CSCI 4181 / CSCI 6802 Assignment 4

Phylogenetics (March 28, 2022)

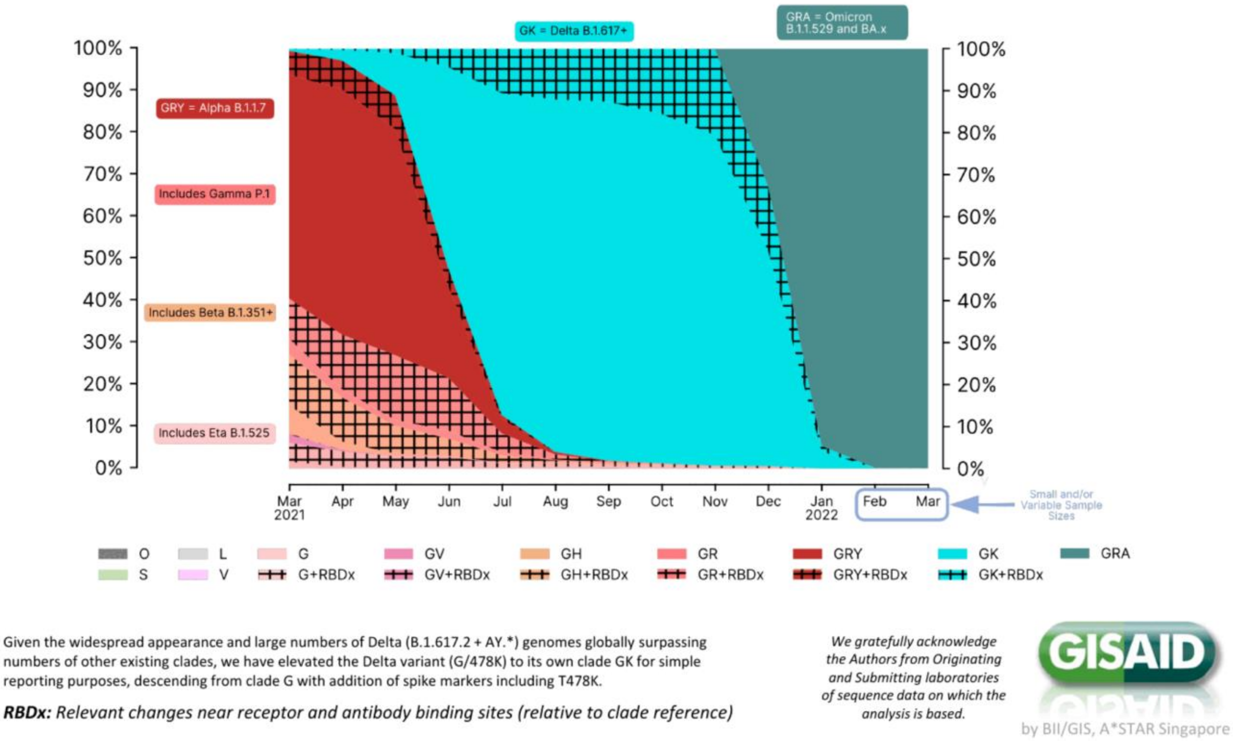
Contents

Overview

This is the last of four assignments in the course that will give you practical experience with biological data. The goal of this tutorial is to introduce you to phylogenetic analysis from retrieval of sequences to alignment, tree construction and finally interpretation. The due date for this assignment is **Monday, April 11, 2022, by midnight**. Please submit to the submission dropbox on CSCI 6802 Brightspace: Assessments -> Assignment 4**.** 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 COVID-19 pandemic began in late 2019, and almost immediately the first genome of the causative virus (SARS-CoV-2) was sequenced and published. In the intervening 27 or so months, the number of sequenced genomes in the GISAID database (https://www.gisaid.org/) has grown to over 9.7 million. 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.

Phylogenetic trees have been assembled from over half of these genomes: a Newick-formatted tree containing 6,295,310 high-quality sequences is available for download in GISAID. Good luck loading that into a tree viewer. 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 between March 2021 and today. The changes we have seen in that time are striking: in March 2021 we had a few co-circulating variants, but these were wiped out by the Delta and then much more dramatically by the Omicron variant.



Given the widespread appearance and large numbers of Delta (B.1.617.2 + AY.*) genomes globally surpassing numbers of other existing clades, we have elevated the Delta variant (G/478K) to its own clade GK for simple reporting purposes, descending from clade G with addition of spike markers including T478K.

***RBDx:** Relevant changes near receptor and antibody binding sites (relative to clade reference)*

We gratefully acknowledge the Authors from Originating and Submitting laboratories of sequence data on which the analysis is based.

Soooooo we're not going to be building phylogenetic trees of 9.7 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.

A nice overview of recombination as it pertains to SARS-CoV-2 is provided in this Twitter thread: https://twitter.com/PeacockFlu/status/1504158873938272269. One of the key lineages referenced is the new "XD" recombinant, which is mostly derived from Delta but with the critical gene encoding the Spike protein acquired from an Omicron lineage. Discussion of the discovery and verification of XD can be seen at https://github.com/cov-lineages/pango-designation/issues/444 if you're interested - note that it is highly technical though with concepts that are way beyond anything I teach in this course.

The assignment will be carried out with Web-based versions of some of the tools I talked about in class. If you wanted to build many trees and/or very large trees you would ideally want to use these tools at the command line, but we're going to keep it simple 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. I took as XD representative one of the IDs (hCoV-19/France/HDF-IPP04947/2022) given in the issue 444 discussion mentioned above. As a comparison set, I chose 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. I was not systematic in choosing these representatives: I selected them by browsing GISAID and choosing recent representatives from an assortment of countries. To this I added two additional genomes: the reference Wuhan isolate and a representative Alpha variant. These will serve as outgroups in our analysis.

> **Q1 (1 point). What is the purpose of using outgroup sequences in a maximum-likelihood phylogenetic analysis?**

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

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. If we want to build trees then we need to do a couple of things: 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/, 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. Notice the run of '0's for the Wuhan sequence: most of these values represent differences from the reference, and Wuhan **is** the 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 with itself! But you should see plenty of 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 clade, or XD and one of the two Omicron clades? 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 download tab (the one with the little down arrow, you may need to scroll to the far right to see it), 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.

Note that you can click on the little tree icon next to the download icon 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 very helpfully gives us 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/).

So the next step is to build phylogenetic trees of our three sets of sequences. Once again we could use a command-line version of RAxML or IQ-TREE (and please feel free to do so if you'd like to!) but we will use the IQTREE server at http://iqtree.cibiv.univie.ac.at/. You will see an option to upload your sequence alignment, and a bewildering array of options. One interesting option is "DNA -> AA", which converts your DNA (er, RNA, because coronavirus) into protein sequences before performing the phylogenetic analysis.

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

So, er, we're going to keep that option set to "Autodetect". IQTREE also gives you an impressive array of substitution models including Jukes-Cantor, Kimura 2-parameter, GTR, etc.

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

We'll leave this at autodetect and see what comes out the other end. The last piece of the puzzle is "Branch Support Analysis". You will see options for the nonparametric and ultrafast bootstrap methods.

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

We'll stick with the default values here - UF bootstrap and SH-aLRT tests. The results of these will be reflected in the support values on your tree.

You can optionally give your email and then non-optionally hit Submit to submit your job (you'll do this for each of your three alignment files separately). I have intentionally kept the number of sequences small-ish so the runs should not take more than a few minutes. You'll find your results by going to the "Analysis results" tab and hitting "Download selected jobs".

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 IQTREE 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 labeling / colouring, resizing, rerooting, and so on. Here we will keep it simple and use only a few commands.

- First, root the tree by hovering over the Wuhan lineage, clicking to bring up the menu, then selecting "Tree Stucture" -> "Re-root the tree here".
- 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 change "Position on branch" to a higher value if you like.
- 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 format, or whatever else you like as long as you get a legible tree.

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

That's almost it! I should mention that the patterns we see in this assignment are consistent with those I have seen when I looked at the phylogenies of bacteria and archaea that diverged and shared genes hundreds of millions of years ago. I find it striking that the same patterns I saw in ancient divergences of heat- and acid-loving extremophiles are plain to see in viral lineages that diverged less than two years ago.

Which leads me 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?