RESEARCH

Introduction to Focus Areas in Bioinformatics Project Week 14

Sina Glöckner*, Christina Kirschbaum, Swetha Rose Maliyakal Sebastian and Gokul Thothathri

*Correspondence:
sina.gloeckner@fu-berlin.de
Institute for Informatics, Freie Universität Berlin, Takustr. 9, Berlin, DE
Full list of author information is available at the end of the article

Abstract

Goal of the project: The goal was to create a phylogenetic tree for the given alignment with a Maximum Likelihood and a Neighbor-Joining approach. The resulting trees are used to compare the methods.

Main results of the project: Both methods were successful and resulted in trees with some similar features. However, outliers, several clades, and the general structure differed.

Personal key learnings:
Sina: handle and analyze phylogenetic trees
Christina: work with RAxML and analyze phylogenetic trees
Gokul & Swetha: understood how to construct a phylogenetic tree using Neighbor-Joining and the Geneious tool

Estimation of the time:
Sina: 5h
Christina: 6h
Gokul & Swetha: 3h

Project evaluation: 2

Number of words: 841

1 Scientific Background

Phylogenies are a fundamental bioinformatic tool. A phylogenetic tree is mainly used to infer evolutionary dynamics and define genetic diversity [1, 2]. Multiple approaches can be utilized to create such a tree.

On one hand, there is the popular Maximum Likelihood method, first introduced by Felsenstein [3]. It attempts to determine the most probable tree for a given alignment. To this end, several trees are generated. Then, the likelihood of each site in the tree is computed and multiplied to get an evaluation of the complete tree. This likelihood is dependent on a distance matrix for the sequence alignment. Different models calculate such a matrix, the most complex and accurate being the GTR substitution model [4].

Another approach to phylogenetic tree construction is Neighbor-Joining [5]. Starting with a starlike tree, operational taxonomic units are searched, that minimizes the total branch length, and with that the evolutionary steps. For this purpose, it is assumed, that distances between sequences are computed using an additive metric.

In the following, we used a sequence alignment of *Escherichia coli* (E. coli) to create phylogenetic trees with both methods.

2 Methods

In the first approach, we executed the Maximum Likelihood method with RAxML [6]. One of the abilities of RAxML is to compute phylogenetic trees based on the Maximum Likelihood inference. To run the Bootstrapping with RAxML, the fasta file was converted into an interleaved file with the R package ape [7]. Further, it was specified to use the GTR nucleotide substitution model with the Γ model of rate heterogeneity, combined with non-parametric bootstrapping. With these settings ten bootstraps were calculated.

Listing 1: Command for Bootstrapping with RAxML

```
raxmlHPC -m GTRGAMMA -p 12345 -# 10 -s mydata.phy -n T13
```

In the second approach, a distance matrix describing the distance between each pair of taxa is used as input for a Neighbor-Joining algorithm. For the construction of the phylogenetic tree using the Neighbor-Joining method, a sequence analysis tool called Geneious was utilized. Geneious Basic has been designed to be an easy-to-use and flexible desktop software application framework for the organization and analysis of biological data, with a focus on molecular sequences and related data types [8].

Both output files are retrieved in newick file format and visualized and highlighted with FigTree [9].

3 Results

After visualizing the results, we received the tree in Figure 1 for the Maximum Likelihood approach with RAxML and the tree in Figure 2 for the Neighbor-Joining approach with Geneious.

Some outstanding features were highlighted in both figures to make the comparison easier. The green clades are outliers in the RAxML tree and the dark blue clades contain outliers in the Geneious tree. The red highlighting shows an identical clade of both trees, where the sequences are very close. Moreover, the light blue and purple highlighting display two big, identical clades.

4 Discussion

Even though the two trees we received were similar, there are also varying clades and structures. In such cases, it is not possible to make a statement which result is better. To be aware of this circumstance, it is important to build trees with different algorithms, like Maximum Likelihood and Neighbor-Joining in our project.

The comparison between the trees of different algorithms holds the advantage that it helps to evaluate them and it gives reliability especially for the clades that are equal with different methods.

Using more than one algorithm is very common, a combination of Maximum Likelihood and Neighbor-Joining can be seen for example in the works of Shan et al. [10], Smith et al. [11], Ni et al. [12], and Aslanyan et al. [1].

This project about phylogeny showed the importance of bioinformatics in evolutionary science. To operate and analyze a big amount of data is a great challenge,

it requires efficient algorithms and is impossible without bioinformatical methods, such as the ones we tested. This introductory example on a small data set is helpful to get a first insight into such a complex matter.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
Gokul Thothathri and Swetha Rose Maliyakal Sebastian searched for the second method, created a tree with Geneious and wrote the Methods for it. Sina Glöckner visualized the phylogenetic trees in FigTree and wrote the Scientific Background and the Results. Christina Kirschbaum worked with RAxML and wrote the Methods part for it as well as the Discussion.

References

1. Aslanyan L, Avagyan H, Karalyan Z. Whole-genome-based phylogeny of African swine fever virus. Veterinary World. 2020 Oct;13(10):2118–2125.
2. Szabo KV, O'Neill CE, Clarke IN. Diversity in Chlamydial plasmids. PloS One. 2020;15(5):e0233298.
3. Felsenstein J. Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution. 1981 Nov;17(6):368–376. Available from: https://doi.org/10.1007/BF01734359.
4. Waddell PJ, Steel MA. General time-reversible distances with unequal rates across sites: mixing gamma and inverse Gaussian distributions with invariant sites. Molecular Phylogenetics and Evolution. 1997 Dec;8(3):398–414.
5. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution. 1987 Jul;4(4):406–425. Available from: https://doi.org/10.1093/oxfordjournals.molbev.a040454.
6. Stamatakis A. RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. Bioinformatics (Oxford, England). 2014 01;30.
7. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019;35:526–528.
8. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: An Integrated and Extendable Desktop Software Platform for the Organization and Analysis of Sequence Data. Bioinformatics (Oxford, England). 2012 04;28:1647–9.
9. FigTree: Graphical Viewer of Phylogenetic Trees;. Available from: http://tree.bio.ed.ac.uk/software/figtree/.
10. Shan Y, Li XQ. Maximum gene-support tree. Evolutionary Bioinformatics Online. 2008 May;4:181–191.
11. Smith GJD, Bahl J, Vijaykrishna D. Genetic analysis. Methods in Molecular Biology (Clifton, NJ). 2012;865:207–227.
12. Ni ZX, Cui JM, Zhang NZ, Fu BQ. Structural and evolutionary divergence of aquaporins in parasites (Review). Molecular Medicine Reports. 2017 Jun;15(6):3943–3948.

Figures

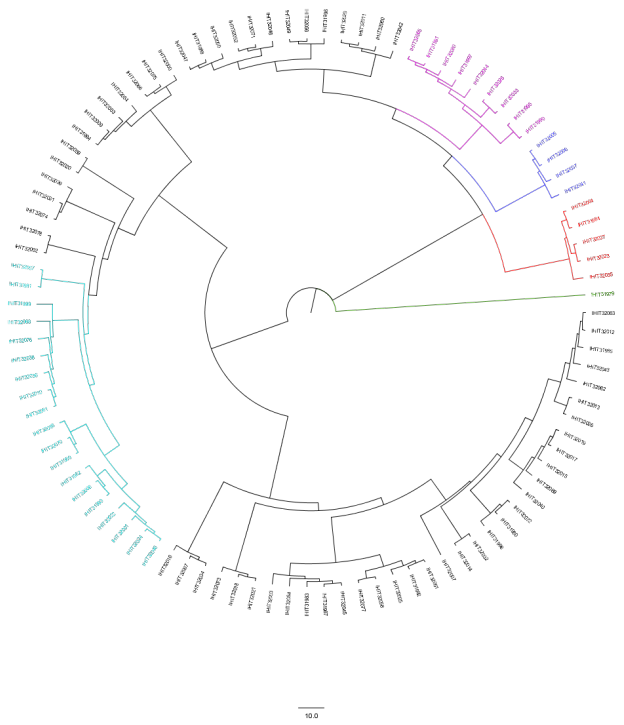

Figure 1: **Phylogenetic Tree with Maximum Likelihood (RAxML)**

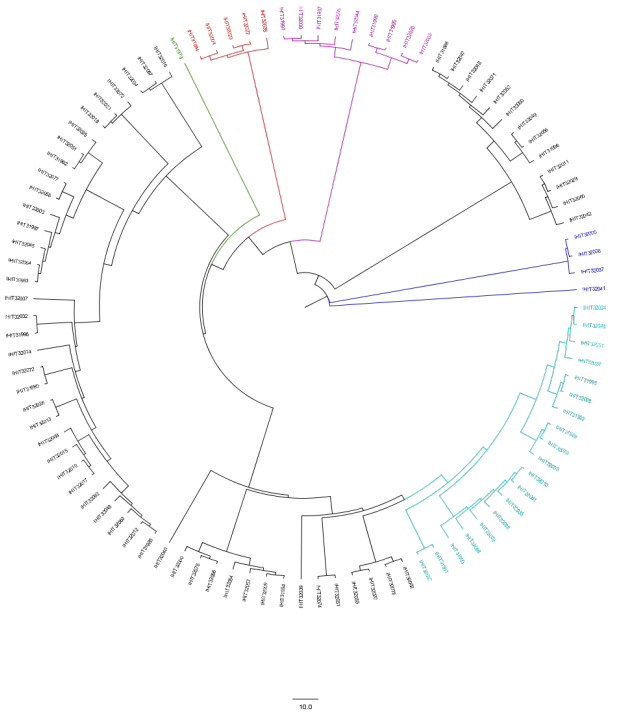

Figure 2: **Phylogenetic Tree with Neighbor Joining (Geneious)**