M. Engg. Project

PhyloQon: A Tool for Accurate Phylogenetic Tree Reconstruction from Quartets

By
Sharifa Rania Mahmud

Submitted to
Department of Computer Science and Engineering

in partial fulfilment of the requirements for the degree of
Master of Engineering in Computer Science and Engineering

Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology (BUET)
Dhaka-1000

March, 2019

Dedicated to my loving family

AUTHOR'S CONTACT

Sharifa Rania Mahmud
Lecturer
Department of Computer Science and Engineering
Military Institute of Science and Technology (MIST)
Email: sraniamahmud@gmail.com

The thesis titled "**PhyloQon: A tool for accurate phylogenetic tree reconstruction from quartets**", submitted by Sharifa Rania Mahmud, Roll No. 1009052061, Session October 2009, to the Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, has been accepted as satisfactory in partial fulfilment of the requirements for the degree of Master of Engineering in Computer Science and Engineering and approved as to its style and contents. Examination held on March 30, 2019.

Board of Examiners

1. ______________________________

Dr. M. Sohel Rahman — Chairman
Professor — (Supervisor)
Department of CSE, BUET, Dhaka.

2. Rifat ______________________________

Dr. Rifat Shahriyar — Member
Associate Professor
Department of CSE, BUET, Dhaka.

3. ______________________________

Dr. Md. Shamsuzzoha Bayzid — Member
Assistant Professor
Department of CSE, BUET, Dhaka.

Candidate's Declaration

This is to certify that the work presented in this thesis titled "**PhyloQon: A Tool for Accurate Phylogenetic Tree Reconstruction from Quartets**" is the outcome of the investigation carried out by me under the supervision of Professor Dr. M. Sohel Rahman in the Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology (BUET), Dhaka. It is also declared that neither this project nor any part thereof has been submitted or is being currently submitted anywhere else for the award of any degree or diploma.

Sharifa Rania Mahmud
Candidate

Contents

Board of Examiners **i**

Candidate's Declaration **ii**

List of Figures **vi**

List of Tables **viii**

List of Abbreviations **ix**

Acknowledgements **x**

Abstract **xi**

1 Introduction **1**

1.1 Phylogenetic Tree 1

1.1.1 Types of Phylogenetic Tree 2

1.1.2 Limitations 2

1.2 Applications of Phylogenies 3

1.3 Motivation Behind this Project 4

1.4 Objective of this Project 4

1.5 Main Contribution 5

1.6 Scope 6

1.7 Thesis Organization 6

2 Phylogenetic Tree Reconstruction Methods **8**

2.1 Distance-Based Methods 8

2.2 Character-Based Methods 9

2.3 Supertree-Based Tree Construction 9

2.4 Quartet 10

2.5 Quartet-based Phylogenetic Tree Reconstruction 11

2.5.1 Quartet Max-Cut (QMC) 13

2.5.2 Quartet Fiduccia Mattheyses (QFM) 14

2.5.3 Singular Value Decomposition Quartets (SVDquartets) 15

3 Existing Visualzation Tools 17

3.1 Web-based Tools 17

3.1.1 EvolView v2 17

3.1.2 PHYLOViZ Online 18

3.1.3 IcyTree 19

3.1.4 AQUAPONY 20

3.1.5 Phylo.io 21

3.1.6 Phylogeny.fr 22

3.1.7 T-Rex 23

3.2 Desktop-based Tools 24

3.2.1 Bio::Phylo 24

3.2.2 TreeGraph 2 25

3.2.3 PHYLOViZ 26

3.2.4 Dendroscope 26

3.2.5 FigTree 27

3.2.6 ETE 3 27

3.3 Libraries 28

3.3.1 jsPhyloSVG 28

3.3.2 PhyD3 28

3.3.3 phylotree.js 30

3.3.4 GGTREE 30

4 PhyloQon 32

4.1 Software 32

4.2 Quartet-based Approaches 33

4.3 Features 33

4.3.1 Input Formats and Approaches 33
4.3.2 Comprehensive and Intuitive User-Interface 34
4.3.3 Tree Visualization 36
4.3.4 Data Import and Export 37
4.3.5 Application Programming Interface (API) 38

5 Experimental Analysis 40

5.1 Robinson and Foulds Distance 40
5.2 Datasets 41
5.3 Computational Analyses 41

6 Conclusions and Future Work 43

6.1 Future Directions 44

List of Figures

2.1 The quartet tree *((Human, Chimpangees), (Gorillas, Orangutans))*. 10
2.2 A set of quartets. 11
2.3 Phylogenetic tree reconstructing from a set of quartets shown in Figure 2.2. . . 12
2.4 Set of six quartets. 13
2.5 Reconstructed phylogenetic tree $((((1, 2), 3), 4), (5, 6))$ using QMC over set of six quartets shown in Figure 2.4. 14
2.6 Reconstructed phylogenetic tree $((5, 6), (4, ((1, 2), 3)))$ using QFM over set of six quartets shown in Figure 2.4. 15

3.1 The snapshot of the interface of EvolView v2. 18
3.2 The snapshot of the interface of PHYLOViZ Online. 19
3.3 Interface of IcyTree. 20
3.4 The snapshot of the interface of AQUAPONY. 21
3.5 The snapshot of the interface of Phylo.io. 22
3.6 The snapshot of the interface of Phylogeny.fr. 23
3.7 The snapshot of the interface of T-Rex. 25
3.8 The snapshot of the interface of PhyD3. 29
3.9 The snapshot of the interface of phylotree.js. 31

4.1 A snapshot to demonstrate the input format and approaches choose options. . . 34
4.2 Different sample input data format. 35
4.3 A snapshot of **PhyloQon**. 36
4.4 A snapshot of **PhyloQon** for tree visualization. 37
4.5 A screenshot of the history of a user. 37
4.6 snapshot after using the API URL in browser. 38

5.1 RF distance for QFM and QMC approaches. 42

List of Tables

4.1 Different approaches available in **PhyloQon** API. 39

5.1 Comparison of QFM and QMC. RF distance and running time in minutes are compared. 42

List of Abbreviations

API	Application Programming Interface
ETE	Environment for Tree Exploration
FM	Fiduccia and Mattheyses
HGT	Horizontal Gene Transfer
HTML	Hypertext Markup Language
ML	Maximum-Likelihood
MP	Maximum-Parsimony
MRP	Matrix Representation with Parsimony
NJ	Neighbor Joining
QFM	Quartet Fiduccia Mattheyses
QJ	Quartet Joining
QMC	Quartet Max Cut
QP	Quartet Puzzling
RF	Robison and Foulds
SNP	Single Nucleotide Polymorphism
SQP	Short Quartet Puzzling
SVDquartets	Singular Value Decomposition Quartets
SVG	Scalable Vector Graphics

Acknowledgements

First of all, I would like to declare that all the appraisals belong to the Almighty **ALLAH**. He has blessed me with the capability to continue and complete this research work. I deeply express my gratitude for all His kindness.

I would like to thank my supervisor Professor Dr. M Sohel Rahman for introducing me to the fascinating and prospective field of phylogeny. I have learned from him how to carry on a project work, how to write, speak and present well. My heartiest gratitude goes to him for his patience in reviewing my so many inferior drafts, for correcting my proofs and language, suggesting new ways, leading to the right way and motivating me to continue my work. I again express my indebtedness, sincere gratitude and profound respect to him for his constant supervision, suggestions and guidance throughout the progress of this work.

I would also want to thank the members of the Board of Examiners for their valuable suggestions especially to Dr. Md. Shamsuzzoha Bayzid for guiding with the proper solution whenever I was in problem.

I convey my heartfelt reverence to my parents and other family members for giving their endless support and wishes throughout my work.

Abstract

Evolution and the reconstruction of phylogenetic tree or phylogeny reconstruction are classical topics in biology. Various computational techniques have been applied in the computational biology literature for phylogeny reconstruction. One such method, quartet-based phylogenetic tree reconstruction method, has been accepting broad consideration over than last few years. Quartet is assumed to be the most basic information unit in the context of phylogeny and is an unrooted tree having 4 (four) taxa. The goal is to combine a set of quartets into a single tree that can represent the evolutionary relation of all the species in the entire set of quartets. The most recent and high-flying quartet-based phylogenetic tree reconstruction approaches are, Quartet Max-Cut (QMC) [1], Quartet FM (QFM) [2] and SVDquartets [3]. Reconstruction based phylogenetic tree visualization tools has became very essential and desirable subjects for biological research. There are large numbers of phylogenetic tree vizualization tool available, some are web-based and some are desktop based. Most of the them visualize already constructed phylogenetic tree, that means a tool takes a already constructed phylogenetic tree as input and visualise that tree in different format. Very few number of tools are available which construct, visualise and analyze phylogenetic trees, but the situation is, all of them are using sequence-based method to reconstruct the trees.

We have developed a web-based tool **PhyloQon** which can construct, visualise and analyze phylogenetic trees using quartet-based methods. We incorporate QFM, QMC and SVDquartets approaches to reconstruct the trees, where QFM and QMC both take a set of quartets as input and construct phylogenetic tree, but the format of the quratets are different. QFM takes input quartets in Newick format and constructs a unrooted phylogenetic tree as output. SVDquartets takes sequences as input in PHYLIP format, and it only generates a set of quartets as output. Next, it applies QFM or QMC for reconstructing the phylogenetic tree and recommends QMC or QFM for better result. Our developed tool has also integrated with another option for taking

input, is gene tree. Several tree visualization options (i.e. circular and/or rectangular pattern) for viewing the constructed phylogenetic trees are available. Very few number of tools are available, that can visualise more than one trees at a same view. In our tool we have facilities to reconstruct trees with single or multiple approaches and also view them side-by-side. It will help the users to visually compare the trees which are constructed using different approaches with same input datasets. User can also import data and export data and graphical view of the constructed trees. We have developed an API of this tool, which enables us to get the data without using the graphical user interface of the web page.

Chapter 1

Introduction

Bioinformatics is an interdisciplinary field that develops ways and software package tools for understanding biological knowledge. The study of evolution is prime to the investigation of a large array of biological queries. A phylogenetic tree could be a branching diagram or tree showing the evolutionary relationships among numerous biological species or different entities. It is a diagram setting out the genealogy of a species. Phylogenetic trees are hypotheses, not definitive facts. Phylogeny is the evolutionary history and line of descent of a species. The purpose of phylogeny is to reconstruct the correct genealogical ties between related objects and also to estimate the time of divergence between them since they last shared a common ancestor. Charles Darwin [4] is credited with the earliest representation of a phylogenetic tree published in his book *The Origin of Species*. Over a century later, evolutionary biologists still use tree diagrams to depict evolution. One of the foremost bold goals in evolution is to discover the relationships among all the species on Earth, the *Tree of Life*.

1.1 Phylogenetic Tree

A tree is a formal structure of the representation of the process of evolution. The leaves represent the species under study, the interior nodes represent virtual ancestors and the edges represent the evolutionary events. In biology this tree is called a phylogenetic tree. A phylogenetic tree is a diagram that represents evolutionary relationships among organisms. The pattern of branching in a phylogenetic tree reflects how species or different groups evolved from a series of common ancestors. In phylogenetic trees, two species are more related if they have a more recent common ancestor and less related if they have a less recent common ancestor. It can be drawn in various equivalent patterns. Rotating a tree about its branch points doesn't change the information it carries.

1.1.1 Types of Phylogenetic Tree

A rooted tree is a tree in which one of the nodes is stipulated to be the root, and thus the direction of ancestral relationships is determined. The root of a tree is considered as the oldest point in the tree which represents the last common ancestor of all groups included in the tree. Hence, a rooted tree shows the direction of evolutionary time. Since the rooted tree depicts the direction of evolutionary time, it is easy to find the older or newer groups it has. A rooted tree can be used to study the entire groups of organisms. An unrooted phylogenetic tree is a phylogenetic diagram which lacks a common ancestor or a basal node. This type of a tree does not indicate the origine of evolution of the groups of interest. It depicts only the relationship between organisms irrespective of the direction of the evolutionary timeline. It make an illustration about the leaves or branches.

A phylogenetic tree that bifurcates has a maximum of two descendants arising from each of the interior nodes. The multi-furcates has multiple descendants arising from each of the interior nodes. Some phylogenetic tree shows their edge lengths, some doesn't. The distance of the lines is used to determine how long ago they may have had a common ancestor.

1.1.2 Limitations

Phylogenetic trees produced on the premise of sequenced genes or genomics data in different species can provide evolutionary insight, but these analyses have some important limitations. The trees that they generate are not necessarily correct. They do not necessarily accurately represent the evolutionary history of the included taxa. Trees are meant to provide insight into a research question and not intended to represent an entire species history. Several factors, like gene transfers, may affect the output placed into a tree.

1.2 Applications of Phylogenies

Tree of Life reconstruction is one of the great challenges of science, which is the evolutionary history of all organisms on Earth. Besides evolutionary history phylogeny become more popular in many applications. Population history, rates of evolutionary change, origins of diseases, prediction of sequence function are some of the uses of phylogeny. It can be applied to organisms, sequences, viruses, languages etc. In this section, we give a brief overview of applications of phylogeny.

- **Conservation Biology:** Phylogeny can help to inform conservation policy when conservation biologists have to make tough decisions about which species they try to prevent from becoming extinct. For example, illegal whale hunting.

- **Classification:** Phylogeny based on sequence data provides us with more accurate descriptions of patterns of relatedness than was available before the advent of molecular sequencing. This is now informs the classification of new species.

- **Forensics:** Phylogeny is used to assess DNA evidence presented in court cases to inform situations. For example, dental practice HIV transmission, if someone has committed a crime, when food is contaminated, or where the father of a child is unknown. Evolution of diseases: In a more practical vein, phylogenies can be used to track the

- **Evolution of Diseases:** The inference of phylogenies with computatinal methods has many applications in medical biological research, such as drug discovery and development. It can be used to design drugs and vaccines that are more likely to be effective against the currently dominant strains. The most prominent example of this use is the flu vaccine, which is altered from year to year as medical experts work to keep track of the influenza types most likely to dominate in a given flu season.

- **Bioinformatics and Computing:** Many of the algorithms developed for phylogenetics have been used to develop software in other fields.

1.3 Motivation Behind this Project

Most of phylogenetic tree reconstruction methods are sequence-based and can only be applied on small to moderate sized datasets. That means, if anyone want to provide results having an acceptable level of accuracy within a moderate amount of time, then they have to take small datasets. If we take big datasets (hundreds of taxa), these methods need several weeks or months to provide results with an acceptable level of accuracy. Scientists are facing new computational challenges to analyze large amount of data. In this circumstances, supertree methods is reliable, where smaller trees on overlapping groups of species are combined together to get a single larger tree. So, the quartet-based reconstruction methods are introduced and tools are needed which can use the quartet-based phylogenetic tree reconstruction method.

TA large number of phylogenetic tree vizualization tools are available, some are web-based and some are desktop based. Most of the tools are visualize already constructed phylogenetic tree, that means they have take a already constructed phylogenetic tree as input and visualise that tree in different format. There are very few tools are available which construct, visualise and analysis phylogenetic trees, but the situation is, all of them are using sequence-based method to reconstruct the trees. We have developed a web-based tool **PhyloQon** which can construct, visualise and analysis phylogenetic trees, and the tool used some accurate quartet-based methods to reconstruct those trees.

Very few tools are available, those can visualise more than one trees at a same view. In our tool we have facility to reconstruct trees with single or multiple approaches and also view them side-by-side. It will help the users to visually compare the trees which are constructed using different approaches with same input datasets.

1.4 Objective of this Project

Our main objective is to develop a web-based tool which can construct, visualise and analysis phylogenetic trees, where the tool will use the quartet-based approaches for reconstruction the trees. Some main objectives of this project are pointed below:

- We want to develop a web based tool (PhyloQon) that reconstructs and visualizes phylogenetic trees implementing the state of the art quartet based approaches, namely, QMC [1], QFM [2] and SVDquartets [3].

- Incorporate the capability to graphically compare phylogenetic trees produced using different approaches.
- Analyze the constructed phylogenetic trees based on visualization.
- Conduct an experimental analysis of the different approaches based on the developed tool (PhyloQon).
- Developed an Application Programming Interface (API), which enables to get the data without using the graphical user interface of the web page. The API is convenient if a user need to programmatically access some information but still do not want to download the entire datasets.

1.5 Main Contribution

In this project, we address the problem that there are no such visualization tool available to visualize phylogenetic tree, which reconstruct phylogenetic tree by quartet-based methods. Our contributions are summarized below.

i. We have developed a web based tool (PhyloQon) that reconstructs and visualizes phylogenetic trees implementing the state of the art quartet based approaches, namely, QMC , QFM and SVDquartets. We know that QFM and QMC take input as quartet sets, but in different format. To make our tool user-friendly, **PlyloQon** can take input quartet sets in same format and that is Newick format.

ii. There are options for taking input as gene tree and sequences as well.

iii. The tool is able to construct the trees using various approaches. If a user choose multiple options, then the constructed trees are visualise side-by-side. User can easily analyze the constructed phylogenetic trees based on visualization.

iv. Since the results may not be found in real time, **PhyloQon** has the facility to notify the user via e-mail that the result is ready for viewing.

v. There are several tree visualization options (i.e. circular and/or rectangular pattern) for viewing the constructed phylogenetic trees.

vi. There are options for export the visualise trees, the input ant output files. from this tool, we can export the quartet sets generated by SVDquartets method.

vii. The implemented quartet-based approaches has implemented in a suitable language that is compatible to the platform selected for the web-site.

viii. We have conduct an experimental analysis of the different approaches based on the developed tool **PhyloQon**.

ix. We have developed an API, which enables to get the data without using the graphical user interface of the web page.

1.6 Scope

This thesis, from broader perspective, covered the general research field including biochemical characteristic such as molecular phylogeny, infectious disease, biotransformation. Comparisons of the plant species or gene sequences in a phylogenetic context can provide the most important and meaningful sagacity into biology. Most of the time it is very difficult to work with different raw phylogenetic reconstruction methods and their input formats. Through a web-based phylogenetic reconstruction tool we can easily use those methods, compare and analysis the trees.

1.7 Thesis Organization

The rest of the thesis is organized as follows. In Chapter 2 we have discussed the different phylogenetic tree reconstruction methods, where three major quartet-based approaches are incorporated in our tool. Chapter 3 describes different existing web-based and desktop-based visualization tools. There also described different libraries for proper visualization of trees. Chapter 4 thoroughly described the features, uses of our tool **PlyloQon**. Chapter 5, its deals with our experimental works, datasets, experimental setup, result summary and detailed analysis on results. Finally, We conclude in Chapter 5 with some future directions.

Chapter 2

Phylogenetic Tree Reconstruction Methods

By analyzing the molecular sequences of different species, phylogenetic tree reconstruction can be regarded as the *sequence-based* reconstruction of the phylogeny. Sequence-based phylogenetic methods are basically of two types: (a)distance-based methods, such as Neighbor Joining (NJ) [5], which has very and (b) character-based methods, such as Maximum-Likelihood (ML) [6] or Maximum-Parsimony (MP) [7].

2.1 Distance-Based Methods

Distance-based methods, use measures of the overall differences between all pairs of sequences in the alignment, which represented as a matrix of pairwise genetic distances. This type of methods rely on the calculation of genetic distances between each pair of sequences in a dataset with phylogenetic trees being constructed from the resulting distance matrix using a clustering algorithm. Neighbor Joining [5] is one of most popular distance-based method. The primary advantage of such distance-based methods is their computational speed. These methods are therefore ideally suited to the initial exploration of evolutionary relationships between sequences in a dataset. Many of these methods also directly help speed up slower but more accurate character-based tree construction methods by directing these methods to preferably search for highly likely phylogenetic trees that resemble distance-based trees.

2.2 Character-Based Methods

Character-based methods, directly use the individual columns of aligned nucleotides or amino acids. Distance-based methods compress phylogenetic information within a set of sequences into a pairwise-distance matrix, character-based methods take advantage of all the information available in sequences at each homologous site. The most widely used methods are the Maximum Parsimony (MP) [7] and Maximum Likelihood (ML) [6] methods.

- **Maximum-Likelihood (ML):** ML is a statistical concept that has been widely used in many areas of biology such as population genetics and ecological modelling, and which has applied to the field of phylogenetics. The basic concept of likelihood is relatively simple to comprehend, that the probability of obtaining some data, under a model of evolution, with parameters. We can consider the data as nucleotide or amino acid sequences, model of evolution as the mutation process from one base to another. The set of parameters can be the tree topology, tree branch lengths and the substitution model parameters. The maximum likelihood estimates of the parameter values included correspond to the set of values that maximize this probability. The principle is easily understandable, but the calculation of the likelihood function can be mathematically complex, and this method can be very computationally expensive.

- **Maximum-Parsimony (MP):** The MP method is one of the most widely used for phylogenetic inference. This method initially developed for the analysis of morphological traits. Its main idea can be states that, when several hypotheses with different degrees of complexity are proposed to explain the same phenomenon, one should choose the simplest hypothesis. In a phylogenetic context, this principle proposes that the most believable or parsimonious phylogenetic tree will be the tree that invokes the smallest number of evolutionary changes during the divergence of the sequences it represents.

2.3 Supertree-Based Tree Construction

The above described methods are sequence-based and can only be applied on small to moderate sized datasets. That means, if we want to provide results having an acceptable level of accuracy within a moderate amount of time, then we have to take small datasets. If we take big datasets (hundreds of taxa), these methods need several weeks or months to provide results with an acceptable level of accuracy. As the amount of molecular data is accumulating exponentially with the continuous advancement in sequencing technologies, scientists are facing new computational challenges to analyze these enormous amount of data. In this situation we can rely on supertree methods, where smaller trees on overlapping groups of species are combined together to get a single larger tree. In first phase of supertree-based tree construction, many small trees on overlapping subsets of taxa are constructed using a sequence-based method, in the second phase the small trees are summarized into a complete tree over the full set of taxa. The efficient design of supertree methods may considerto work on very larger datasets more accurately and easily. The most widely used supertree method is called the Matrix Representation with Parsimony (MRP) [8,9].

2.4 Quartet

The most persuasive approaches in supertree construction is to decompose the input sub-trees into the most basic information units, which is, quartets, and subsequently combine them into a single tree over the entire set. This is called the *quartet amalgamation* problem and in itself a special case of the supertree problem, that is, all input trees are of size four. Quartet is assumed to be the most basic information unit in the context of phylogeny and is an unrooted tree having 4 (four) taxa (species). The smallest informative piece of phylogenetic information is called a quartet tree or just a quartet,atree over four leaves with a single, central, internal edge. The meaning of a *((Human, Chimpangees), (Gorillas, Orangutans))* quartet is that there is an edge separating *Human* and *Chimpangees* from *Gorillas* and *Orangutans*. Figure 2.1 shows a single quartet.

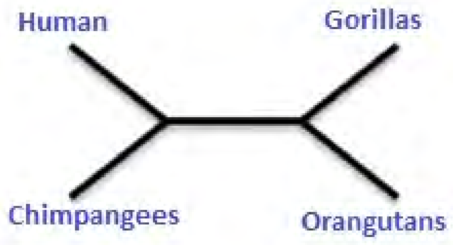


Figure 2.1: The quartet tree *((Human, Chimpangees), (Gorillas, Orangutans))*.

There are many format for representing a quartet. Newick is the most popular format for representing a quartet. For example, *((a,b),(c,d))* is the Newick representation of four taxa, *a*, *b*, *c* and *d*. Another representation of quartet for this four taxa set can be $a \mid b, c \mid d$. This representation has been used in QMC method.

2.5 Quartet-based Phylogenetic Tree Reconstruction

Quartet-based phylogenetic tree reconstruction method, has been receiving extensive attention for more than two decades. For quartet-based phylogenetic tree reconstruction the goal is here to combine a set of quartets into a single tree that can represent the evolutionary relation of all the species in the entire set (of quartets). Taxa that are closely related should be more similar to each other than taxa that are more distantly related, so, when the phylogenetic tree is constructed that put similar taxa on nearby branches. Figure 2.2 shows a set of quartets and Figure 2.3 shows a phylogenetic tree reconstructing from these set of quartets.

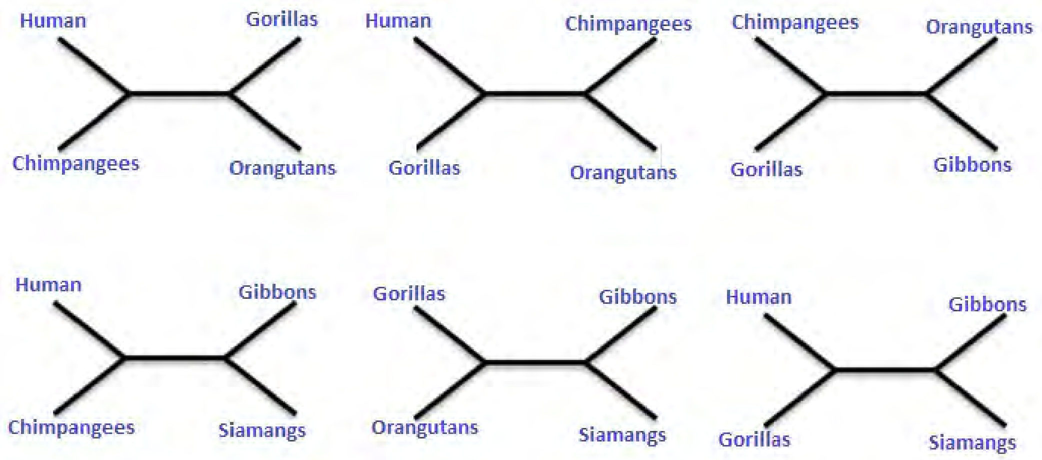


Figure 2.2: A set of quartets.

There are many quartet-based phylogenetic reconstruction methods. Quartet-based phylogenetic tree reconstruction has been receiving extensive attention in the bioinformatics and as well as in literature for more than two decades. Different approaches have been proposed and improved time to time. Among these, the most prominent approaches are, Quartet Puzzling (QP) [10], Quartet Joining (QJ) [11] and Quartet Max-Cut (QMC) [1, 12, 13]. QP construct the phylogeny of n sequences using a weighting mechanism. First of all, it computes the maximum-

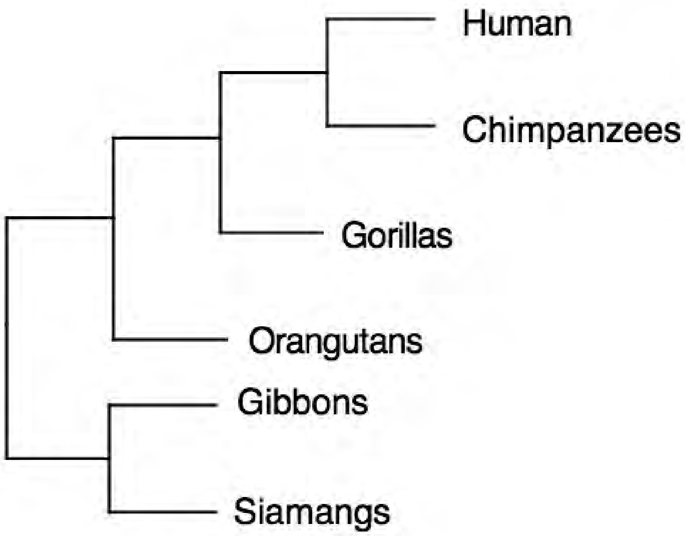


Figure 2.3: Phylogenetic tree reconstructing from a set of quartets shown in Figure 2.2.

likelihood values for the three topologies on every 4 taxa and uses these values to compute the corresponding probabilities. By using these probabilities as weights, the puzzling step has constructed a collection of trees over n taxa. Finally it returns a consensus tree over n taxa. QJ provides the theoretical guarantee to generate the accurate tree if a complete set of consistent quartets is present. On average QJ outperforms QP and its performance is very close to the performance of NJ, but QJ outperforms NJ on quartet sets with low quartet consistency rate.

The most popular quartet-baser accurate phylogenetic tree reconstruction methods are QFM and QMC. In our project we have developed a tool **PhyloQon** for reconstructing, visualizing and analysing phylogenetic tree, where we use different quartet-based phylogenetic tree reconstruction methods. We incorporate QFM, QMC and SVDquartets approaches to reconstruct the trees, where QFM and QMC both take set of quartets as input and construct phylogenetic tree, but the format of the quratets are different. QFM takes input quartets in Newick format and construct a unrooted phylogenetic tree as output. SVDquartets take sequences as input in PHYLIP format, and it only generates a set of quartets as output. Then apply QFM or QMC for reconstruction the phylogenetic tree and recommended QMC for better result. These three approaches will be described thoroughly later in this chapter.

2.5.1 Quartet Max-Cut (QMC)

In 2008, Snir and Rao has [12] presented a new quartet-based phylogeny reconstruction algorithm Short Quartet Puzzling (SQP) and demonstrated the improved topological accuracy of the new method over MP and NJ, disproving the conjecture of Ranwez and Gascuel [14]. They have also showed a improvement over QP. This study shows that quartet methods are not as limited in performance as was previously conjectured, and opens the possibility to further improvements through new algorithmic designs. It differs from the previous techniques in that it does not require all three topologies of the quartets on every 4 taxa. It is able to construct the output tree from a subset of all possible quartets as input. At first this method has used the randomized technique for selecting input quartets from all possible 4-trees (estimated using ML), and then finally it has used Quartet Max Cut (QMC) [12, 13] technique for combining quartets into a single tree. The experimental study conducted by Swenson et al. [15] concludes that QMC performs better than the other supertree methods and MRP for smaller (100-taxon and 500-taxon) and high scaffold density datasets. But MRP outperforms QMC and other supertree methods on larger and low scaffold density datasets [15].

In this mannar, Snir and Rao has described a adaptable implementation of QMC [1], where they experimented and reported the improvement of QMC over MRP in terms of accuracy and running time. Although MRP is the mostly used supertree method in practice, the studies of [1, 15] suggest that QMC is so far the best quartet-based supertree method. Figure 2.5 shows the reconstructed phylogenetic tree using QMC over six quartet set shown in Figure 2.4. 2.4.

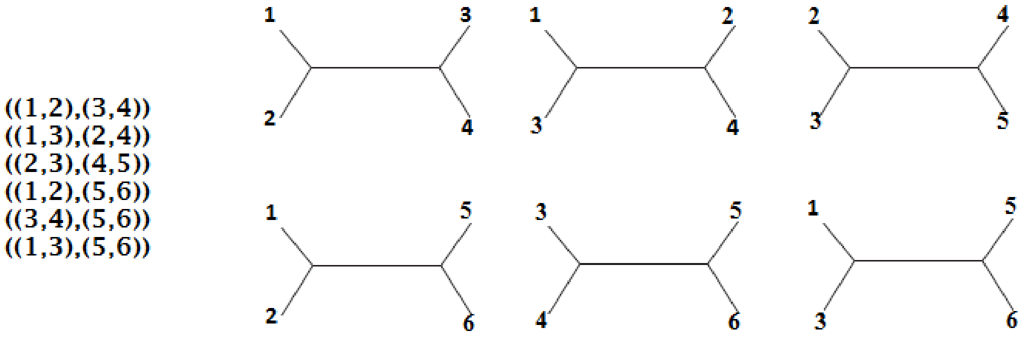


Figure 2.4: Set of six quartets.

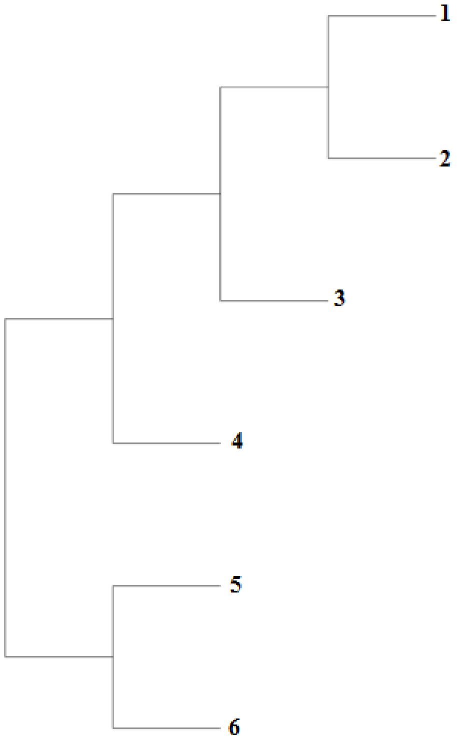


Figure 2.5: Reconstructed phylogenetic tree $((((1,2),3),4),(5,6))$ using QMC over set of six quartets shown in Figure 2.4.

2.5.2 Quartet Fiduccia Mattheyses (QFM)

Reaz et al. [2] has presented a novel and highly accurate quartet-based phylogeny reconstruction algorithm, Quartet Fiduccia Mattheyses (QFM). In their method thay have used a bipartition technique inspired from the famous Fiduccia and Mattheyses (FM) algorithm for bipartitioning a hyper graph minimizing the cut size [16]. They have reported that QFM is highly accurate and scalable to large datasets, upto several hundreds of taxa. They have also demonstrated the accuracy of QFM by analyzing its performance on both simulated and biological datasets. The authors have compared QFM on simulated datasets with QMC, and showed the superiority of QFM method over QMC in terms of the accuracy of the estimated trees.

In experimental study they have comparing QFM against QMC under different model conditions by varying different parameters. In almost all model conditions considered, QFM performs at least equal but in most cases better than QMC. Both QFM and QMC can reconstruct very accurate trees with $n^{2.8}$ quartets, indicating that it is possible to reconstruct an accurate supertree from large number of quartets, even with high amount of noise in the input data. QFM has also been tested on real biological datasets and has been shown to perform pretty well. QFM is known to be the best quartet amalgamation method to date. Figure 2.6 shows the reconstructed phylogenetic tree using QFM over the same quartet sets shown in Figure 2.4. We can observe that, Figure 2.5 showed the reconstructed phylogenetic tree using QMC method and Figure 2.6 showed the reconstructed phylogenetic tree using QMC method, and they are almost same, only there are topological differences.

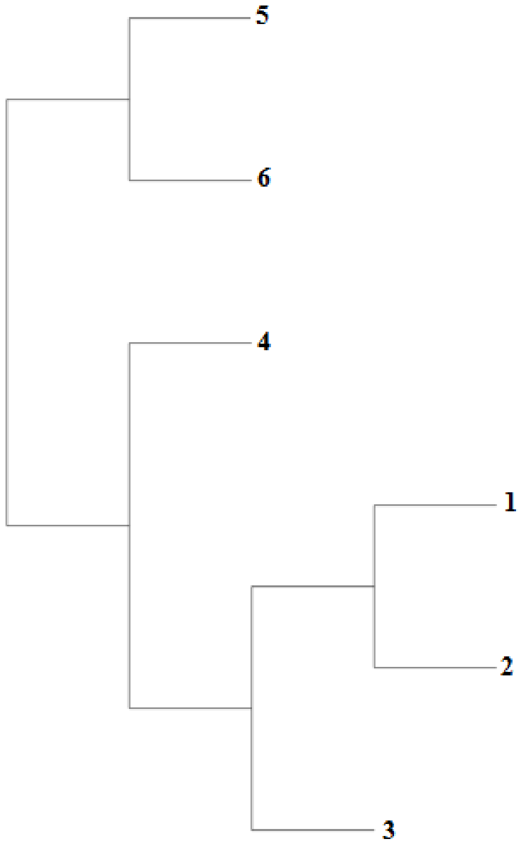


Figure 2.6: Reconstructed phylogenetic tree $((5,6),(4,((1,2),3)))$ using QFM over set of six quartets shown in Figure 2.4.

2.5.3 Singular Value Decomposition Quartets (SVDquartets)

Chifman and Kubatko [17] introduced a single-site method SVDquartets, which can deal with nucleotide information. They have proved that under the multi-species coalescent and with the assumption of a strict molecular clock (i.e., that the rate of sequence evolution per unit time is constant throughout the model gene tree), an unrooted species tree on four taxa is generically identifiable from site pattern probabilities at the leaves of the tree [17]. Chifman and Kubatko have proved it using algebraic statistics and singular value decomposition. The SVDquartets algorithm takes unlinked multi-locus data for a set of four taxa as input and assigns a score *SVD* *score*, to each of the three possible quartet topologies. The quartet topology with the lowest *SVD score* is selected as the true topology for that quartet [3].

The method SVDquartets, just computes a set of quartets. A quartet amalgamation method is needed to combine these set of quartets on every four species into a species tree on the full set of taxa. Chifman and Kubatko [17] suggested QMC [1] approach for reconstructing phylogenetic tree. However, SVDquartets has also been implemented in PAUP* [18], which uses a variant of QFM [2] to combine the quartet sets into a species tree or phylogenetic tree, and is the implementation currently recommended by the developers of SVDquartets.

Recently Vachaspati and Warnow [19] present SVDquest, which has improved SVDquartets using exact optimization within a constrained search space. It is a new method for coalescent-based phylologenetic tree estimation. This method has used site patterns and is robust to gene tree estimation error. SVDquartets have been only use the site patterns to estimate the phylogenetic tree, and so are not impacted by gene tree estimation issues. SVDquest has constructed phylogenetic trees using site patterns that is guaranteed to produce phylogenetic trees that satisfy at least as many quartets as SVDquartets. It has dominated SVDquartets implemented in PAUP$*$ for criterion scores and also find the optimal solutions to its optimization problem.

Chapter 3

Existing Visualzation Tools

There are many web based and desktop based phylogenetic tree analysis tools. In this chapter we have discussed some most recent and extensively used phylogenetic tree visualization tools that have been developed over the past few years. The presented tools in this chapter are selected to cover the different functionalities and features essential for analysis and visualization of phylogenetic trees. We have also commented on the limitations and the specific strengths of these tools, while the discussed tools are all broadly applicable.

We have discussed both desktop-based and web-based tools. Some tools visualize the already constructed phylogenetic trees. They take already constructed phylogenetic tree as input and visualize the graphical view of that tree. Some tools are both construct and visualize the phylogenetic trees. The tools which can construct and visualize the trees, all of them are using sequence-based phylogenetic methods.

3.1 Web-based Tools

3.1.1 EvolView v2

EvolView v2 [20] introduced a comprehensive web application for visualizing, annotating and managing phylogenetic trees. This tool also incorporated an evolutionary dataset system that allows users not only to upload biologically relevant contents to be displayed graphically on the phylogenetic trees, but also to control all major aspects of the graphical annotations. The tool can be freely accessed at `http://www.evolgenius.info/evolview`. Figure 3.1

shows the snapshot of this tool. Some features of EvolView are as follows:

- It takes phylogenetic trees in several formats as input and display trees as phylograms and cladograms, each in either rectangular or circular layout.
- It has a management system for phylogenetic trees and associated datasets.
- The user-interface is comprehensive and intuitive .
- It also has an unique dataset system for graphical annotations.

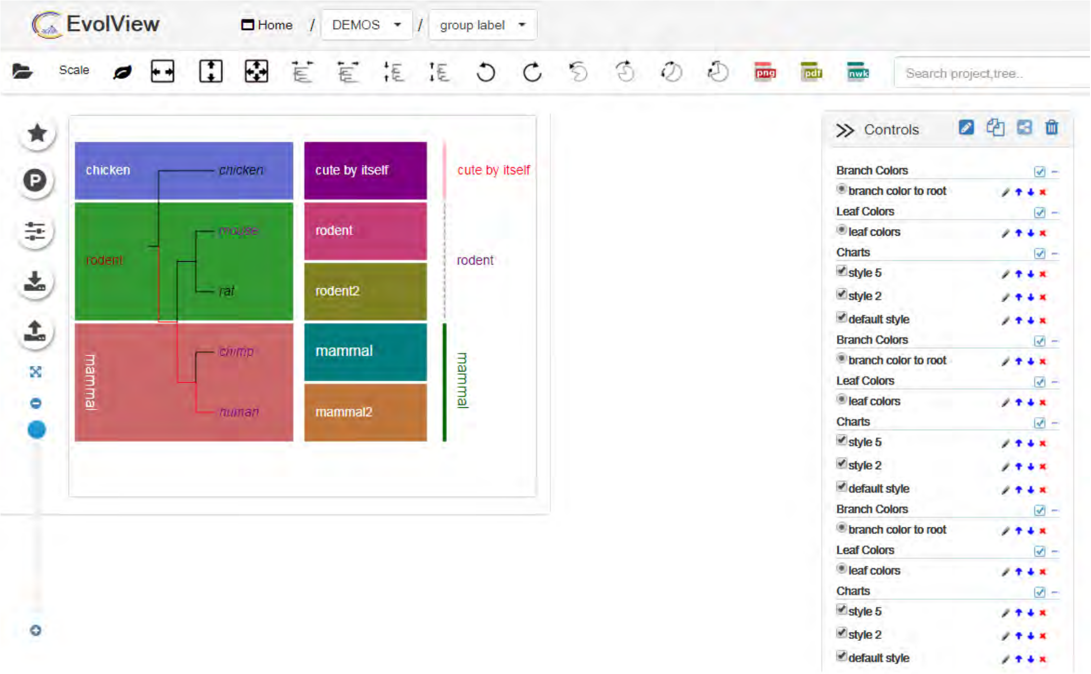


Figure 3.1: The snapshot of the interface of EvolView v2.

3.1.2 PHYLOViZ Online

PHYLOViZ Online [21] was developed as a web application for profile-based data analysis, visualization and sharing, also allowing the application of visual analytic processes on trees through traditional phylogenetic methods. The tool can be freely accessed at https://online.phyloviz.net. The snapshot of this tool is showed in Figure 3.2. Some features of this online tool are mentioned below:

- Takes input in different format, visualize the output tree and also displays input data in a tabular format.

- Provides users with a dynamic interactive distance matrix that can be constructed depending on the nodes selected.
- The application also offers a visual representation of multi-sequence alignment for FASTA input files.
- One goal of PHYLOViZ Online [21] is to fill the existing software gap for sharing and reproducing phylogenetic inference using MST-like approaches. All datasets uploaded by unregistered users can also be shared, but these will only be available for a limited of time.
- Also makes available an authenticated API, providing users with programmatic access to public data or to registered user data.

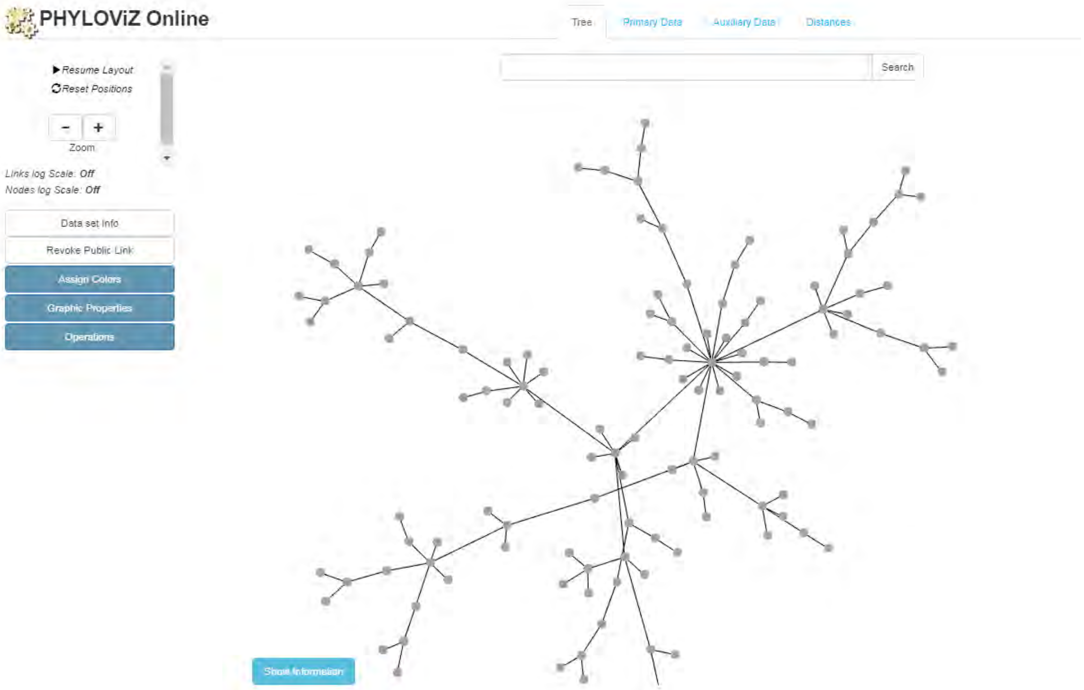


Figure 3.2: The snapshot of the interface of PHYLOViZ Online.

3.1.3 IcyTree

Application IcyTree [22] is used to visualize variety of phylogenetic trees and networks, having a responsive user interface, supporting phylogenetic networks (ancestral recombination graphs in particular), and efficiently drawing trees that include information such as ancestral locations or trait values. It is written entirely in JavaScript and runs in modern web browsers. The tool can be freely accessed at http://tgvaughan.github.com/icytree. The snapshot of IcyTree is showed in Figure 3.3. Features of this online tool are mentioned below:

- The SVG graphical format allows to display trees containing thousands or even tens of thousands of taxa while remaining responsive. It provides intuitive panning and zooming utilities that make exploring large phylogenetic trees of many thousands of taxa feasible.
- It has focused on drawing rooted phylogenetic time trees such as those inferred by Bayesian phylogenetic inference packages.
- It is capable of viewing trees stored in a number of different formats, including Newick, NEXUS, PhyloXML, NeXML and Extended Newick format.

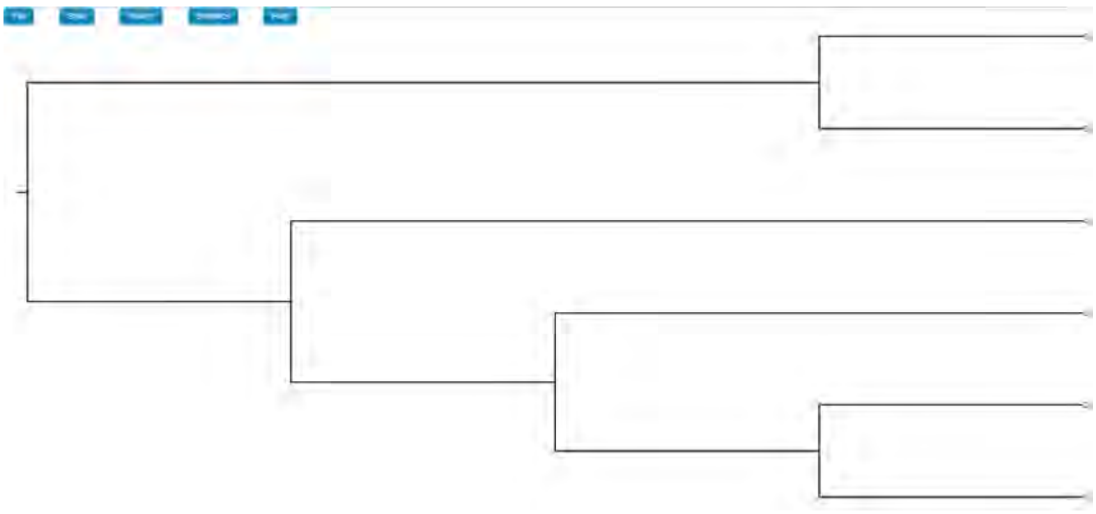


Figure 3.3: Interface of IcyTree.

3.1.4 AQUAPONY

The web-based tool AQUAPONY [23] visualize evolutionary tree with ancestral annotations, the user can easily control the display of ancestral information on the entire tree or a subtree. It is a Javascript tree viewer for Beast [24]. It is coded in JavaScript and HTML. It instantaneously updates the tree visualizations even for large trees, which can be exported as image files. The tool can be accessed directly at http://www.atgc-montpellier.fr/aquapony. The snapshot of the tool is showed in Figure 3.4. Features of this online tool are mentioned below:

- It displays the traits obtained by parsing the input.The interface allows pointing each node on the tree visualization.
- It also contains some fields to control display.
- There are two figure panels offers a dynamically linked, whole tree and subtree visualizations.
- It displays two alternative phylogeographic scenarios upon selection of a leaf or group of leaves in the tree.

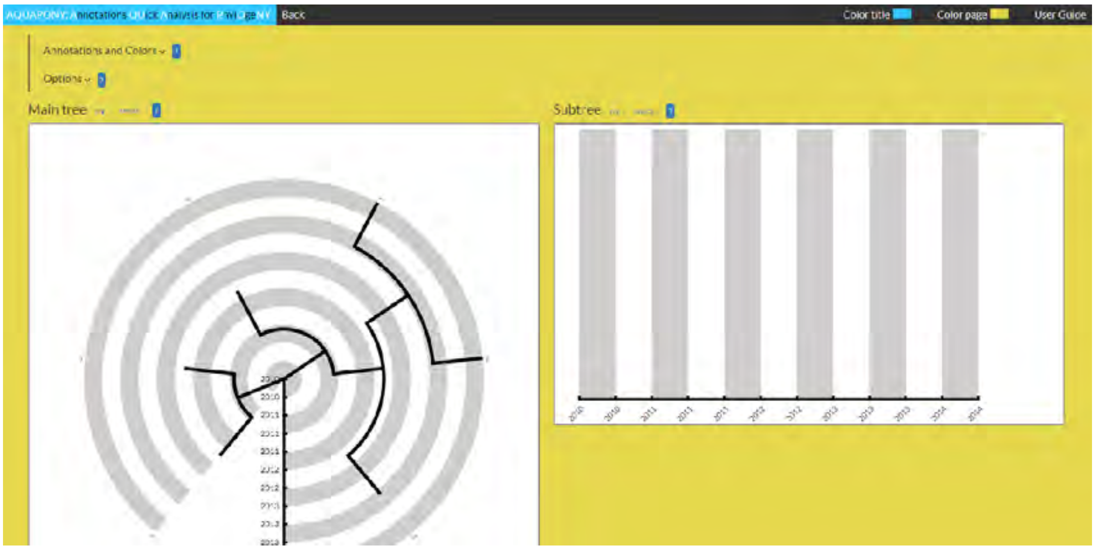


Figure 3.4: The snapshot of the interface of AQUAPONY.

3.1.5 Phylo.io

The web application Phylo.io [25] has visualized and compared phylogenetic trees side-by-side. It highlight the similarities and differences between two trees. It has also automatically identify of the best matching rooting and leaf order, scalability to large trees, high usability, multiplatform support via standard HTML5 implementation, and possibility to store and share visualizations. The tool has built using HTML5, CSS, Ajax, jQuery, and the D3 JavaScript visualization library. The tool can be freely accessed at `http://phylo.io`. The screenshot of this tool is showed in Figure 3.5. Features of this online tool are mentioned below:

- Using this tool the user can choose between two modes using the newick format as input. The *view* mode makes it possible to display a single tree and the *compare* mode, two trees are displayed side-by-side.

- It improves the legibility of large trees by estimating an optimal collapsing depth using the number of leaves and size of viewing area and displaying automatically a collapsed version of the tree.
- The user can compute the best corresponding rooting and match leaves in the tree to find the best corresponding internal node in the opposing tree.
- For interactive analysis of specific parts of the tree, the user can select a node and highlight it. It has allowed the user to interactively match structures within the compared trees.
- It has also allowed users to share tree visualizations using the GitHub Gist API.

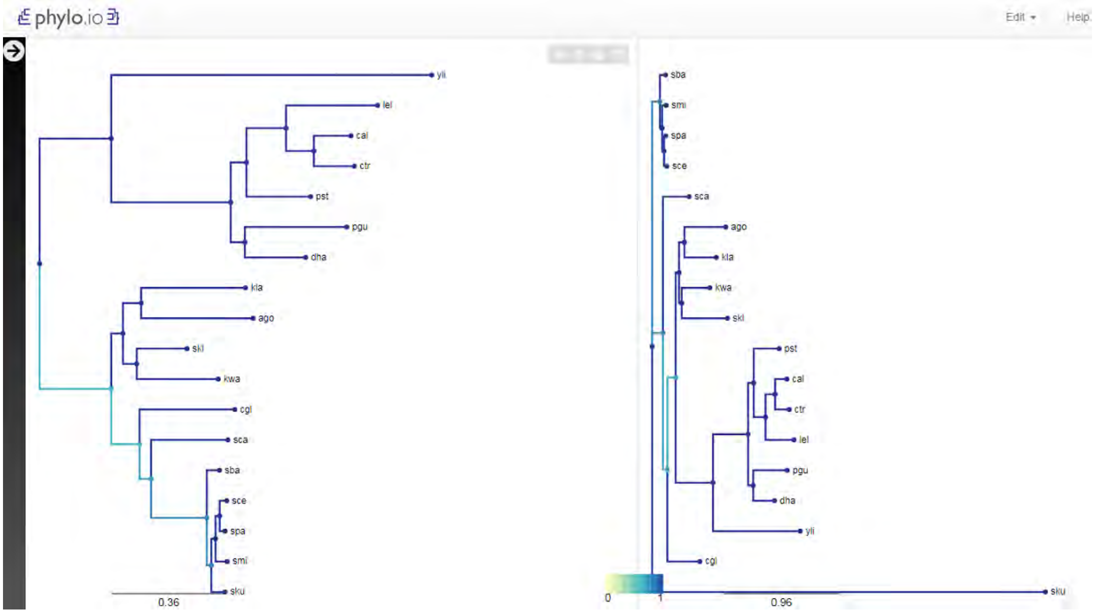
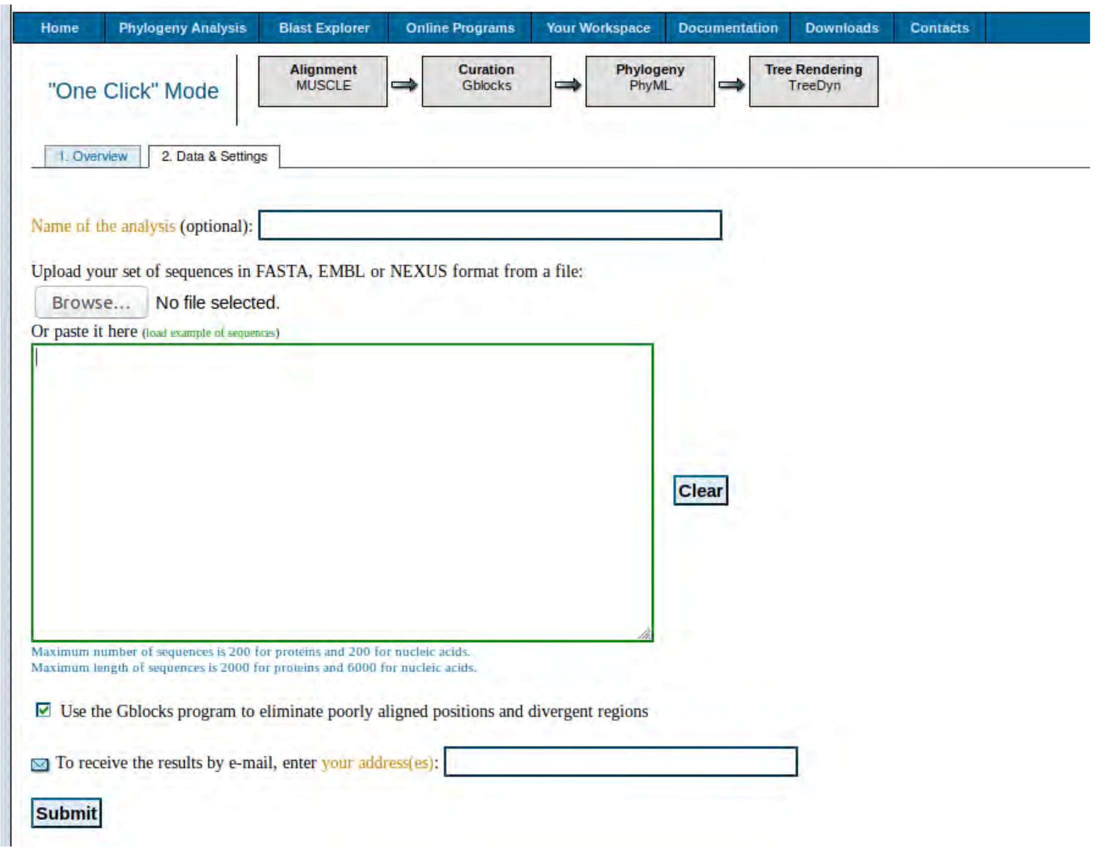


Figure 3.5: The snapshot of the interface of Phylo.io.

3.1.6 Phylogeny.fr

Above described web-based tools are only visualize the constructed phylogenetic tree. That means they take already constructed phylogenetic tree as input and visualize that tree. Phylogeny.fr [26] is a web-based tool which have reconstruct and visulize the phylogenetic trees. Primarily it has designed for biologists with no experience in phylogeny to analyze their data in a simple and robust way. It can also meet the needs of specialists, while the specialists will be able to easily build and run sophisticated analyses. This tool offers three main modes. Through *One Click* mode users can paste their set of sequences and let the software make decisions on their behalf. User can manually set parameters for the various steps in *Advanced* mode. The third mode is *A la Carte* mode, through this mode users can create their own phylogeny workflow using more programs available. The pipeline to reconstruct a phylogenetic tree from a set of sequences through an automated process that successively performs multiple sequence alignment, alignment refinement, phylogenetic reconstruction and, finally, a graphical representation of the resulting tree. Phylogeny.fr is available at `http://www.phylogeny.fr`. The screenshot of the *One Click* mode interface of this tool is showed in Figure 3.6.



Name of the analysis (optional):

Upload your set of sequences in FASTA, EMBL or NEXUS format from a file:

No file selected.

Or paste it here (load example of sequences)

Maximum number of sequences is 200 for proteins and 200 for nucleic acids.
Maximum length of sequences is 2000 for proteins and 6000 for nucleic acids.

☑ Use the Gblocks program to eliminate poorly aligned positions and divergent regions

To receive the results by e-mail, enter your address(es):

Figure 3.6: The snapshot of the interface of Phylogeny.fr.

3.1.7 T-Rex

Another web-based tool which have reconstruct and visulize the phylogenetic trees is T-REX (Tree and reticulogram REConstruction) [27]. It is a web server dedicated to the reconstruction of phylogenetic trees, reticulation networks and to the inference of Horizontal

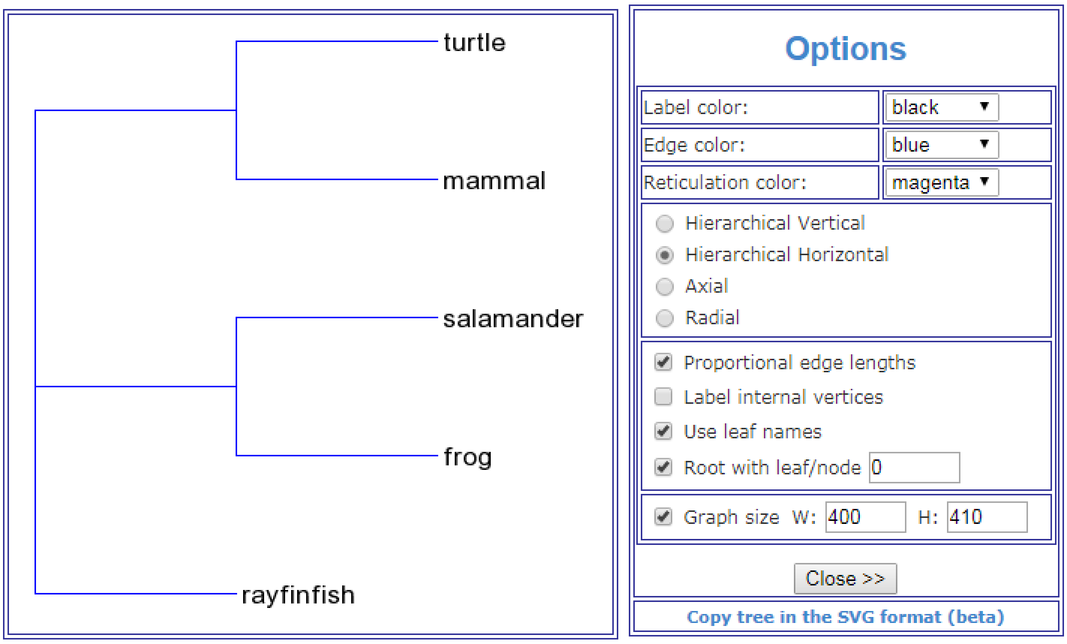
Gene Transfer (HGT) events. T-REX includes several popular bioinformatics applications, random phylogenetic tree generator and some well-known sequence to distance transformation models. The web server of this tool is available at www.trex.uqam.ca. The snapshot of the interface of this tool is showed in Figure 3.7. The latest web server version of this tool includes the following applications:

- Methods for the visualization and interactive manipulation of phylogenetic trees.
- An application for drawing phylogenetic trees.
- Methods for inferring and validating phylogenetic trees using distances.
- Methods for reconstructing phylogenetic trees from a distance matrix containing missing values.
- A method for inferring reticulograms from distance matrices.
- Complete and partial HGT detection and validation methods.
- Widely used multiple sequence alignment tools, are available with slow and fast pairwise alignment options.
- Most common Sequence to Distance transformations.
- Computation of the Robison and Foulds (RF) [28] topological distance.
- Newick to Distance matrix and Distance matrix to Newick format conversion.
- Random phylogenetic tree generation program.

3.2 Desktop-based Tools

3.2.1 Bio::Phylo

Bio::Phylo [29] is a desktop-based Perl5 toolkit for phyloinformatic analysis. It has been deployed successfully on a variety of computer architectures, including various Linux distributions, Mac OS X versions, Windows, Cygwin and UNIX-like systems. It implements



Options

Label color: black ▾

Edge color: blue ▾

Reticulation color: magenta ▾

○ Hierarchical Vertical
● Hierarchical Horizontal
○ Axial
○ Radial

☑ Proportional edge lengths
□ Label internal vertices
☑ Use leaf names
☑ Root with leaf/node 0

☑ Graph size W: 400 H: 410

Figure 3.7: The snapshot of the interface of T-Rex.

classes and methods that are compatible with the well-known BioPerl toolkit, but is independent from it. It provides a richer API and a data model that is better for managing the complex relationships between different fundamental data and metadata objects in phylogenetics. It supports commonly used file formats for phylogenetic data, which allows rich annotations of phylogenetic data to be stored and shared. It can access the file formats that BioPerl supports. Many methods for data simulation, transformation and manipulation, the analysis of tree shape, and tree visualization are provided. It is available as open source software from `http://search.cpan.org/dist/Bio-Phylo`.

3.2.2 TreeGraph 2

TreeGraph 2 [30] is a application to produce ready-to-publish trees. It can display any number of annotations in several ways, and permits easily importing and combining them. A huge number of editing and formatting operations is available in this tool. It is freely available at `http://treegraph.bioinfweb.info`. Some key features of this tool in mentions below:

- It can read trees in different format like, Newick, Nexus format, NeXML or PhyloXML.
- It import annotations from text files or combine information from different phylogenetic analysis.

- This tool can present unlimited number of numerical or textual annotations on every branch.
- It export trees to various vector and pixel graphic formats (e.g. PDF, SVG, EMF or PNG).
- It has many global and element specific formats like line width or color and text formats.
- Using this tool user can automatically setting branch widths or colors according to the value of any attached data.
- Also has editing operations like rerooting, ladderizing or moving and collapsing nodes or copying or manually creating whole clades.
- It can generate commands and import data for ancestral state reconstruction.

3.2.3 PHYLOViZ

PHYLOViZ 2.0 [31] is a user-friendly software that allows the combined analysis of multiple data sources for microbial epidemiological and population studies. It is a platform independent Java tool that allows phylogenetic inference and data visualization for large datasets of sequence based typing methods, including Single Nucleotide Polymorphism (SNP) and whole genome/core genome analysis. This tool incorporates new data analysis algorithms and new visualization modules, as well as the capability of saving projects for subsequent work or for dissemination of results. It is freely available at http://www.phyloviz.net.

3.2.4 Dendroscope

Dendroscope 3 [32] is a software for working with rooted phylogenetic trees and networks. It provides a number of methods for drawing and comparing rooted phylogenetic networks, and for computing them from rooted trees. The program is written in Java and can be used interactively or in command-line mode. The software is free and can be downloaded from www.dendroscope.org. Below some main features of this tool are pointed,

- The aim of this tool is to provide a user-friendly tool for working with rooted phylogenetic trees and networks.

- Phylogenetic trees and networks can be loaded from a file in Newick format, in the case of trees, extended Newick format, in the case of networks, or in Nexus format.
- It can be configured to show a grid of n×m trees or networks simultaneously, thus making it easier to work with data sets that contain multiple trees or networks, such as obtained from multiple genes or by using multiple methods.
- This tool provides a number of distance calculations for comparing two rooted phylogenetic trees or networks based on the contained clusters or trees and other concepts.
- The tool contains a very general algorithm that can compute a tanglegram [33] for two rooted trees or networks that may have different taxon sets.

3.2.5 FigTree

FigTree [34] is designed as a graphical viewer of phylogenetic trees and as a program for producing publication-ready figures. As with most of my programs, it was written for my own needs so may not be as polished and feature-complete as a commercial program. In particular it is designed to display summarized and annotated trees produced by BEAST [24]. It is available from `https://github.com/rambaut/figtree/releases`.

3.2.6 ETE 3

The above presented desktop-based tools are only visualize the already constructed phylogenetic tree. The Environment for Tree Exploration (ETE) v3 [35] is a desktop-based computational framework that simplifies the reconstruction, analysis, and visualization of phylogenetic trees and multiple sequence alignments. The tool has provided a comprehensive Python programming library (API) that allows researchers to automate common tasks in comparative genomics. This tool is freely available at `http://etetoolkit.org`. Some most important features of this tool are,

- Building gene-based and supermatrix-based phylogenies using a single command.
- Testing and visualizing evolutionary models.
- calculating distances between trees of different size or including duplications.

- Providing seamless integration with the NCBI taxonomy database.

Above mentioned web-based and desktop-based tools, analysis the phylogenetic tree using sequence-based phylogenetic methods. There are no tool available which use quartet-based approaches. Besides these above discussed tools there are many others (see Wikipedia [36]), each tool has its own advantages and disadvantages.

3.3 Libraries

Many software packages have been developed to address the need for generating phylogenetic trees intended for print. There are different popular libraries are used to visualize and analysis phylogenetic trees. Most of them are JavaScript libraries. These libraries are used by different applications and those applications can view and interact with phylogenetic trees. The effects of such interactions can be captured and communicated to other package components, making it possible to engineer complex and responsive applications that include phylogenetic trees. This section present some most recent and popular JavaScript libraries.

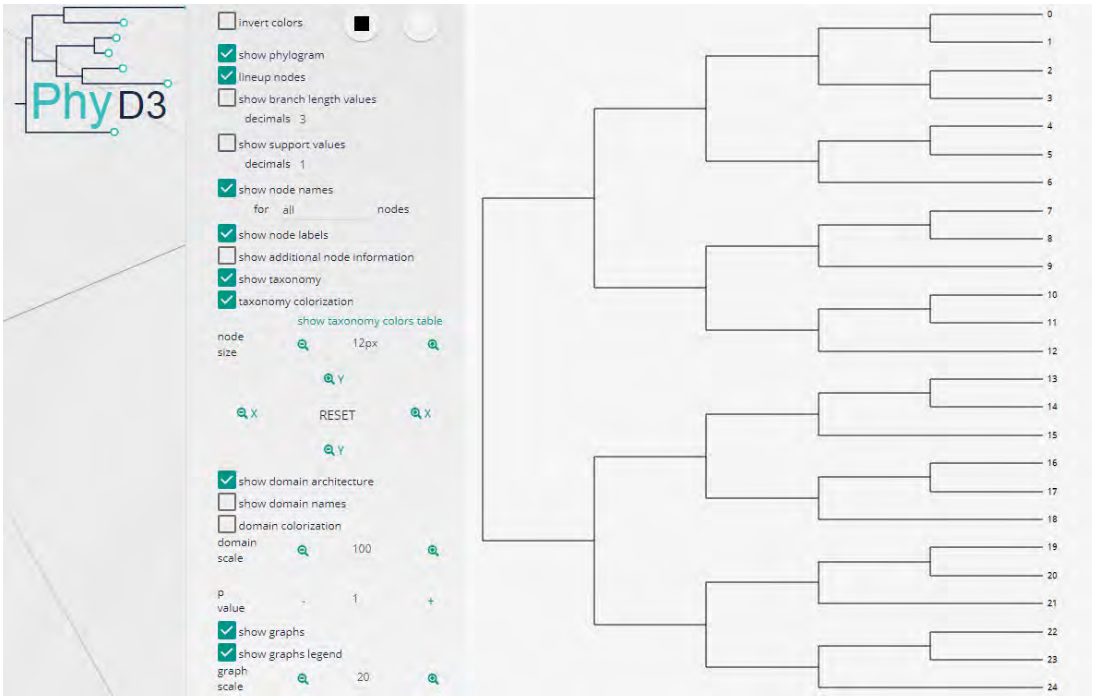
3.3.1 jsPhyloSVG

jsPhyloSVG [37] is a flexible, lightweight JavaScript library for building interactive and complex phylogenetic trees in a web-based environment with the broadest range of accessibility. It construct interactive phylogenetic trees from raw Newick or phyloXML formats directly within the browser in Scalable Vector Graphics (SVG) format. It is designed to work across all major browsers and renders an alternative format for those browsers that do not support SVG. The library provides tools for building rectangular and circular phylograms with integrated charting. Interactive features may be integrated and made to respond to events such as clicks on any element of the tree, including labels.

3.3.2 PhyD3

Mainly PhyD3 [38] is a phylogenetic tree viewer which use the JavaScript library **D3.js** [39] to visualise phylogenetic tree in the web browser. It has presented an extension of the widely used PhyloXML standard with several new options to accommodate functional genomics or annotation datasets for advanced visualization. The main target was for developing this tool was to rich open source solution using current web technologies is was not available previously. This tool can be accessed directly at https://phyd3.bits.vib.be. The screenshot of this tool is showed in Figure 3.8. Features of this online tool are mentioned below:

- It has displayed the basic information, like node and taxonomy names, branch length and support values, next to the respective nodes with detailed information.
- In this tool nodes can be annotated with structural and numerical information, displayed in the form of various graphs.
- Tree nodes are automatically hidden when they do not fit in the space between nodes to prevent overlapping of text and graphs; tree colours can be inverted for greater readability; node texts can be coloured according to node taxonomy.
- It has included extra display options to swap tree nodes within a sub-tree as well as colouring options has added.
- It provides import and export tools to facilitate greater ability.



☐ invert colors

☑ show phylogram

☑ lineup nodes

☐ show branch length values

decimals 3

☐ show support values

decimals 1

☑ show node names

for all nodes

☑ show node labels

☐ show additional node information

☑ show taxonomy

☑ taxonomy colorization

show taxonomy colors table

node size 12px

Y

X RESET X

Y

☑ show domain architecture

☐ show domain names

☐ domain colorization

domain scale 100

p value - 1 +

☑ show graphs

☑ show graphs legend

graph scale 20

Figure 3.8: The snapshot of the interface of PhyD3.

3.3.3 phylotree.js

phylotree.js [40] is a library that extends the popular data visualization framework d3.js [39], and is suitable for building JavaScript applications where users can view and interact with phylogenetic trees. It implements several abstractions in addition to features, and comes with a documented application programming interface, thus promoting interoperability and extensibility. An example application is developed to visualize and annotate phylogenetic trees. phylotree.js [40] is a useful tool and application module for a variety of computational biology software applications. This web application is developed for comparative sequence analysis, a structural viewer that interacts with a large phylogenetic tree, and an interactive tanglegram. The tool can be accessed directly at `http://phylotree.hyphy.org`. The snapshot of this tool is showed in Figure 3.9. Features of this online tool are mentioned below:

- Multiple types of selection categories are supported to facilitate comparative analysis. It include the ability to select clades, paths to the root node, individual branches, external or internal branches, and branches that are nearby on the screen.

- Also supports an algorithmic abstraction which allows developers to traverse the tree in either pre or post order and compute associated metadata as they proceed.

- It has come with a variety of features that make up common use cases. Cladogram and radial layouts are available. Support for Newick format and certain ad hoc extensions.

- It has variety of built-in features and several core abstractions, has a demonstrable ability to allow to users to select portions of a tree in a wide variety of ways and interface these selections with downstream components.

3.3.4 GGTREE

GGTREE [41] is a R package, which provides programmable visualization and annotation of phylogenetic trees. It can read different tree file formats including newick, nexus, NHX, phylip and jplace formats, and support visualization of phylo, multiphylo, phylo4, phylo4d, obkdata and phyloseq tree objects defined in other R packages. It can also extract the data from the analysis outputs of different softwares, and allows using these data to annotate the tree.

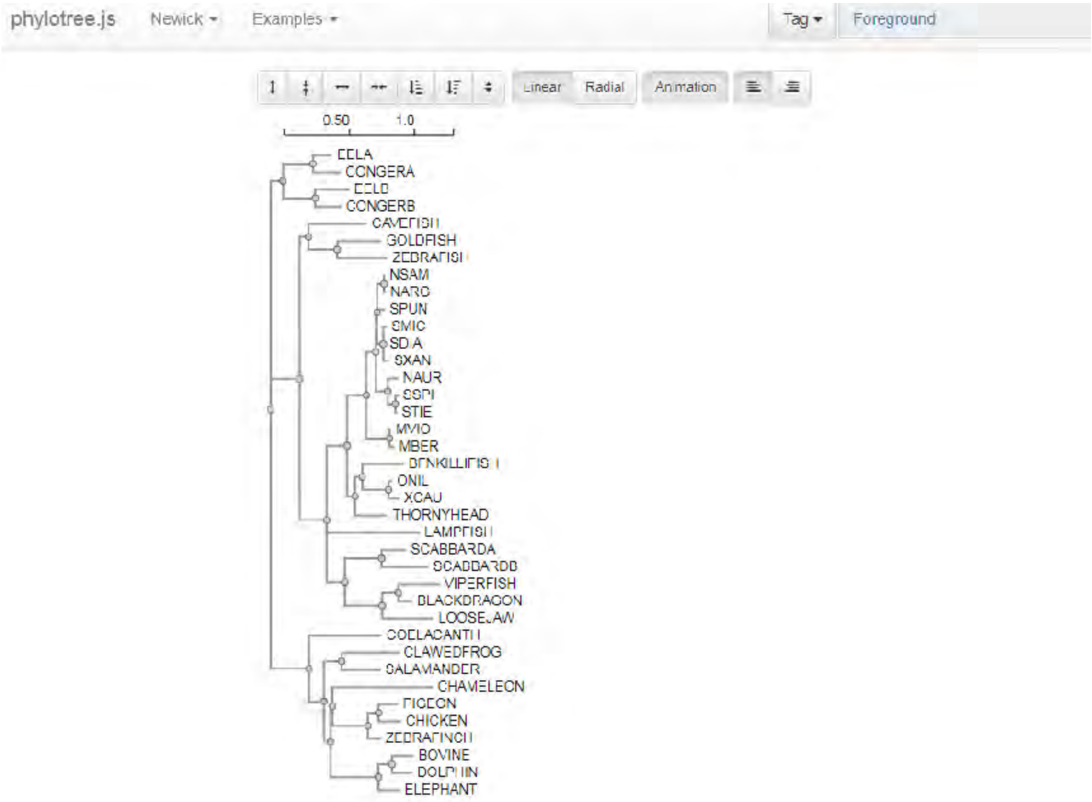


Figure 3.9: The snapshot of the interface of phylotree.js.

This package allows colouring and annotation of a tree, highlighting user selected clades or operational taxonomic units and exploration of a large tree by zooming into a selected portion. By using this package a two-dimensional tree can be drawn by scaling the tree width based on an attribute of the nodes.The features of GGTREE package are summarized below:

- This package can import evolutionary data from different tree file formats and analysis programs as well as other associated data from experiments, so that various sources and types of data can be displayed on a tree for comparison and further analyses. these increase it's interoperability.

- It can present complex phylogeny, such as two-dimensional tree and graph/image-associated trees.

- It has highly flexible graphic system, and also supports visualization of tree objects defined by other R packages so that it can be easily integrated into their packages.

Chapter 4

PhyloQon

Quartet-based phylogeny reconstruction methods have been receiving considerable attentions in last few years, such as QMC [1], QFM [2] etc. These are the most popular approaches for reconstruction of accurate phylogenetic trees. As we have already seen that there are lots of tools available both web-based and desktop-based, which can visualize phylogenetic trees. Among these existing tools, most of them are visualize the already constructed phylogenetic tree. That means the tools can take already constructed phylogenetic tree as input and visualize that tree in different format and color. There are very few tools are available which can both construct and visualize phylogenetic tree, such as Phylogeny.fr [26], T-Rex [27], ETE v3 [35] etc. But the tools those can both construct and analysis phylogenetic tree, all of them are using sequence-based phylogenetic method. No tools are there which are using quartet-based approaches. We have developed a web-based tool **PhyloQon** for construct, visualize and analysis phylogenetic trees using most accurate quartet-based methods. **PhyloQon** is a web-based tool, therefore, it does not need to be installed and is instantly accessible and usable on all modern web browsers.

4.1 Software

PhyloQon is developed by different applications like, HTML, CSS, PHP, JavaScript, jQuery. For database we have used phpMyAdmin which is a free web application that provides a convenient GUI for working with the MySQL database management system. The constructed phylogenetic trees are visualize using jsPhyloSVG [37] JavaScript visualization library. The jsPhyloSVG library takes advantage of JavaScript for phylogenetic tree display, making rich interactive figures for local display. It is a javascript library for visualizing interactive and vector-based phylogenetic trees on the web.

4.2 Quartet-based Approaches

Among different types of quartet-based approaches we implement three most accurate quartet-based approaches in our tool **PhyloQon**,

- QFM (Quartet Fiduccia and Mattheyses) [2]
- QMC (Quartet Max-Cut) [1]
- SVDquartets (Singular Value Decomposition Quartets) [3]

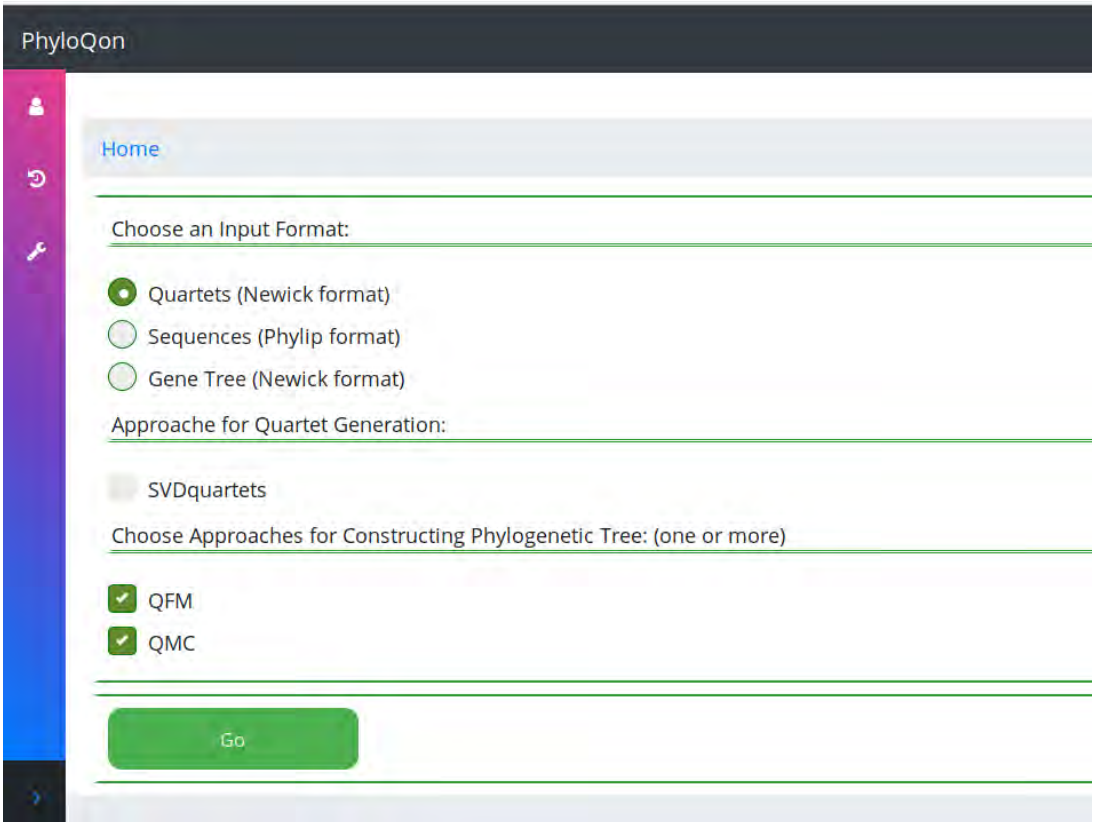
Here QFM and QMC both take set of quartets as input and construct phylogenetic tree, but the format of the quratets are different. QFM takes input quartets in Newick format and construct a unrooted phylogenetic tree as output. QFM and QMC both take quartet sets as input but in different format. To make our tool user-friendly in our developed tool both methods can take input in same format, that is Newick. SVDquartets take sequences as input in PHYLIP format, and it only generates a set of quartets as output. Then apply QFM or QMC for reconstruction the phylogenetic tree and recommended QMC for better result.

4.3 Features

We have developed a user-friendly tool **PhyloQon**, which allow users to do data analyses without software installation and to enable easy accessing of data and analyses results from any internet enabled computer. It has different features which enrich the interoperability of this tool.

4.3.1 Input Formats and Approaches

User can choose different input format for constructing phylogenetic tree, such as, quartets, sequences and gene tree. The input quartets should be in Newick format and the sequences should be in PHYLIP format. Three approaches are implemented in this tool, all are quartet-based. There are facilities to choose single or multiple approaches. A snapshot of **PhyloQon** for choosing input format and approaches are showed in Figure 4.1.



PhyloQon

Home

Choose an Input Format:

● Quartets (Newick format)
○ Sequences (Phylip format)
○ Gene Tree (Newick format)

Approache for Quartet Generation:

□ SVDquartets

Choose Approaches for Constructing Phylogenetic Tree: (one or more)

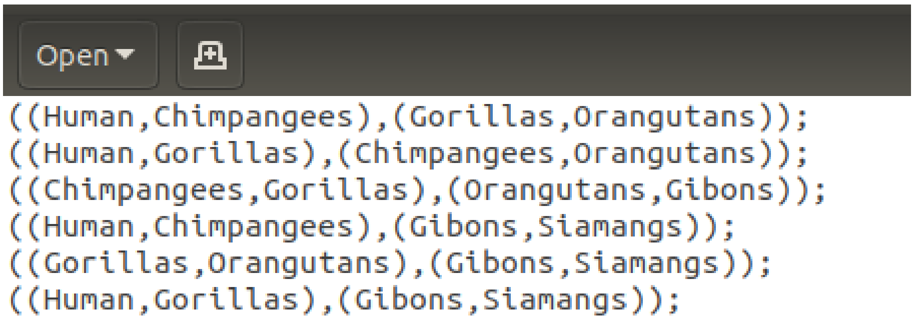
☑ QFM
☑ QMC

Figure 4.1: A snapshot to demonstrate the input format and approaches choose options.

For reconstructing phylogenetic tree, QMC [1] and QFM [2] both approaches take set of quartets as input, but in different format. **PhyloQon** facilitate with taking input quartets set in same format (that is Newick format) for both QFM and QMC approach. For SVDquartets [3] user have to take sequences as input, and the sequences should be in PHYLIP format. Developed tool has integrated with another option for taking input and that is gene tree. **PhyloQon** reconstruct a phylogenetic tree from gene tree. User can upload large sized input file. Three sample input files are shown in Figure 4.2.

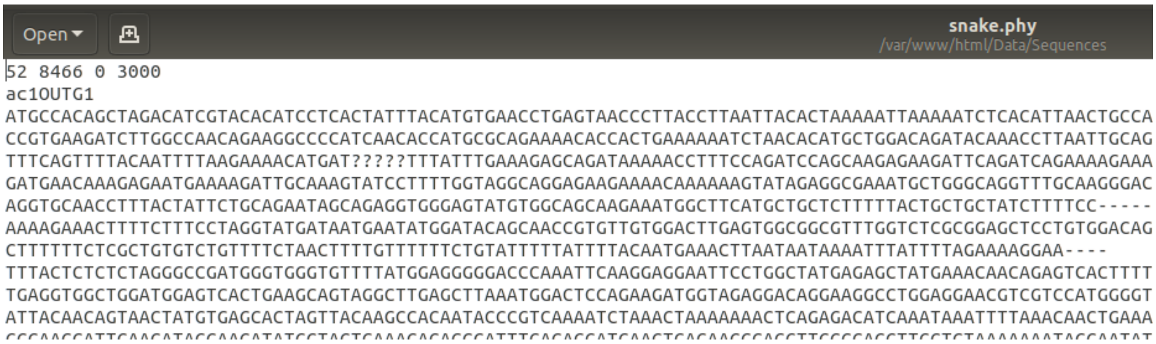
4.3.2 Comprehensive and Intuitive User-Interface

PhyloQon has a very comprehensive and intuitive user-friendly interface. For accessing this tool each user have to create an account, then login. If password forgot password, that can be retrieved easily by few easy steps. Most of the options are clickable and there are radio button



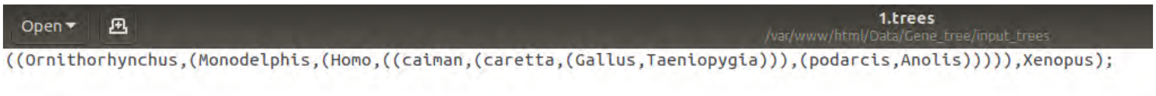
```
((Human,Chimpangees),(Gorillas,Orangutans));
((Human,Gorillas),(Chimpangees,Orangutans));
((Chimpangees,Gorillas),(Orangutans,Gibons));
((Human,Chimpangees),(Gibons,Siamangs));
((Gorillas,Orangutans),(Gibons,Siamangs));
((Human,Gorillas),(Gibons,Siamangs));
```

(a) Input set of quartets in Newick format.



```
52 8466 0 3000
ac1OUTG1
ATGCCACAGCTAGACATCGTACACATCCTCACTATTTACATGTGAACCTGAGTAACCCTTACCTTAATTACACTAAAAATTAAAAATCTCACATTAACTGCCA
CCGTGAAGATCTTGGCCAACAGAAGGCCCCATCAACACCATGCGCAGAAAACACCACTGAAAAAATCTAACACATGCTGGACAGATACAAACCTTAATTGCAG
TTTCAGTTTTACAATTTTAAGAAAACATGAT?????TTTATTTGAAAGAGCAGATAAAAACCTTTCCAGATCCAGCAAGAGAAGATTCAGATCAGAAAAGAAA
GATGAACAAAGAGAATGAAAAGATTGCAAAGTATCCTTTTGGTAGGCAGGAGAAGAAAACAAAAAAGTATAGAGGCGAAATGCTGGGCAGGTTTGCAAGGGAC
AGGTGCAACCTTTACTATTCTGCAGAATAGCAGAGGTGGGAGTATGTGGCAGCAAGAAATGGCTTCATGCTGCTCTTTTTACTGCTGCTATCTTTTCC-----
AAAAGAAACTTTTCTTTCCTAGGTATGATAATGAATATGGATACAGCAACCGTGTTGTGGACTTGAGTGGCGGCGTTTGGTCTCGCGGAGCTCCTGTGGACAG
CTTTTTTCTCGCTGTGTCTGTTTTCTAACTTTTGTTTTTTCTGTATTTTTATTTTACAATGAAACTTAATAATAAAATTTATTTTAGAAAAGGAA----
TTTACTCTCTCTAGGGCCGATGGGTGGGTGTTTTATGGAGGGGGACCCAAATTCAAGGAGGAATTCCTGGCTATGAGAGCTATGAAACAACAGAGTCACTTTT
TGAGGTGGCTGGATGGAGTCACTGAAGCAGTAGGCTTGAGCTTAAATGGACTCCAGAAGATGGTAGAGGACAGGAAGGCCTGGAGGAACGTCGTCCATGGGGT
ATTACAACAGTAACTATGTGAGCACTAGTTACAAGCCACAATACCCGTCAAAATCTAAACTAAAAAAACTCAGAGACATCAAATAAATTTTAAACAACTGAAA
CCCAACCATTCAACATACCAACATATCCTACTCAAACACACCCATTTCACACCATCAACTCACAACCCACCTTCCCCACCTTCCTCTAAAAAAATACCAATAT
```

(b) Input Sequences in PHYLIP format.



```
((Ornithorhynchus,(Monodelphis,(Homo,((caiman,(caretta,(Gallus,Taeniopygia))),(podarcis,Anolis))))),Xenopus);
```

(c) Input gene tree in Newick format.

Figure 4.2: Different sample input data format.

and checkbox for choose the input format and approaches. There are some hints and example option for user for easy to understand the developed tool. After uploading the input the user can visualize his expected phylogenetic tree by using only one click. Since the results may not be found in real time, a system must be in place to notify the user via e-mail that the result is ready for viewing. Figure 4.3 shows a snapshot of the tools where user can give input and construct the phylogenetic tree.

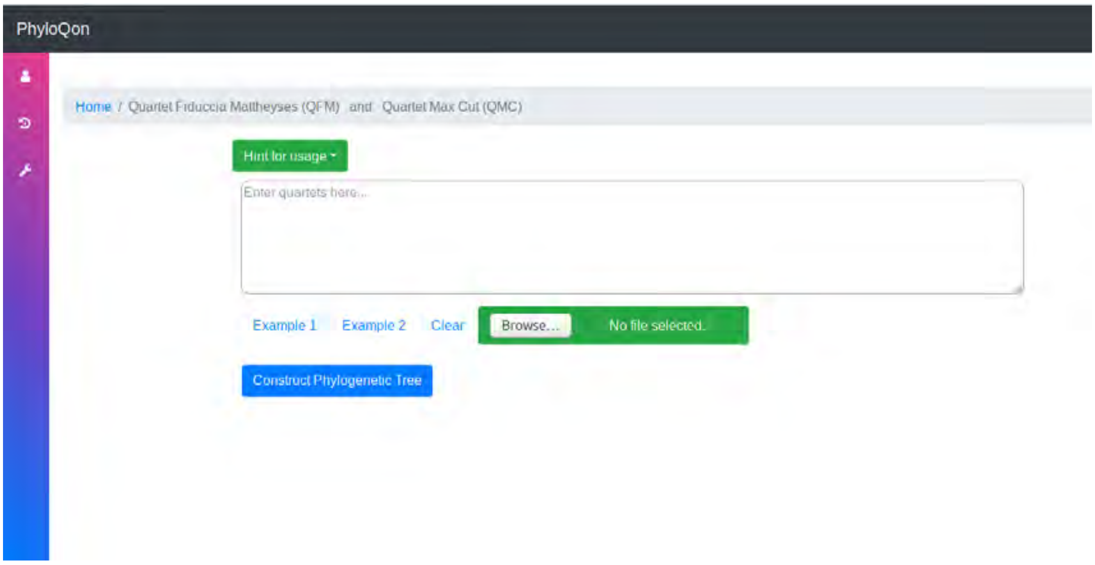
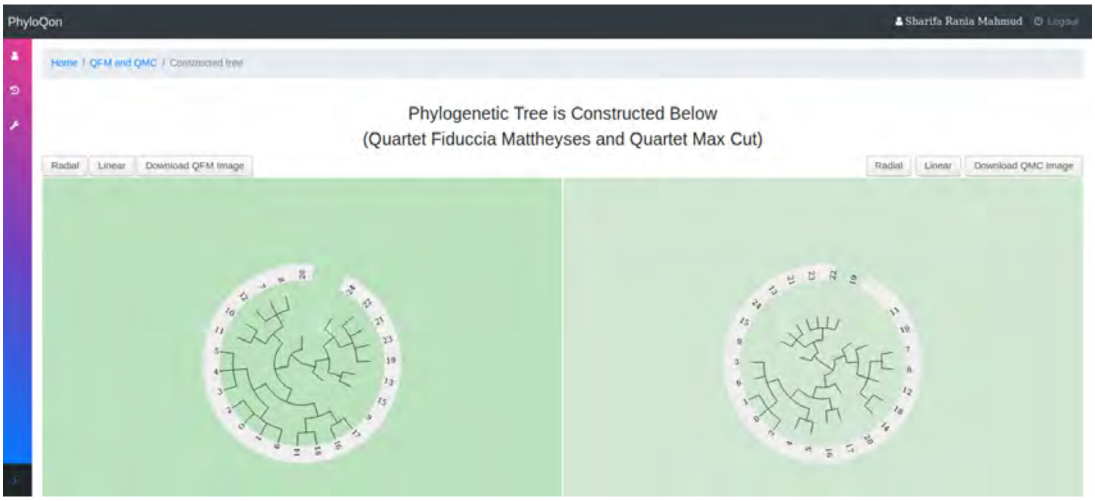

PhyloQon
Home / Quartet Fiduccia Mattheyses (QFM) and Quartet Max Cut (QMC)
Hint for usage
Enter quartets here...
Example 1
Example 2
Clear
Browse...
No file selected.
Construct Phylogenetic Tree

Figure 4.3: A snapshot of **PhyloQon**.

4.3.3 Tree Visualization

This tool can construct tree by using single and multiple approaches. The constructed phylogenetic trees are visualize using jsPhyloSVG [37] JavaScript visualization library. jsPhyloSVG is an open-source solution for rendering dynamic phylogenetic trees. It is capable of generating complex and interactive phylogenetic trees across all major browsers without the need for plugins. We have introduced **PhyloQon**, a web application to visualize and compare phylogenetic trees side-by-side. If we using multiple approaches the constructed trees are visualize side-by-side. So that user can easily see the differences and visually compare the constructed tree with different approaches. There added several tree visualization options (i.e. circular and/or rectangular pattern). User can download the visualized image in SVG format. How this tool visualize the constructed trees are shown in Figure 4.4.

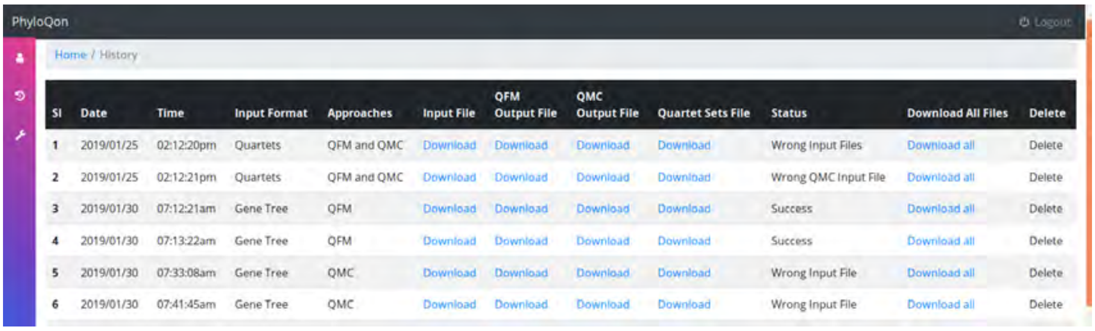


Phylogenetic Tree is Constructed Below
(Quartet Fiduccia Mattheyses and Quartet Max Cut)

Figure 4.4: A snapshot of **PhyloQon** for tree visualization.

4.3.4 Data Import and Export

User can upload input file for reconstruction of phylogenetic tree. In input files there can be set of quartets in Newick format or sequences in PHYLIP format or gene tree. The uploaded input files and also the constructed phylogenetic trees are stored in the user profile. Later if a user wants to see these information he can see these from his history and also can download the previous input and output files in text format. User can individually download the input/output files or can export compressed folder for input, output files for the same project. For SVDquartets [3] user can export the set of quartets file from history. Also can export set of quartets file, if user has gave gene tree as input. Figure 4.5 shows the screenshot of the history of a user. User can download the visualized image of the constructed trees in SVG format.



Sl	Date	Time	Input Format	Approaches	Input File	QFM Output File	QMC Output File	Quartet Sets File	Status	Download All Files	Delete
1	2019/01/25	02:12:20pm	Quartets	QFM and QMC	Download	Download	Download	Download	Wrong Input Files	Download all	Delete
2	2019/01/25	02:12:21pm	Quartets	QFM and QMC	Download	Download	Download	Download	Wrong QMC Input File	Download all	Delete
3	2019/01/30	07:12:21am	Gene Tree	QFM	Download	Download	Download	Download	Success	Download all	Delete
4	2019/01/30	07:13:22am	Gene Tree	QFM	Download	Download	Download	Download	Success	Download all	Delete
5	2019/01/30	07:33:08am	Gene Tree	QMC	Download	Download	Download	Download	Wrong Input File	Download all	Delete
6	2019/01/30	07:41:45am	Gene Tree	QMC	Download	Download	Download	Download	Wrong Input File	Download all	Delete

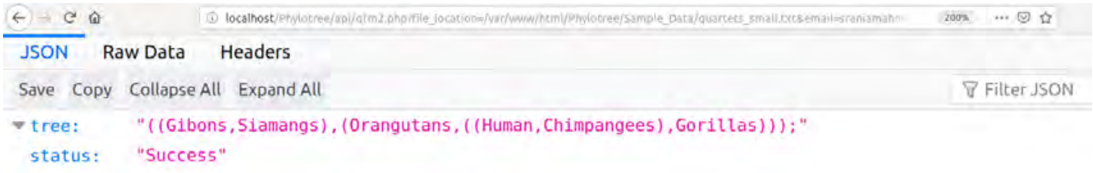
Figure 4.5: A screenshot of the history of a user.

4.3.5 Application Programming Interface (API)

We have developed an Application Programming Interface (API), which enables to get the data without using the graphical user interface of the web page. The API is convenient if a user need to programmatically access some information but still do not want to download the entire datasets. There are several scenarios when it is practical to use it. For example, user might need to access some interaction from their own scripts or want to incorporate **PhyloQon** in their web page. As the API works like normal HTTP request, user can access it like any other webpage. Just copy/paste the following URL into the browser to get the output phylogenetic tree. Put a input file location in URL or URI format in the positon $location\_of\_input\_file$ and give a registered email ID in $registered\_email\_ID$ in the following URL.

```
http://localhost/Phylotree/api/[link_for_different_approaches]?file_
location=[location_of_input_file]&email=[registered_email_ID]
```

User must have an registered user account in the PhyloQon web tool before using it's different approaches. There are several quartet-based phylogenetic (re)construction approaches available in **PhyloQon** API are listed in Table 4.1. Figure 4.6 shows a snapshot after using the API URL in browser.


localhost/Phylotree/api/qfm2.php?file_location=/var/www/html/Phylotree/Sample_Data/quartets_small.txt&email=sraniamah
200%
JSON
Raw Data
Headers
Save
Copy
Collapse All
Expand All
Filter JSON

```
▾ tree:     "((Gibons,Siamangs),(Orangutans,((Human,Chimpangees),Gorillas)));"
  status:   "Success"
```

Figure 4.6: snapshot after using the API URL in browser.

We have the developed tool thoroughly and the beta version of the tool will be released to limited audience for testing. The final version of the tool will be released and experiments on different approaches will be carried out.

Table 4.1: Different approaches available in **PhyloQon** API.

Approaches	**Input Type**	**API method URL**	**Description**
QFM	Quartets	`/api/qfm2.php?`	Take a set of quartets (Newick) as input and (re)contruct phylogenetic tree using QFM approach
QMC	Quartets	`/api/qmc2.php?`	Take a set of quartets (Newick) as input and (re)contruct phylogenetic tree using QMC approach
QFM and QMC	Quartets	`/api/QFM_QMC.php?`	Take a set of quartets (Newick) as input and (re)contruct phylogenetic tree using both QFM and QMC approaches
SVDquartets with QFM	Sequences	`/api/SVDQ_QFM.php?`	Take a sequences (PHYLIP) as input and (re)contruct phylogenetic tree using SVDquartets with QFM approach
SVDquartets with QMC	Sequences	`/api/SVDQ_QMC.php?`	Take a sequences (PHYLIP) as input and (re)contruct phylogenetic tree using SVDquartets with QMC approach
SVDquartets with QFM and QMC	Sequences	`/api/SVDQ_QFM_QMC.php?`	Take a sequences (PHYLIP) as input and (re)contruct phylogenetic tree using SVDquartets with both QFM and QMC approaches
QFM	Gene tree	`/api/geneTree_QFM.php?`	Take a gene tree (Newick) as input and (re)contruct phylogenetic tree using QFM approach
QMC	Gene tree	`/api/geneTree_QFM.php?`	Take a gene tree (Newick) as input and (re)contruct phylogenetic tree using QMC approach
QFM and QMC	Gene tree	`/api/geneTree_QFM.php?`	Take a gene tree (Newick) as input and (re)contruct phylogenetic tree using both QFM and QMC approaches

Chapter 5

Experimental Analysis

Phylogenetic tree is a formal structure of the representation of process of evolution. It is important to be able to compare trees obtained from the observed real data because different fitting methods may provide different trees even for the same initial dataset. When two datasets are considered, the knowledge of the comparative indices, like the Robinson and Foulds (RF) distance [28] can give us some ideas about the similarity or dissimilarity of the evolutionary processes corresponding to these data. In this chapter we have compared the topological accuracy of two approaches, QFM and QMC by using our developed tool **PhyloQon**.

Here, we compare the RF distance between QFM and QMC. We ignore the comparison with SVDquartets, because, this approach generate only the set quartets and then apply QFM or QMC to reconstruct phylogenetic tree. We also compare the running time of QFM and QMC.

5.1 Robinson and Foulds Distance

The reconstructed tree was compared using the RF symmetric difference distance to the model tree. The RF distance between two trees is computed by the removal of an edge in a phylogenetic tree partitions the tree into two subtrees. The removed edge by the partition it induces over the taxa set. The RF count between two trees is the number of edges (partitions) that exist in exactly one tree. Since all external edges are always present in every tree, the difference is obtained only in internal edges and therefore count only internal edges.

5.2 Datasets

For inspecting the performance of QFM and QMC, we have generated quartet sets, taken uniformly at random from model trees, by varying the number of taxa (n) and the number of quartets (q). The consistency (c) of these quartet sets are 100%. We have generated model species trees with n=25, 50, 100, 200 and 300 taxa. For, n = 25, 50, 100, we have generated $n^{1.5}$, n^2 and $n^{2.8}$ quartets. We have not generated more quartets because $n^{2.8}$ quartets have been empirically shown to be enough to construct very accurate phylogenetic trees, proved by [1]. For $n^{2.8}$ we can see almost perfect reconstruction of the model tree. Although $n^{1.5}$ is a small number, we have chosen this size to test the performance of both methods on a comparatively smaller number of quartets as well. For 200 and 300 taxon model trees, we have generated datasets with $q = n^{1.5}$ and $q = n^2$. For each size (q), we have evaluated QFM and QMC on the noise-free model conditions. Here, all the quartets are accurate and we report the average RF distance over these datasets.

5.3 Computational Analyses

We have calculated the RF distance by the tool developed in [27]. This tool computes the RF topological distance, which is a well-known measure of the tree similarity, between the first tree and all the following trees specified by the user. The trees can be supplied in the Newick or Distance matrix formats. Here, we use the Newick format. The tool used an optimal algorithm described in [42] to compute the RF metric. In our calculation, at first we reconstruct all the phylogenetic trees for each cases. Then we take the model tree and the reconstructed phylogenetic tree to measure the RF topological distance for each cases. Table 5.1 demonstrates the results under the parameters (n,q) with c = 100%. It shows the results of this experiment under an Intel Quad core machine of 8GB RAM with 64-bit operating system. RF distance and running time in minutes has been compared. For RF distance, among the 11 model conditions are analyzed, QFM has been found to be better than QMC, and the improvements are statistically significant. QMC is better than QFM in only few cases but the differences are not statistically significant. There are also some cases where QFM and QMC have identical accuracy. When we consider the running time, QMC is faster than QFM. Figure 5.1 showed the RF distance for each method in a line chart, where x-axis define the $n - q$ value and y-axis define the RF distance.

Table 5.1: Comparison of QFM and QMC. RF distance and running time in minutes are compared.

n	q	RF distance		Running time (min)	
		QFM	QMC	QFM	QMC
25	125	**16**	**20**	0.04	0.01
25	625	**21**	**23**	0.04	0.01
25	8208	**0**	**1**	0.12	0.01
50	354	**57**	**60**	0.01	0.01
50	2500	15	11	0.11	0.02
50	57164	**0**	**8**	6.28	0.02
100	1000	**102**	**119**	0.10	0.02
100	10000	49	46	0.52	0.02
100	398108	89	71	200	0.03
200	40000	**144**	**155**	10.40	0.07
300	5197	**329**	**380**	3.35	0.09

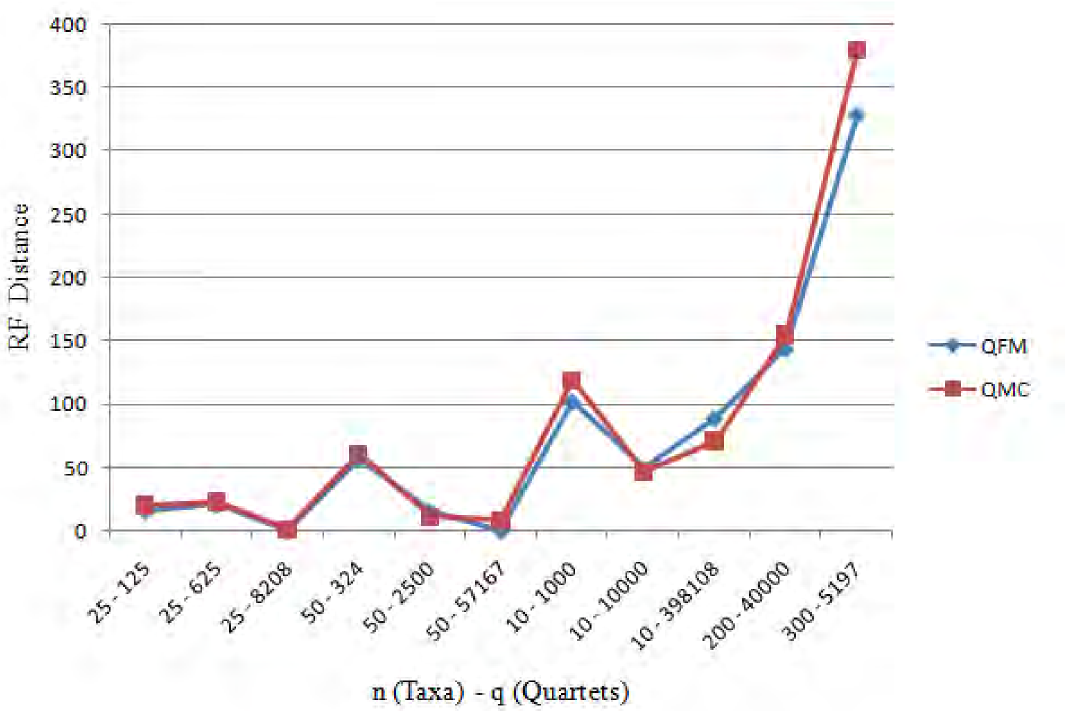


Figure 5.1: RF distance for QFM and QMC approaches.

Chapter 6

Conclusions and Future Work

Phylogenetic tree (re)constructio based visualization tools has became very essential and most-wanted subjects for biological research. In this thesis, we introduced a user-friendly tool **PhyloQon**, which is a comprehensive web application for constructing, visualizing, comparing and analysing phylogenetic trees. We incorporate QFM, QMC and SVDquartets approaches to reconstruct the trees, where QFM, QMC both take set of quartets as input in Newick format and SVDquartets take sequences as input in PHYLIP format for constructing phylogenetic tree. The constructed trees have several tree visualization options (i.e. circular and/or rectangular pattern). The developed tool has integrated with another option for taking input is gene tree for constructing phylogenetic tree. In our tool we have facility to reconstruct trees with single or multiple approaches and also view them side-by-side. It will help the users to visually compare the trees graphically, which are constructed using different approaches with same input datasets. User can also import and export data and graphical view of the constructed trees. We have developed an API of this tool, which enables to get the data without using the graphical user interface of the web page. The developed tool thoroughly and the beta version of the tool will be released to limited audience for testing. The final version of the tool will be released and experiments on different approaches will be carried out.

We have started with an introductory overview on phylogeny in Chapter 1. In that chapter we have discussed various practical applications and limitations of phylogeny. We also gave our motivation, described our objective, main contribution and scope of this project.

In Chapter 2 we have described an overview of some phylogenetic tree reconstruction methods. We have discussed on quartet-based accurate phylogenetic tree reconstruction approaches. Introduced some basic concepts and terminology related to phylogeny. In particular we focused on the concepts related to quartet based phylogeny.

In Chapter 3 we gave the literature review of some existing popular web-based and desktop-based visualization tools for phylogeny analysis.

In Chapter 4 we have presented our developed tool **PhyloQon**. We have discussed the uses, features and facilities of the tool. In Chapter 5 we have done experimental analysis. We have showed the comparative analysis of QFM and QMC methods with same datasets.

6.1 Future Directions

We will try to improve the user interface and better performance. In the future, we aim to keep **PhyloQon** up-to-date and state-of-the-art by supporting new annotation types that meet the developing needs of our broad range of user.

Bibliography

[1] S. Snir and S. Rao, "Quartet maxcut: a fast algorithm for amalgamating quartet trees," *Molecular phylogenetics and evolution*, vol. 62, no. 1, pp. 1–8, 2012.

[2] R. Reaz, M. Bayzid, and M. Rahman, "Accurate phylogenetic tree reconstruction from quartets: A heuristic approach," *PLoS ONE*, vol. 9, no. 8, p. 104008, 2014.

[3] J. Chifman and L. Kubatko, "Quartet inference from snp data under the coalescent model," *BIOINFORMATICS*, pp. 1–8, 2014.

[4] C. Darwin, *On the origin of species, 1859.* Routledge, 2004.

[5] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees.," *Molecular biology and evolution*, vol. 4, no. 4, pp. 406–425, 1987.

[6] J. Felsenstein, "Evolutionary trees from dna sequences: a maximum likelihood approach," *Journal of molecular evolution*, vol. 17, no. 6, pp. 368–376, 1981.

[7] W. M. Fitch, "Toward defining the course of evolution: minimum change for a specific tree topology," *Systematic Biology*, vol. 20, no. 4, pp. 406–416, 1971.

[8] B. R. Baum and M. A. Ragan, "The mrp method," in *Phylogenetic supertrees*, pp. 17–34, Springer, 2004.

[9] M. A. Ragan, "Matrix representation in reconstructing phylogenetic relationships among the eukaryotes," *Biosystems*, vol. 28, no. 1-3, pp. 47–55, 1992.

[10] K. Strimmer and A. Von Haeseler, "Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies," *Molecular Biology and Evolution*, vol. 13, no. 7, pp. 964–969, 1996.

[11] L. Xin, B. Ma, and K. Zhang, "A new quartet approach for reconstructing phylogenetic trees: Quartet joining method," in *International Computing and Combinatorics Conference*, pp. 40–50, Springer, 2007.

[12] S. Snir, T. Warnow, and S. Rao, "Short quartet puzzling: A new quartet-based phylogeny reconstruction algorithm," *Journal of Computational Biology*, vol. 15, no. 1, pp. 91–103, 2008.

[13] S. Snir and S. Rao, "Quartets maxcut: a divide and conquer quartets algorithm," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 7, no. 4, pp. 704–718, 2010.

[14] V. Ranwez and O. Gascuel, "Quartet-based phylogenetic inference: improvements and limits," *Molecular biology and evolution*, vol. 18, no. 6, pp. 1103–1116, 2001.

[15] M. S. Swenson, R. Suri, C. R. Linder, and T. Warnow, "An experimental study of quartets maxcut and other supertree methods," *Algorithms for Molecular Biology*, vol. 6, no. 1, p. 7, 2011.

[16] C. M. Fiduccia and R. M. Mattheyses, "A linear-time heuristic for improving network partitions," in *19th Design Automation Conference*, pp. 175–181, IEEE, 1982.

[17] J. Chifman and L. Kubatko, "Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites," *Journal of theoretical biology*, vol. 374, pp. 35–47, 2015.

[18] D. L. Swofford, "Paup*: Phylogenetic analysis using parsimony (and other methods) 4.0. b5," 2001.

[19] P. Vachaspati and T. Warnow, "Svdquest: Improving svdquartets species tree estimation using exact optimization within a constrained search space," *Molecular phylogenetics and evolution*, vol. 124, pp. 122–136, 2018.

[20] Z. He, H. Zhang, S. Gao, M. J. Lercher, W.-H. Chen, and S. Hu, "Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees," *Nucleic acids research*, vol. 44, no. 1, pp. 236–241, 2016.

[21] B. Ribeiro-Gonçalves, A. P. Francisco, C. Vaz, M. Ramirez, and J. A. Carriço, "Phyloviz online: web-based tool for visualization, phylogenetic inference, analysis and sharing of minimum spanning trees," *Nucleic acids research*, vol. 44, no. W1, pp. W246–W251, 2016.

[22] T. G. Vaughan, "Icytree: rapid browser-based visualization for phylogenetic trees and networks," *Bioinformatics*, vol. 33, no. 15, pp. 2392–2394, 2017.

[23] B. Cazaux, G. Castel, and E. Rivals, "Aquapony: visualization and interpretation of phylogeographic information on phylogenetic trees," 2017.

[24] A. J. Drummond, M. A. Suchard, D. Xie, and A. Rambaut, "Bayesian phylogenetics with beauti and the beast 1.7," *Molecular biology and evolution*, vol. 29, no. 8, pp. 1969–1973, 2012.

[25] O. Robinson, D. Dylus, and C. Dessimoz, "Phylo. io: interactive viewing and comparison of large phylogenetic trees on the web," *Molecular biology and evolution*, vol. 33, no. 8, pp. 2163–2166, 2016.

[26] A. Dereeper, V. Guignon, G. Blanc, S. Audic, S. Buffet, F. Chevenet, J.-F. Dufayard, S. Guindon, V. Lefort, M. Lescot, *et al.*, "Phylogeny. fr: robust phylogenetic analysis for the non-specialist," *Nucleic acids research*, vol. 36, no. 2, pp. 465–469, 2008.

[27] A. Boc, A. B. Diallo, and V. Makarenkov, "T-rex: a web server for inferring, validating and visualizing phylogenetic trees and networks," *Nucleic acids research*, vol. 40, no. 1, pp. 573–579, 2012.

[28] D. F. Robinson and L. R. Foulds, "Comparison of phylogenetic trees," *Mathematical biosciences*, vol. 53, no. 1-2, pp. 131–147, 1981.

[29] R. A. Vos, J. Caravas, K. Hartmann, M. A. Jensen, and C. Miller, "Bio:: Phylo-phyloinformatic analysis using perl," *BMC bioinformatics*, vol. 12, no. 1, p. 63, 2011.

[30] B. C. Stöver and K. F. Müller, "Treegraph 2: combining and visualizing evidence from different phylogenetic analyses," *BMC bioinformatics*, vol. 11, no. 1, p. 7, 2010.

[31] M. Nascimento, A. Sousa, M. Ramirez, A. P. Francisco, J. A. Carriço, and C. Vaz, "Phyloviz 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods," *Bioinformatics*, vol. 33, no. 1, pp. 128–129, 2016.

[32] D. H. Huson and C. Scornavacca, "Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks," *Systematic biology*, vol. 61, no. 6, pp. 1061–1067, 2012.

[33] C. Scornavacca, F. Zickmann, and D. H. Huson, "Tanglegrams for rooted phylogenetic trees and networks," *Bioinformatics*, vol. 27, no. 13, pp. 248–256, 2011.

[34] p. Molecular evolution and epidemiology, "FigTree." http://tree.bio.ed.ac.uk/software/figtree/. [Online; accessed 08-February-2019].

[35] J. Huerta-Cepas, F. Serra, and P. Bork, "Ete 3: reconstruction, analysis, and visualization of phylogenomic data," *Molecular biology and evolution*, vol. 33, no. 6, pp. 1635–1638, 2016.

[36] Wikipedia, "List of phylogenetic tree visualization software." https://en.wikipedia.org/wiki/List_of_phylogenetic_tree_visualization_software. [Online; accessed 05-February-2019].

[37] S. A. Smits and C. C. Ouverney, "jsphylosvg: a javascript library for visualizing interactive and vector-based phylogenetic trees on the web," *PloS one*, vol. 5, no. 8, p. 12267, 2010.

[38] Ł. Kreft, A. Botzki, F. Coppens, K. Vandepoele, and M. Van Bel, "Phyd3: a phylogenetic tree viewer with extended phyloxml support for functional genomics data visualization," *Bioinformatics*, vol. 33, no. 18, pp. 2946–2947, 2017.

[39] M. Bostock, V. Ogievetsky, and J. Heer, "D^3 data-driven documents," *IEEE Transactions on Visualization & Computer Graphics*, no. 12, pp. 2301–2309, 2011.

[40] S. D. Shank, S. Weaver, and S. L. K. Pond, "phylotree. js-a javascript library for application development and interactive data visualization in phylogenetics," *BMC bioinformatics*, vol. 19, no. 1, p. 276, 2018.

[41] G. Yu, D. K. Smith, H. Zhu, Y. Guan, and T. T.-Y. Lam, “ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data,” *Methods in Ecology and Evolution*, vol. 8, no. 1, pp. 28–36, 2017.

[42] V. Makarenkov and B. Leclerc, “Comparison of additive trees using circular orders,” *Journal of Computational Biology*, vol. 7, no. 5, pp. 731–744, 2000.