University of TEHRAN

Algorithms in Bioinformatics

Project 3

Learning phylogenetic trees from multiple sequence alignment of the COVID data

Due date: Tir 20, 1402

Teaching assistant: Sajedeh Bahonar

Instructors: Alireza Fotuhi Siahpirani and Hesam Montazeri

Department of Bioinformatics, IBB, University of Tehran

Objective

The objective of this project is to investigate the evolutionary history of a virus strain by performing a phylogenetic analysis. To accomplish this, you will use the "MAFFT" and "RAxML-NG" tools to estimate the topology and parameters of the phylogenetic tree using the FASTA file data of Omicron, Delta, or Alpha strain genomes.

Data

The dataset is a portion of the GISAID dataset, comprising of SARS-CoV-2 strains collected from humans in various cities across Iran between 2021 and 2023. Each student will choose a subset of the data from the table below and study the evolution of a particular COVID-19 strain in Iran during a designated time period.

Data Nr.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
strain	Omicron	Omicron	Omicron	Omicron	Delta	Delta	Alpha
Time interval	Jan 2022-Apr 2022	Jun 2022-Sep 2022	Aug 2022-Nov 2022	Sep 2022-now	Jan 2021-Nov 2022	Jan 2022-now	Oct 2021-now

Project steps

This project will involve the following steps:

1. **Data preparation**
   I. Download the selected subset of the GISAID dataset.
   II. align the collected dataset using the MAFFT tool and check the MSA for any standard format issues, such as duplicate taxon names, invalid characters in taxon names, or duplicate sequences.
2. **Phylogenetic analysis**
   I. Estimate the optimal nucleotide substitution model using maximum likelihood or Bayesian methods by selecting the model with the lowest AIC or BIC score (you may use phangorn package in R).
   II. Calculate the pairwise distance between the sequences in the alignment.
   III. Generate a first tree topology based on pairwise distances.
   IV. Optimize the phylogenetic tree topology, branch lengths, and nucleotide substitution model.
   V. To assess the reliability of the phylogenetic tree, generate multiple bootstrap replicates of the dataset and estimate the phylogenetic tree for each replicate.
   VI. Infer bootstrap support for branches in the optimized tree.

   Hint: you may use the RAxML-NG tool for steps II - VI.
3. **Data visualization**

   Use the phylogenetic tree and bootstrap support values to visualize the evolutionary history of the selected strain in Iran. (you may use ggtree package in R)

Instructions on how to perform MSA and the initial steps of using RAxML-NG have been given in phyloTutorial.txt as a guide.

Important note: The final deliverable for this project will be a report summarizing the findings from the analysis, including the optimized phylogenetic tree, bootstrap support values for each edge, and any interpretations or analysis of the results. Report all the results in a directory named ***Project3.***