CS 4000

# Homework # 4: Phylogenies, Pandemics, and OpenMPI

due Monday March 29th, 2021 11:59 pm

(70 pts.)

## Introduction

This assignment is inspired by the recent global COVID-19 pandemic caused by the coronavirus SARS-CoV-2. The virus, while deadly, is a relatively simple organism that consists of an RNA sequence that is less than 30,000 characters long. We have already sequenced the genetic data that makes up the virus, along with many, many versions of it. You can get the data for variants from around the world here.

Scientists around the world are studying the variants of SARS-CoV-2 to try to learn

1. how the virus is evolving,
2. how the virus is spreading, and
3. Where and when did the virus originate?

The tools are our disposal to answer these questions come from computer science.The first tool is sequence comparison. While there are many ways to compare two strings, the method for this assignment is to find the *longest common subsequence* between two strings. A subsequence of a string $x_1x_2 \ldots x_n$ is a substring of the first string $x_{l1}x_{l2} \ldots x_{lk}$ such that $l1 < l2 < \ldots l_k$. The longest common substring between two strings $x$ and $y$ is a string $z$ that is a subsequence of both $x$ and $y$ of the longest length. For example, the longest common subsequence between `trexdinosaur` and `regularexpression` is `rexio`. In the case of SARS-CoV-2, we are trying to find the longest RNA sequence that is shared between two variants of the virus.

The second tool that we have at our disposal is tree building. In particular, we will use string comparison to build a *putative* (def: generally considered or reputed to be) genetic tree. In particular, we will build a *phylogenetic tree* by using the longest common sequence algorithm to group related variants of SARS-CoV-2. The root of the tree will be a sequence that is a common subsequence of all of the variants that we will study. The tree and its root might be of interest to scientists studying this virus.

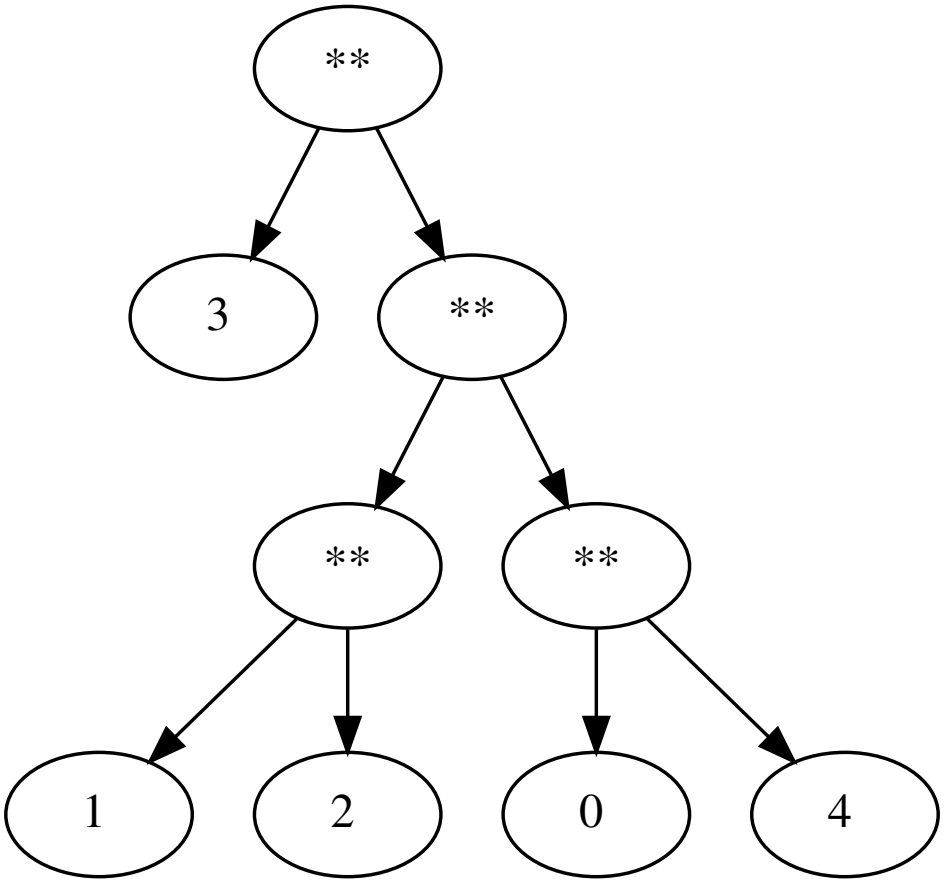
## The Data

I have downloaded complete genomic data for many, many variants of SARS-CoV-2. I have built two test cases consisting of 5 (covid5.dat) and 10 (covid10.dat) variants of SARS-CoV-2. Each line in these two files consist of the RNA sequence of one of the variants.

# Tree Building

An natural way to build phylogentic tree is to proceed as follows. Given a set of $k$ strings, examine each pair $(i, j)$ of these $k$ strings, and find the pair whose common subsequence $A$ is the longest. These two strings will become children of their common ancestor $A$. The strings $i$ and $j$ will be removed from the list of strings and replaced by $A$. This process continues until only one string remains. This string is the root of the phylogenetic tree.

I have created a program `phylogeny_tree.cc` that, when given a file with $k$ strings, creates a phylogenetic tree from those strings. For example, for `covid5.dat`, it might create the tree `(3,((1,2),(0,4)))`.




# Your Assignment

The program `phylogeny_tree.cc` is relatively slow. To build a phylogenetic tree (phylogeny) with 5 strings takes over 2 minutes. Building a phylogenytree for a list of 20 strings might take over 2 hours. Your task is write an OpenMPI version of my program that can solve the problem more quickly using parallel processing. In particular, you should be able to solve a problem with 20 SARS-CoV-2 sequences in around 3 minutes using 40 computers.

a. (20 pts.) I will create a test case with 20 sequences (covid20.dat). Submit the phylogenetic tree for this input and the root sequence.

b. (50 pts.) Submit your OpenMPI code.

## Submission

Submit your parallel solution to this problem via Blackboard. Provide any additional documentation that you find necessary. Submit your code as a single `.cc` file.