An in silico experimental study of DNA sequence evolution simulation models

Aramis E. Matos[1],
[1]Department of Mathematics, University of Puerto Rico, Mayaguez, PR

Background

- A phylogeny is a model that represents the relatedness in between species, resulting in a tree
- Phylogenies calculate relatedness by genetic distance or by trait similarities
- This process is done by computer models, such as the Generalized Time Reversal Model (GTR) (Tavare, S, 1986)[1]
- These computer models require certain values as input, such as the substitution rate for pairs of nucleotide bases
- In the case of the GTR model, it requires a list of 6 substitution rates called a substitution rate vector

Introduction

- The GTR model receives a phylogenic tree, a substitution rate vector and a base frequency vector as input and returns the estimated DNA sequence of each taxa in the tree
- This estimation consumes time on the computer that is performing it, this is called runtime
- **Research Question**: If changing the values in the substitution rate vector affects the runtime of the model
- **Expectations**: If exceedingly small or exceedingly large values increase runtime

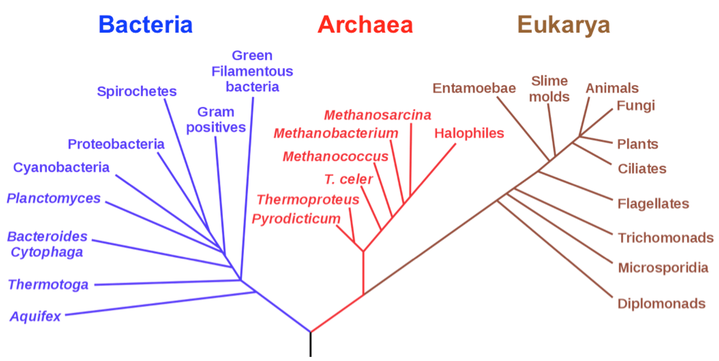

Figure 1. Phylogeny of Life (Phylogenetic trees and geologic time. Organismal Biology ,n.d.)

Methods

- Evaluated 3 sets of numbers
  - Close to Zero Set (CTZ): [1/(10^1), 1/(10^2) … 1/(10^100)]
  - Very Large Values Set (VLV): : First 100 values generated by 2^{63}-100+x for x = 1,2,3… [2^{63}-99, 2^{63}-98 … 2^{63}-1]
  - Control Group (CG): 0.5
- A value from a set is placed into the GTR model, setting each value besides the 6th one in the substitution rate vector to it and run the simulation 1000 times
- Average the runtimes of all runs to generate the set's descriptive statistics
- Generate the descriptive statistics of all three groups
- Run paired sample T-Tests in between VLV and CG, CTZ and CG and VLV and CTZ

Results

Runtime Mean	VLV	CTZ	Control
Time (s)	0.000148	0.000145	0.000146
P-Values of Paired T-tests (95%)	**VLV-CG**	**CTZ-CG**	**VLV-CTZ**
Two-Tailed P-Value	0.002	0.001	<0.001

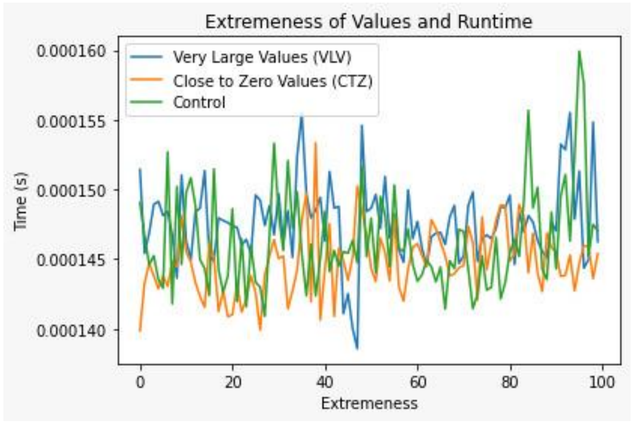

Figure 2. Runtimes and Extremeness (aramis matos, July 2022)

Discussion

- Given that the p-values do not exceed 5%, it cannot be said that exceedingly large or exceedingly small values in the substitution rate vector have a significant effect of runtime. Thus, our hypothesis was rejected
- Given that GTR model was ran in a computer simulation, the results are not surprising given because:
  - All numbers of the same type are processed in the same manner; thus, it should not matter what kinds of values are placed into the substitution rate vector
  - The values in the substitution rate vector act as multipliers and executed identically, irrespective of what the specific value is

Conclusion

- In summary, exceedingly large or exceedingly small values in the substitution rate vector for the GTR model do not have a significant effect on runtime performance
- These finding echo the findings in the literature.
- Future work can involve assessing estimated tree accuracy under differing substitution rates or other parameters such as base frequency under the GTR model or even other models and frameworks

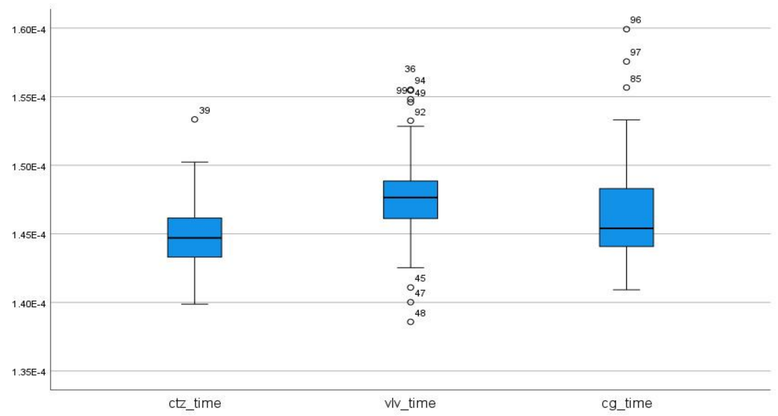

Figure 3. Box-plot for All 3 Groups

Acknowledgments

I would like to thank the following people for their help on this project: Dr. Kevin Liu and Julia Zheng as well as the SROP staffs and facilitators.

Contact

Aramis E. Matos
University of Puerto Rico, Mayaguez
Aramis.matos1@gmail.com

References

1. Tavare, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. Providence, R.I. American Mathematical Society, c1986., 57–86.
2. *Phylogenetic trees and geologic time*. Organismal Biology. (n.d.). Retrieved July 7, 2022, from https://organismalbio.biosci.gatech.edu/biodiversity/phylogenetic-trees/
3. aramis matos. (2022, July). aramis-matos/seq-gen-script. Retrieved 2022-07-03, from https://github.com/aramis-matos/seq-gen-script (originaldate: 2022-06-30T23:03:57Z)