Comparative Genomics Using Phylogenetic Trees

A Comparative Study of Phylogenetic Tree Construction Methods: A python Implementation of Maximum Likelihood, Maximum Parsimony, and Kmer Alignment-Free Algorithms.

Aly, A.H., Hesam
University of Science and Technology at Zewail City, s-aly.hesam@zewailcity.edu.eg
Rahma, R.A., Abbas
University of Science and Technology at Zewail City, s-rahma.abbas@zewailcity.edu.eg
Menna-Tallah, M.S., Shakran
University of Science and Technology at Zewail City, s-mennatallah.shakran@zewailcity.edu.eg

A phylogenetic tree, known as an evolutionary tree or a "Phylogeny" or rarely as “Dendrogram”, is a branching diagram representing the evolutionary relationships between different living organisms through estimator means. It is a powerful tool for understanding the relationships among different species, and it is used in evolutionary biology, ecology, medicine as well as many other fields. The species of interest are found in a phylogenetic tree at the tips of the tree's branches. The pattern in which the branches connect represents the hypothetical understanding of how said species have evolved from a series of common ancestors. Maximum Likelihood, Maximum Parsimony and Alignment free Methods are the most common methods employed to construct and assess phylogenetic trees.

CCS CONCEPTS • Phylogenetics Analysis • Distance-based • Maximum parsimony • Maximum Likelihood

1 Introduction

A phylogenetic tree, known as an evolutionary tree or a "Phylogeny" or rarely as “Dendrogram”, is a branching diagram representing the evolutionary relationships between different living organisms through estimator means. It is a powerful tool for understanding the relationships among different species, and it is used in evolutionary biology, ecology, medicine as well as many other fields [1].

In a phylogenetic tree, the species of interest are found in a phylogenetic tree at the tips of the tree's branches. The pattern in which the branches connect represents the hypothetical understanding of how said species have evolved from a series of common ancestors. Each branch point called an internal node represents a point of divergence where a group or species has split apart from an ancestral group into two descendant groups. Each branch length in a tree is a representative of a series of ancestors that lead up to the descendant. while the root of a tree is a representative of a series of ancestors that lead up to the most recent common ancestor of all the species portrayed in the tree [2].

They are divided into mainly two types: Rooted and Unrooted trees. A rooted tree is a directed tree with a unique node corresponding to the most recent common ancestor of all the entities at the leaves of the tree while an unrooted tree displays the relatedness of the species without assigning common ancestry. Most unrooted trees could be generated from rooted ones by omitting the root, and a root cannot be inferred from an unrooted tree without the identification of common ancestry [3].

Both types of trees could be either Bifurcating,where it has two descendants only arising from each node, or Multifurcating where three or more descendants rise from the same node. Moreover the trees could be labeled, each branch is assigned a different characteristic, or unlabeled usually known as tree topologies [3].

The main use of phylogenetic trees is to understand the evolutionary history of different groups of organisms by studying the branching patterns, through which scientists can infer the relationships between different species, and trace the evolution of different characteristics [4].

Besides the main use, phylogenetic trees are also applied in medical research to identify the origins of diseases and track their spread through populations. For example, by analyzing the genetic sequences of different strains of a virus or a bacterium, scientists can construct a phylogenetic tree that shows how the different strains are correlated to each other. This method helps researchers understand how the disease is spreading and develop strategies to control its spread [4].

In the following report Maximum Likelihood, Maximum Parsimony and Alignment free Methods are used , employed and assessed in Constructing phylogenetic trees [1].

1.1 Problem Definition

Comparative genomics is a rapidly growing subject that has produced outstanding findings. With several fully sequenced genomes readily available, comparative genomic analysis has become possible. The availability of numerous fully sequenced genomes boosts the predictive capacity in unlocking the secrets of genome design, function, and evolution. Comparisons of whole genomes amongst organisms enable global insights on genome evolution. Consequently, a genomic landscape's comparison of human genes with genes from other genomes could aid in determining novel gene functions for genes that are still unannotated [5].

We propose that by comparing genomics, we can know the percentage of similarity between humans and other species. So, if we want to transfer a specific protein into a human, we can extract it from the species that have a high similarity with humans.

1.2 Related Work and Survey

There are several methods and software tools available for creating phylogenetic trees, each with its own set of pros and cons. In this section, we will evaluate the characteristics and capabilities of some of the most extensively used software tools for phylogenetic tree construction, such as RAxML, MrBayes, and MEGA.

RAxML is a popular software tool for creating phylogenetic trees. It is based on the maximum likelihood algorithm, which includes determining which tree is most likely to have produced a given collection of sequences. RAxML provides several features, including the capacity to handle big datasets and perform quick bootstrap analysis. It is, however, computationally intensive [15].

MrBayes is another popular software tool for constructing phylogenetic trees. It is based on the Bayesian inference method, which estimates the evolutionary links between sequences using a probabilistic methodology. MrBayes is very effective in determining the evolutionary history of huge, complicated datasets. It can, however, be computationally intensive and may need a large number of processing resources.

MEGA (Molecular Evolutionary Genetics Analysis) is another popular software tool for constructing phylogenetic trees. It is a user-friendly program that offers several approaches for inferring evolutionary connections, including maximum likelihood, maximum parsimony, and distance-based methods. MEGA has the capacity to handle numerous types of data, including DNA, RNA, and protein sequences. It also includes several molecular evolution models and allows the user to run various statistical tests to check the robustness of the inferred tree. In comparison to the other software solutions listed above, MEGA provides a broader range of algorithms and has a more user-friendly interface. However, in terms of processing capacity, it may not be as sophisticated as other software tools, and it may not be suited for very big datasets [14].

Out of all three mentioned softwares, MEGA is a great alternative for phylogenetics researchers who are new to the field, and it provides a complete set of features that make it simple to use. So, we preferred to use it to test our algorithms.

2 Proposed Methods:

Our proposed methods for constructing phylogenetic trees, two alignment based algorithms , Maximum parsimony and Maximum Likelihood and alignment free algorithm called Alignment free distance based approach.

2.1 List of questions your experiments are designed to answer, description of your testbed.

1. When we apply this algorithm on the whole genome, not only genes, how much time would it take to build a tree?
2. If this is applied to four genes only, what could happen if the number of genes was increased to 10 genes?
3. In generating phylogenetic trees from multi-fasta files, how accurate is the suggested k-mer-based alignment-free technique compared to known methods such as Maximum Likelihood and Maximum Parsimony?
4. In terms of accuracy and computing efficiency, what is the ideal value of k for the k-mer method?
5. What effect does the length of the k-mer have on the accuracy and computing efficiency of the alignment-free k-mer technique for phylogenetic tree reconstruction?
6. What is the performance of the k-mer algorithm on datasets of varied sizes and complexities?
7. How can the k-mer technique be enhanced to make it a more useful tool for rebuilding phylogenetic trees in the future?

Testbed: The tests are carried out on a computer equipped with a 3.4 GHz Intel Core i7 CPU and 16 GB of RAM. The suggested k-mer-based method's software implementation is created in Python and tested on multi-fasta files including gene sequences from diverse species. The method's evaluation is verified by comparing the generated phylogenetic trees to those created using other methods and by understanding evolutionary relationships.

2.2 Algorithms

2.2.1 Maximum Likelihood

Intuition: The maximum likelihood (ML) method is a probabilistic method of estimating phylogenetic trees based on nucleotide sequence, it is a method that has been advocated and

used by many scientists as it is believed to have the statistical property of consistency. The method was proposed, in the beginning due to the increase in the number of large data sets, as a fast and reliable phylogeny reconstruction method [6].

It utilizes the likelihood equation, to calculate the likelihood that two or more sequences are related to each other based on a proposed probabilistic model. If the outcome of the equation matched the observed data then the method is reliable, if not we choose a different sequence. The algorithm takes mainly gene sequences, genomes and protein. But throughout this project DNA sequences stored in fasta files were used.

Proposed approach: Two approaches were made, one using MUSCLE and PhyML command lines which are widely used tools to construct ML phylogenetic trees while the other used needle-Wunsch alignment.

ML is known to use pairwise alignment, and so to deviate, Needle-Wunsch was used as it provided a more numerical insight as well as global alignment of the sequences, Needle-Wunsch, NW will be used as denotation, utilizes a scoring matrix which was built for the 4 DNA nucleotides to count the number of matches and mismatches for , and a Traceback function, get_alignment(), to obtain the aligned sequences.

Seeming as Maximum Likelihood takes into account all cases of mutations, as it considers all tree topologies before giving rise to the optimum one, no further functions were obtained other than a get_p_value(), which performs minimal bootstrapping to the sequences to ensure the alignment score was not a fluke, if the alignment score changed then that means that the sequences are not closely related.

The likelihood of the sequences was then calculated and based on the value they were arranged in a newick format, to draw or visualize the tree.

Results: A tree was obtained using the command line setup, however, when attempting to draw the same tree using ete3 package errors were encountered and only the alignment and percentage of similarity were obtained.

2.2.2 Maximum parsimony

Intuition: One of the important methods for the construction of a phylogenetic tree is the Maximum parsimony algorithm. Its intuition is to find the best tree that describes the relationship between groups of organisms and how all these organisms are close to each other. The principle of parsimony suggests that while reconstructing the evolutionary relationships of lineages, we should favor the phylogeny that necessitates the fewest evolutionary changes [7].

Parsimony is the practice of arranging taxa in a way that reduces the number of character-level evolutionary changes that would have been necessary. The reasoning behind this is that all other things being equal, a simple hypothesis (only four evolutionary changes) is more likely to be true than a more complicated hypothesis (15 evolutionary changes) [8].

Biologists employ brute computing effort to identify the most parsimonious tree. The plan is to construct every conceivable tree for the taxa that were chosen, map the characters onto the trees, and then choose the tree that has undergone the fewest number of evolutionary changes. Although the concept is straightforward, the first two phases demand a lot of effort or computational power [8].

The computational complexity of maximum parsimony is a downside. It is doubtful that any algorithm can quickly locate the most parsimonious tree for every possible input sequence since finding the most parsimonious tree is an NP-hard issue [11].

Maximum parsimony did not perform especially well in identifying the genuine phylogenetic tree for the parameter values taken into account. The likelihood that maximal parsimony discovered the correct tree was less than 25% in 24.2% of simulations, less than 50% in 60.2% of simulations, and less than 75% in 85.2% of simulations.

Proposed Approach: At the beginning, the algorithm started to be tried on a gene only. The data used was FASTA files that have the sequence of a desired gene for different organisms that we want to draw the relation between them in terms of this gene. To start construction or implementation of the algorithm there were some steps to be followed and it describes how actually this algorithm works. The first step is to make multiple sequence alignment to these all genes of the different organisms. The method used here for the alignment was Needleman-Wunsch Alignment.

Finding the ideal alignment of sequence pairs is ensured by the dynamic programming-based Needleman-Wunsch algorithm. With this approach, a major problem (the entire sequence) is basically broken down into several smaller problems (brief sequence segments), with the larger problem's solution being built using the answers to the smaller problems. The approach allows for the discovery of gaps in sequence alignment by scoring similarity in a matrix [9].

The next step after the alignment is to detect the parsimonious informative sites between the aligned sequences. A site that is parsimony-informative has at least two distinct character states for the sequence there. For the least amount of evolutionary change in the genome, this location must have at least two different types of nucleotides (or amino acids), and at least two of them must exist with a minimum frequency. These educational positions are most useful for providing details about the evolutionary links between the analyzed genes [10].

Finally, these informative sites should be used to build all possible trees. For example, imagine that our goal is to identify the evolutionary connections between only the four taxa A, B, C, and D. The number of possible relationships between those taxa, as shown below in the figure, jumps to 15 as the number of taxa rises. There are more than 34 million different potential trees for just 10 species. Biologists employ computer programs created for this activity since the enormous number of potential trees is just too great to be handled on paper [8].

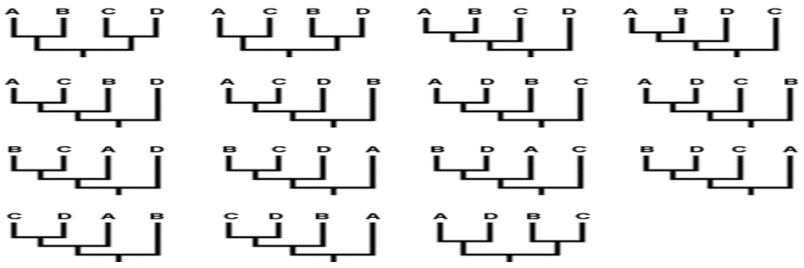


Figure 1: All the different ways that four taxa could be related, [Public Domain] via Google

2.2.3 Alignment-Free Distance based Approach

Intuition: K-mer-based alignment-free algorithm for phylogenetic tree construction has received a lot of interest in recent years because it has various benefits over standard alignment-based methods. One of the primary advantages of the K-mer-based alignment-free approach is that it does not require multiple sequence alignment, which may be time-consuming and error-prone. Instead, this approach compares sequences based on the frequency of k-mer patterns in the sequences, which is a more robust and efficient method.

Another benefit of the K-mer-based alignment-free approach is that it is more resistant to sequencing mistakes and gaps, which can be problematic in standard alignment-based methods. The k-mer algorithm can compensate for sequence variability while still inferring precise evolutionary relationships. Moreover, the K-mer-based alignment-free approach may be used on a broad variety of sequences, including genomic, transcriptome, and proteomic data, and can handle enormous datasets with excellent computing efficiency [12].

Furthermore, the K-mer-based alignment-free approach has been proven in multiple studies to have comparable or higher performance than standard alignment-based methods, owing to its ability to handle big datasets with great computational efficiency and produce robust findings with fewer mistakes. In other words, using multiple alignments to establish the evolutionary connections or ancestry of sequences can be inaccurate when applied to sequences that have diverged significantly. Some alignment techniques focus on aligning sequences that have closer relations first, but this can lead to inaccurate conclusions about the evolutionary relationships. Additionally, using dynamic programming to align sequences can make the computation process more complex and may not be suitable for large datasets. Consequently, alignment-free methods have been proposed and are considered important. The quality of alignment is crucial in determining the evolutionary relationships depicted in a phylogenetic tree [13].

Finally, the K-mer-based alignment-free approach has emerged as a significant tool for phylogenetic tree construction, offering various advantages over classic alignment-based methods such as better computing efficiency, tolerance to sequencing mistakes, and the capacity to handle enormous datasets.

Description of K-mer approach: The study proposes a new alignment-free approach for reconstructing the evolutionary history between species [12]. The methodology is based on the k-mer method, in which a genetic sequence is represented as a frequency vector of fixed-length subsequences. The algorithm begins by reading FASTA files for DNA sequences and performing k-mer decomposition to each sequence. The k-mer length (k) is an essential parameter in the process, with 7 being the optimal number. With this number, there are 4^7= 16384 different k-mer combinations that may be generated from a DNA sequence. But I developed this part to make the possible combination in each sequence as the proposed approach in the paper was to make every possible combination regardless of its existence in the sequences. So, I modified this part to get only the possible k-mer combinations in each sequence.

The approach computes the number of occurrences of all possible k-mers of each input sequence after k-mer decomposition and saves the value in a hashmap. The hashmap is employed in order to compute the distance between two sequences. The distance between two sequences is determined as the sum of the relative differences in their k-mer frequencies divided by the total number of k-mers in each sequence. This is accomplished by employing the following equations:

$$dAB = \sum fAi - fBi / fAi + fBi \quad (1)$$

$$dBA = \sum fBi - fAi / fAi + fBi \quad (2)$$

- dAB = number of k-mers that are present in sequence A but not in sequence B
- dBA = number of k-mers that are present in sequence B but not in sequence A
- fAi = number of occurrences of k-mer i in sequence A
- fBi = number of occurrences of k-mer i in sequence B
- nA = total number of k-mers present in sequence A = lA - k + 1
- nB = total number of k-mers present in sequence B = lB - k + 1

The distance between two sequences is calculated as:

$$distance = (dAB/nA) + (dBA/nB) / 2 \quad (3)$$

Where distance is symmetric, the matrix is based on k-mer frequency. This distance matrix is then used to construct the phylogenetic tree using the neighbor-joining method. The neighbor-joining method is a popular distance-based method for constructing phylogeny.

I could extract the distance matrix efficiently, but the problem was to turn this distance matrix into plotted phylogeny. I tried to get the NEWICK string from the distance matrix through the UPGMA algorithm, but I could not implement the last step of it, which is creating the NEWICK format. but I could implement the clustering and update the distance matrix efficiently. So, to check my k-mer algorithm I regretfully had to use the package “BioPython” to draw the phylogeny tree from the distance matrix.

Evaluation/Experiments/Results: Now to evaluate my work, I used MEGA 11 software on the neighbor end joining mode. I pass the same FASTA file “COX2” to MEGA and to my program. Then, I start comparing the results. I tried to use in my evaluation more than one multi-fasta file, but I will mention the results of only one trial here to avoid the length of the report problem.

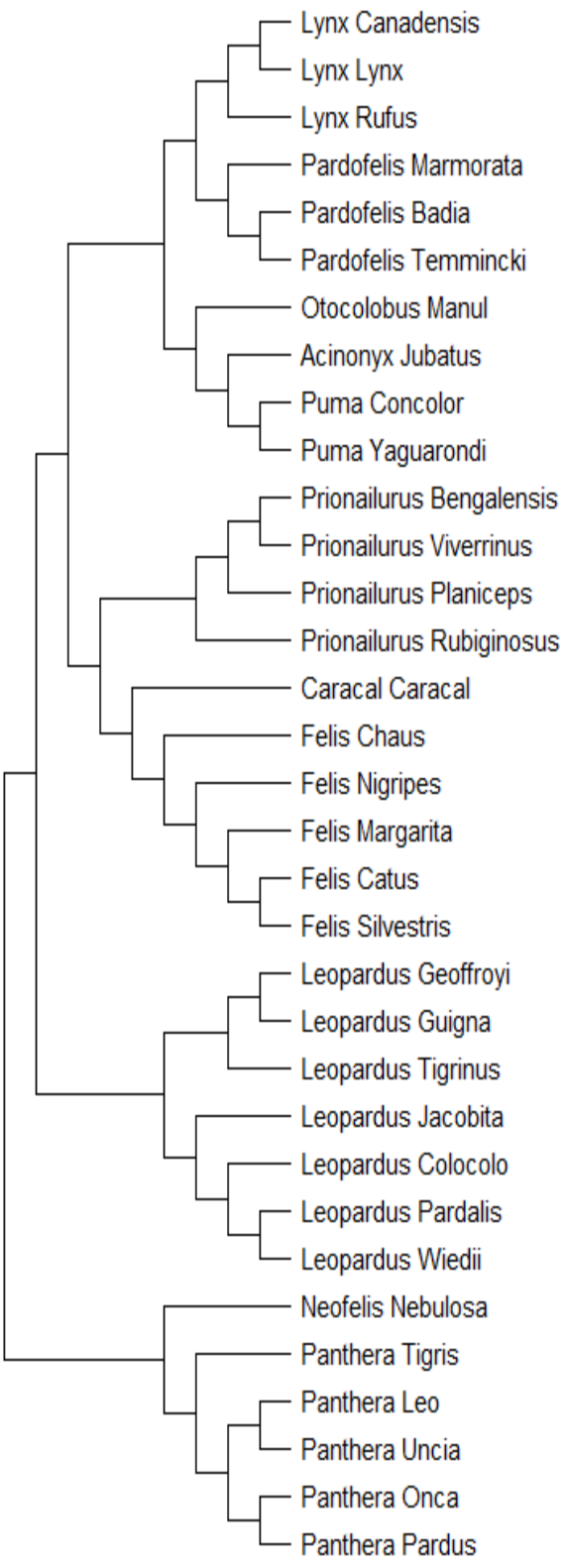
After performing my experiments, I discovered that my results were comparable to those obtained using the MEGA program. However, following further examination, I discovered that the tree created by my algorithm was the reverse of the tree generated by MEGA. This might be related to how the distance matrix is created or how the neighbor-joining mechanism is performed in my approach vs MEGA.

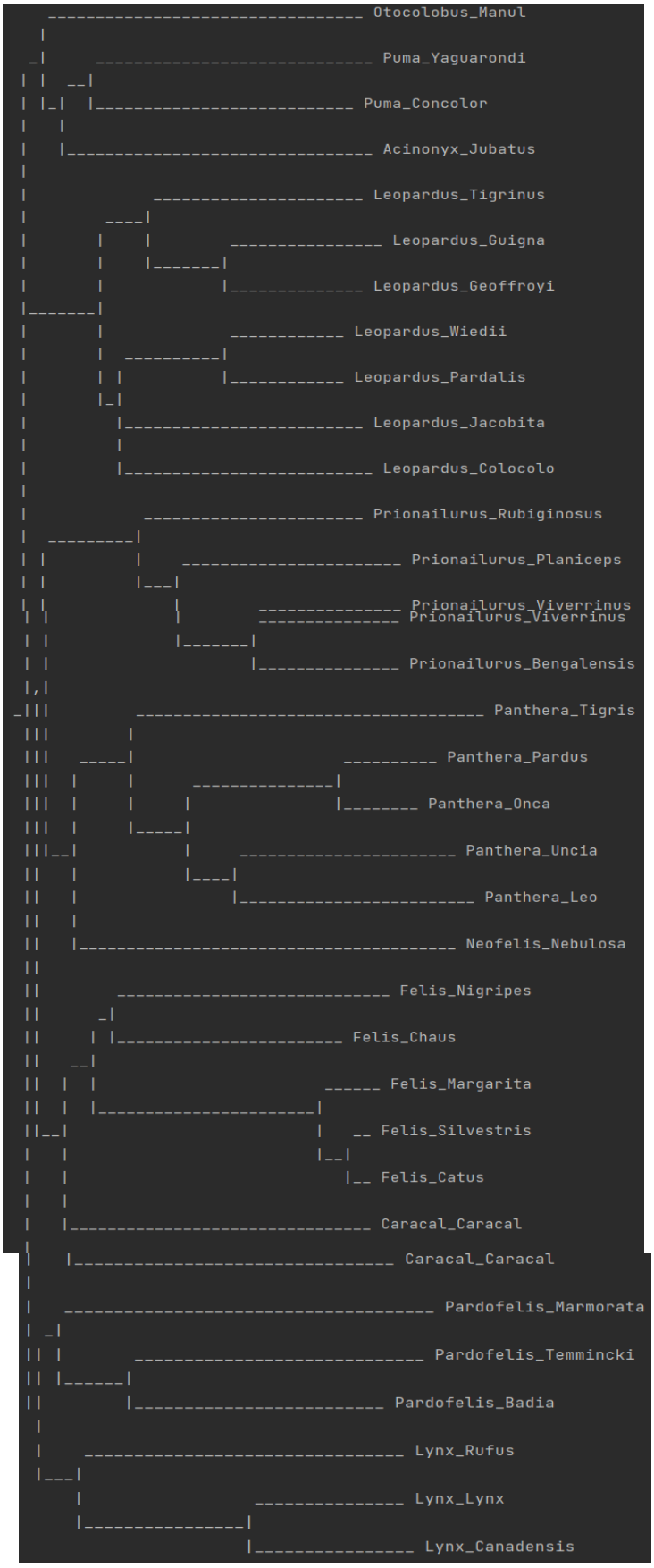
It is crucial to highlight that the orientation of a phylogenetic tree has no effect on its general structure or the connections between species. As a result, while the tree's orientation may change, the conclusions gained from it should stay consistent.

As a future work, I will continue to tune and enhance my method in order to get results equivalent to those produced using known tools such as MEGA. I will also run further tests to discover the causes of the differences in tree orientation and the best parameters for my algorithm.

Finally, While the overall findings of my algorithm are comparable to those achieved by MEGA, the tree orientation is different. This, however, has little bearing on the final finding, and I will continue to tweak my method in order to enhance the results.

For COX 2:


Lynx Canadensis
Lynx Lynx
Lynx Rufus
Pardofelis Marmorata
Pardofelis Badia
Pardofelis Temmincki
Otocolobus Manul
Acinonyx Jubatus
Puma Concolor
Puma Yaguarondi
Prionailurus Bengalensis
Prionailurus Viverrinus
Prionailurus Planiceps
Prionailurus Rubiginosus
Caracal Caracal
Felis Chaus
Felis Nigripes
Felis Margarita
Felis Catus
Felis Silvestris
Leopardus Geoffroyi
Leopardus Guigna
Leopardus Tigrinus
Leopardus Jacobita
Leopardus Colocolo
Leopardus Pardalis
Leopardus Wiedii
Neofelis Nebulosa
Panthera Tigris
Panthera Leo
Panthera Uncia
Panthera Onca
Panthera Pardus


Otocolobus_Manul
Puma_Yaguarondi
Puma_Concolor
Acinonyx_Jubatus
Leopardus_Tigrinus
Leopardus_Guigna
Leopardus_Geoffroyi
Leopardus_Wiedii
Leopardus_Pardalis
Leopardus_Jacobita
Leopardus_Colocolo
Prionailurus_Rubiginosus
Prionailurus_Planiceps
Prionailurus_Viverrinus
Prionailurus_Viverrinus
Prionailurus_Bengalensis
Panthera_Tigris
Panthera_Pardus
Panthera_Onca
Panthera_Uncia
Panthera_Leo
Neofelis_Nebulosa
Felis_Nigripes
Felis_Chaus
Felis_Margarita
Felis_Silvestris
Felis_Catus
Caracal_Caracal
Caracal_Caracal
Pardofelis_Marmorata
Pardofelis_Temmincki
Pardofelis_Badia
Lynx_Rufus
Lynx_Lynx
Lynx_Canadensis

3. Conclusion

In conclusion, building phylogenetic trees is no easy feat, it is a valuable tool that truly if implemented in python without the use of intermediate files or command lines would require additional time for thorough research and knowledge of the great science that is comparative genomics. Unfortunately a comparison between all three algorithms could not be obtained because the main argument the tree could not be constructed, there was information required to build newick string or simply reviews explaining the procedures that could not be found. However the team plans to keep searching and studying the topic more thoroughly to make these algorithms applicable.

ACKNOWLEDGMENTS

We would like to extend our deepest gratitude to our instructor Dr. Eman for her unwavering support and guidance throughout this project. Her invaluable insights and advice have been instrumental in the successful completion of this project. Her dedication and commitment to teaching have been an inspiration to us and have greatly contributed to our personal and professional growth. In addition, we would also like to express our sincere appreciation to the teacher assistant Mrs. Reem for her help, support, understanding, and guidance during the course.

CONTRIBUTION

Aly:

1. Perform mainly all work related to the alignment-free method.
2. Conduct the “related work and survey” section in this report.
3. Obtain the data about Corona Virus and tlr3-tlr8-nlrc3-calr-ikbke.
4. Write the README.txt file.

Menna:

1. All work related to Maximum Likelihood software and in the report
2. All file refinements used in the algorithm
3. Introduction, abstract, conclusion, maximum likelihood sections in this report

Rahma:

1. All work related to Maximum parsimony algorithm.
2. The problem definition and proposed methods, maximum parsimony sections in this report.
3. The data about COX1,COX2,COX3 genes.

REFERENCES

[1] Rohlf, F. J. (2005). J. Felsenstein, Inferring Phylogenies, Sinauer Assoc., 2004, pp. xx + 664. Journal of Classification, 22(1), 139–142. doi:10.1007/s00357-005-0009-4

[2] Carl R. Woese,(2000). Interpreting the universal phylogenetic tree,Proceedings of the National Academy of Sciences, 97(15),8392-8396.doi:10.1073/pnas.97.15.8392

[3] John P. Huelsenbeck, Jonathan P. Bollback, Amy M. Levine, (2002). Inferring the Root of a Phylogenetic Tree, Systematic Biology, 51(1), 32–43, https://doi.org/10.1080/106351502753475862

[4] Marten Winter, Vincent Devictor, Oliver Schweiger,(2013).Phylogenetic diversity and nature conservation: where are we?,Trends in Ecology & Evolution, 28(4), 199-204. doi:https://doi.org/10.1016/j.tree.2012.10.015.

[5] S. Sivashankari and P. Shanmughavel, “Comparative genomics - a perspective,” Bioinformation, vol. 1, no. 9, pp. 376–378, 2007.

[6] Stéphane Guindon, Olivier Gascuel, A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood, Systematic Biology, Volume 52, Issue 5, 1 October 2003, Pages 696–704, https://doi.org/10.1080/10635150390235520

[7] “parsimony - Understanding Evolution,” Jul. 16, 2020. https://evolution.berkeley.edu/glossary/parsimony/ (accessed Jan. 20, 2023).

[8] “Building trees using parsimony - Understanding Evolution,” Jul. 04, 2021. https://evolution.berkeley.edu/the-tree-room/how-to-build-a-tree/building-trees-using-parsimony/

[9] D. Searls, “Computational biology,” Encyclopedia Britannica. https://www.britannica.com/science/computational-biology

[10] O. Saha, Md. S. Hossain, and Md. M. Rahaman, “Genomic exploration light on multiple origin with potential parsimony-informative sites of the severe acute respiratory syndrome coronavirus 2 in Bangladesh,” Gene Reports, vol. 21, p. 100951, Dec. 2020, doi: 10.1016/j.genrep.2020.100951.

[11] D. Hang, E. Torng, C. Ofria, and T. M. Schmidt, “The effect of natural selection on the performance of maximum parsimony,” BMC Evolutionary Biology, vol. 7, no. 1, Jun. 2007, doi: 10.1186/1471-2148-7-94.

[12] A. Kundu, R. T. Usha, N. K. Samia, and M. M. Rahman, “Alignment-free phylogenetic tree estimation,” in *2019 Innovations in Power and Advanced Computing Technologies (i-PACT)*, 2019.

[13] G. Munjal, M. Hanmandlu, and S. Srivastava, “Phylogenetics Algorithms and Applications,” in *Advances in Intelligent Systems and Computing*, Singapore: Springer Singapore, 2019, pp. 187–194

[14] Koichiro Tamura, Glen Stecher, and Sudhir Kumar (2021) MEGA11: Molecular Evolutionary Genetics Analysis version 11. *Molecular Biology and Evolution* 38:3022-3027

[15] RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies". In Bioinformatics, 2014.

[16] Ayres, D.L., A. Darling, D.J. Zwickl, P. Beerli, M.T. Holder, P.O. Lewis, J.P. Huelsenbeck, F. Ronquist, D. L. Swofford, M. P. Cummings, A. Rambaut, and M. A. Suchard. 2012. BEAGLE: an application programming interface for statistical phylogenetics. Syst. Biol. 61:170-173.