In Silico Phylogenetic Analysis and Estimation of Divergence Time for Ebola Virus using Bayesian Inference

Manish Kumar Sinha
Department of Mathematics, Bioinformatics and Computer Applications
Maulana Azad National Institute of Technology, Bhopal, Madhya Pradesh, India

Usha Chouhan
Department of Mathematics, Bioinformatics and Computer Applications
Maulana Azad National Institute of Technology, Bhopal, Madhya Pradesh, India

Abstract - Ebola virus (EBOV, formerly designated Zaire *Ebolavirus*) is one of five known viruses within the genus *Ebolavirus*. Four of the five known *ebolaviruses*, including EBOV cause a severe and often fatal hemorrhagic fever in humans and other mammals, known as Ebola virus disease (EVD). Ebola virus has caused the majority of human deaths from EVD, and is the cause of the 2013–2015 Ebola virus epidemics in West Africa, which has resulted in at least 26,969 suspected cases and 11,135 confirmed deaths. The cure for EVD is still far from development for humans and vaccines are still not properly developed so far. Phylogenetic analysis of the nucleotide sequence can reveal crucial information and knowledge for understanding virus evolution, geographic distribution and host specificity of viruses. In this paper attempt to perform the phylogenetic analysis for the 47 different strains of *Ebolavirus* that outbreak during the centuries. By the Bayesians analysis for the 47 strains phylogenetic tree were constructed for the epidemic outbreak in Guinea. *Ebolavirus* caused many numbers of deaths recently in the year 2014 and caused hemorrhagic fever in human. The strains of the sequences that were found in Guinea formed a separate divergent lineage. Based on our analysis the strains that were evolved in Guinea were not a new strain but the ancestry behind that were the Zaire virus from the Democratic republic of Congo.

Keywords – BEAST, MEGA, NCBI, Ebolavirus, ESS.

I. INTRODUCTION

Ebola virus till date known is very dangerous and it belongs to member of the virus filoviridae which is a RNA virus and belongs to viral family [1]. The characteristic of this virus is long and thin filamentous. This virus was first discovered near Ebola River and hence named on the name of the river. The virus family filoviridae is the only known family which is till date ignored and not given any of that importance until the epidemic caused in the African country. So the thing to be worried for is to know the way in which these things can be kept under control by natural being, and information about these disease is less known and their pathogenesis, and viriological information. This information were cumulated when the epidemic outbreak and studied which was helpful in knowing in detail about the filovirous [2].

Filoviridae virus when comes in contact with human cause's very severe and deadly disease. It causes Hemorrhagic fever. This virus damages the epithelial cell that coat of the inner blood vessel and this would lead in to non coagulation of blood in the diseased individual. The platelets would not be able to coagulate due to the rupture caused by the virus; and the person can go under certain shock as their organ might not work properly as the heart would not be able to pump the requisite amount of blood, or there might be a lowering in the blood pressure. According to the study of epidemic statistics the virus causes 90% of death who are getting once infected through this. After a human being gets infected to this there is very high chance of transfer of this to other human being [3].

II. PROPOSED ALGORITHM

We used different tools such as National Centre for Biotechnology Information (NCBI), Clustal W, MEGA (Molecular Evolution Genetic Analysis), BEAST software (Bayesian Evolutionary Analysis Sampling Trees), Tracer, Fig Tree.

Every scientific research has some protocol mine is not an exception. For our study we downloaded the sequences from the NCBI database and performed different calulation on our retrieved sequences using different tools and softwares[4]. The ouput of the work is summarized and hence used as input to the different tool. The following figure1 shows the procedure for the workflow .

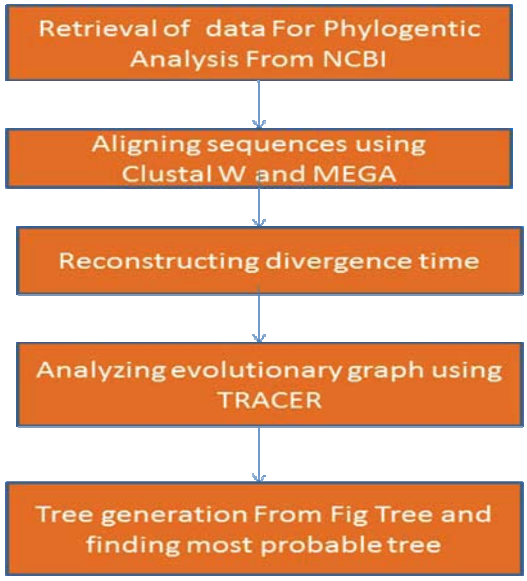

Figure 1. Phylogenetic tree generation Flow Diagram.

i. Retrieving Sequences from the NCBI Database

We have used National Centre for Biotechnology Information (NCBI) database for exploring and downloading sequences of the different strains of species. National Centre for Biotechnology Information (NCBI) is a public repository which collect and without any charge servers free information regarding genomic sequences. The URL for downloading the sequences is htpp://www.ncbi.nic.nih.gov/.

ii. Aligning sequences using MEGA

The alignment explorer is the tool for building and editing multiple sequence alignment in MEGA. MEGA also supports a wide collection of models for the estimating evolutionary distances and we can select any of them as per our requirement. So compared hear evolutionary distance calculated by different models for our study and analysis [12]. Some of the other known example models are, Dayhoff Model, Jones- Taylor- Thornton (JTT) Model, Poisson Model and equal Input Model.

iii. Reconstructing Divergence Time using BEAST (Bayesian Evolutionary Analysis Sampling Trees)

BEAST uses a complex and powerful input format (specified in XML) to describe the evolutionary model. This has advantages in terms of flexibility in that the developers of BEAST do not have to try and predict every analysis that researchers may wish to perform and explicitly provide an option for doing it. However, this flexibility means it is possible to construct models that don't perform well under the Markov chain Monte Carlo (MCMC) [5] inference framework used [6]. We cannot test every possible model that can be used in BEAST.

There are two solutions to this: Firstly, we supply a range of recipes for commonly performed analyses that we know should work in BEAST and provide input files for these (although, the actual data can also produce unexpected behavior)[10]. Secondly, we provide advice and tools for the diagnosis of problems and suggestions on how to fix them: http://beast2.cs.auckland.ac.nz/.BEAST is not a black-box into which you can put your data and expect an easily interpretable answer [13]. It requires careful inspection of the output to check that it has performed correctly and usually will need tweaking, adjustment and a number of runs to get a valid answer [7].

BEAUti is a graphical user-interface (GUI) application for generating BEAST XML files. This application provides a clear way to specify priors, partition data and calibrating the internal nodes [8].

iv. Analyzing Evolutionary Graph Using Tracer

Tracer is graphical tool for visualization and diagnostics of MCMC output. It can read output files from MrBayes and BEAST. The standard deviation and the mean is calculated. It takes into account the effective sample (ESS) size so a small ESS will give a large Stdev[9].

Median: The median value of the sampled trace across the chain (excluding the burn-in).

95% HPD Lower: The lower bound of the highest posterior density (HPD) interval. The HPD is a credible set that contains 95% of the sampled values.

95% HPD Upper: The upper bound of the highest posterior density (HPD) interval. The HPD is a credible set that contains 95% of the sampled values.

v. Tree Generation From FigTree And Finding the most probable Tree

FigTree is a graphical user-interface (GUI) application for viewing phylogenies and producing publication quality figures. It has many features. It provides Cross-platform graphical tree display. Three different tree styles: rectangular, polar and radial trees were obtained. It displays node heights, branch lengths, support values and other annotations. Node height range bars if available and collapses clades into triangles. For clear visualization coloring of branches and tip labels FigTree provides easy way out.

III. Experiment and Result

We have generated alignments of forty seven nucleotide sequences, in the rooted tree [11]. We have included MCMC, Bayesian inference and included constant (molecular clock), auto correlated, and uncorrelated rates.

BEAUti which is a graphical user-interface (GUI) application for generating the BEAST XML files. The XML file was taken as an input to the BEAST and the log output for the states were obtained. We run the program for longer time so that the number of states is more to get to a desired result.

Divergence times for all branching points in the topology were calculated with the RelTime method using the branch lengths contained in the inferred tree. From the time tree we infer that Zaire and Sudan Diverge 1200-1900 years ago. The Sudan and Reston diverged 1000- 2000 years ago. And Bundibugyo and cote diverged around 900- 1700 years ago. So we conclude that the strains of Ebola virus diverged several hundreds of years ago which is around 1000 to 2000 year ago.

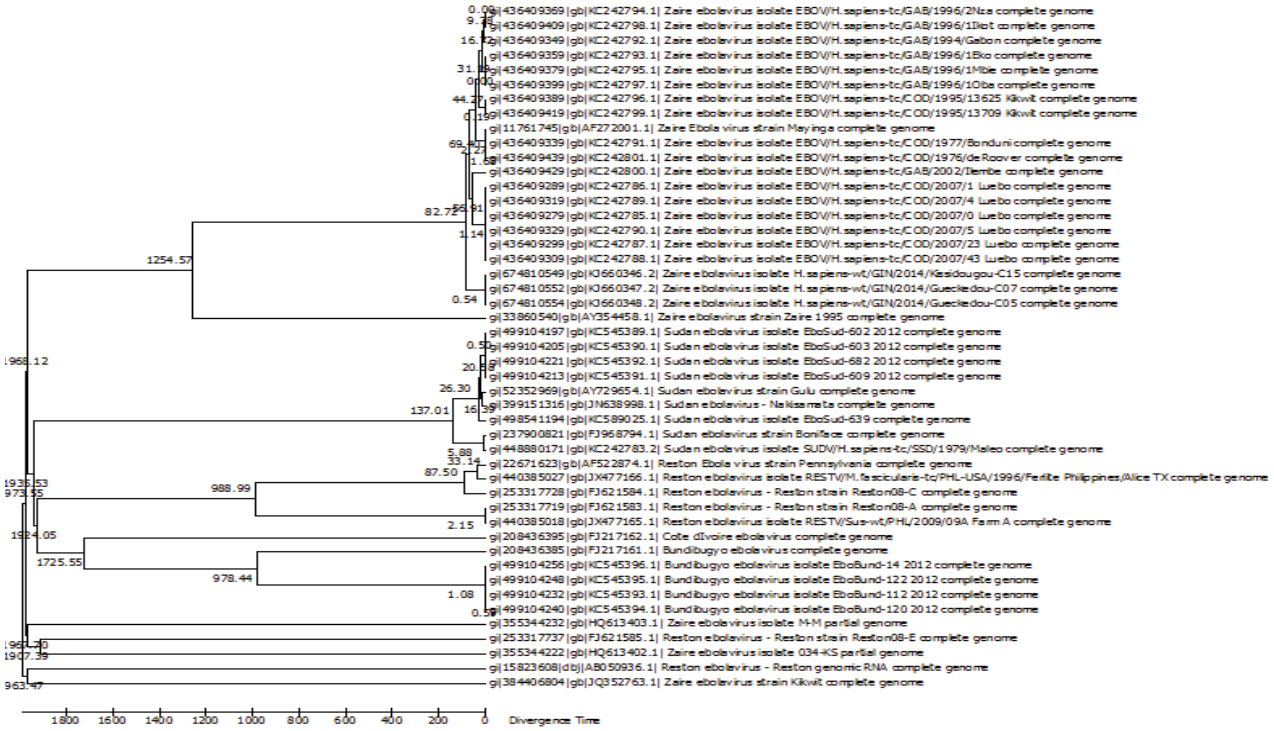

Figure 2. Time tree for Estimating Divergence Time

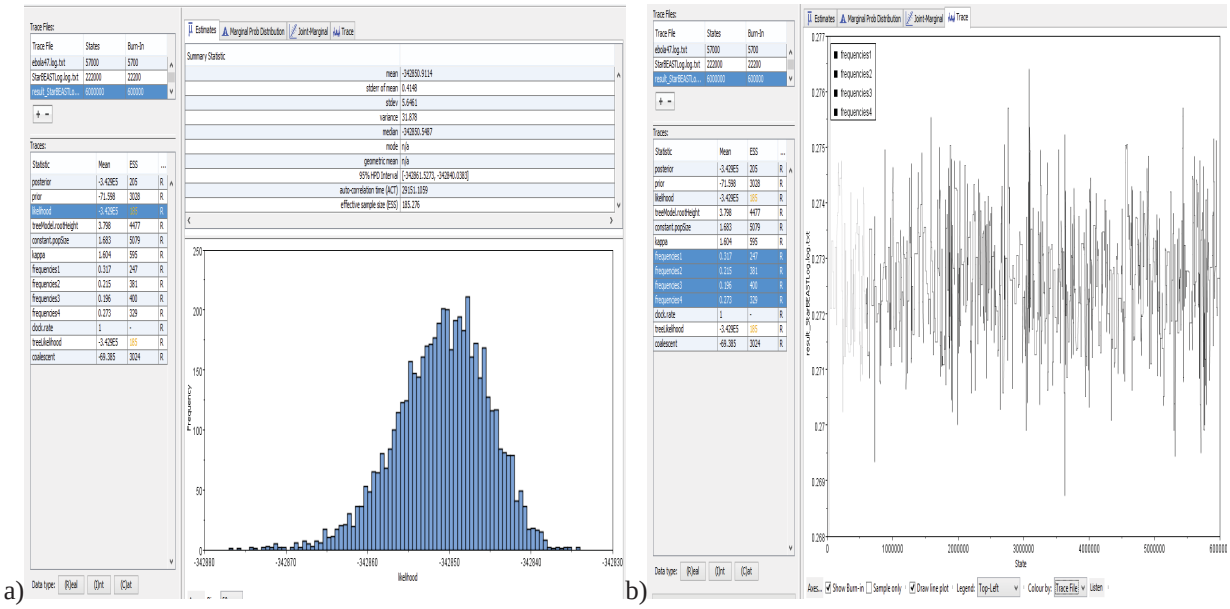

Figure 3. a) Likelihood Estimate for the increased estimates b) Tracer results showing the required frequency with less variation

If the ESS of a parameter is small then the estimate of the posterior distribution of that parameter will be poor. In Tracer we calculated the standard deviation of the estimated mean of a parameter. If the ESS is small then the standard deviation will be large. This is exactly the same as the sample size of an experiment consisting of measurements. The mean estimate is 342850 years ago with a 95% HPD of (342861, 342840). We have standard deviation value 5.6, in the figure 4 which is reduced from the value of nearly 350. The figure 3 thus shows the required frequency with less variation.

The tree obtained from Bayesians was refined using BEAST software and TreeAnnotator. The tree was graphically visualized using FigTree tool.

The BEAST uses the Bayesian Inference which is considered to be somewhat better than the traditional methods. So we can assume that this phylogenetic tree generated by using BEAST is more accurate than the trees generated using traditional methods.

The figure 4 shows the best phylogenetic tree obtained from after the Bayesian inference which included the molecular clock. Five different colors represent strains for five different specie of the EBOV which come from different regions of the African country.

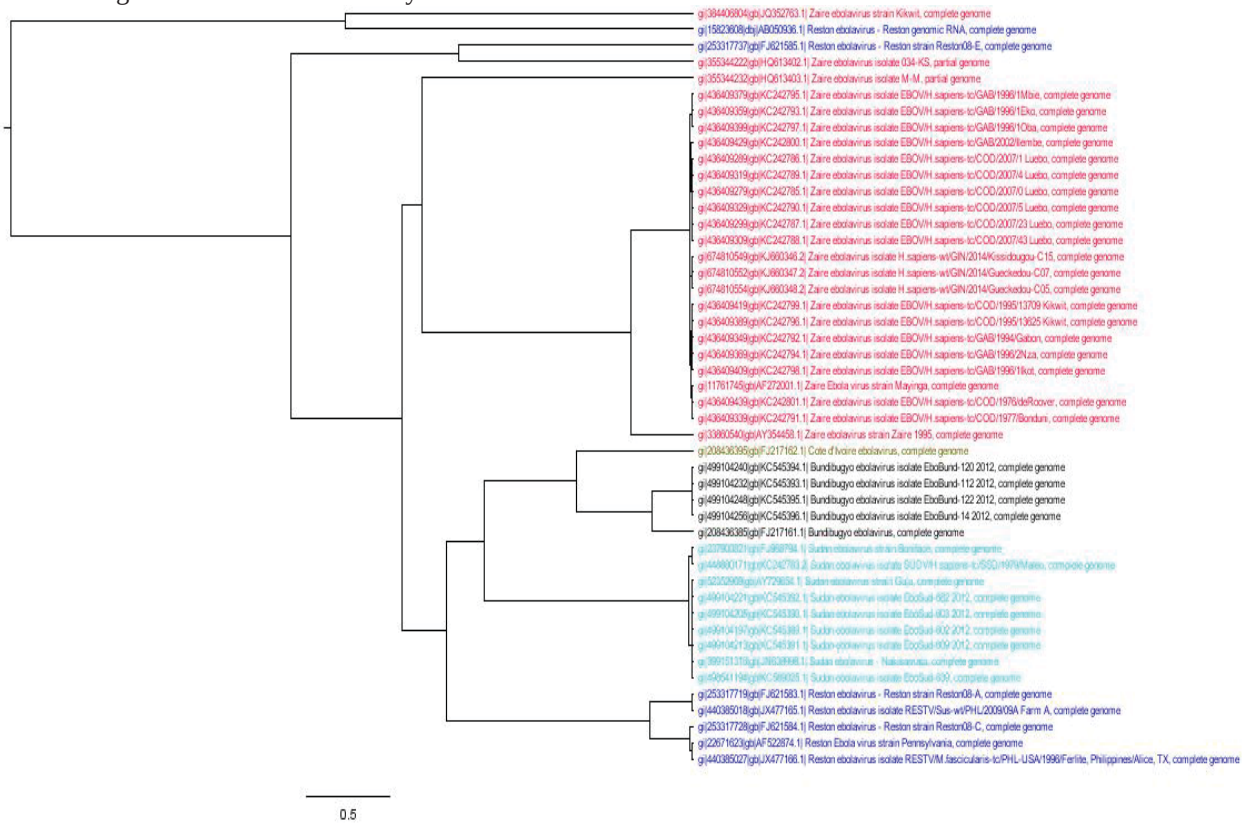

Figure 4. Phylogenetic Tree obtained from BEAST

The two strains with the close relation with Guinea out break are gi|15823608|dbj|AB050936.1| Reston ebolavirus and gi|253317737|gb|FJ621585.1| Reston ebolavirus which is seen in the phylogenetic tree in figure 5. These are represented in blue color and present at top of the tree. The clear view is seen in the map of Africa.

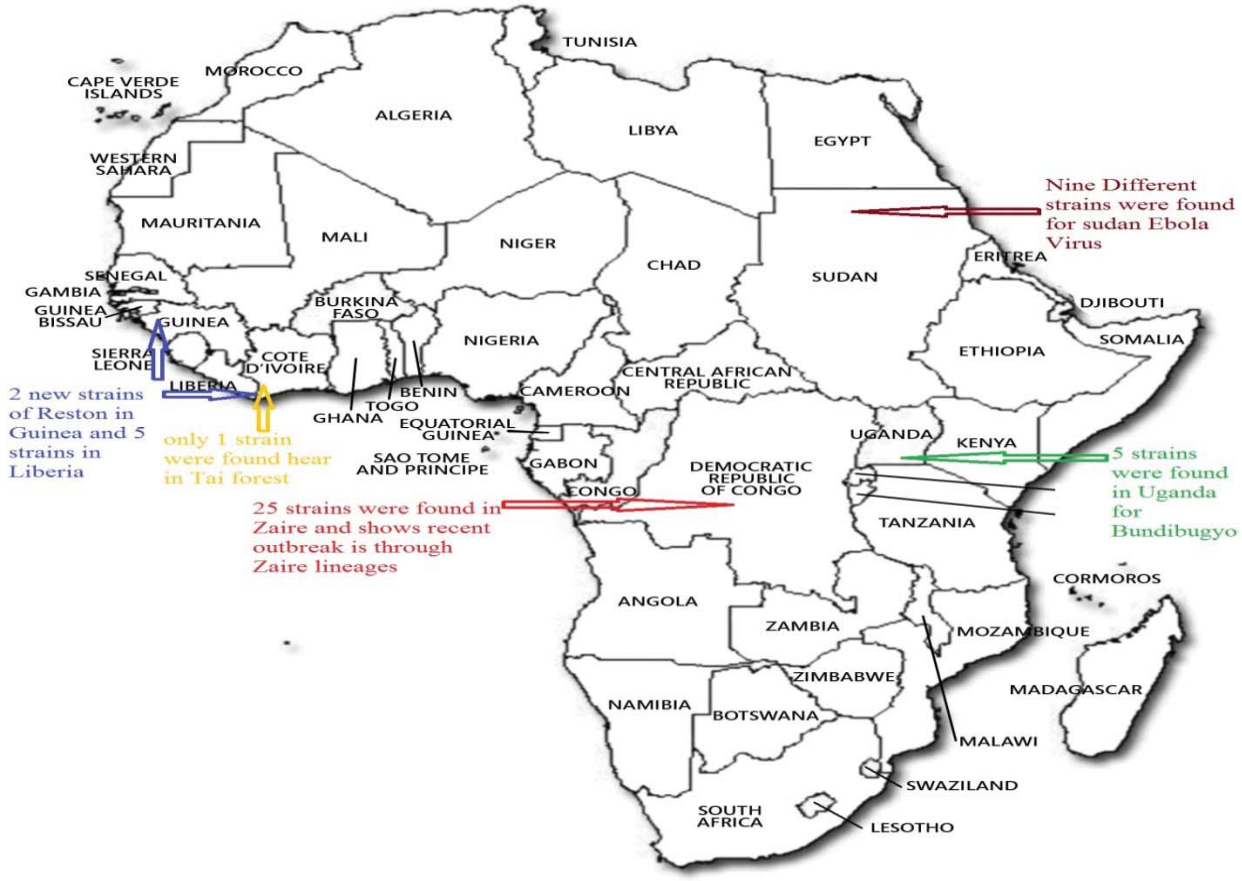

Figure 5. Map of Africa showing the regions of different Ebov strains

From the figure 5 which illustrates the regions for the occurrence of Ebov. The two new strains of Reston which were found guinea were the one closely related with the Zaire specie we figured this out from the final tree. The map shows that the new strains found in Guinea are not any new specie but they have close relation with the strains of Democratic Republic of Congo. So the epidemic outbreak that was seen in Guinea was due the Zaire strains only which evolved during the time. The topological distributions of the strains are seen on the Map of Africa.

IV.CONCLUSION

DNA sequence data can provide information on when species have diverged thus we inferred Branching order and relative timing of these branching Events. We considered our work under “Molecular -clock”, which in practice means substitution rate is constant along all the lineages and thus we get Ultra metric tree (i.e. a tree with same root to tip path length for all the lineages). We inferred that new sequences that evolved were closely related to Zaire EBOV. Thus topologically they are closely related evolved from Zaire country.

By the Neighbor-Joining method evolutionary history is inferred. And the bootstrap consensus tree by using 500 replicates was taken into account to represent the evolutionary history of the taxa from output displayed tree. The branches corresponding to partitions which reproduced less than 50% bootstrap replicates were collapsed. P-distance method was used to calculate the evolutionary

REFERENCES

[1] Carroll, Serena A., et al. "Molecular evolution of viruses of the family Filoviridae based on 97 whole-genome sequences." *Journal of virology* 87.5 (2013): 2608-2616.

[2] Tamura, Koichiro, et al. "Estimating divergence times in large molecular phylogenies." Proceedings of the National Academy of Sciences 109.47 (2012): 19333-19338.

[3] Baize, Sylvain, et al. "Emergence of Zaire Ebola virus disease in Guinea." New England Journal of Medicine 371.15 (2014): 1418-1425.
[4] Sanchez, Anthony, et al. "Sequence analysis of the Ebola virus genome: organization, genetic elements, and comparison with the genome of Marburg virus." Virus research 29.3 (1993): 215-240.
[5] Wu, Chieh-Hsi, and Alexei J. Drummond. "Joint inference of microsatellite mutation models, population history and genealogies using trans dimensional Markov Chain Monte Carlo." *Genetics* 188.1 (2011): 151-164.
[6] Yang, Ziheng, and Bruce Rannala. "Branch-length prior influences Bayesian posterior probability of phylogeny." Systematic Biology 54.3 (2005): 455-470.
[7] Aris-Brosou, Stéphane, and Ziheng Yang. "Bayesian models of episodic evolution support a late Precambrian explosive diversification of the Metazoa."Molecular Biology and Evolution 20.12 (2003): 1947-1954.
[8] Drummond, Alexei J., and Marc A. Suchard. "Bayesian random local clocks, or one rate to rule them all." BMC biology 8.1 (2010): 114.
[9] Heled, Joseph, and Alexei J. Drummond. "Bayesian inference of population size history from multiple loci." BMC Evolutionary Biology 8.1 (2008): 289.
[10] Drummond, Alexei J., et al. "Bayesian phylogenetics with BEAUti and the BEAST 1.7." Molecular biology and evolution 29.8 (2012): 1969-1973.
[11] Lemmon, Alan R., and Emily C. Moriarty. "The importance of proper model assumption in Bayesian phylogenetics." *Systematic Biology* 53.2 (2004): 265-277.
[12] Tavaré, Simon. "Some probabilistic and statistical problems in the analysis of DNA sequences." Lectures on mathematics in the life sciences 17 (1986): 57-86.
[13] Minin, Vladimir N., Erik W. Bloomquist, and Marc A. Suchard. "Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics." Molecular biology and evolution 25.7 (2008): 1459-1471.