MolBiol 3I03 Final Report

Siddharth Reed
April 21 2020

1 Code

The repository is at `https://github.com/DJSiddharthVader/thesis_SidReed`
For several steps in the pipeline add checks to see if results are already calculated and continue from there, i.e. if the script outputs files{1..10} but files{1..5} were already produced skip regenerating those and only produce files{6..20}. Also changed the network building steps to be run in parallel using the doSNOW package from R. I re-wrote parts of the `network_analysis.py` code to re-use certain functions more effectively and to calculate the closeness vitality and centrality. I also wrote a script `make_results_table.py` to parse the output of `network_analysis.py` and the output of markophylo into a csv file where rows are genera and column sare the network statistics. Finally after producing all of the data I re-created the figures using this new data in the jupyter notebook `PlottingResults.ipynb` using ggplot. I also updated the manuscript with the new figures, but not the discussion section yet.

2 Data Produced

For the 210 genera that I ran the pipeline on the following sets of files were produced

- species tree from whole genome data
- species tree from 16S rRNA genes (for some larger genera there were no 16S genes common to all members so no 16S species tree was produced)
- gene trees for all eligible genes up to a maximum of 3000 tree (eligible meaning the gene must be present in $\geq 40\%$ of the members of the genus)
- 1000 HGT networks produced from a WGS species tree and 50 randomly sampled gene trees
- 1000 HGT networks produced from a 16S species tree and 50 randomly sampled gene trees (if a 16S species tree was produced
- Network statistics calculated for each genus as detailed in the `network_analysis.py` file
- markophylo estimates of the gene indel rates (only 128 genera produced data points without error and not on the boundary)

All this data is on infoserv under `/home/sid/thesis_SidReed/data/genus_data` where each genus has it's own directory with all of the data produced. Another note is that for all statistics the mean and standard error of that statistic is calculated across all 1000 replicate networks. In all cases where the error was calculated it would produced negligible error bars compared to the mean values (i.e the top and bottom bars would overlap eachother) and thus are not shown. One final note is that markphylo was not able to generate any results using the WGS species trees but produced data for 128 genera using the 16S species trees. So currently the markophylo data is created using the 16S species trees but all of the network statistics were calculated using the WGS trees.

3 Future Work

- Update the discussion section to reflect the new results
- Check how different the 16S and WGS species trees, likely using distance metrics (R package)
- Continue looking into the network data