Genetic Algorithm Based Phylogenetic Tree Estimation from Aligned Sequences

Md Alamin
Department of Computer Science and Engineering
Michigan State University
alaminmd@msu.edu

ABSTRACT

In this paper we implemented Genetic Algorithm to estimate a phylogenetic tree from a given sequence alignment. Inferring a phylogenetic tree can be extremely computationally intensive. For example, there can be more than 34 million possible rooted trees for only 10 taxa. We used a Maximum Likelihood based phylogenetic search using genetic algorithm to reduce the computational effort to a substantial amount compared to a conventional heuristic search method. The result shows that based on the number of taxa, the fitness value becomes consistent after a certain number of generations and we get a phylogenetic tree which is almost as accurate as a tree generated from the most widely used methods. The source code can be found here.

KEYWORDS

Genetic Algorithm, Phylogenetic Tree, Substitution Model, RF Distance.

ACM Reference format:

Md Alamin. 2020. Genetic Algorithm Based Phylogenetic Tree Estimation from Aligned Sequences. In *Proceedings of ACM GECCO conference, Cancun, Mexico, July 2020 (GECCO'20)*, 5 pages.

1 Introduction

Millions of species living on the earth show similarities in bio–chemical, morphological and gene sequence data, which suggest genetic relation among themselves. This relationship, when portrayed through an evolutionary tree, represents the phylogeny of organisms. Phylogeny is the history of evolution of species and a phylogenetic tree is a diagrammatic representation showing these evolutionary interrelations of a group of organisms. Fig. 1 shows a real phylogenetic tree between 4 species including humans.

Figure 1: Phylogenetic tree. A phylogenetic tree relating four species: humans, chimpanzees, gorillas and orangutans [1]

Reconstruction of phylogenetic trees is considered as an NP-complete problem [2]. The total number of possible phylogenetic trees becomes exponentially large; as a result, the space of topologies cannot be searched exhaustively. There are different model-based methods which are used to reconstruct phylogenetic trees with most accuracy. The complexity becomes higher when the Maximum Likelihood (ML) method is used as the optimality criterion for these model based methods. However, despite its computational cost, the ML approach for phylogeny inference is advantageous for many reasons regarding nucleotide sequence evolution. Genetic algorithms can increase the efficiency of the heuristic search, based on the maximum likelihood criterion.

For this project, we will be given an alignment of nucleotide sequences for a specific number of taxa. The goal is to construct a phylogenetic tree that will depict the evolutionary relationship among those taxa. A maximum likelihood based GA will search for a tree that is most consistent with the observed data in terms of a proposed evolutionary model. These models are used to roughly describe the way that we believe the sequences have evolved. Models can be sophisticated, based on the number of free parameters. There are different evolutionary models: Jukes and Cantor model (JC 1969), Kimura 2 Parameter model (K2P 1980), Felsenstein model (F81 1981), Hasegawa, Kishino and Yano model (HKY 1985) and Generalised time reversible model (GTR 1986). The models here are listed according to the number of free

parameters with increasing order. Therefore, JC69 is the simplest model whereas GTR is the most sophisticated model here.
In this project, the 5-parameter HKY85 model [3] is used. The substitution model is given in Fig. 2 where purine-to-purine or pyrimidine-to-pyrimidine mutations are called transition and purine-to-pyrimidine or pyrimidine-to purine mutations are called transversion. All the base frequencies are considered 0.25 for this project.

$$Q_{HKY} = \begin{pmatrix} \cdot & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & \cdot & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & \cdot & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & \cdot \end{pmatrix}$$

Figure 2: HKY substitution model. π_A, π_T, π_G, π_c parameters denote the frequencies of the nucleotide bases and the κ parameter denotes the transition/transversion rate ratio.

One other goal of this project is to optimize the κ parameter using the maximum likelihood criteria.

2 EXPERIMENTAL AND COMPUTATIONAL DETAILS

This project is implemented using different phylogeny tools that are available in Python based on the existing work of [4].

2.1 Representation

The individual for the GA for this project is represented as a tree in Newick format. The branch lengths and the κ value for each individual is also included in the representation. Fig. 3 is an example of an individual:

$\kappa = 4.0$ (((B:0.05, D:0.05):0.5,C:1.0):1.0, (A:1.0, E:1.0):1.0);

Figure 3: An individual of the population with 5 taxa in Newick format. Here A, B, C, D, E are the name of the taxa; associated numbers are the branch lengths.

The Newick formatted tree can be analyzed easily with the help of different Python modules for phylogeny analysis.

2.2 Fitness Function

The fitness of a tree (an individual of the population) is the natural log likelihood of the tree for the input sequences. We used Felsenstein's pruning algorithm [5], which is a dynamic programming approach to calculate the likelihood of a tree. We can calculate the likelihood of a single site of the sequences using this algorithm. This is a post order traversal algorithm, so we start from assigning the probability of the site at the leaves and then carry on the calculation with a bottom-up fashion. The probability of the leaf nodes can be assigned directly from the observed data. There are 4 possible states for each of the sites: A, T, G and C. The following equation [5] is used for calculating the probability of a transition from one state to other through a particular branch in an internal node *v*:

$$FPA(v,x) = \sum_{a\in\Sigma}\left[\Pr(v = x|w_1 = a) * FPA(w_1 = a)\right] * \sum_{a\in\Sigma}\left[\Pr(v = x|w_2 = a) * FPA(w_2 = a)\right]$$

Here, *FPA(v, x)* is the likelihood of state *x* for the internal node *v*. Σ is the set of states (A, T, G, C), w_1 and w_2 are the left and right child respectively of the current node *v*. Pr is the probability distribution for the state changes and is calculated using the substitution model from the following equation:

$$Pr = e^{t*Q_{HKY}}$$

Here, *t* is the branch length of the node *v* from its parent and Q_{HKY} is the HKY substitution model. The calculation is done for all of the sites independently and then the final likelihood of the whole tree is calculated by multiplying all of the likelihood probabilities of all the sites.

2.3 Selection

Rank-based selection is implemented on the basis of the fitness values of the individuals. 20 % of the offspring are generated by copying the fittest individual to the next generation. For the rest offspring, roulette wheel selection is applied based on the ranks of the parent generation instead of the absolute fitness values.

2.4 Mutation and Recombination

The individual having the highest rank is protected from mutation and recombination. Three types of mutations are implemented on the offspring: branch length mutation, κ mutation and topology mutation.
All of the individuals go through the branch length mutation. However, all the branches will not be mutated, instead only a portion of each offspring will go through the process. This portion value will be generated for each of the individuals randomly. Different mutation rates will be used for both the κ mutation and the topology mutation.
For both branch length and κ mutation:

$$new\ value := old\ value * multiplicative\ factor$$

Here, the $multiplicative\ factor$ is drawn from a gamma distribution. Gamma distributed random variables are used here because this distribution is generally used to predict the waiting time until a future event occurs and in phylogeny waiting time between different speciation events are important.
For topology mutation, *Subtree Pruning and Regrafting* (SPR) is used. Here we detach a subtree from one node and then reattach it

to another node. In this project topology mutation is applied after the branch length mutation. We can see how SPR mutation is implemented from Fig. 4. Topology mutation will be applied with a probability, so all the individuals of the population will not go through this mutation.

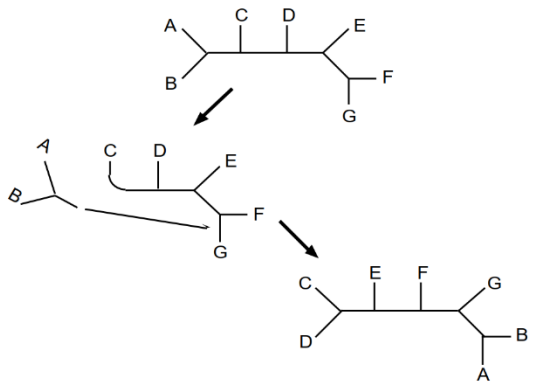

Figure 4: An example of "subtree pruning and regrafting" (a) the starting tree; (b) tree after pruning the subtree containing terminal taxa A and B; (c) tree after regrafting this subtree to the peripheral branch incident to terminal taxon G [6]

Recombination occurrs with a predefined recombination rate between a parent and an offspring. We remove a subtree from the offspring and attach it to the parent to get the recombined offspring. However, to avoid leave duplication, tips from the subtree are pruned from the parent individual first. Fig. 5 depicts the process of recombination between a parent and an offspring. While the same individual from the parental population can assume the role of first or second parent in a number of recombination events, no offspring individual can be the product of more than one recombination event.

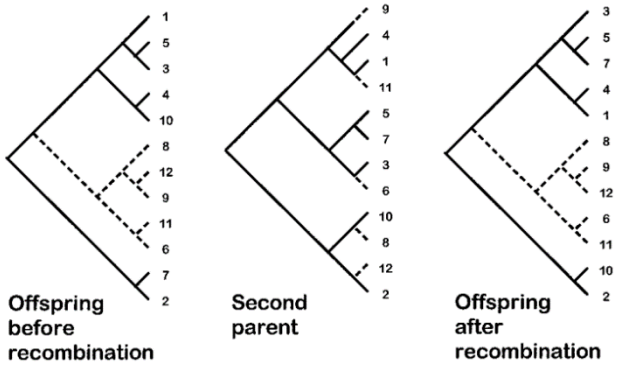

Figure 5: An example of recombination [4]

3 RESULTS AND DISCUSSION

Simulated sequence alignments with two different number of taxa (5 and 8) have been used to test the algorithm. All the sequences of the 5 taxa alignment are of length 100 sites whereas the sequences in the 8 taxa alignment are of length 200 sites. The simulated sequence alignments are generated using the INDELible [7] software with HKY model and κ value 2.5. The software uses a model tree to generate the sequence alignment. This model tree is the reference to calculate the accuracy of the estimated tree in terms of Robinson Foulds distance [8]. RF distance is the sum of false positive and false negative number of partitions between the two phylogenetic trees.

The summary of the GA used in this project is given in Table 1:

Table 1: GA summary for the phylogenetic tree estimation

Initialization	Random
Topology Mutation Rate	0.2
Kappa Mutation Rate	0.1
Branch Length Mutation Rate	100% (with random proportion of a topology)
Recombination Rate	0.2
Population Size	50
Selection	Rank based selection with elitism
Termination	500 generations or solution remains unchanged for 50 generations

In Fig.6 and Fig. 7, we can find the model tree to generate the sequence alignment for 5 taxa in INDELible software and the estimated tree for that alignment respectively. The RF distance between these two trees is 2.

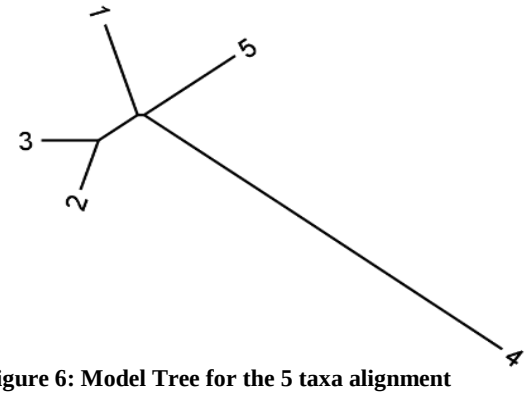

Figure 6: Model Tree for the 5 taxa alignment

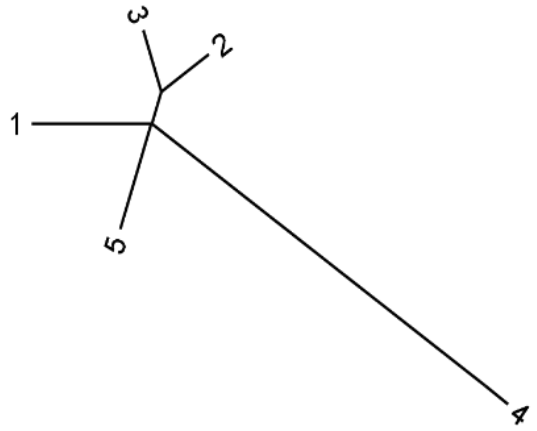

Figure 7: Estimated Tree for the 5 taxa alignment

Fig. 8 and Fig. 9 shows the same trees for the 8 taxa alignment The RF distance between these two trees is 4.

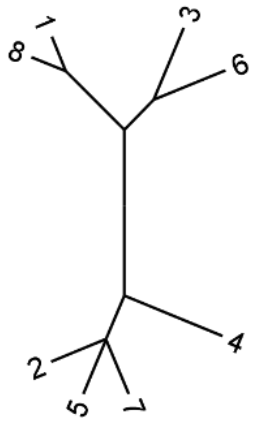

Figure 8: Model Tree for the 8 taxa alignment

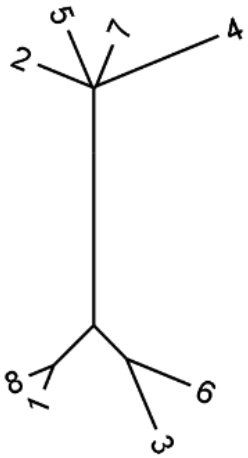

Figure 9: Estimated Tree for the 8 taxa alignment

Fig. 10 and Fig. 11 shows the progression of the fitness function (Natural Log Likelihood) for the two alignments . We find that for the 5 taxa alignment, the fitness score becomes consistent after 200 generations whereas for the 8 taxa alignment it's still improving after 500 generations.

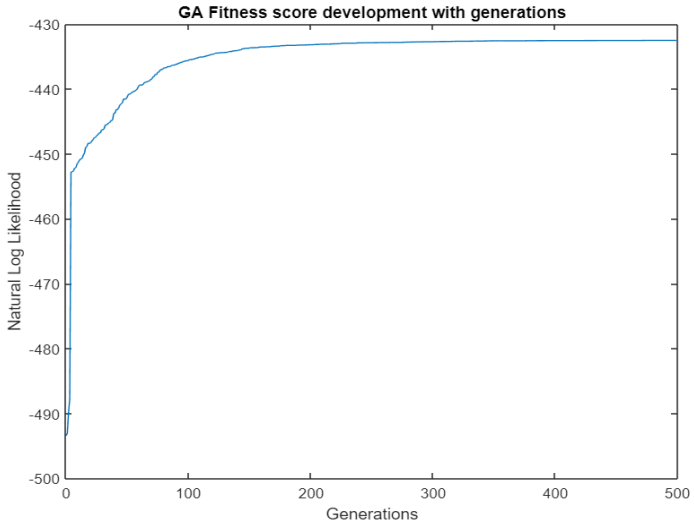

Figure 10: Fitness score progression for the best individual per generation for the 5 taxa alignment.

Figure 11: Fitness score progression for the best individual per generation for the 8 taxa alignment.

This also reflects in the RF distance comparison. The comparison of the RF distances for these two sequence alignments is conducted between the GA and two other popular phylogenetic tree construction methods: RAxML [9] and Neighbor Joining [10]. The result is given in Table 2:

Table 2: RF distance comparison between Genetic Algorithm, RAxML and Neighbor Joining methods.

Method	RF distance for 5 taxa alignment	RF distance for 8 taxa Alignment
GA	2	4
RAxML	2	2
NJ	2	2

From the comparison, it appears that, the GA worked well for the smaller data set but its performance was not up to the mark for the 8 taxa alignment. This is because we stopped the algorithm after 500 generations. If it could run for more generations, the fitness score could become consistent and it is expected that the GA could produce a more accurate tree. However, it took almost 3.5 hours in a core i3 7th Gen local machine to run the 500 generations of the 8 taxa alignment.

Fig. 12 and Fig. 13 shows the progression of the κ value for the 5 taxa alignment and the 8 taxa alignment respectively. We used the kappa value 2.5 while simulating both of the input sequence alignments using the INDELible software. As expected, we can find that the progression of the kappa value of the best individual in each generation is towards the optimal value.

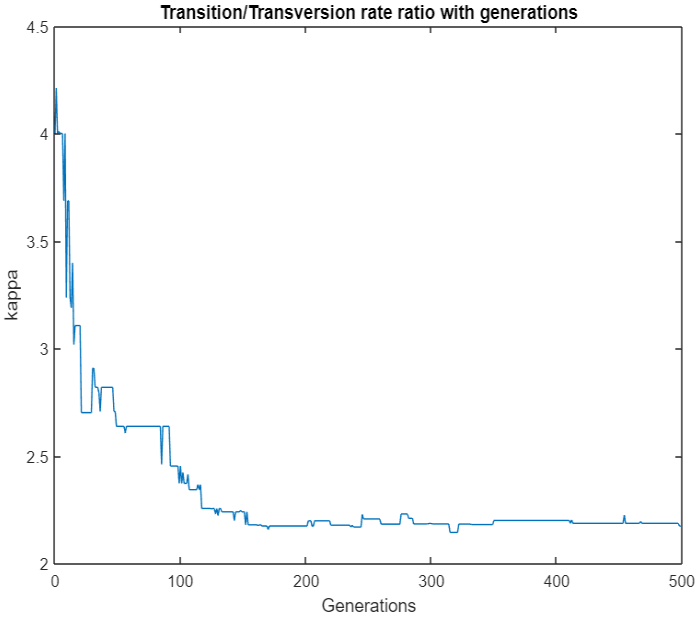

Figure 12: κ value progression for the best individual per generation for the 5 taxa alignment.

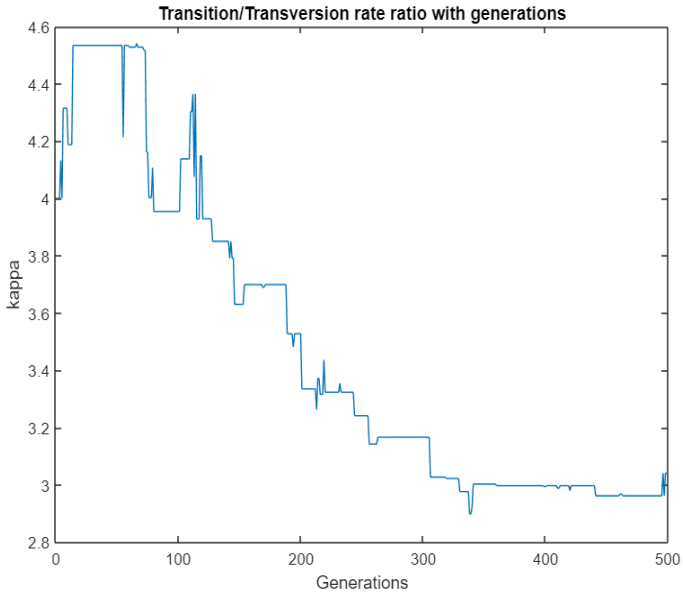

Figure 13: κ value progression for the best individual per generation for the 8 taxa alignment.

4 Conclusion

We have shown that Genetic Algorithm can be used successfully to reconstruct a phylogenetic tree with good accuracy. However, the computational cost for calculating the maximum likelihood value makes it too slow for larger sets.

Some other approaches can be tried in the future to improve the method. For example, instead of a random initialization, a stepwise addition method can be used to generate the initial population. Two other topological mutation strategies named Nearest Neighbor Interchange (NNI) and Tree Bisection/Reconnection (TBR) can be tested. Also the empirical values of the nucleotide base frequencies from the input sequences can be used in the HKY model instead of the fixed 0.25 value.

References

[1] M. S. Bayzid, "Estimating Species Trees from Gene Trees Despite Gene.," UT Electronic Theses and Dissertations, 2016.

[2] W. Day, "Computational complexity of inferring phylogenies from dissimilarity matrices," *Bulletin of Mathematical Biology*, pp. 461-467, 1987.

[3] Hasegawa, Kishino and Yano, "Dating of the Human-Ape Splitting by a molecular Clock of Mitochondrial DNA,"

Journal of Molecular Evolution, pp. 160-174, 1985.

[4] P. O. Lewis, "A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide," *Molecular Biology and EvolutionMolecular Biology and Evolution,* pp. 277-83, 1998.

[5] T. Warnow, Computational Phylogenetics: An Introduction to Designing Methods for Phylogeny Estimation., Cambridge University Press, 2017.

[6] SWOFFORD and BEGLE, "PAUP: phylogenetic analysis using parsimony.," 1993.

[7] W. Fletcher and Z. Yang, "INDELible: A Flexible Simulator of Biological Sequence Evolution.," *Molecular Biology and Evolution,* pp. 1879-1888, 2009.

[8] D. F. ROBINSON and L. R. FOULDS, "Comparison of Phylogenetic Trees," *MATHEMATICAL BIOSCIENCES,* pp. 131-147, 1981.

[9] A. Stamatakis, "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models," *Bioinformatics,* pp. 2688-2690, 2006.

[10] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees," *Molecular Biology and Evolution,* pp. 406-25, 1987.