



الجمهورية الجزائرية الديمقراطية الشعبية
People's Democratic Republic of Algeria
وزارة التعليم العالي والبحث العلمي
Ministry of Higher Education and Scientific Research



University of Constantine 1 Mentouri Brothers
Faculty of Natural and Life Sciences

جامعة قسنطينة 1 الإخوة منتوري
كلية علوم الطبيعة والحياة

Department: Applied Biology

قسم : البيولوجيا التطبيقية

Dissertation presented with a view to obtaining the Master's Diploma

Domain: Natural and Life Sciences

Sector: Biotechnology

Speciality: Bioinformatics

Order number:

Series number:

Title:

Comparison of phylogenetic inference methods.

Presented by: Chettah Imène

Chial Affef

On: 10th june ,2024.

Evaluation jury:

President: Dr.Medjroubi Med ElArbi (**MCB** in University of Mentouri Brother's Constantine 1).

Supervisor: Dr. DAAS Mohamed Skander (**MCA** in University of Mentouri Brother's Constantine 1).

Examiner: Dr. Gherboudj Amira (**MCA** in University of Mentouri Brother's Constantine 1).

Academic year : 2023 – 2024.

Acknowledgments

Acknowledgments

Alhamdulillah , no work can be accomplished without the help of the most powerful God. So, and first of all, we thank our Creator who gave us the strength and health to be here today.

*We would like to take this opportunity to extend our deep thanks and our deep appreciation to: Our supervisor of this memory, **Dr. DAAS Mohamed Skander**, who would like us to thank sincerely for his valuable advice and help throughout the period of work, and without whom this memory would never have been born.*

*We also thank the members of the jury, President **Dr. Medjroubi Mohamed ElArbi** and Examiner **Dr. Gherboudj Amira**, for giving us the honour to preside over and judge our work.*

To all the professors of the Bioinformatics specialty ,who taught us and who by their skills supported us in the continuation of our studies.

To our parents, the whole family and our friends who through their Doua'a and their encouragement, we were able to overcome all obstacles.

To any person who has participated near or far in the execution of this modest work.

Thanks to all of you.

Dedicate

Dedicate

Alhamdulillah who gave us the strength and courage to complete this work. With the expression of our gratitude, we dedicate this modest work to those who, whatever the terms embraced, we would never be able to express our sincere love for them.

To my dear mother Meriem

*You have given me the life, the tenderness and the courage to succeed in everything that I can offer you will not be able to express the love and appreciation that I carry to you, May this modest work be the fulfillment of your wishes, the fruit of your countless sacrifices. May God, the Most High, grant you health, happiness and long life,
I love you.*

To my dear father Ahmed

*My father, who can be now proud and find here the result of long years of sacrifices and deprivations to help me move forward in life,
I love you.*

To my lovely sisters:

Zayneb, Oum Koulthoum and Rokaya .

To my nephews and nieces:

*Nizar Nedjm-Elddine , Amir, Aya Tassnim, Amdjed Bahaa , Belkis
Ibtihel , Mazen and the new baby coming.*

To my project partner Affef

For her patience, sympathy, creativity, diligence, and unwavering dedication to every aspect of our work together. Our collaboration has been a real pleasure, I wish her a life full of happiness and success.

*All my friends of promo 2024 speciality bioinformatics and especially
Laib Choumaïssa.*

Imène

Dedicate

To my dear mother Mounia,

The heart and soul of our family. You raised us with love, instilling strong principles that guide us every day. You are our life. May Allah keep you for us. I love you, Mama.

To my dear father Djamel,

you always push us to be the best. Because of you, I am here now.

I love you, Papa

To my siblings Hadjer and Taki Eddine,

Hadjer and Taki Eddine, growing up together, sharing deep conversations have brought me comfort, joy, and inspiration.

May Allah bless you both a bright future I love you.

To my family and specially my aunt Dounia and my cousin Rofia thank you for everything. I am so grateful to have you in my life.

To my dear friends,

Maram, Lyna, and Nahla, your help and support have been invaluable. I am blessed to have you in my life.

To my project partner, Imene,

working with you has been a wonderful journey. I hope life brings you all the amazing experiences you deserve.

Finally, to Chahra, Moncef, thank you for filling our master's journey with laughter, joy, and special experiences.

Affef

Table of contents

Table of contents

List of figures	I
List of tables	II
List of abbreviations	III
Introduction	1

Chapter 01: Theoretical Foundation.

1. What is Phylogenetics?	3
1.1 Why Phylogenetics Is Used?.....	3
1.2 Molecular Phylogeny.....	3
1.3 Phylogenetic Tree.....	3
1.3.1 Rooted & Unrooted Trees.....	3
1.3.2 Key Concepts of Phylogenetic Trees.....	4
1.3.2.1 Operational Taxonomic Units (OTUs) or Leaf Nodes.....	4
1.3.2.2 Hypothetical Taxonomic Unit (HTU) Inferred Ancestors.....	4
1.3.2.3 Root Node.....	4
1.3.2.4 Ancestor.....	4
1.3.2.5 Branch.....	4
1.3.2.6 Branch Length.....	5
2. Monophyletic, Paraphyletic, Polyphyletic Groups	5
2.1 Monophyletic Group.....	5
2.2 Paraphyletic Group.....	5
2.3 Polyphyletic Group.....	5
3. Different Graphical Representations for Trees	6
3.1 Dendrogram.....	6
3.2 Cladogram.....	6
3.3 Phylogram.....	6
3.4 Phenograms.....	7

Table of contents

3.5 The Newick Format.....	7
4. Methods for Phylogenetic Tree Construction.....	8
4.1 Distance-based Methods.....	8
4.1.1 Unweighted Pair Group Method with Arithmetic Mean.....	8
4.1.2 Minimum Evolution Method.....	9
4.1.3 Neighbor Joining Algorithm.....	9
4.2 Character-based Methods.....	10
4.2.1 Maximum Parsimony Method.....	10
4.2.2 Maximum Likelihood Method.....	10

Chapter 02: Methodology.

1. Choice of Software.....	12
1.1 RStudio.....	12
1.2 IQ-TREE.....	12
1.3 MEGA Software.....	13
2. Database Generation and Protocol.....	13
3. Evaluation Using RF Distance and Mean Branch Length.....	17

Chapter03:Results and Discussion.

1. Results of Robinson Foulds distance.....	21
2. Results of Mean Branch Length distance.....	24
3. RF distance histograms	27
3.1Variation in length of sequences	27
3.2Variationin number of sequences.....	28

Table of contents

3.3Variation In Insertion-Deletion rate	30
4. Mean Branch Length distance histograms	32
4.1Variation in length of sequences	32
4.2Variation in number of sequences	34
4.3Variation in Insertion-Deletion rate.....	35
5. General performance evaluation	37
Conclusion	40
References.....	44

List of figures

List of figures

Figure 01: Example of a Phylogenetic tree.....	5
Figure 02: Monophyletic, Paraphyletic, Polyphyletic groups.....	5
Figure 03: A: Cladogram, B: Rectangular Cladogram.....	6
Figure 04: Phylogram.....	7
Figure 05: Example of a Newick format.....	7
Figure 06: Methods for Phylogenetic Tree-Construction.....	8
Figure 07: Code of TreeSim package in RStudio.....	14
Figure 08: Tree_n8 generated by TreeSim.....	14
Figure 09: Fasta format of Tree_n8 (reference tree).....	15
Figure 10: Aligned sequences.....	16
Figure 11: Construction of phylogenetic trees with MEGAX software by different methods (ML, MP, NJ, UPGMA and ME).....	16
Figure 12: Construction results of the phylogenetic tree_n8 with a sequence length of 100, indel rate of 0.001, using the Minimum Evolution Method, exported in Newick format.....	17
Figure 13: Python Code of metric Robinson-Foulds (RF) distance.....	18
Figure 14: Python code of metrics Mean and Standard Error distances.....	19
Figure 15: Histogram represents Robinson-Foulds (RF) distance in function of sequences length.....	27
Figure 16: Histogram represents Robinson-Foulds (RF) distance in function of number of sequences.....	28
Figure 17: Histogram represents Robinson-Foulds (RF) distance in function of Insertion-Deletion rate.....	30
Figure 18: Histogram represents Mean Branch Length distance in function of sequences length.....	32
Figure 19: Histogram represents Mean Branch Length distance in function of number of sequences.....	33
Figure 20: Histogram represents Mean Branch Length distance in function of Insertion-Deletion rate.....	35

List of tables

List of tables

Table 01: RF distance calculated for 8 taxa with all variations of indels and sequence lengths.....	21
Table 02: RF distance calculated for 16 taxa with all variations of indels and sequence lengths.....	22
Table 03: RF distance calculated for 24 taxa with all variations of indels and sequence lengths.....	22
Table 04: RF distance calculated for 32 taxa with all variations of indels and sequence lengths.....	23
Table 05: RF distance calculated for 40 taxa with all variations of indels and sequence lengths..	23
Table 06: Mean Branch Length distance calculated for 8 taxa with all variations of indels and sequence lengths.....	24
Table 07: Mean Branch Length distance calculated for 16 taxa with all variations of indels and sequence lengths.....	24
Table 08: Mean Branch Length distance calculated for 24 taxa with all variations of indels and sequence lengths.....	25
Table 09: Mean Branch Length distance calculated for 32 taxa with all variations of indels and sequence lengths.....	25
Table 10: Mean Branch Length distance calculated for 40 taxa with all variations of indels and sequence lengths.....	26

List of abbreviations

List of abbreviations

1. **CMD**: **Command Line**.
2. **DNA**: **Deoxyribonucleic Acid**.
3. **FASTA**: **Federal Acquisition Streamlining Act**.
4. **HTU's**: **Hypothetical Taxonomic Units**.
5. **IDE**: **Integrated Development Environment**.
6. **IQ-TREE**: **Intelligent Quick Tree**.
7. **ME**: **Minimum Evolution**.
8. **MEGA**: **Molecular Evolutionary Genetics Analysis**.
9. **ML**: **Maximum Likelihood**.
10. **MP**: **Maximum Parsimony**.
11. **NJ**: **Neighbor-Joining**.
12. **OTU's**: **Operational Taxonomic Units**.
13. **RF**: **Robinson-Foulds**.
14. **RStudio**: **Integrated Development Environment for R and Python**.
15. **UPGMA**: **Unweighted Pair Group Method with Arithmetic Mean**.

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

Introduction

Introduction

Phylogenetics is a vital subfield of evolutionary biology, examines the relationships among different species or organisms. By analyzing molecular data such as DNA and protein sequences, scientists construct phylogenetic trees to visualize evolutionary pathways, understand mechanisms of diversification, and identify common ancestors. These trees are crucial not only in biology but also in ecology, medicine, and biodiversity conservation. They facilitate the tracing of genetic and genomic evolutionary histories, infer ancestral characteristics, and predict gene functions (Felsenstein, 2004). Phylogenetic analysis is fundamental in various fields, including systematics and genomics. It aids in tracing the origins of species, studying gene evolution, understanding the evolutionary basis of diseases and traits, and tracking the origins and transmission pathways of pathogens.

The main goal of this research is to evaluate and compare the effectiveness of various phylogenetic tree construction methods. By using reference dataset, we aim to identify which method yields the most accurate and reliable trees in terms of reconstructing evolutionary relationships. The specific objectives of this study are: first, to compare the performance of different phylogenetic tree construction methods using two key metrics: Robinson-Foulds (RF) distance and Mean Distance distance; Second ,to determine the strengths and weaknesses of each method regarding accuracy and robustness; and to provide recommendations for best practices in constructing phylogenetic trees across various contexts.

There are several approaches to generating phylogenetic trees, each with unique assumptions, algorithms, and applications. Among the most used methods are:

Distance-based methods: are UPGMA (Unweighted Pair Group Method with Arithmetic Mean), Neighbor-Joining (NJ) and Minimum Evolution (ME) use distance matrices to construct trees (Saitou & Nei, 1987).

Character-based methods: These methods include Maximum Parsimony (MP) and Maximum Likelihood (ML), which directly use sequence data to infer trees (Felsenstein, 1981).

Each method has its advantages and limitations, and the choice of method often depends on the nature of the data, research questions, and available resources. This study aims to thoroughly examine these methods and provide empirical comparisons based on reference datasets.

Chapter 01:

Theoretical foundation

1. what is Phylogenetics ?

The word "phylogeny" comes from the two Greek words "phûlon" which means "tribe, clan, family" etc. or "species", and "g genesis" meaning "creation or origin". The so-called "word" phylogenesis was coined in 1866 by the German biologist Ernst Haeckel. It is the history of the evolution of a genetically related group of organisms(Haeckel, 1860).

1.1 Why Phylogenetics Is Used?

The goal of phylogenetics is to illustrate or deduce links between organisms, establishing or deducing evolutionary or ancestral links. It involves calculating the time since an organism's last common ancestor. The tree of life, or phylogeny, is a branching structure used in phylogenetics research to illustrate relationships.

1.2 Molecular phylogeny :

Molecular phylogeny is the study of evolutionary or ancestral links between organisms, groups of organisms, or genes. Molecular data, such as DNA and protein sequences, are used for reconstructing or inferring these linkages(T.Dandekar and M.kunz ,2011).

1.3 Phylogenetic Tree:

A phylogenetic tree is a graph with only one path connecting any two nodes, visually depicting the evolutionary or ancestral relationships between genes or organisms(T.Dandekar and M.kunz ,2011).

1.3.1 Rooted & Unrooted Trees:

- **Rooted Trees:** A tree's molecular phylogeny has the oldest point, the root, determining the elongated time on each branch. The feeling of temporal evolution on all branches is defined by this moment.
- **Unrooted Trees:** However, molecular phylogeny techniques can only reconstruct an unrooted tree without the extended time sense. It is impossible to reconstruct a rooted tree. An unrooted tree is measured according to the traditional graphism in which the time on each branch is not clear. But each branch has a definite length. Determine, without resolution, any unrooted tree. This will represent part of the unrooted tree's area. It is biologically defended. Find the separation between leaves. The benefit of the way it is presented is that it can identify any unrooted tree. But the strong theory of the molecular clock is not always applicable (G. Del age and al., 2021).

Theoretical foundation

1.3.2 key concepts of phylogenetic trees:

1.3.2.1 Operational Taxonomic Units (OTUs) or Leaf Nodes:

Leaf nodes, also known as Operational Taxonomic Units (OTUs), are the outermost terminal nodes of a phylogenetic tree. They represent the chemicals or creatures being compared during the tree-building process.

1.3.2.2 Hypothetical Taxonomic Unit (HTU) Inferred Ancestors or Internal Nodes:

Internal nodes of the tree indicate hypothesized shared ancestors who introduced two independent lineages at some point in the past. These hypothetical structures imply a common ancestor among the studied organisms or substances.

1.3.2.3 Root Node:

The root node represents the last common ancestor of all species or molecules being compared, forming the base of the phylogenetic tree. Identifying the correct root node often requires additional historical and physical evidence beyond molecular data.

1.3.2.4 Ancestor:

An ancestor is an assumed common progenitor from which two or more independent lineages originated, typically represented by interior nodes within the evolutionary tree.

1.3.2.5 Branch:

A branch in a phylogenetic tree represents an evolutionary pathway or relationship between nodes, visually connecting them and indicating the divergence or divergence that occurred over time. Branches convey information about the duration and magnitude of evolutionary changes between nodes (T.Dandekar and M.kunz, 2011).

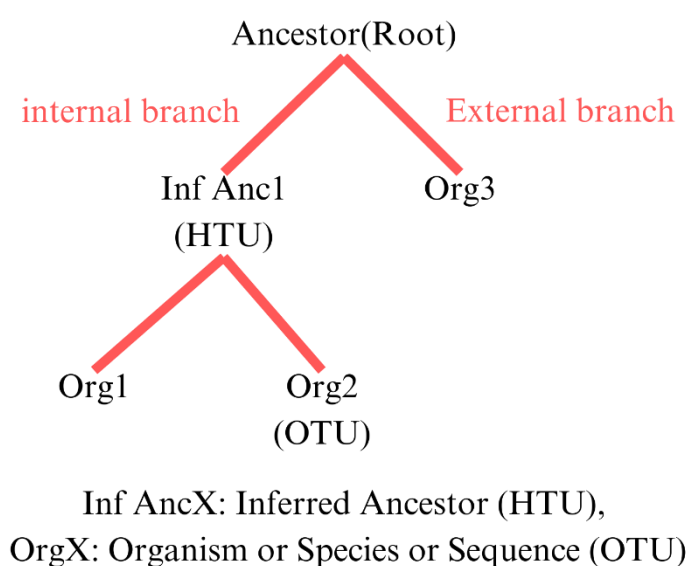


Figure 01: Example of a Phylogenetic tree.

Theoretical foundation

Branches represent the molecular evolution.

- The internal branch: link two OTU
- The external branch: link HTU and OTU

1.3.2.6 Branch length:

The number of ways to replace a residue with a different one at each site in the molecule studied along this branch (unit: substitution number per site).

The branch length is proportional to the evolutionary distance between the sequences and their ancestry.

2. Monophyletic, Paraphyletic, Polyphyletic groups:

2.1. Monophyletic group:

A monophyletic group or a clade is a taxon consisting of two or more species, including an ancestral species and all descendants of that ancestral species (Figure 02).

2.2. Paraphyletic group:

Paraphyletic groups are incomplete groups, which one or more descendants of a common ancestor do not belong to in group (Figure 02).

2.3. Polyphyletic group:

Polyphyletic groups are composed of descendants of ancestors not included in the group definitively (E. O. Wiley and B.S. Lieberman, 2011).

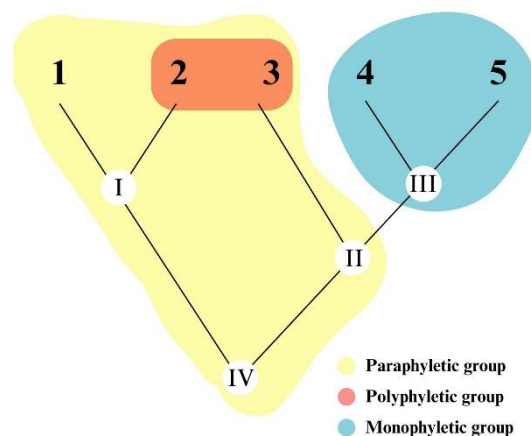


Figure 02: Monophyletic, Paraphyletic, Polyphyletic groups.

3. Different graphical representations for trees:

3.1. Dendrogram:

The term "dendrogram" refers to a vertical cluster organization in which similar objects are grouped together based on predetermined criteria.

A dendrogram therefore illustrates the relationships between the various groupings. Dendrograms are also used outside of the field of phylogenetics, possibly even outside of biology.

3.2. Cladogram:

The cladogram is a hierarchical tree with branches that illustrates the relationships between the classes. The cladograms are not nested (Figure 03).

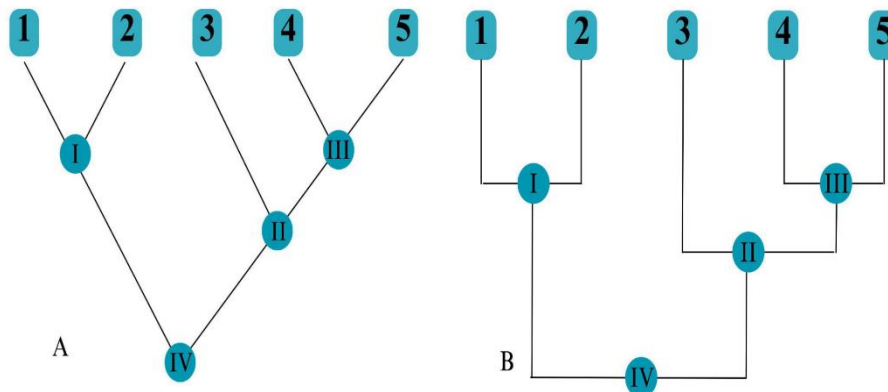


Figure 03:A:Cladogram, B:Rectangular Cladogram .

3.3. Phylogram:

Phylogenetic trees at scale, or phylogenetic trees, have branch lengths that correspond to the degree of evolutionary divergence (Figure04). For example, the number of nucleotide changes that occur between connected branch sites can be used to determine the length of a branch(Choudhuri, 2014).

Theoretical foundation

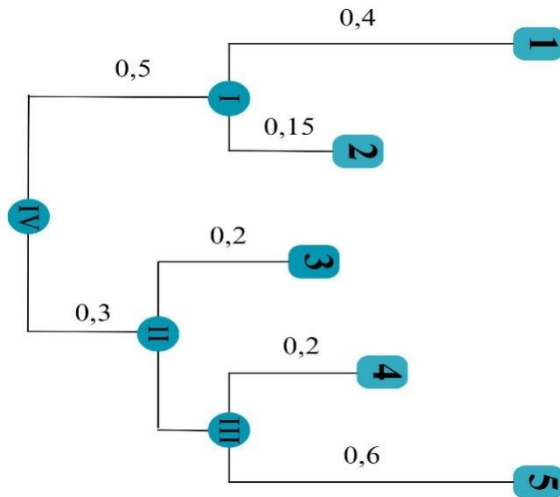


Figure 04: phylogram.

3.4. Phenograms :

These are trees that represent the parent-child relationships between molecules. Based on phenotypic methods, they are a type of dendrogram produced by digital taxonomy, where the relationships between taxa represent the global similarity degrees (Tahiri, 2012) .

3.5. The Newick format :

The Newick format is a common computer format for writing phylogenetic trees. The format's name comes from the tiny group of researchers that developed it in a New England town (Figure 05).

Two brotherly groups, A and B, make comprise the basic structure of a tree (A, B). If the two species C and D are found in the second group, it is written (A. (C, D)).

Lengthening the terminal branches la, lc, and ld as well as the ancestral branch at group C+D, lcd, gives the following : (A: la,(C: lc,D: ld): lcd); in which case the terminal point-virgule is required and indicates the end of the tree(Del eage and al., 2021).

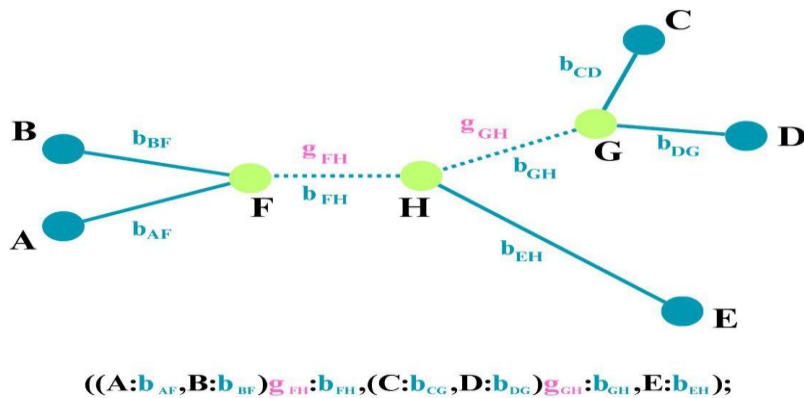


Figure 05: Example of a Newick format .

4. Methods for Phylogenetic Tree-Construction:

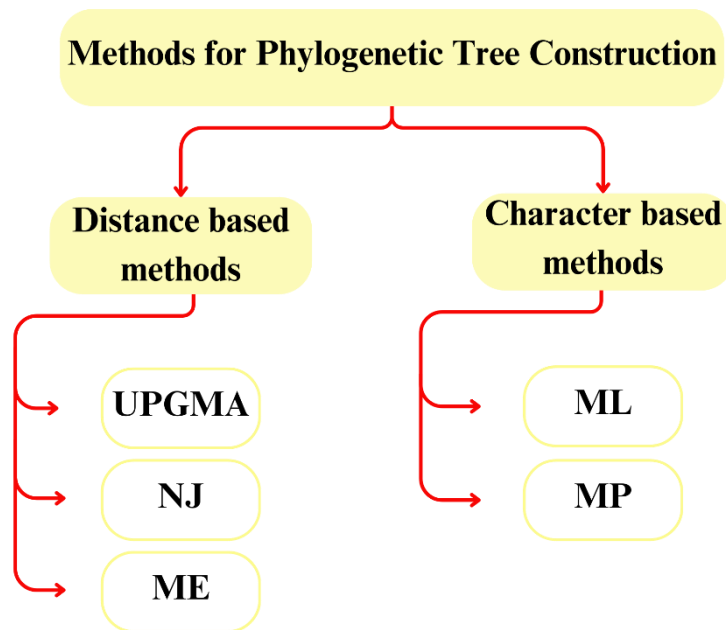


Figure 06:Methods for Phylogenetic Tree-Construction

4.1. Distance-based Methods:

Several tools for reconstructing phylogenetic trees from sequence data use the evolutionary distances between sequences as intermediate data. These tools allow for the two-step building of a tree. First, the evolutionary distances between each pair of aligned sequences are calculated. The tree is then calculated using distances, no longer requiring the use of sequences.

- **Measure of Similarity & Distinction Matrix:**

"Distance" is the number of differences between two sequences when they are aligned, and it is the most basic way to calculate the molecular similarity between two sequences. Additionally, a matrix format known as the Distance Matrix is used to display the pairwise distances between any two sequences of a collection of sequences that have been aligned by multiple sequence alignment.

4.1.1. Unweighted Pair Group Method with Arithmetic Mean (UPGMA):

The Unweighted Pair Group Method with Arithmetic Mean (UPGMA) is possibly the most widely used and straight forward distance-based hierarchical clustering approach for creating phylogenetic trees, based on the evolutionary distances between each pair of sequences.

Theoretical foundation

It is sufficient to look for the smallest distance among all of these (T.Dandekar and M.kunz,2011).

4.1.2. Minimum Evolution method (ME):

The minimum evolution method (ME), or minimum evolution score method (MES), is a distance-based tree construction method that seeks to establish tree topology that minimizes the sum of lengths of branches (or the total amount of evolutionary change) necessary to explain the observed distances between taxa (G.Delèage and al., 2021).

4.1.3. Neighbor Joining Algorithm (NJ):

The minimal evolution method computes for all conceivable tree topologies, which makes it impossible to handle large number of leaf trees in a reasonable amount of time. The Neighbor-Joining method offers a very accurate approximation of the minimal evolution method that is mathematically sound enough to allow for the analysis of hundreds of sequences. The main idea is to just consider a few very specific tree topologies and apply the minimal evolution method to them (T.Dandekar and M.kunz,2011).

4.2. Character-based Method:

Character-based methods consider the accumulated mutation events on the sequences, thus avoiding loss of information. It easily brings information about homoplasy and ancient states. It generates trees more accurately than distance methods. Maximum Parsimony and Maximum Likelihood are the two most commonly used methods (Patwardhan and al, 2014).

4.2.1. Maximum Parsimony (MP) Method:

Probably the most often used character-based cladistic approach is Maximum Parsimony. Using this strategy, one can anticipate the phylogenetic tree that yields the fewest character changes (mutations) required to account for the observed variances in the sequences. Thus, out of all the options, the maximum parsimony approach creates the most parsimonious tree in terms of mutation (change) score (T.Dandekar and M.kunz ,2011).

4.2.2. Maximum Likelihood (ML) Method:

Maximum Likelihood (ML) is the most established and generalized character-based method for the inference of phylogeny, this technique uses a clear evolutionary model to reconstruct a phylogeny. Rather than just counting the mutations, the maximum likelihood

Theoretical foundation

method gives quantitative probability (from the substitution rate matrix) to mutational events. Maximum probability constructs the tree similarly to maximum parsimony, but it determines branch length by considering the likelihood of the hypothesized mutational events. The process looks for the tree that has the highest likelihood or probability (T.Dandekar and M.kunz,2011).

Chapter 02:
Methodology

Introduction

In this chapter, we explore the software and methodologies used for simulating and analyzing phylogenetic trees to achieve a comprehensive comparison of phylogenetic inference tools, this section details the systematic approach adopted for the selection of software and tools, generation of databases, and the evaluation criteria used to compare the performance of the selected tools using metrics such as **Robinson-Foulds** and **Mean Branch Length** distances.

1. Choice of software:

1.1 RStudio:

RStudio is an integrated development environment (IDE) for R and Python. It includes a console, syntax-highlighting editor that supports direct code execution, and tools for plotting, history, debugging, and workspace management. RStudio is available in open source and commercial editions and runs on the desktop (Windows, Mac, and Linux) (website 1).

Simulation of Phylogenetic Trees with TreeSim:

In this study, we used the TreeSim package in R to simulate phylogenetic trees. TreeSim is a versatile tool that allows for the simulation of trees under various birth-death process scenarios. For our analysis, we specifically used the “sim.bd.taxa()” function of TreeSim.

1.2 IQ-TREE:

In our study, we used IQ-TREE, a robust and efficient software for phylogenetic tree inference. IQ-TREE has been under continuous development since its introduction in 2011 and is recognized for its effectiveness and precision in phylogenetic analyses. As an open-source tool, it is extensively used in evolutionary biology to indicate evolutionary relationships among species based on their genetic data, which includes DNA or protein sequences. Our research specifically involved the use of IQ-TREE for inferring phylogenetic trees (website 2).

AliSim: Integrated Sequence Simulation Tool

AliSim is a feature integrated within IQ-TREE for simulating phylogenetic sequences. It is designed to simulate biologically realistic sequence alignments under a broad spectrum of complex evolutionary models. AliSim achieves high performance by implementing an

adaptive approach that combines the commonly-used rate matrix and probability matrix approach.

1.3 Mega software:

In this study, we employed the MEGA (Molecular Evolutionary Genetics Analysis), it's a robust and user-friendly software suite designed for the comparative analysis of DNA and protein sequences from species and populations. Developed in C++ (Website 4), MEGA provides a wide range of tools for conducting molecular evolutionary genetics analyses, including phylogenetic tree construction, sequence alignment, and evolutionary distance estimation. It is widely used in the fields of evolutionary biology and bioinformatics to estimate evolutionary distances and reconstruct phylogenetic trees (website 3).

2.Database Generation and protocol of construction the Phylogenetic Trees:

In our study, we employed a comprehensive methodology to simulate and evaluate phylogenetic trees using a combination of advanced software tools and analytical techniques. Initially, we used TreeSim package in RStudio for simulating phylogenetic reference trees based on various birth-death process scenarios. Specifically, we employed the `sim.bd.taxa()` function. Here are the main features of this function:

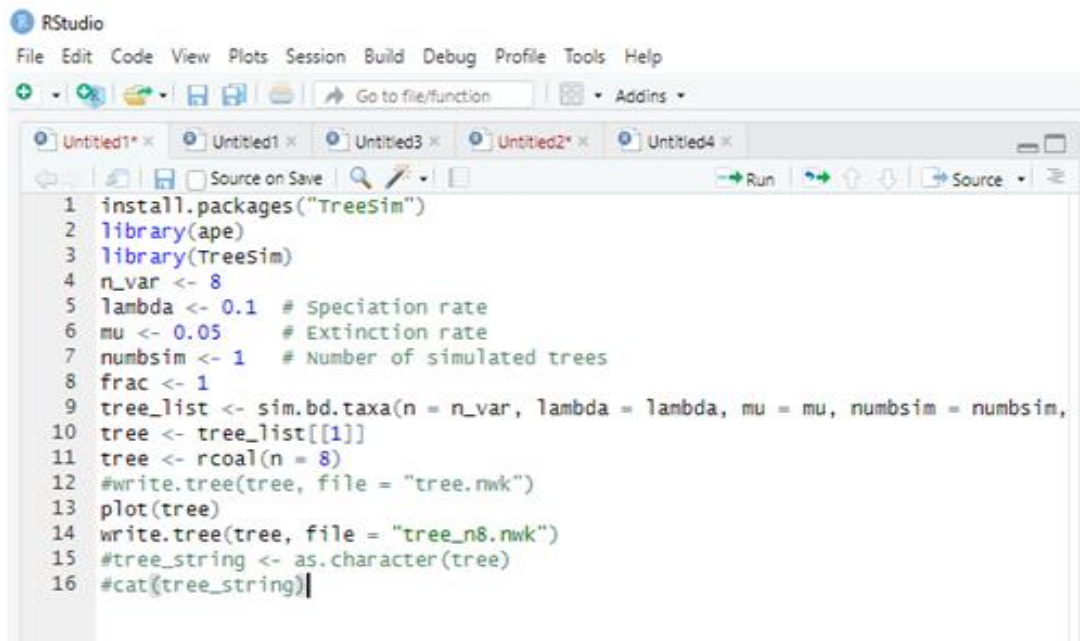
- **Conditioning on the number of taxa:** “`sim.bd.taxa()`” simulates trees where the process stops once the specified number of tips is reached.
- **Function parameters:**
 - **n:** the desired number of taxa.
 - **lambda:** the speciation rate.
 - **mu:** the extinction rate.
 - **Numbsim:** number of simulations repeats.

Simulation Process:

1. **Parameter Definition:** We defined the speciation and extinction rates based on preliminary data and literature estimates (Figure 07).
2. **Tree Generation:** Using “`sim.bd.taxa(n, lambda, mu, numbsim)`” we generated a set of phylogenetic trees for different numbers of tips (Figure 08).

Methodolgy

3. **Saving Trees:** Each generated tree is saved in Newick format using the `write.tree()` function from the `ape` package, with filenames formatted as "tree_n8.newick", "tree_n16.newick"... Etc (Figure 09).



```
1 install.packages("TreeSim")
2 library(ape)
3 library(TreeSim)
4 n_var <- 8
5 lambda <- 0.1 # Speciation rate
6 mu <- 0.05 # Extinction rate
7 numbsim <- 1 # Number of simulated trees
8 frac <- 1
9 tree_list <- sim.bd.taxa(n = n_var, lambda = lambda, mu = mu, numbsim = numbsim,
10 tree <- tree_list[[1]]
11 tree <- rcoal(n = 8)
12 #write.tree(tree, file = "tree.mwk")
13 plot(tree)
14 write.tree(tree, file = "tree_n8.mwk")
15 #tree_string <- as.character(tree)
16 #cat(tree_string)]
```

Figure 07: Code of TreeSim package in RStudio.

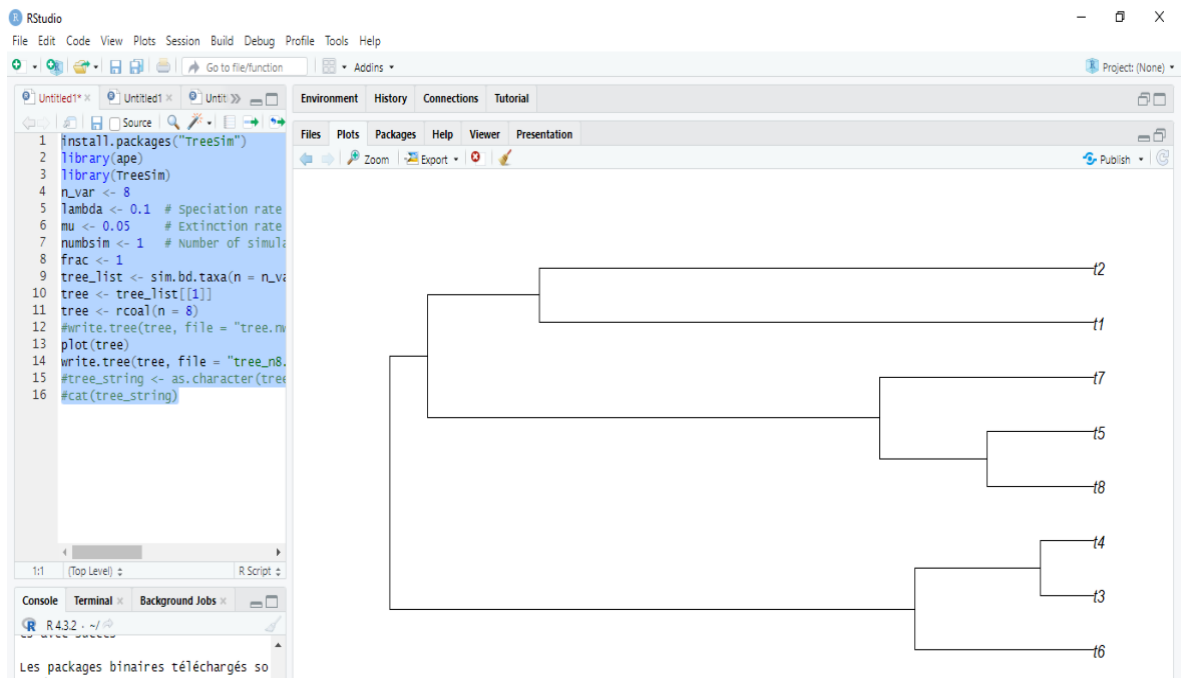


Figure 08: Tree_n8 generated by TreeSim .

Methodolgy

This allowed us to generate trees with a predetermined number of taxa. The simulated trees were saved in Newick format (**Figure 09**).

```
((t6:0.1262418752, (t3:0.03769733294, t4:0.03769733294) :0.088544
5423) :0.3700964053, ((t8:0.07541802444, t5:0.07541802444) :0.075
24559276, t7:0.1506636172) :0.3186205766, (t1:0.3906757701, t2:0.3
906757701) :0.07860842363) :0.02705408679);
```

Figure 09: Fasta format of Tree_n8 (reference tree).

To generate the data sets for constructing phylogenetic trees, we used the IQ-TREE tool. We simulated trees with varying numbers of taxa (8, 16, 24, 32, 40) and different sequence lengths (100, 200, 300, 400, 500). Each sequence length was subjected to five different indel rates (0.001, 0.005, 0.010, 0.015, 0.020). These parameters enabled the creation of a diverse set of data for robust analysis.

IQ-TREE, renowned for its precision and efficiency, facilitated the simulation of sequence alignments through its integrated AliSim feature, we employed the following command in CMD:

```
iqtree2—alisim N8_L100_INDEL0.001 -m LG-t tree8_n8.nwk --length 100 --indel
0.001,0.001 --out-format- fasta
```

This command allows for the specification of:

- **N8_L100_INDEL0.001:** This is a custom prefix for the output file names, indicating the simulation parameters.
- **-m:** Specifies the evolutionary model to be used for the simulation.
- **LG:** substitution model for protein sequences.
- **-t:** Specifies the input tree file in Newick format.
- **tree_n8.nwk:** The name of the reference tree file used as the basis for the simulation.
- **--length 100:** Specifies the length of each sequence in the simulated alignment (100 base pairs).
- **--indel 0.001,0.001:** Indicates the same rate for both insertions and deletions (0.001).
- **--out-format fasta:** Specifies the format of the output alignment file (FASTA format).

Methodolgy

The results of this command provide both aligned and unaligned sequences; in our study, we used only the aligned sequences (Figure 10).

```
Open + N8_L100_Indel0.001.fa Save ...
~/Desktop/traes/tree_n8/L100/N8_L100_Indel0.001
>t8
2 KKQCDYPYKNFMNFGQDSVLMKFTQMYQIIMMSQLSAVGRKVPVAVIADSAAVMLITAYELYQSGQQQKDKLLYYVLETSQSDGK-YNILARYEPAPGSN
3 >t7
4 SRQCQNPKYKNVLYLNVDLCLQQAKRYKFLKMSKTTACGRKTPAVMVPDEAGAVILQNGTQVYEVGTSKSGKPFMYVLQSHPNTP-YDTIARQDDGPNGI
5 >t6
6 HKECDNPKYKLLNINQDDALLKAKRYSFLAMTQ.TNAVGDKAPGVMVADAAGAYILQNGTQVYEVGTFNVGKPLFYFLLAHPSEP-YETLARESDDGPGGN
7 >t1
8 YKR.TDNPFKNLLNLMPECLLVQAKRYSFLQ.MSTTTVTGRKTPSMMVAQAAGAKLFQNGTQVYEVGVNNAHRPLFYFLLAHP.TDP.YQALARERNGPGGN
9 >t3
10 TKEVKR.TWRMMV.TL.GSDNAALDKYHFPAILINL.VEGARHPAIIIMDEEAGMAVVQIVL.TTYRDGETQDEELVYFVLK.LH.TGPGDVLQ.LYKHLIRAVD
11 >t4
12 HQ.QETK.VWYK.VIHL.GLENAT.LASYQR.NAMGLIMD.QIYTARK.PAIVIMHEERAMAV.VRIFFIYR.GGEEKDDKAV.FFTMRLS.SSPGDVLELYKHLV.RVAT
13 >t2
14 GREAQRSFKYRIKLGRENALLAQ.YEQ.YAIAVLMRLVEV.FRRPAVIMRNIVELVQ.TRIVFIVYRSGEDQDDKLYFMIMILN.TGPGDVLALYKDLV.RVAV
15 >t5
16 GREAQRTFKWRVKLGKENALLAQ.YEQ.DAVAVLMRLVEVARQP.VAVIMRNIVMDMVQ.TRLV.FVVYRSGEEQDDKLYFMIMILD.TGPGDVLALYKDLV.RVAD
```

Figure 10: Aligned sequences.

After we configured the simulations with specific parameters, sequence lengths, and indel rates, ensuring biologically realistic alignments. The resulting aligned sequences were then analyzed using the Molecular Evolutionary Genetics Analysis (MEGA) software.

The detailed process in MEGA was as follows:

1. Uploading Aligned Sequences: We uploaded our aligned sequences using the "File" button in MEGA.
2. Selecting Protein sequences: We selected this option to analyze our protein sequences.
3. Constructing Trees with Different Methods: We constructed phylogenetic trees using several methods available in MEGA (ML, MP, NJ, UPGMA, and ME).
4. Saving Trees: Finally, we saved the constructed trees in Newick format for further analysis and visualization.

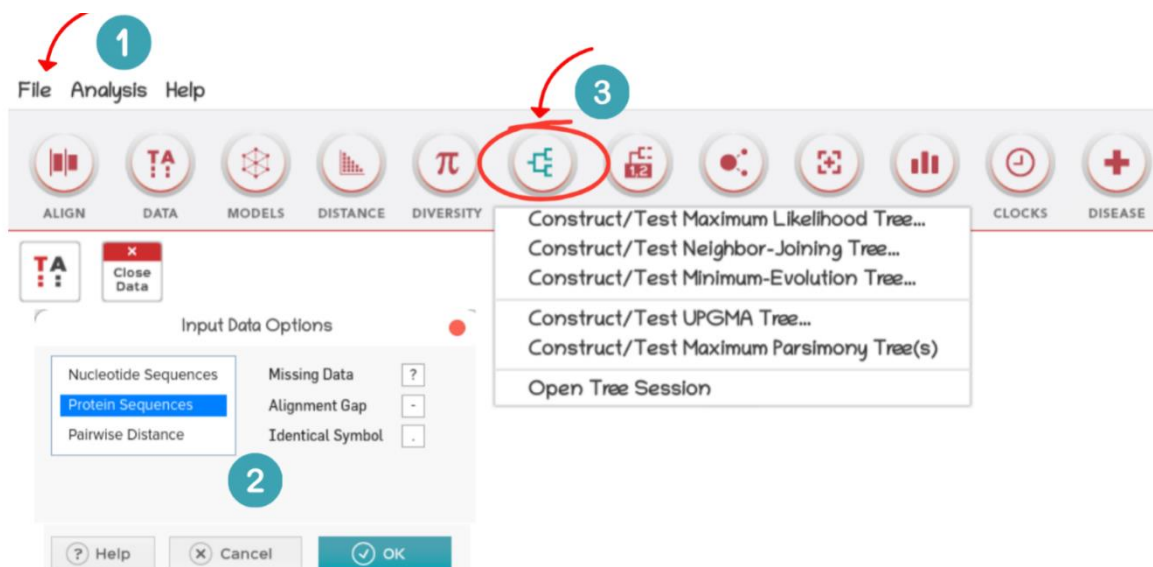


Figure11: Construction of phylogenetic trees with MEGAX software by different methods (ML, MP, NJ, UPGMA and ME).

Methodology

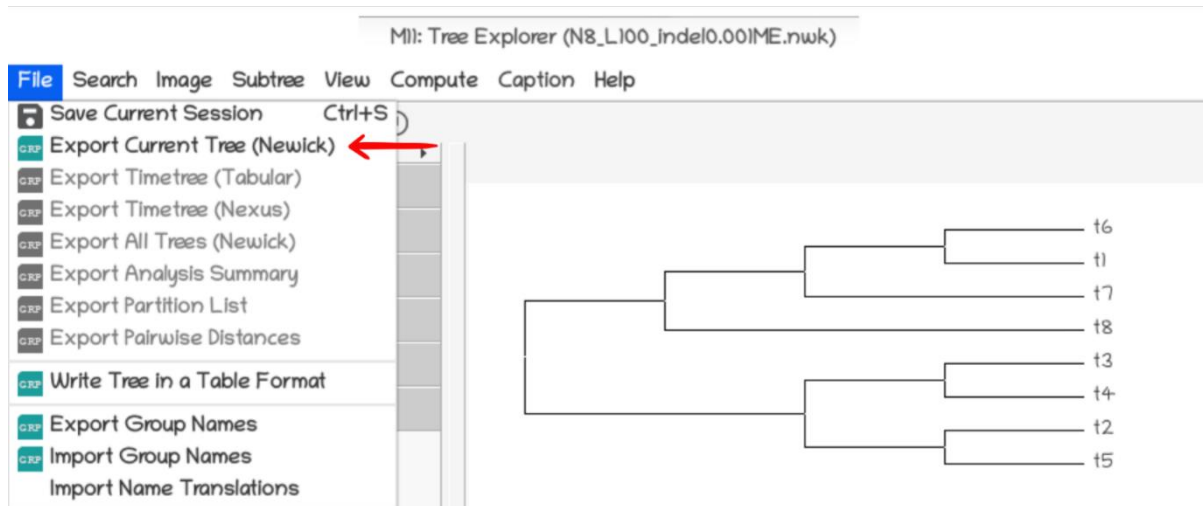


Figure12: Construction results of the phylogenetic tree_n8 with a sequence length of 100, indel rate of 0.001, using the Minimum Evolution Method, exported in Newick format.

To assess the performance of these phylogenetic tree construction methods, we employed metrics such as Robinson-Foulds (RF) distance (Figure13) and Mean distance (Figure14). These metrics provided a comprehensive framework for evaluating the topological dissimilarity, branch length variations in the phylogenetic relationships inferred by different methods.

3. Evaluation of Phylogenetic Trees Using Robinson-Foulds, and Mean branch Length Distances:

To compare the reference tree with each one of the trees generated by IQ-TREE, we used two key metrics to assess the performance of different phylogenetic tree construction methods: Robinson-Foulds (RF) Distance and Mean Branch Length Distance.

Robinson-Foulds (RF) distance:

(Robinson and Foulds, 1981), was employed as a key metric to assess the topological dissimilarity between two phylogenetic trees. This measure quantifies the number of partitions or splits in one tree that do not appear in the other, effectively capturing the structural differences between the trees. By counting these discrepancies in branching patterns, the RF distance provides an indication of how similar or different the trees are. A lower RF distance denotes a greater similarity between the compared trees, indicating that their overall topologies are more closely matched (Figure13).

Methodolgy

```
In [1]: import dendropy

def calculate_rf_distance(generated_tree_file, reference_tree_file):
    # Parse reference tree and generated tree
    taxon_namespace = dendropy.TaxonNamespace()
    reference_tree = dendropy.Tree.get(path=reference_tree_file, schema="newick", taxon_namespace=taxon_namespace)
    generated_tree = dendropy.Tree.get(path=generated_tree_file, schema="newick", taxon_namespace=taxon_namespace)
    # Ensure consistent taxon namespace
    generated_tree.encode_bipartitions()
    reference_tree.encode_bipartitions()
    # Calculate Robinson-Foulds distance
    rf_distance = dendropy.calculate.treecompare.symmetric_difference(reference_tree, generated_tree)
    return rf_distance

# List of file paths
file_paths = [
    "/home/affef/Desktop/trees/tree_n8/L100/N8_L100_inde10.001/N8_L100_inde10.001ME.nwk",
]

# Reference tree file path
reference_tree_file = "/home/affef/Desktop/trees/tree_n8/tree_n8.nwk"

# Iterate through file paths
for file_path in file_paths:
    # Extract the file name
    file_name = file_path.split('/')[-1]
    # Calculate RF distance
    rf_distance = calculate_rf_distance(file_path, reference_tree_file)
    # Print the file name and its RF distance
    print(f"Tree Name: {file_name}, RF Distance: {rf_distance}")

Tree Name: N8_L100_inde10.001ME.nwk, RF Distance: 0
```

Figure 13: Python Code of metric Robinson-Foulds (RF) distance.

The Mean Branch Length Distance : provided insight into the average differences in branch lengths between pairs of trees. By calculating the average of these differences, we could determine how much the branch lengths of one tree deviated from those of another.

This Metric offered a detailed understanding of variations in branch lengths, facilitating comparisons between tree structures in terms of their evolutionary distances.

These metrics provided a comprehensive framework for assessing the performance and reliability of phylogenetic tree construction methods. By considering both structural dissimilarities and branch length variations ,we aimed to offer a holistic evaluation of the comparative effectiveness of these methods in reconstructing accurate evolutionary relationships (Figure14).

Methodolgy

```
In [2]: import dendropy
def calculate_branch_length_distances(generated_tree_file, reference_tree_file):
    # Parse reference tree and generated tree
    taxon_namespace = dendropy.TaxonNamespace()
    reference_tree = dendropy.Tree.get(path=reference_tree_file, schema="newick", taxon_namespace=taxon_namespace)
    generated_tree = dendropy.Tree.get(path=generated_tree_file, schema="newick", taxon_namespace=taxon_namespace)
    # Ensure consistent taxon namespace
    generated_tree.encode_bipartitions()
    reference_tree.encode_bipartitions()
    # Get postorder edge iterators
    reference_edges = reference_tree.postorder_edge_iter()
    generated_edges = generated_tree.postorder_edge_iter()
    # Calculate branch length distances
    branch_length_distances = []
    for reference_edge, generated_edge in zip(reference_edges, generated_edges):
        reference_length = reference_edge.length
        generated_length = generated_edge.length
        # Skip if any of the lengths is None
        if reference_length is None or generated_length is None:
            continue
        # Calculate absolute difference in branch lengths
        branch_length_distances.append(abs(reference_length - generated_length))
    # Check if branch length distances is empty
    if not branch_length_distances:
        return None, None
    # Calculate mean and standard deviation of branch length distances
    mean_distance = sum(branch_length_distances) / len(branch_length_distances)
    std_error_distance = (sum((x - mean_distance) ** 2 for x in branch_length_distances) / len(branch_length_distances)) ** 0.5
    return mean_distance, std_error_distance

# List of file paths
file_paths = [
    "/home/affef/Desktop/trees/tree_n8/L100/N8_L100_indel0.001/N8_L100_indel0.001ME.nwk"
]

# Reference tree file path
reference_tree_file = "/home/affef/Desktop/trees/tree_n8/tree_n8.nwk"
# Iterate through file paths
for file_path in file_paths:
    # Extract the file name
    file_name = file_path.split('/')[-1]
    # Calculate branch length distances
    mean_distance, std_error_distance = calculate_branch_length_distances(file_path, reference_tree_file)
    if mean_distance is not None:
        # Output results
        print(f"Tree Name: {file_name}, Mean Branch Length Distance: {mean_distance}, Standard Error Distance: {std_error_distance}")
    else:
        print(f"Tree Name: {file_name}, No branch lengths present in one of the trees.")
```

Figure 14: Python code of metrics Mean and Standard Error distances.

Chapter 03: Results and Discussion

Results and discussion

Introduction

In this chapter, we present and analyse the results obtained from our phylogenetic tree simulations and constructions. The performance of various phylogenetic tree construction methods was assessed using Robinson-Foulds (RF) distance and Mean Branch Length distance. These metrics were calculated for trees with varying numbers of taxa (N), sequence lengths(Len) , and insertion-Deletion rates (InsDel). The results are illustrated through histograms and detailed in Excel tables, providing a comprehensive overview of the topological dissimilarities and branch length variations in the inferred phylogenetic relationships. This analysis offers critical insights into the effectiveness and reliability of the different methods used in our study.

1. Results of Robinson Foulds distance:

Parameters			RFD (Robinson-Foulds Distance)				
N	Len	InsDel	ML	MP	NJ	UPGMA	ME
8	100	0.001	0	2	2	2	0
8	100	0.005	0	0	4	0	2
8	100	0.010	0	0	6	0	0
8	100	0.015	2	2	6	0	0
8	100	0.020	4	4	6	0	0
8	200	0.001	2	2	0	0	2
8	200	0.005	2	2	2	2	2
8	200	0.010	4	0	8	0	0
8	200	0.015	0	0	4	0	0
8	200	0.020	2	2	6	0	2
8	300	0.001	2	0	6	0	2
8	300	0.005	0	2	8	0	0
8	300	0.010	0	0	4	0	0
8	300	0.015	2	2	6	0	0
8	300	0.020	2	0	6	0	2
8	400	0.001	0	0	4	0	0
8	400	0.005	0	0	6	0	0
8	400	0.010	0	2	6	0	0
8	400	0.015	2	0	4	0	0
8	400	0.020	2	2	4	0	2
8	500	0.001	0	0	0	2	0
8	500	0.005	25	0	0	0	0
8	500	0.010	25	2	0	0	0
8	500	0.015	25	0	2	0	2
8	500	0.020	25	0	0	0	0

Table 01: RF distance calculated for 8 taxa with all variations of indels and sequence lengths.

Results and discussion

Parameters			RFD (Robinson-Foulds Distance)				
N	Len	InsDel	ML	MP	NJ	UPGMA	ME
16	100	0.001	2	6	4	4	8
16	100	0.005	8	8	6	6	4
16	100	0.010	4	6	8	4	10
16	100	0.015	10	8	6	6	4
16	100	0.020	8	8	6	6	10
16	200	0.001	6	6	2	2	4
16	200	0.005	0	2	2	0	2
16	200	0.010	0	4	2	4	2
16	200	0.015	8	4	4	4	10
16	200	0.020	6	4	4	2	4
16	300	0.001	2	0	2	0	4
16	300	0.005	2	2	6	2	4
16	300	0.010	2	0	2	4	2
16	300	0.015	2	2	4	2	4
16	300	0.020	4	2	2	0	2
16	400	0.001	2	2	4	2	4
16	400	0.005	0	4	0	2	2
16	400	0.010	2	2	2	2	0
16	400	0.015	6	2	4	4	6
16	400	0.020	2	0	2	2	4
16	500	0.001	2	0	2	2	4
16	500	0.005	0	0	2	0	0
16	500	0.010	4	4	6	2	6
16	500	0.015	4	4	2	2	0
16	500	0.020	2	4	2	0	2

Table 02: RF distance calculated for 16 taxa with all variations of indels and sequence lengths.

Parameters			RFD (Robinson-Foulds Distance)				
N	Len	InsDel	ML	MP	NJ	UPGMA	ME
24	100	0.001	8	10	4	8	8
24	100	0.005	4	6	4	6	6
24	100	0.010	4	10	4	4	4
24	100	0.015	6	6	6	0	4
24	100	0.020	12	10	12	2	12
24	200	0.001	2	2	4	0	2
24	200	0.005	2	6	2	0	4
24	200	0.010	4	8	4	6	4
24	200	0.015	2	4	4	2	4
24	200	0.020	4	0	4	6	4
24	300	0.001	8	6	6	8	6
24	300	0.005	4	4	4	8	6
24	300	0.010	0	0	0	2	0
24	300	0.015	0	0	2	0	2
24	300	0.020	2	4	2	4	2
24	400	0.001	0	0	2	2	2
24	400	0.005	0	0	4	2	4
24	400	0.010	2	4	2	4	6
24	400	0.015	0	0	2	6	2
24	400	0.020	2	2	2	0	4
24	500	0.001	0	0	0	2	4
24	500	0.005	2	2	2	4	2
24	500	0.010	2	0	6	2	6
24	500	0.015	0	0	0	0	0
24	500	0.020	0	0	0	0	0

Table 03: RF distance calculated for 24 taxa with all variations of indels and sequence lengths.

Results and discussion

Parameters			RFD (Robinson-Foulds Distance)				
N	Len	InsDel	ML	MP	NJ	UPGMA	ME
32	100	0.001	12	14	10	4	10
32	100	0.005	12	12	14	10	8
32	100	0.01	14	12	8	10	10
32	100	0.015	12	8	16	6	14
32	100	0.02	8	8	10	14	10
32	200	0.001	6	2	4	6	6
32	200	0.005	6	8	8	10	8
32	200	0.01	10	10	10	10	12
32	200	0.015	4	4	4	2	4
32	200	0.02	4	6	2	6	4
32	300	0.001	2	4	4	4	2
32	300	0.005	6	0	4	2	4
32	300	0.01	10	10	6	2	4
32	300	0.015	4	4	4	2	2
32	300	0.02	8	8	6	2	6
32	400	0.001	4	4	2	0	4
32	400	0.005	2	4	8	6	10
32	400	0.01	4	6	2	4	2
32	400	0.015	2	0	2	2	2
32	400	0.02	4	4	6	2	6
32	500	0.001	4	6	6	0	4
32	500	0.005	2	4	2	2	2
32	500	0.01	0	2	4	4	4
32	500	0.015	2	4	6	2	2
32	500	0.02	4	0	0	2	2

Table 04: RF distance calculated for 32 taxa with all variations of indels and sequence lengths.

Parameters			RFD (Robinson-Foulds Distance)				
N	Len	InsDel	ML	MP	NJ	UPGMA	ME
40	100	0.001	20	32	30	26	32
40	100	0.005	22	38	20	24	28
40	100	0.010	18	24	12	14	22
40	100	0.015	20	18	14	20	28
40	100	0.020	22	14			
40	200	0.001	20	14	10	8	16
40	200	0.005	20	12	16	6	16
40	200	0.010	16	16	12	10	16
40	200	0.015	16	18	8	8	12
40	200	0.020	20	14	14	4	16
40	300	0.001	8	22	12	10	14
40	300	0.005	8	20	14	6	18
40	300	0.010	26	12	14	4	22
40	300	0.015	18	10	18	14	26
40	300	0.020	10	8	12	12	16
40	400	0.001	12	28	14	12	18
40	400	0.005	10	22	16	6	12
40	400	0.010	12	10	14	4	14
40	400	0.015	14	8	14	18	20
40	400	0.020	10	8	16	6	14
40	500	0.001	8	20	10	8	16
40	500	0.005	6	14	12	10	18
40	500	0.010	8	8	12	8	16
40	500	0.015	6	6	12	4	18
40	500	0.020	6	6	12	10	16

Table 05 :RF distance calculated for 40 taxa with all variations of indels and sequence lengths.

Results and discussion

2.Results of Mean Branch Length distance:

Parameters			MeanDist (Mean Branch Length Distance)				
N	Len	InsDel	ML	MP	NJ	UPGMA	ME
8	100	0.001	0.406472		0.406472	15.713659	0.363477
8	100	0.005	0.460837		0.460837	22.938404	0.371733
8	100	0.010	0.420847		0.420847	72.106274	0.351643
8	100	0.015	0.496013		0.496013	35.446899	0.387839
8	100	0.020	0.430780		0.430780	5.762884	0.364055
8	200	0.001	0.423944		0.423944	14.066806	0.377734
8	200	0.005	0.423943		0.423943	9.980810	0.363409
8	200	0.010	0.482670		0.482670	20.453064	0.384409
8	200	0.015	0.445145		0.445145	10.428137	0.360646
8	200	0.020	0.453239		0.453239	14.066806	0.401275
8	300	0.001	0.433320		0.433320	11.304036	0.377603
8	300	0.005	0.428185		0.428185	7.820107	0.368536
8	300	0.010	0.393544		0.393544	6.814297	0.341243
8	300	0.015	0.406847		0.406847	9.884216	0.370516
8	300	0.020	0.424858		0.424858	7.710052	0.377524
8	400	0.001	0.421171		0.421171	8.433006	0.361069
8	400	0.005	0.650369		0.650369	15.650681	0.367085
8	400	0.010	0.438538		0.438538	16.288001	0.378677
8	400	0.015	0.432766		0.432766	11.332973	0.350308
8	400	0.020	0.443854		0.443854	11.662446	0.384565
8	500	0.001	0.399581		0.399581	0.282915	0.360030
8	500	0.005	0.675645		0.675645	0.258599	0.349527
8	500	0.010	0.675645		0.675645	0.266761	0.363788
8	500	0.015	0.675645		0.675645	0.251401	0.338973
8	500	0.020	0.675645		0.675645	0.256636	0.354542

Table 06: Mean Branch Length distance calculated for 8 taxa with all variations of indels and sequence lengths.

Parameters			MeanDist (Mean Branch Length Distance)				
N	Len	InsDel	ML	MP	NJ	UPGMA	ME
16	100	0.001	0.2124199		0.168643	0.146226	0.200019
16	100	0.005	0.201240		0.187098	0.165883	0.171094
16	100	0.010	0.204700		0.205183	0.122631	0.188068
16	100	0.015	0.239640		0.177166	0.152945	0.183541
16	100	0.020	0.1981955		0.177616	0.141542	0.179955
16	200	0.001	0.193131		0.195267	0.148225	0.193020
16	200	0.005	0.198568		0.183244	0.138884	0.184641
16	200	0.010	0.193049		0.188618	0.151439	0.189531
16	200	0.015	0.201576		0.180622	0.151737	0.178161
16	200	0.020	0.202445		0.181215	0.144048	0.179338
16	300	0.001	0.201700		0.178279	0.142680	0.183180
16	300	0.005	0.211158		0.188540	0.137572	0.186298
16	300	0.010	0.200837		0.184250	0.144153	0.175727
16	300	0.015	0.220844		0.194918	0.149459	0.199235
16	300	0.020	0.168319		0.175855	0.146459	0.186785
16	400	0.001	0.210555		0.183609	0.141267	0.183555
16	400	0.005	0.203598		0.184730	0.148572	0.183054
16	400	0.010	0.200837		0.184250	0.144153	0.179290
16	400	0.015	0.220844		0.184037	0.149459	0.180699
16	400	0.020	0.168319		0.175855	0.146459	0.185348
16	500	0.001	0.207208		0.181872	0.145855	0.183520
16	500	0.005	0.205260		0.174721	0.144316	0.187948
16	500	0.010	0.207060		0.182916	0.147274	0.180167
16	500	0.015	0.198735		0.180390	0.141766	0.179525
16	500	0.020	0.214992		0.192274	0.145616	0.191070

Table 07 : Mean Branch Length distance calculated for 16 taxa with all variations of indels and sequence lengths.

Results and discussion

Parameters			MeanDist (Mean Branch Length Distance)				
N	Len	InsDel	ML	MP	NJ	UPGMA	ME
24	100	0.001	0.155537		0.146852	0.150185	0.152633
24	100	0.005	0.159867		0.149802	0.152474	0.149103
24	100	0.010	0.163078		0.151130	0.153776	0.150243
24	100	0.015	0.171693		0.152256	0.155901	0.145460
24	100	0.020	0.172186		0.152499	0.157091	0.152256
24	200	0.001	0.157804		0.141499	0.144324	0.144993
24	200	0.005	0.161353		0.144445	0.146784	0.146572
24	200	0.010	0.165280		0.145968	0.152611	0.141499
24	200	0.015	0.166696		0.148103	0.157657	0.148727
24	200	0.020	0.168446		0.151100	0.168093	0.147475
24	300	0.001	0.156864		0.144773	0.149812	0.145599
24	300	0.005	0.159649		0.145625	0.151308	0.149052
24	300	0.010	0.159754		0.146078	0.151472	0.149141
24	300	0.015	0.161887		0.147993	0.155315	0.144773
24	300	0.020	0.166352		0.148317	0.155707	0.145650
24	400	0.001	0.162430		0.145503	0.151402	0.146888
24	400	0.005	0.162796		0.145604	0.151545	0.148385
24	400	0.010	0.165089		0.146978	0.151732	0.146063
24	400	0.015	0.165292		0.148922	0.153045	0.153025
24	400	0.020	0.166648		0.152917	0.153322	0.144817
24	500	0.001	0.158796		0.144966	0.149226	0.149376
24	500	0.005	0.159309		0.146617	0.150295	0.149559
24	500	0.010	0.162274		0.148476	0.154223	0.148475
24	500	0.015	0.162618		0.148607	0.155718	0.145832
24	500	0.020	0.163485		0.150193	0.156164	0.145896

Table 08 : Mean Branch Length distance calculated for 24 taxa with all variations of indels and sequence lengths.

Parameters			MeanDist (Mean Branch Length Distance)				
N	Len	InsDel	ML	MP	NJ	UPGMA	ME
32	100	0.001	0.136637		0.122444	0.121466	0.121452
32	100	0.005	0.133360		0.117374	0.113297	0.118824
32	100	0.010	0.127801		0.114310	0.117594	0.114205
32	100	0.015	0.134546		0.118737	0.121098	0.116213
32	100	0.020	0.129231		0.121732	0.122411	0.118404
32	200	0.001	0.130238		0.116929	0.118984	0.116047
32	200	0.005	0.130990		0.104640	0.123773	0.105664
32	200	0.010	0.129332		0.117451	0.118712	0.117388
32	200	0.015	0.129036		0.113811	0.117296	0.114730
32	200	0.020	0.129038		0.118125	0.118455	0.119241
32	300	0.001	0.125800		0.113407	0.115101	0.113554
32	300	0.005	0.128885		0.115936	0.115031	0.115491
32	300	0.010	0.126552		0.116256	0.116553	0.118630
32	300	0.015	0.128388		0.128388	0.118806	0.113808
32	300	0.020	0.117632		0.116260	0.117635	0.118063
32	400	0.001	0.130639		0.118497	0.119935	0.114004
32	400	0.005	0.125900		0.113153	0.117162	0.111992
32	400	0.010	0.129387		0.118625	0.119078	0.118752
32	400	0.015	0.126732		0.114137	0.114107	0.114387
32	400	0.020	0.122451		0.111992	0.112822	0.115161
32	500	0.001	0.132827		0.119188	0.121449	0.113619
32	500	0.005	0.127526		0.114444	0.116669	0.112635
32	500	0.010	0.130164		0.115770	0.118504	0.116514
32	500	0.015	0.127134		0.113687	0.116539	0.117786
32	500	0.020	0.123963		0.113227	0.112879	0.115537

Table 09: Mean Branch Length distance calculated for 32 taxa with all variations of indels and sequence lengths.

Results and discussion

Parameters			MeanDist (Mean Branch Length Distance)				
N	Len	InsDel	ML	MP	NJ	UPGMA	ME
40	100	0.001	0.180873		0.150107	0.117375	0.153486
40	100	0.005	0.199539		0.150528	0.120331	0.149034
40	100	0.010	0.188422		0.156958	0.127533	0.150748
40	100	0.015	0.200056		0.147400	0.112382	0.15104
40	100	0.020	0.197780				
40	200	0.001	0.156617		0.144339	0.111717	0.144162
40	200	0.005	0.179768		0.153417	0.102657	0.153349
40	200	0.010	0.174246		0.149048	0.103427	0.147798
40	200	0.015	0.172163		0.149921	0.105131	0.149701
40	200	0.020	0.202629		0.144333	0.106322	0.147617
40	300	0.001	0.169317		0.149510	0.106749	0.152268
40	300	0.005	0.179745		0.144107	0.099183	0.149611
40	300	0.010	0.209108		0.151691	0.104683	0.154295
40	300	0.015	0.176055		0.151412	0.101103	0.152628
40	300	0.020	0.165169		0.143532	0.120294	0.142853
40	400	0.001	0.168909		0.142846	0.102428	0.141663
40	400	0.005	0.172374		0.145529	0.102136	0.142925
40	400	0.010	0.177606		0.144400	0.100833	0.14432
40	400	0.015	0.169098		0.146196	0.121384	0.145801
40	400	0.020	0.185516		0.149142	0.103144	0.150654
40	500	0.001	0.176548		0.154171	0.103070	0.153326
40	500	0.005	0.166433		0.144186	0.104225	0.142256
40	500	0.010	0.174461		0.145712	0.118577	0.145105
40	500	0.015	0.167458		0.151299	0.107060	0.151513
40	500	0.020	0.170068		0.148800	0.122878	0.150373

Table 10: Mean Branch Length distance calculated for 40 taxa with all variations of indels and sequence lengths.

3. RF Distance Histograms:

3.1 Variation in length of sequences :

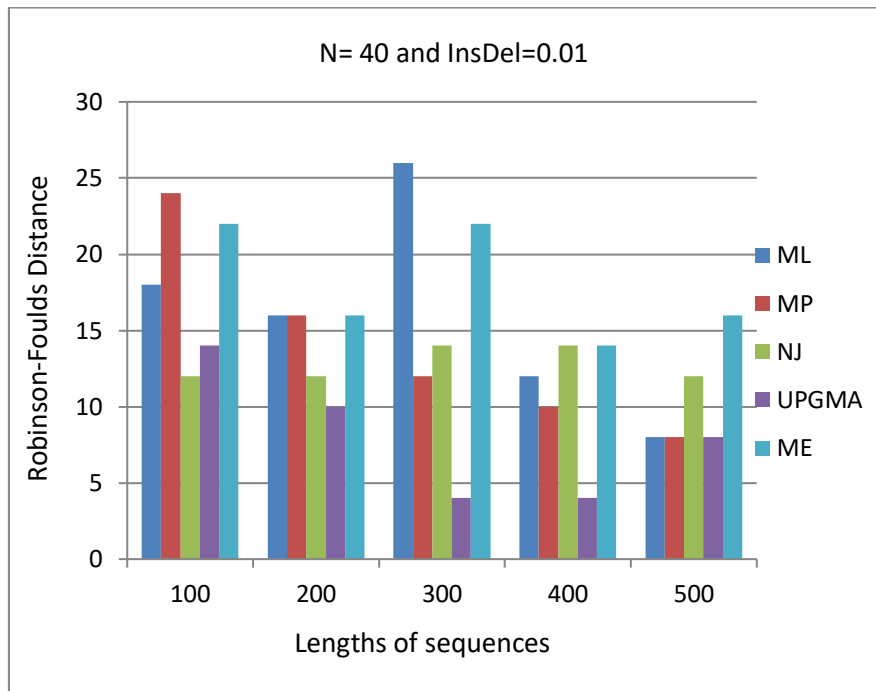


Figure 15: Histogram represents Robinson-Foulds (RF) distance in function of sequences length.

Overall-Trend:

For most methods, the RF distance varies with sequences length, indicating differences in topological accuracy depending on the method used.

Method-Specific observations:

- **Maximum Likelihood (ML):** Exhibits relatively high RF distances, indicating greater topological differences from the reference tree.
- **Maximum Parsimony (MP):** Shows high RF distances for shorter sequences and decreases as the sequence length increases, suggesting better accuracy with longer sequences.
- **Neighbor-Joining (NJ):** Maintains medium RF distances with a slight increase for longer sequences.
- **UPGMA:** Consistently shows low RF distances across all sequence lengths, indicating stable topological accuracy.
- **Minimum Evolution (ME):** Similar to ML, shows higher RF distances for longer sequences, indicating reduced accuracy in these scenarios.

Results and discussion

Discussion:

The variation in RF distances highlights the importance of selecting the appropriate phylogenetic reconstruction method. **UPGMA** consistently demonstrates better performance, making it a reliable choice for accurate topology estimation in these scenarios.

ML and **ME** show less precise performance .

MP, unlike other methods, improves its accuracy with longer sequences.

NJ shows intermediate performance, with medium RF distances and a slight increase for longer sequences.

3.2.Variation in number of sequences :

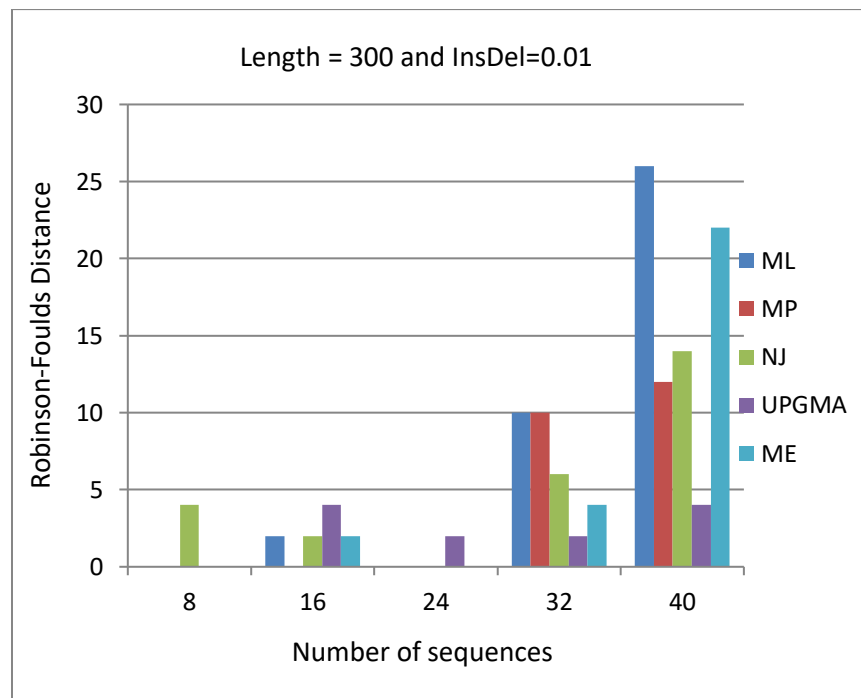


Figure 16 : Histogram represents Robinson-Foulds (RF) distance in fuction of number of sequences .

Overall Trend:

The bar graph displays the Robinson-Foulds Distance across five different phylogenetic methods (ML, MP, NJ, UPGMA, ME) for various numbers of sequences (8, 16, 24, 32, 40). The results indicate that the Robinson-Foulds Distance, which measures the topological difference between the reconstructed and the true tree, varies with the number of sequences, highlighting differences in the accuracy of these methods.

Results and discussion

Method Comparison:

- **Maximum Likelihood (ML) :**
 - Shows a significant increase in the Robinson-Foulds Distance at 40 sequences, peaking notably at this point.
 - ML demonstrates variable performance across the different sequence numbers, with lower values at 8, 16, and 24 sequences but much higher values at 32 and 40 sequences.
- **Maximum Parsimony (MP) :**
 - Exhibits consistent performance with moderate Robinson-Foulds Distances, peaking at 40 sequences.
- MP shows moderate values at 40 sequences and lower values at other sequence numbers .
- **Neighbor-Joining (NJ) :**
 - Demonstrates a gradual increase in the Robinson-Foulds Distance as the number of sequences increases.
 - NJ has noticeable peaks at 32 and 40 sequences, indicating higher topological differences at these points.
- **UPGMA :**
 - Shows lower Robinson-Foulds Distances at most sequence numbers, indicating potentially better topological accuracy.
- **Minimum Evolution (ME) :**
 - Exhibits high Robinson-Foulds Distances at 40 sequences, similar to ML.
 - Shows moderate performance at other sequence numbers, with relatively lower values at 8, 16, and 24 sequences but higher at 32 and 40 sequences.

Discussion :

The analysis reveals that the choice of phylogenetic reconstruction method significantly affects topological accuracy, as indicated by the Robinson-Foulds Distance. The results show varying levels of topological accuracy across different sequence numbers:

- **UPGMA** generally demonstrates better topological accuracy with lower Robinson-Foulds Distances across most sequence numbers.

Results and discussion

- **ML** and **ME** methods show significant increases in Robinson-Foulds Distances at higher sequence numbers (40), suggesting decreased accuracy in these scenarios.
- **MP** and **NJ** show intermediate performance with peaks at higher sequence numbers, indicating variability in their topological accuracy.

This variability emphasizes the importance of selecting the appropriate method based on dataset characteristics. UPGMA might be preferred for consistent topological accuracy, while methods like ML and ME require careful consideration for larger datasets.

3.3.Variation in Insertion-Deletion rate :

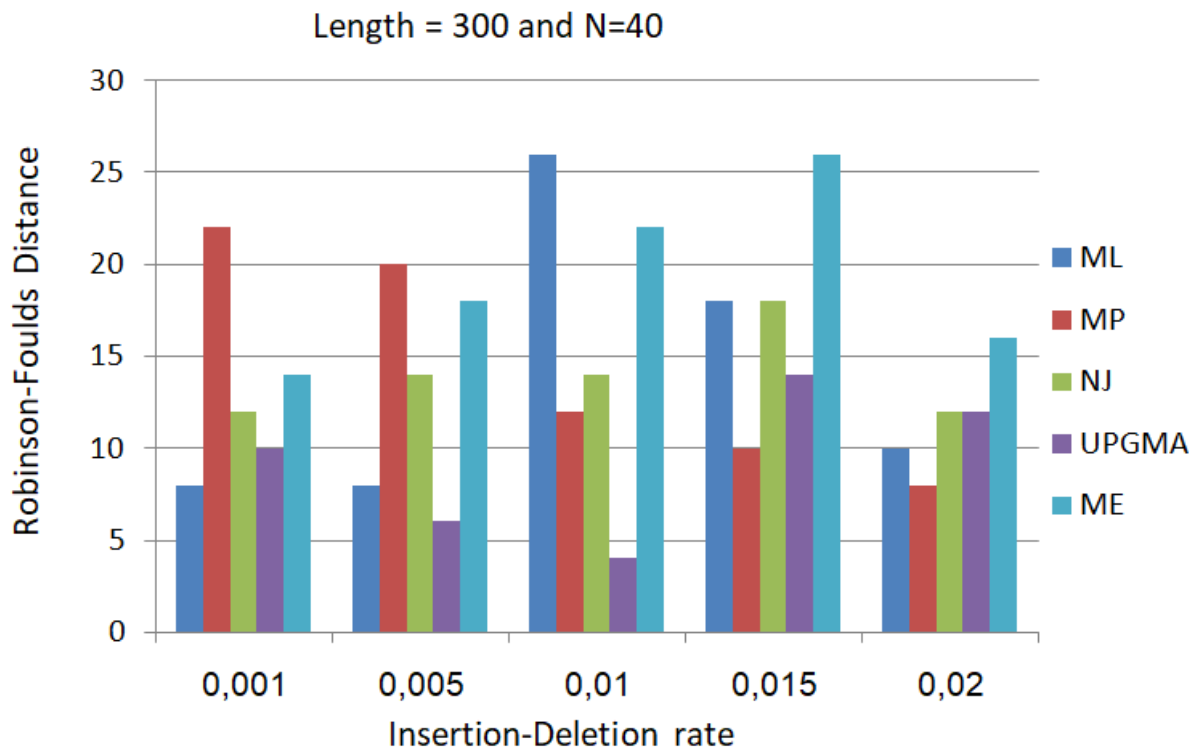


Figure 17: Histogram represents Robinson-Foulds (RF) distance in function of Insertion-Deletion rate.

Overall Trend:

The figure illustrates how the Robinson-Foulds (RF) distance varies with different insertion and deletion (insDel) rates. The RF distance measures the topological differences between the reconstructed trees and a reference tree, indicating the accuracy of the phylogenetic methods under varying insDel rates.

Results and discussion

Method Comparison:

- **UPGMA:** Shows lower RF distances across varying insDel rates, indicating better performance and robustness in maintaining tree accuracy despite changes in insDel rates.
- **Neighbor-Joining (NJ) and Minimum Evolution (ME):** Display moderate RF distances, indicating that these methods are somewhat sensitive to insDel rates but generally perform better than others except UPGMA.
- **Maximum Likelihood (ML):** Exhibits higher RF distances, indicating that this method is more affected by higher insDel rates and may produce less accurate tree topologies in such conditions.

Discussion:

The analysis reveals that the accuracy of phylogenetic tree reconstruction, as measured by RF distances, is impacted by insDel rates. UPGMA consistently performs well even as insDel rates increase, making it a reliable method under varying conditions. NJ and ME also show relatively good performance but are more affected by higher insDel rates than UPGMA. ML, on the other hand, appears to struggle more with increased insDel rates, leading to higher RF distances and less accurate trees.

4. Mean Branch length distance histograms:

4.1. Variation in length of sequences :

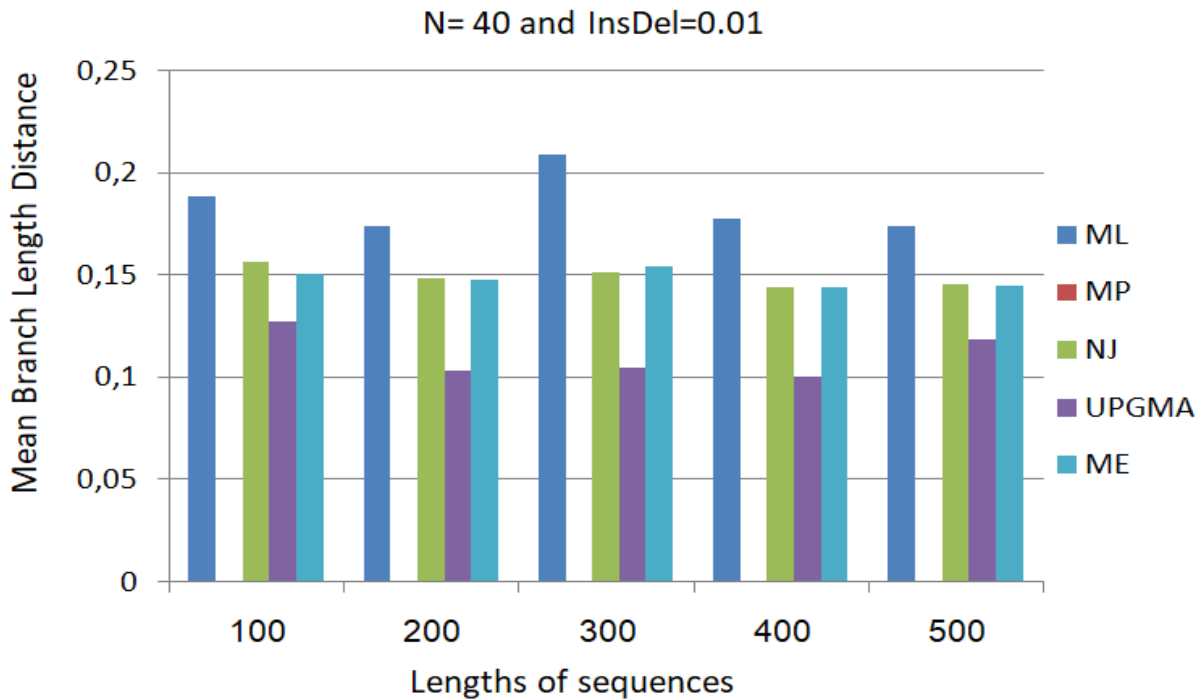


Figure 18 : Histogram represents Mean Branch Length distance in function of sequences length .

Overall Trend:

The mean branch length distance varies with the number of sequences, indicating differences in the accuracy of branch length estimations by different methods. There is no clear negative impact of the number of sequences on mean branch length distance; instead, the accuracy appears to fluctuate with sequence length.

Method Comparison:

- **UPGMA:** Displays lower mean branch length distances across all sequence lengths, indicating the best accuracy in branch length estimation.
- **Neighbor-Joining (NJ) and Minimum Evolution (ME):** Exhibit similar performance, generally better than other methods except UPGMA.
- **Maximum Likelihood (ML):** Shows higher mean branch length distances, indicating less accurate branch length estimations compared to other methods in these scenarios.

Results and discussion

Discussion:

The choice of phylogenetic tree reconstruction method significantly impacts the accuracy of branch length estimation, as indicated by mean branch length distances. **UPGMA** demonstrates the most consistent performance with the lowest mean branch length distances across different sequence lengths. **NJ** and **ME** methods also perform well but are outperformed by **UPGMA**. **ML** and **MP** methods show higher variability and generally less accurate branch length estimations, particularly with larger sequence datasets.

4.2.Variation in number of sequences :

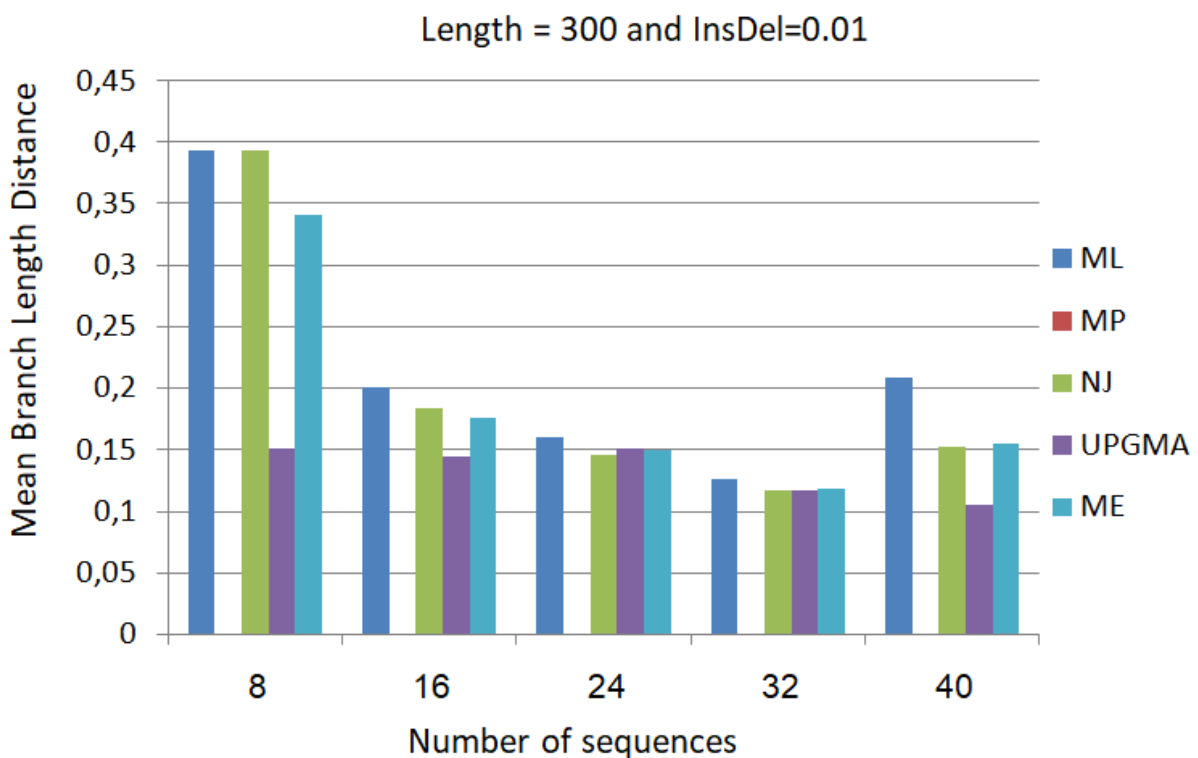


Figure 19: Histogram represents Mean Branch Length distance in function of number of sequences .

Overall Trend :

The bar graph represents the Mean Branch Length Distance across five different phylogenetic methods (ML, MP, NJ, UPGMA, ME) for various numbers of sequences (8, 16, 24, 32, 40). The mean branch length distance is impacted negatively by the number of sequences.

Results and discussion

Method Comparison:

- **Maximum Likelihood (ML) :**
 - Shows high Mean Branch Length Distances at 8 sequences, which decrease significantly as the number of sequences increases.
 - ML exhibits relatively low distances at 16, 24, and 32 sequences, but increases again at 40 sequences.
- **Neighbor-Joining (NJ) :**
 - Displays high Mean Branch Length distances at 8 sequences, decreasing notably at higher sequence numbers.
 - NJ exhibits moderate distances from 16 to 40 sequences, indicating improved accuracy with more sequences.
- **UPGMA :**
 - Shows lower Mean Branch Length distances across all sequence numbers, indicating better accuracy in branch length estimation.
 - UPGMA demonstrates the lowest values at 32 and 40 sequences, indicating highly accurate branch length estimations.
- **Minimum Evolution (ME) :**
 - Exhibits high Mean Branch Length Distances at 8 sequences, which decrease at higher sequence numbers.
 - ME shows moderate to low distances from 16 to 40 sequences, with improved accuracy as the number of sequences increases.

Discussion :

The analysis indicates that the choice of phylogenetic reconstruction method significantly impacts the accuracy of branch length estimations, as shown by the Mean Branch Length Distances. The results reveal varying levels of accuracy across different sequence numbers:

- **UPGMA** consistently demonstrates the best performance with the lowest Mean Branch Length Distances, indicating highly accurate branch length estimations across all sequence numbers.

Results and discussion

- **ML** and **ME** shows higher distances at 8 sequences, suggesting initial inaccuracies, but improve with more sequences, although **ML** increases again at 40 sequences.
- **NJ** exhibit high distances at 8 sequences but stabilize and show improved accuracy with higher numbers of sequences.

This variability highlights the importance of selecting the appropriate method based on the dataset size and characteristics. **UPGMA** appears to be the most reliable method for consistent accuracy in branch length estimation, while **ML** and **ME** show variability that requires consideration, especially for larger datasets.

4.3. Variation in Insertion-Deletion rate :

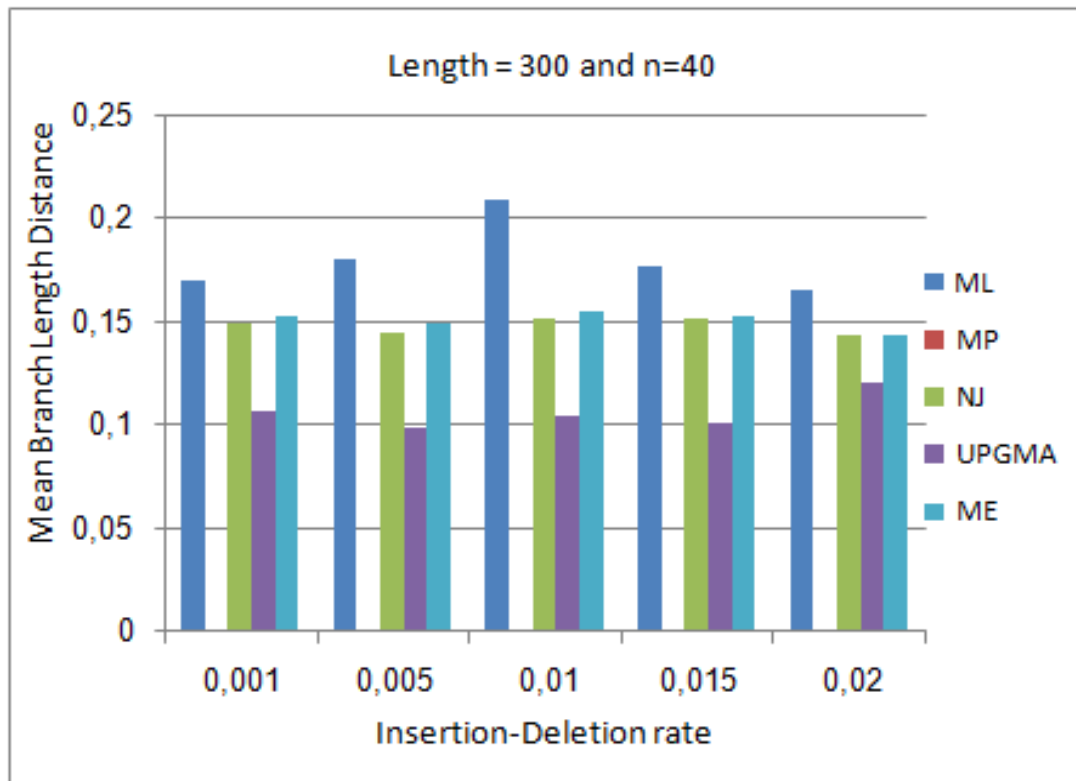


Figure 20: Histogram represents Mean Branch Length distance in function of Insertion-Deletion rate.

Overall Trend

The mean branch length distance varies with the insertion-deletion rate, indicating differences in the accuracy of branch length estimations by different phylogenetic methods. Across different insertion-deletion rates, the mean branch length distance shows notable

Results and discussion

variability, suggesting that the accuracy of branch length estimation is influenced by the insertion-deletion rate.

Method Comparison:

- **Maximum Likelihood (ML)** : ML exhibits relatively higher mean branch length distances across most insertion-deletion rates, indicating less accurate branch length estimations compared to other methods. However, its performance is consistent across different insertion-deletion rates.
- **Neighbor-Joining (NJ)** : NJ displays consistent Mean Branch Length distances across various insertion-deletion rates, generally performing better than ML but not as well as UPGMA and ME. NJ's stability across different conditions makes it a reliable method.
- **Unweighted Pair Group Method with Arithmetic Mean (UPGMA)** : UPGMA has lower mean branch length distances across all insertion-deletion rates, indicating the best accuracy in branch length estimation among the methods tested. Its performance is particularly notable at mid-range insertion-deletion rates.
- **Minimum Evolution (ME)** : ME shows competitive performance, with mean branch length distances similar to those of NJ.

Discussion:

The choice of phylogenetic tree reconstruction method significantly impacts the accuracy of branch length estimation, as indicated by mean branch length distances. UPGMA demonstrates the most consistent and accurate performance, with low variability in mean branch length distances across different insertion-deletion rates. NJ and ME perform reliably, though not as accurately as UPGMA . ML shows higher variability and less accuracy, particularly at higher insertion-deletion rates.

The results suggest that methods like UPGMA is better suited for scenarios where accurate branch length estimation is critical, especially in the presence of varying insertion-deletion rates. The consistent performance of UPGMA highlight its robustness in phylogenetic analysis.

5. General Performance of Phylogenetic Reconstruction Methods:

The evaluation of phylogenetic reconstruction methods revealed significant variations in terms of topological accuracy and branch length estimation accuracy. The results demonstrated that the choice of method has a substantial impact on the quality of the reconstructed trees.

1. UPGMA (Unweighted Pair Group Method with Arithmetic Mean):

- **Topological Accuracy:** UPGMA consistently showed lower RF distances, indicating better topological accuracy.
- **Branch Length Estimation Accuracy:** UPGMA also exhibited lower mean branch length distances, demonstrating excellent precision in branch length estimation.
- **Time Performance:** UPGMA was one of the fastest methods, allowing efficient reconstruction of phylogenetic trees.

2. Neighbor-Joining (NJ):

- **Topological Accuracy:** NJ showed moderate performance with intermediate RF distances, particularly sensitive to larger datasets.
- **Branch Length Estimation Accuracy:** NJ performed relatively well, though slightly less accurately than UPGMA and ME.
- **Time Performance:** NJ was also fast, comparable to UPGMA in terms of computation time.

3. Maximum Parsimony (MP):

- **Topological Accuracy:** MP exhibited low RF distances for a smaller number of sequences and moderate RF distances for larger datasets.
- **Time Performance:** MP was relatively fast, enabling efficient reconstruction for moderate-sized datasets.

4. Minimum Evolution (ME):

- **Topological Accuracy:** ME displayed variability similar to ML, with higher RF distances for larger datasets.
- **Branch Length Estimation Accuracy:** ME showed competitive mean branch length distances, close to those of NJ.
- **Time Performance:** ME was also fast, comparable to UPGMA and NJ.

Results and discussion

5. Maximum Likelihood (ML):

- **Topological Accuracy:** ML exhibited higher RF distances, particularly for larger datasets, indicating reduced topological accuracy.
- **Branch Length Estimation Accuracy:** ML showed higher mean branch length distances, indicating less precision in branch length estimation.
- **Time Performance:** ML was the slowest method, requiring substantial computation time, especially for larger datasets.

Note : Maximum Parsimony (MP method did not provide branch length values during the tree construction in MEGA, hence the mean branch length could not be calculated.

Conclusion

Conclusion

Our study, titled "Comparison of Phylogenetic Inference Methods," aimed to evaluate and compare the performance of five commonly used phylogenetic reconstruction methods: UPGMA, Neighbor-Joining (NJ), Maximum Parsimony (MP), Minimum Evolution (ME), and Maximum Likelihood (ML). The evaluation encompassed both topological accuracy and branch length estimation accuracy across various dataset sizes and insertion/deletion (ins/del) rates.

The results indicated clear distinctions in the performance of the methods. **UPGMA** consistently emerged as the method with the highest topological accuracy, maintaining stable performance even with increasing dataset sizes. Its ability to produce precise branch length estimations further solidified its position as a reliable choice for phylogenetic reconstruction.

Neighbor-Joining (NJ) demonstrated moderate performance, particularly suitable for datasets with a larger number of sequences. While not as accurate as UPGMA, NJ still provided dependable results across different conditions.

Minimum Evolution (ME) showed competitive performance, especially in branch length estimation accuracy, making it a viable alternative to UPGMA in certain scenarios. The ME method's focus on minimizing the total branch length of the tree often results in accurate and biologically meaningful trees.

Maximum Parsimony (MP) exhibited good performance with smaller datasets but demonstrated variability in accuracy as dataset sizes increased. Despite this, MP remained a valuable option for phylogenetic reconstruction, particularly when computational resources are limited. Its simplicity and non-parametric nature make it a suitable choice for initial exploratory analyses.

Maximum Likelihood (ML), while capable of providing detailed results, showed higher variability and less accuracy compared to other methods, especially for larger datasets. Its computational intensity also posed challenges, making it less practical for analyses requiring speed and efficiency. However, ML's model-based approach allows for more nuanced reconstructions, particularly useful when evolutionary models are well understood.

Conclusion

In conclusion, our study "Comparison of Phylogenetic Inference Methods" emphasizes the importance of method selection for achieving accurate and reliable phylogenetic reconstructions. **UPGMA** stands out as a robust and efficient method, suitable for various datasets and analytical scenarios. However, **NJ**, **ME**, and **MP** also offer valuable alternatives depending on the specific requirements of the analysis. Despite the challenges associated with **ML**, its detailed approach remains crucial for certain applications, particularly where model accuracy is paramount.

This study highlights the nuanced performance of different phylogenetic inference methods and underscores the necessity for researchers to carefully choose the method that best suits their specific dataset and research objectives.

Bibliographic References

Bibliographic References

Bibliographic References

1. **Choudhuri, S. (2014).** Bioinformatics for Beginners: Genes, Genomes, Molecular Evolution, Databases and Analytical Tools. Academic Press.
2. **Dandekar, T., & Kunz, M. (2011).** Bioinformatics Primer (An Introductory Handbook for Bioinformatics Practitioners) (pp. 119-122, 126-131). Bio-Bio-1 Team.
3. **Delèage, G., Bourguignon, P. Y., Miadana, E., & Parrinello, H. (2021).** Phylogénétique Moléculaire. Éditions Quae (pp. 61-63, 88).
4. **Felsenstein, J. (1981).** Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6), 368-376.
5. **Felsenstein, J. (2004).** *Inferring Phylogenies (Vol. 2)*. Sunderland, Massachusetts: Sinauer Associates.
6. **Haeckel, E. H. P. A. (1860).** Fernere Abbildungen und Diagnosen neuer Gattungen und Arten von lebenden Radiolarien des Mittelmeeres. *Monatsberichte der Königlichen Preuss. Akademie der Wissenschaften zu Berlin*, 835-845.
7. **Patwardhan, A., Ray, S., & Roy, A. (2014).** Molecular markers in phylogenetic studies - a review. *Journal of Phylogenetics & Evolutionary Biology*, 2(2), 1-9.
8. **Robinson, D. F., & Foulds, L. R. (1981).** Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2), 131-147.
9. **Saitou, N., & Nei, M. (1987).** The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4), 406-425.
10. **Tahiri, N. (2012).** Un nouvel algorithme pour retrouver les relations phylogénétiques entre la distribution géographique des espèces comme exigence partielle des espèces et leur composition génétique. Mémoire, Université du Québec, Montréal, 152 p.
11. **Wiley, E. O., & Lieberman, B. S. (2011).** *Phylogenetics: Theory and Practice of Phylogenetic Systematics*. John Wiley & Sons (pp. 09-10).

Bibliographic References

List of Web sites

<https://www.r-studio.com/fr/>

<http://www.iqtree.org/>

<https://www.megasoftware.net/>

<https://pubmed.ncbi.nlm.nih.gov/8019868/>

Abstract

Abstract : Phylogenetics, a key area of evolutionary biology, studies relationships among species through DNA and protein sequence analysis, creating trees that reveal evolutionary paths and common ancestors. These trees are essential in biology, ecology, medicine, and conservation, helping trace genetic histories, infer ancestral traits, and predict gene functions. This research aims to compare the effectiveness of different phylogenetic tree construction methods. Using a reference dataset, we evaluate methods based on two metrics: Robinson-Foulds (RF) and Mean Branch Length Distances, identifying each method's strengths and weaknesses to recommend best practices for phylogenetic tree construction.

Keywords : Phylogenetics, protein sequence analysis, Evolutionary paths, Common ancestors and Phylogenetic tree construction methods .

Résumé : La phylogénétique, un domaine clé de la biologie évolutive, étudie les relations entre les espèces grâce à l'analyse des séquences d'ADN et de protéines, créant des arbres qui révèlent les chemins évolutifs et les ancêtres communs. Ces arbres sont essentiels en biologie, écologie, médecine et conservation, aidant à tracer les histoires génétiques, à inférer les traits ancestraux et à prédire les fonctions des gènes. Cette recherche vise à comparer l'efficacité des différentes méthodes de construction des arbres phylogénétiques. En utilisant un jeu de données de référence, nous évaluons les méthodes basées sur deux métriques : la distance de Robinson-Foulds (RF) et la moyenne des branches, en identifiant les forces et faiblesses de chaque méthode pour recommander les meilleures pratiques de construction des arbres phylogénétiques.

Mots clés : Phylogénétique, Séquences de protéines, Arbres évolutifs, Ancêtres communs et Méthodes de construction d'arbres phylogénétique .

ملخص : علم تطور السلالات هو مجال رئيسي في علم الأحياء التطوري، يدرس العلاقات بين الأنواع من خلال تحليل تسلسلات الحمض النووي والبروتين، مما يخلق أشجار تكشف المسارات التطورية والأسلاف المشتركة. هذه الأشجار ضرورية في علم الأحياء والبيئة والطب والحفاظ على التنوع البيولوجي، حيث تساعد في تتبع التاريخ الجيني، واستنتاج الصفات الأسلافية، والتنبؤ بوظائف الجينات. يهدف هذا البحث إلى مقارنة فعالية طرق بناء أشجار تطور السلالات المختلفة، باستخدام مجموعة بيانات مرجعية، نقيم الطرق بناءً على مسافة روبنسون-ومتوسط طول الفروع، مع تحديد نقاط القوة والضعف لكل طريقة لتقديم توصيات بأفضل الممارسات في بناء أشجار تطور السلالات.

الكلمات المفتاحية : علم تطور السلالات، تحليل تسلسلات البروتين، مسارات تطورية، أسلاف مشتركة و طرق بناء أشجار تطور السلالات.

Academic year: 2023 – 2024

Presented by: Chial Affef
Chettah Imene

Comparison of phylogenetic inference methods

Thesis for the Master's degree in Bioinformatics

Abstract :

Phylogenetics, a key area of evolutionary biology, studies relationships among species through DNA and protein sequence analysis, creating trees that reveal evolutionary paths and common ancestors. These trees are essential in biology, ecology, medicine, and conservation, helping trace genetic histories, infer ancestral traits, and predict gene functions. This research aims to compare the effectiveness of different phylogenetic tree construction methods. Using a reference dataset, we evaluate methods based on two metrics: Robinson Foulds (RF) and Mean Branch Length Distances, identifying each method's strengths and weaknesses to recommend best practices for phylogenetic tree construction.

Keywords : Phylogenetics, protein sequence analysis, Evolutionary paths, Common ancestors and Phylogenetic tree construction methods .

President of jury: Dr. Medjroubi Med ElArbi (MCB in University of Mentouri Brother's Constantine 1).

Supervisor: Dr. DAAS Mohamed Skander (MCA in University of Mentouri Brother's Constantine 1).

Examiner: Dr. Gherboudj Amira (MCA in University of Mentouri Brother's Constantine 1).