets, which holds true for the experiments run because GTR allows for a richer model with more parameters.

For Neighbour Joining using FastME, improvements in performace going from p-distance to JC69 were observed for both datasets. However, LogDet performed worse than both on 1000M1 and worse than JC69 on 1000M4. Also, for both JC69 and LogDet there were cases where the pairwise distance could not be calculated and were capped at 5.0, especially for 1000M1 dataset. This is because of $p_{i,j} > 0.75$ which leads to log not being defined in calculation of JC69 distances and because of the determinant not existing for

the LogDet distances.
FastTree performs better than FastME (NJ). However, FastTree is slower, which reflects the running time and accuracy trade-off. This set of experiments points out some important points including how performance drops with more gaps in the alignment. Also, we need to careful when working with JC69 and LogDet distances because they might not be defined on some datasets as noted above.

Software Details

All the relevant scripts can be found on GitHub

FastTree

Version 2.1.11 [3] [4]
For JC69: FastTree -nt inputfile > outputfile
For GTR: FastTree -gtr -nt inputfile > outputfile
Relevant script: fasttreem1

FastME

Version 2.1.5 [2]
For p-distance: fastme-2.1.5-linux64 -mN -dp -i inputfile -o outputfile
For JC69: fastme-2.1.5-linux64 -mN -dJ -i -inputfile -o outputfile
For LogDet: fastme-2.1.5-linux64 -mN -dL -i -inputfile -o outputfile
Relevant script: fastmem1

The input files need to be converted from fasta to phylip format for FastME which was done using AlignIO from Biopython [1] (version 1.78) (Relevant scripts f2p and fasta2phylip.py). False positive and false negative rates were calculated with reference to the rose.tt tree for each case. The script used for the calculation is provided on github (treecomapre.py (due Erin Molloy) and fnfprate.py), along with the rest of the scripts used in this study.

References

[1] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[2] Vincent Lefort, Richard Desper, and Olivier Gascuel. Fastme 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Molecular biology and evolution*, 32(10):2798–2800, 2015.

[3] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular biology and evolution*, 26(7):1641–1650, 2009.

[4] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Fasttree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3):e9490, 2010.