
CS 590 Introduction to Bioinformatics Homework Assignment (9)

Objectives

- Query and read list of sequence data.
- Produce and analyze multiple sequence alignments.
- Construct, plot, and analyze Phylogenetic trees.

Description

As a continuation for our work on the NCBI sequence data for the Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). *Write R code to perform each of the following tasks :*

1. **(10 points)** Retrieve/download the following DNA sequences from GenBank:
SARS coronavirus MA15 ExoN1 (FJ882953)
SARS coronavirus Frankfurt (AB257344)
SARS coronavirus ZJ0301 (DQ182595)
SARS coronavirus strain CV7 (DQ898174)
SARS-CoV-2 (MT334547)
2. **(20 points)** Build and plot an unrooted phylogenetic tree from the sequence data using the neighbor-joining algorithm. *For better display results you may set the type of the phylogenetic tree using type=phylogram*
3. **(10 points)** Which are the most closely related sequences based on the tree?
4. **(10 points)** Based on the bootstrap values in the tree, how confident you are of the groups formed in this tree ?
5. **(30 points)** Build a rooted phylogenetic tree of the SARS virus sequences. Note: the Bovine viral diarrhea virus (GenBank accession JX297515) is related to Corona viruses and so can be used as an outgroup in this case.
6. **(10 points)** Which are the most closely related SARS virus sequences, based on the tree constructed in Q5?
7. **(10 points)** What extra information does the tree in Q5 tell you, compared to the unrooted tree in Q6?