Computational Experiments to Investigate Biological Questions Using ms

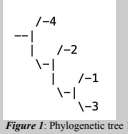
By. Maria Pacifico

Abstract

Programs have been developed for comparative genomics and the creation of phylogenetic trees using algorithms. Within the program ms (Hudson, 2002), we are studying population genetics and phylogenetics. Ms generates a simple sample model according to the Wright-Fisher neutral model based on chosen parameters. Parameters may include recombination, population size, migration, and gene conversion. For this experiment, we are using ms and seq-gen (Rambaut & Grassly, 1997) to examine the distances of the random outputs and analyzing the differences. Working with bioinformatics can help further advancements in biomedicine and our understanding of evolutionary biology since programs like ms and seq-gen are continuously being updated to be more accurate and efficient.

Introduction

Genetic theories can be applied with the use of standard softwares. One prominent software, focused on in this project is called ms which is used to generate independent replicate samples under chosen parameters (Hudson, 2002). Ms assumes the simplest Wright-Fisher neutral model in finite population with no recombination or selection (Techimsa & Innan, 2009). For the purposes of this project, we are replicating a multispecies coalescent model, which treats each species as a population of individuals with each individual having a set of alleles for each gene (Warnow, 2017). We are investigating how different the trees created by ms are from each other, despite having the same parameters. To measure this, we're using the normalized Robinson-Foulds distance.



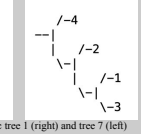


Figure 1: Phylogenetic tree 1 (right) and tree 7 (left) when taxa = 4. The Normalized Robinson-Foulds distance = 0.0

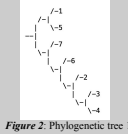


Figure 2: Phylogenetic tree 1 (right) and tree 20 (left) when taxa = 7. The Normalized Robinson-Foulds distance = 1.0

The normalized Robinson-Foulds distance has a range between 0.0 and 1.0 where 0.0 is the smallest amount of error (**Figure 1**) and 1.0 is the largest amount of error (**Figure 2**) (Warnow, 2017).

Methods

- Used ms to create an arbitrary tree using an input such as:
  - ./ms 4 1 -T
    - 4 = number of taxa
    - 1 = number of trees produced
    - -T = output a tree
- With the tree created above, we used a function created in python to create 30 multispecies coalescent trees (**Figure 3**) to create an executable for ms. The output of this function was put back into ms to create the trees such as those in **Figure 1** and **Figure 2**.
- To find the normalized Robinson-Foulds Distance, we used ETE3 in python and compared all of the outputted trees from ms.

Figure 3: Multispecies-Coalescent function in python.

```
def multispecies_coalescent_ms_arguments(species_tree, hownany=1):
    """
    Generate arguments for a multispecies coalescent simulation under ms,
    provided a species tree and sampling one gene per species.
    Note: this modifies the input tree object's internal node names
    """
    num_samples = len(species_tree) # how many taxa are there
    arguments = [f"{num_samples} {hownany}", # First two arguments of ms
                 "-T", # output the trees
                 f"-I {num_samples}{' 1' * num_samples}"]
    # Traverse the tree
    for node in species_tree.traverse(strategy="postorder"):
        children = node.get_children()
        if children: # If there are children to this node
            left = children[0] # Arbitrarily assign the children to be the left or right
            right = children[1]
            # Right subpopulation always merges into the left one
            ej_arg = f"-ej {node.get_farthest_leaf()[1]} {right.name} {left.name}"
            node.name = left.name # Keep track of which subpopulation this is
            arguments.append(ej_arg)
    return " ".join(arguments)
```

Results

- There was a 212.8% increase of average normalized Robinson-Foulds distance between taxa 4 and 10
  - 148% increase between taxa 4 and 7
  - 26.1% increase between taxa 7 and 10
- Output of multispecies coalescent model function (**Figure 3**) when taxa = 4:
  - ./ms 4 30 -T -I 4 1 1 1 1 -ej 0.039 3 1 -ej 0.415 1 2 -ej 1.6380000000000001 2 4 (**Figure 1**)

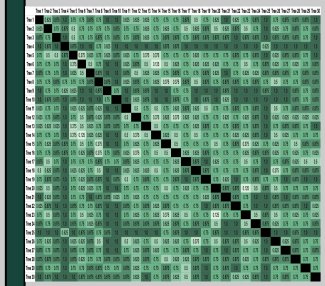


Figure 4: Calculated Normalized Robinson-Foulds distance for tree's 1-30 when taxa = 10 with an average distance of 0.782.

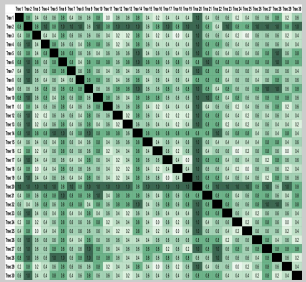


Figure 5: Calculated Normalized Robinson-Foulds distance for tree's 1-30 when taxa = 7 with an average distance of 0.62.

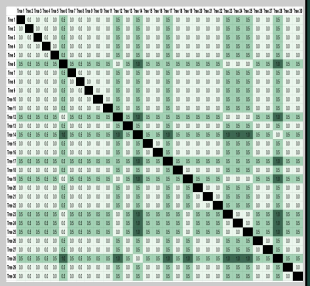


Figure 6: Calculated Normalized Robinson-Foulds distance for tree's 1-30 when taxa = 4 with an average distance of 0.25

Discussion

- The higher the taxa, the more randomized the trees are
  - **Figure 4** had the highest average distance and the highest taxa
  - **Figure 6** had the lowest taxa and average distance
- Since there is more variations between trees when there is a higher number of taxa, it is harder to find an accurate common ancestor within the coalescent model.
- Sources of error:
  - Each trial did have different -ej arguments because the different amount of taxa
    - Trees in each trial were created with the same parameters thus this doesn't affect the distance measurements.

Acknowledgments

- Special thanks to **Rei Doko** (Michigan State University) for providing Multispecies-Coalescent function (**Figure 3**)

References

Hudson, R. R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337-338.

Rambaut, A. and Grassly, N. C. (1997) Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.* **13**: 235-238.

Teshima, K.M., Innan, H. mbs: modifying Hudson's ms software to generate samples of DNA sequences with a biallelic site under selection. *BMC Bioinformatics* 10, 166 (2009). https://doi.org/10.1186/1471-2105-10-166

Warnow, T. (2017). *Computational Phylogenetics: An Introduction to Designing Methods for Phylogeny Estimation*. Cambridge: Cambridge University Press