# CSCI 4181 / CSCI 6802 Assignment 3

# Phylogenetic Analysis
# (due Wednesday March 25, 2026)

# Overview

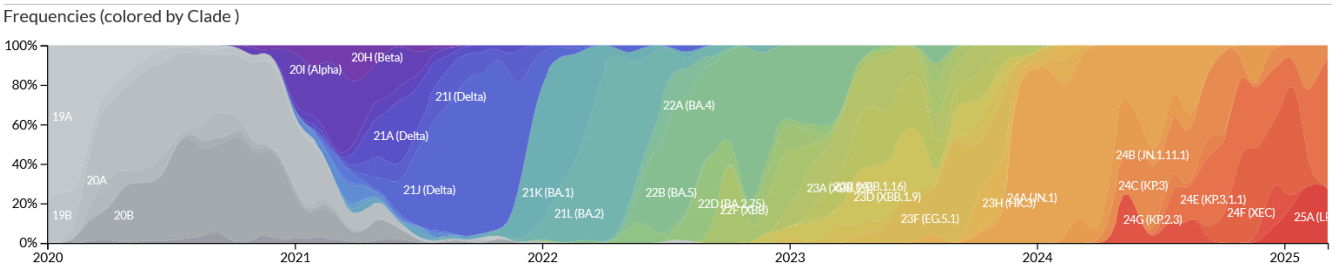
- Please submit your completed assignment to **CSCI 4181/6802 Brightspace: Assessments -> Assignment 3- Phylogenetics.**
- Please submit your assignment as **one file in PDF format**.
- Please name your file following this naming scheme: **BannerID_LastName_Assignment3.pdf**.

Questions that you need to answer are numbered and marked in **boldface**. The point value of each question is indicated next to the question. Be sure to read the questions thoroughly and address each part.

The goal of this assignment is to test your theoretical knowledge of phylogenetics and give you some practical experience with maximum-likelihood methods.

The COVID-19 pandemic began in late 2019, and almost immediately the first genome of the causative virus (SARS-CoV-2) was sequenced and published. The GISAID database (https://www.gisaid.org/) has grown to 219 countries and territories with 16,843,380 viral genome sequences from human cases of COVID-19 since 10 January 2020. The surveillance of the novel coronavirus surpasses anything that has been done in the past, but experience with previous outbreaks including the 2009 Influenza A H1N1 pandemic and the 2014-2015 West African Ebola outbreak was extensive and built a lot of the infrastructure that is now in use for SARS-CoV-2.

Trees have informed our understanding of the geographic spread of the virus and the emergence of key mutations; lineage designations (like B.1.1.7) and Greek variant names (like Alpha) are based on interpretations of phylogenetic trees. In the figure below you can see how new variants have emerged, become relatively common, and then been replaced with newer variants during the height of the pandemic. The changes we have seen in that time are striking: in March 2020 we had a few co-circulating variants (not detailed in the figure), but these were wiped out by Alpha & Delta and then much more dramatically by the Omicron variant.




*Source and Interactive Version: https://nextstrain.org/ncov/open/global/all-time*

We won't be building phylogenetic trees of ~15 million sequences or anything close to that in this assignment, as this is somewhat impractical and requires a customized set of tools to handle datasets of this scale. We will instead use phylogenetic techniques to study a sometimes-alarming evolutionary process that

takes place in many (most) genomes – recombination. One of the key lineages referenced is the "XD" recombinant, which is mostly derived from Delta but with the critical gene encoding the Spike protein acquired from an Omicron lineage. We chose this recombinant because the recombination for XD is between more distantly related lineages so is easier to see.

Although most steps of this analysis will be carried out with Web-based tools, we will perform the phylogenetic-analysis step at the command line using IQ-TREE 2. This will help familiarize you with the command structure and prepare you for the sequence assembly assignment to come later.

## Academic Integrity

All assignments in this course are to be completed **individually**. The following are specifically forbidden and will lead to a submission to the Academic Integrity Officer:

- Submission of any responses (text or otherwise) that were in any way produced or partially produced by generative AI.
- Any text reproduced in part or in whole from online or other sources without attribution. Answers must be in your own words unless otherwise specified.
- Collaboration on any assignment.

I am happy to answer any questions you may have about academic integrity. If you are unsure, please ask.

## Running the assignment

The instructions for IQ-TREE 2 in this tutorial assume that you are running it on the CS "timberlea.cs.dal.ca" server. You **may** run it somewhere else if you prefer; installation instructions can be found here.

## Building the Dataset

Originally conceived as a repository for influenza virus sequences, the GISAID database is now also the primary database of SARS-CoV-2 genome sequences. GISAID is not as open as, say, most of NCBI, since you need to demonstrate your credentials and agree to their terms before receiving an account. But once in, they have various search, analysis, and visualization functions to apply to their massive repository of sequences. We have taken the sequence with ID 'hCoV-19/France/HDF-IPP04947/2022', which is referenced in this discussion (https://github.com/cov-lineages/pango-designation/issues/444).

As a comparison set, we have chosen members of the Delta clade and the two main Omicron sub-clades (BA.1 and BA.2) since they are most likely to represent the parental strains of our recombinant. We selected them by browsing GISAID and choosing representatives from an assortment of countries. To this, we added two additional genomes: the reference Wuhan isolate and a representative Alpha variant. These will serve as outgroups in our analysis.

**Q1 (1 point)** Why is an outgroup necessary for rooting a maximum-likelihood phylogenetic tree?

These sequences are all contained in the file "XD_genome_new.fasta", which is the only starting file you need for the assignment.

**This file can be downloaded from the course website:**

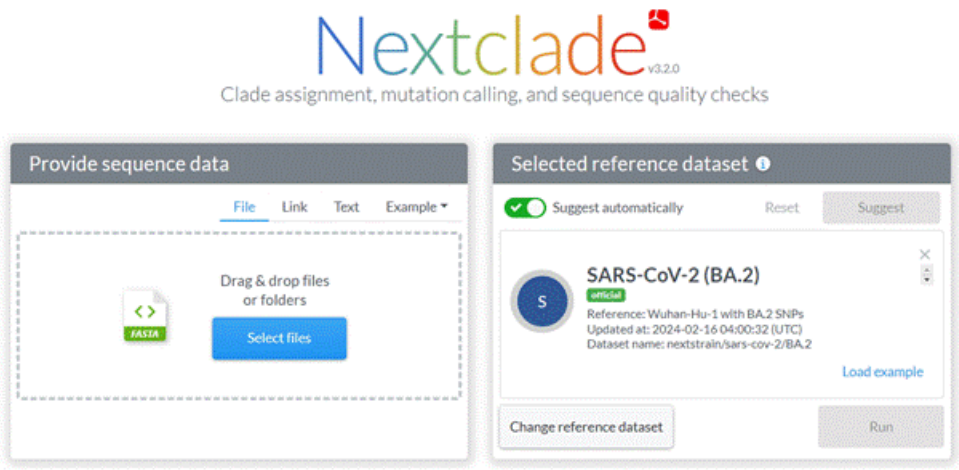
**https://maguire-lab.github.io/bioinformatics_algorithms_2026/static_files/assignments/XD_genome_new.fasta**

# The actual assignment

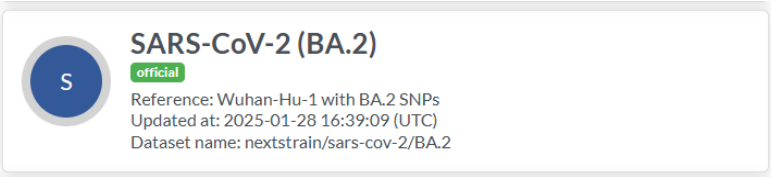
## Step 1: Learning more about your lineages

The viral sequence file is in FASTA format, so each record is represented by a header line (typically the name of that viral isolate) and then the sequence. To build trees we need to first identify where the genes (and thus the corresponding encoded proteins) are in the genome, then do alignments and phylogenies for each set of sequences we're interested in.

Gene predictions and lineage assignments can be done using the Nextclade server, which uses a series of representative genome sequences for each clade (including our friend XD) to do assignments and gene predictions. Navigate to https://clades.nextstrain.org/




If you don't see the above reference dataset in your window, click on “Select Reference Dataset” to choose the below reference:




Drop that genome file into the “Drag & Drop” window and hit “Run”. After a few seconds, you should see a results page that includes the names of the isolates, various quality-control measures, clade and lineage assignments, and a host of other information. Change the coverage (column title ‘Cov.’) to descending order by clicking the top arrow. Most of these values represent differences from the Wuhan reference.

To the right, you will see colourful maps showing mutational differences with respect to the Wuhan isolate –again the Wuhan row is clear because it is identical to itself! But you should see many sites where the other isolates differ from the reference genome. Note that the default displayed sequence is that of the ‘S’ or Spike protein.

> **Q2 (1 point)** Does the pattern of mutations in the S protein suggest a closer relationship between XD and the Delta clades (AY.*; under the ‘Pango lineage (Nextclade)’ column), **or** XD and the Omicron clades (BA.*; under the ‘Pango lineage (Nextclade)’ column)? Give an example of a mutation (position, and specific nucleotide / amino-acid change) that supports your argument.

To carry out the rest of the assignment we will need three files produced by Nextclade: one containing the entire nucleotide alignment of our genomes, and the amino-acid alignments of the ‘S’ and ‘ORF1a’ proteins. The ORF1a gene is furthest from the S gene in the SARS-CoV-2 genome, so it is most likely to show a different phylogenetic pattern if XD is a recombinant.

Go to the **export tab** on the top bar and download two files: 'nextclade.aligned.fasta' and 'nextclade.peptides.fasta.zip'. From the latter zip file, you will want the alignments for the "ORF1a" and "S" proteins which you can navigate to in the "**Genetic Features**" tab in the top right panel.

Note that you can click on the **Tree** tab next to the **export** tab to see a tree that places these sequences in the broader context of the entire outbreak phylogeny. We won't make use of this in the assignment, but it is interesting to look at!

## Step 2: Building phylogenetic trees

Nextclade conveniently provides us with aligned versions of the sequences we want, so we do not need to align these in a separate step. If we needed to do this step ourselves, we could download command-line versions of widely used tools such as MUSCLE or MAFFT, or use the corresponding Web servers (e.g., https://mafft.cbrc.jp/alignment/server/).

**2a. Pre-requisites to IQ-TREE**

However, we will use a command-line version of IQ-TREE with an array of parameters. IQ-TREE is pre-installed on the CS research server 'timberlea.cs.dal.ca'.

Using `scp` in the terminal on your local computer where your files are stored to copy them onto the server.

```
scp PATH/TO/nextclade.aligned.fasta
      ssnmurthy@timberlea.cs.dal.ca:/users/grad/ssnmurthy
```

(replacing with your email address and home directory)

Note that this is a single command, do not hit Enter to separate the two parts.

*Enter your password when prompted and press Enter.*

Once you login to the server [as explained in Appendix 1], you can check the IQ-TREE 2 installation and all parameter options by typing in:

```
iqtree2 -h
```

which will display the IQ-TREE 2 help manual with all its parameter options and defaults.

> **Q3 (2 points)** Does it make more sense to use nucleotide or amino-acid sequence data to build the SARS-CoV-2 tree? Why? [there can be arguments made either way, the key is "Why"]

**2b. Running IQ-TREE**

We need only one command to infer a maximum-likelihood tree from a sequence alignment using IQ-TREE 2. IQ-TREE 2 uses ModelFinder to choose a best-fitting phylogenetic model. We use the Nextclade output which was downloaded previously as input aligned fasta files. The command is as follows:

```
iqtree2 -s nextclade.aligned.fasta -pre XD_genome -m TEST -alrt 1000 -bb
  1000
```

where

`nextclade.aligned.fasta` is the input obtained from Nextclade
`-m` is the automatic model selection parameter
`-alrt` specifies the number of replicates (>=1000) to perform SH-like approximate likelihood ratio test
`-bb` specifies the number of bootstrap replicates (>=1000)

We will repeat this process for the alignments of the "ORF1a" and "S" proteins.

```
iqtree2 -s nextclade.cds_translation.ORF1a.fasta -pre ORF1a -m TEST -alrt
  1000 -bb 1000
```

```
iqtree2 -s nextclade.cds_translation.S.fasta -pre Spike -m TEST -alrt 1000
  -bb 1000
```

We're going to keep the *Sequence type* [`-st`] option set to "Autodetect", as we have nucleotide as well as amino-acid data. IQTREE also gives you an impressive array of substitution models including Jukes-Cantor, Kimura 2-parameter, GTR, etc. We'll leave this at autodetect [`-m TEST`] and see what comes out on the other end.

**Q4 (1 point).** What are the key differences between the Jukes-Cantor and GTR models? Why is J-C inappropriate in many situations?

The last piece of the puzzle is Branch Support Analysis. You will see options for the nonparametric (i.e., standard) and ultrafast bootstrap methods. We'll stick with the default values here - bootstrap values [`-bb`] and SH-aLRT [`-alrt`] tests. The results of these will be reflected in the support values on your tree.

We have intentionally kept the number of sequences small-ish so the runs should not take more than a few minutes. The contents of your directory can be viewed using the **`ls`** command on Linux. Since we gave the option **`-pre`** to specify the prefix of output files, we would see neatly organized output file names for each run based on the prefix set. In the next section of the assignment we will look closely at the ".treefile" from the IQ-TREE output.

**Q5 (1 point)** How does the standard nonparametric bootstrap work? Why does it take so long to run?

## Step 3: Investigating your trees

Now that we have generated the tree using IQTREE, we can look at it and see what sort of inferences we can make. There are many different tree-viewing tools out there, both local and Web-based, and we will use the Interactive Tree of Life (IToL) server at https://itol.embl.de/. Click on "Upload a Tree" under the Annotate header; on the linked page you can paste your tree (that's the ".treefile") from the IQ-TREE 2 output) or upload it. Once you have completed this you will be taken into the tree viewer.
IToL can do all sorts of cool stuff with your tree, like custom node and leaf labelling/colouring, resizing, rerooting, and so on. Here we will keep it simple and use only a few commands:

1. Root the tree by hovering over the Wuhan lineage, clicking to bring up the menu, then selecting "Tree Structure"-> "Re-root the tree here".
2. Under "Advanced", set "Bootstraps/metadata" to "Display". Change the display from "Symbol" to "Text". The default positioning of the bootstraps halfway up each branch is kind of weird, you can change "Position on branch" to a higher value if you like.
3. Apart from that, make whatever adjustments you see fit (larger font, thicker lines, colouring interesting branches) to make the tree as legible as possible.

   **Q6 (1 point)** Paste copies of your trees here with the recombinant genome (and its corresponding proteins) highlighted in all three trees. You can do a screen capture, Export as SVG or other formats, or whatever else you like as long as you get a legible tree.

   **Q7 (3 points)** Looking at the branching order and bootstrap supports in the trees, do you see evidence to support the recombinant origin of the XD genome? Explain why or why not.

That's almost it! The patterns we see in this assignment are consistent with those phylogenies of bacteria and archaea that diverged and shared genes hundreds of millions of years ago. It is striking that the same patterns seen in ancient divergences of heat-and acid-loving extremophiles are plain to see in viral lineages that diverged less than two years ago. This leads us to…

**BONUS Q8** (1 point) Describe the position of the XD recombinant in the *genome* tree. Why do you think it branches where it does?

## Appendix: Using Timberlea

*The Linux Environment*

The server we will use, timberlea.cs.dal.ca, is a Linux server that is accessible to students in Computer Science. You will need to run a few Linux commands in this assignment. A couple of key ones are:

`mkdir` - Make a directory

`cd` - Change to a directory

`ls -1` - List directory contents (that's the letter 'l', not the number '1')

`ls -lrt` - As above, but list in order of time ('t'), reversed ('r'), so the most-recently updated file is at the bottom of the listing. This can be fun if you want to obsessively track the progress of a run.

Very very very useful tip: If you start typing the name of a file or directory, you can autocomplete it by hitting the tab key. If there is some ambiguity, you may need to type a couple more letters. This can save a LOT of time. Also, you can cycle through previous commands using the up arrow.

You will need to connect to **timberlea** using the secure shell 'ssh' protocol. You can do this directly from the command line on a Linux or Mac OS system. On Windows I use the 'putty' (https://www.putty.org/) package; if you're unfamiliar with this let me know, and I can help you out.

If I want to copy the file 'blah.txt' from my home directory on timberlea, I would type from my own computer:

```
scp ssnmurthy@timberlea.cs.dal.ca:blah.txt .
```

That dot at the end means 'copy to my current location' on the current computer.

If you're on a Windows machine, I recommend using WinSCP (https://winscp.net/eng/download.php). Getting set up should be straightforward, and it's a graphical interface - again, let us know if you have trouble.

*Connecting to the Server and Setting Up Your Environment*

We can use the Computer Science research server 'timberlea.cs.dal.ca' to carry out our analyses. To access this server, you need to have a CS ID. If you don't have the login credentials, please let us know, and we will ensure you get set up.

You can connect to timberlea using the 'ssh' command from a Linux or Mac OS prompt; on Windows, I use the 'putty' software. To connect, use the following command:

```
ssh <your CS ID>@timberlea.cs.dal.ca
```