Introduction to phylogenetics: Phylogeny of betacoronaviruses
BIOSCI210, Lab 3
By Nick Matzke

Due dates

This lab can be done on-campus, or purely online (e.g. if we are locked down). All assignments will be submitted through Canvas. For due dates, see Canvas.

Lab 3 Worksheet – multiple choice questions you can work through as you do the lab. These questions correspond to the questions marked with "Q:" in the lab text. I am giving you 3 attempts at the Worksheet, as inevitably someone accidently clicks submit before they have finished. No, you will not be told your score between attempts! (Alternatively, you can just mark down your answers as you go through, and then put them into Canvas at the end.)

Link: see Canvas

Lab files (instructions & data files) can be found here:
http://phylo.wikidot.com/betacoronavirus-lab

Introduction

SARS-CoV-2. You may have heard of it. Where did it come from? The exact answer is not yet known, but phylogenies can help give us an idea.

In this exercise, we are going to do a simple phylogenetic analysis to address this question.

This lab will give you a some practical experience:

0. Setup
1. Getting Genbank data
2. Aligning DNA & viewing alignments
3. Estimating a phylogeny
4. Viewing & interpreting the phylogeny to answer scientific questions.

Setup (assumes laptop; for UoA lab computers, just use e.g. Geneious for alignment viewing, and Geneious or FigTree for phylogeny viewing)

To do this lab, you need:

1. An internet connection (we will use several online web services for sequence alignment and phylogeny estimation).
2. Download and install *AliView* (google it) to view DNA alignments
3. Download and install *FigTree* (google it) to view phylogeny files
4. A plain-text editor (see below)

Windows installs are typically .zip or .exe files, Mac installs are typically .dmg files.

The most common issue may be permissions on newer machines. BEFORE ASKING FOR HELP, GOOGLE ANY ERROR MESSAGES YOU GET, ALONG WITH THE NAME OF THE PROGRAM (these issues are very common and usually have simple solutions for each operating system).

Background

Real-life computational biologists tend to be pluralists -- they will use whatever combination of programs, websites, and programming languages "works" for their current tasks.

Either because of Covid restrictions, or doing a lab remotely due to illness, studying from abroad, etc., some students may not have access to the computer lab. In some ways, this is an advantage! By being forced to install some programs on your own computers (with a variety of different operating systems, etc.), and figuring out how to make them work on your computer, you are doing something much closer to the day-to-day work of a computational biologist.

The main thing you need to make this work is: patience. When you get stuck,

1. First, take a deep breath, and see if you can identify the problem.

2. Second, google the problem (e.g., google the error message). If it's an installation problem, google your question along with the name of your operating system. (E.g., "problem installing AliView on Windows" is better than just "problem installing AliView)

3. If you can't figure out the issue after a few minutes, ASK YOUR TA (in lab) or POST YOUR QUESTION TO PIAZZA (working online)

a. Note: Your questions will get better answers if you include sufficient information in your question. "It's not working!" does not give us any ability to help. We instructors can't see what you are seeing, unless you show us!

b. Therefore, when online, include screenshots of error messages, or (better) copy-paste the commands you used, AND the compete error messages that resulted. This way, other people can search and find the questions/answers for their own questions.

4. If you find a solution, please post that! This will also help others.

Advice on navigating the file system

Every operating system (OS) has a different file viewer. This is the best place to see where your folders and files are, and to arrange files/folders so that you can find them.

On Windows, the file viewer is Windows Explorer, and can usually be opened with WindowsKey+E

On Macs, the file viewer is Finder, and can be opened by clicking the little blue faces icon: https://www.lifewire.com/finders-icon-view-options-2260725

On Linux -- if you have Linux, you probably know all about file viewing already.

Advice on filenames / directory names

Short version: STICK TO PLAIN-TEXT, NO SPACES

While modern operating systems often allow filenames and file paths (directory names) to have spaces in them, as well as apostrophes, inverted commas, regular commas, etc., you should avoid these.

This is because R, and many other pieces of scientific software, react badly to filenames with spaces, special characters, etc. This creates endless difficulties.

Therefore, keep it simple:

BAD PATH AND FILENAME: /Nick's cool documents/Nick's lab.R
GOOD PATH AND FILENAME: /Documents/210/lab3/Nicks_lab3.R

Advice on plain-text editors:

In computational biology, data, inputs, code, etc., are typically contained in PLAIN-TEXT (ASCII) files. These files have *no* fonts, text-sizes, invisible formatting, weird character codes, etc.

Plain-text files are best viewed in an ASCII text viewer, so that you can see the difference between spaces and tabs, you can see blank lines at the end of some files, etc.

Note also that THE DIFFERENCE BETWEEN TABS AND SPACES IS IMPORTANT IN THESE TEXT FILES. COMMON PROBLEMS INCLUDE:

Tabs / spaces at the end of a line, after any text. Delete these in the text-editor. A line in the text file that looks blank, but is actually 8 tabs (or whatever). This can be produced by e.g. copying/pasting a series of distance matrices from Excel to text. When you do this, just check the blank lines for tabs, and delete these tabs in the text-editor so the line is actually blank.

Good plain-text editors: TextWranger (mac), BBedit (mac), Notetab (windows), Rstudio or R.app editors, any code editor

Bad (bad Bad BAD!!) programs for plain-text editing: Word/Office (mac/windows), WordPad (windows), TextEdit (Mac)

Marginal: Notepad (windows)

Part 1. Downloading data

1. Create a working directory on your computer. Your life will be easier if you follow this rule: DON'T USE SPACES IN FILENAMES OR DIRECTORY NAMES

2. Download the "reference sequence" for SARS-CoV-2 from GenBank.

a. Go here: https://www.ncbi.nlm.nih.gov/nuccore/1798174254

b. Click FASTA (just below the sequence name). Wait patiently.

FASTA is a simple, plain-text file format for sequences.

It looks like this: >Sequence ID number, name, descriptive information AGTAGTAGTAGATAGAT AGTTAGTAGCGTGACGA ... etc.

c. When the sequence appears on screen, copy-paste it into a PLAIN-TEXT FILE.

d. Save it (in the working directory!) to something obvious like "sars-cov2_Wuhan_refseq.fasta"

(e. As a backup, I have put a download of this file here: http://phylo.wikidot.com/local--files/betacoronavirus-lab/sars-cov2_Wuhan_refseq.fasta)

f. Sometimes, your browser will force an extra extension onto a filename, e.g., you want "sars-cov2_Wuhan_refseq.fasta", but while downloading, the file is relabelled "sars-cov2_Wuhan_refseq.fasta.txt".

* I like to rename these after download to remove the extra extension (use your file viewer, click on the file, right-click or use the File menu to rename – or use F2 in Windows, “Return” in Macs, to re-name.

* Some operating systems hide file extensions in the file viewer. Google e.g. “Show file extensions in Mac” to figure out how to change this.

* It doesn’t really matter what filenames you use, as long as you can find them on your computer. However, it is easiest to use the same filenames I am using in these lab instructions.

3. Go back to the Genbank record: https://www.ncbi.nlm.nih.gov/nuccore/1798174254

Answer these questions:

Q: What is "NC_045512.2"?

Q: What does the ".2" refer to?

Q: What is the mol_type?

Q: What was the collection date?

Q: What was the former name of this sequence?

4. I have put a bunch of homologous sequences here. Download this file to your working directory:

http://phylo.wikidot.com/local--files/betacoronavirus-lab/betacoronavirus_sequences.fasta

(Sources: the original sources were:

- https://nextstrain.org/groups/blab/sars-like-cov , specifically Download data -> Selected metadata -> nextstrain_groups_blab_sars-like-cov_selected_metadata.tsv -> Accessions (then downloaded from Genbank)
- A few from: https://github.com/blab/sars-like-cov/tree/master/data
- Pangolin from: https://www.ncbi.nlm.nih.gov/nucleotide/MT121216.1?report=genbank&log$=nuclalign&blast_rank=1&RID=PG9G5EMN014
- One could hypothetically get MERS, and other more distantly-related coronavirus genomes, here: https://www.ncbi.nlm.nih.gov/datasets/coronavirus/genomes/ (but, as distant relatives, they are more difficult to align)
)

We want to view these DNA sequences. They are in FASTA format, so should be viewable in any sequence viewer. You can use any program you like, but (especially at home) a good choice is AliView.

AliView is free and open-source, and cross-platform (Windows/Mac/Linux).

Download and install it, using the files and instructions here:
http://ormbunkar.se/aliview/#DOWNLOAD

5. Open AliView, then open betacoronavirus_sequences.fasta in AliView (File->Open)

Answer these questions:

Q: Is this second sequence file aligned?

Q: Would this second sequence file be appropriate for phylogenetic analysis, as-is?

Part 2. Alignment

We are going to align our SARS-related genomes against the SARS-CoV-2 refseq.

2.1. Go here:
https://mafft.cbrc.jp/alignment/server/add_fragments.html?frommanual

- IGNORE the first 2 sections,
    - "For SARS-CoV-2, use this version (2022/May)."
    - Use 1/2 threads only (temporary option, 2021/Jan)"

  …these are addressing specific issues with analyses of only the modern SARS-CoV-2 outbreak (where tens of thousands of genomes are available).

2.2.1. For the first box, use "Choose file", and upload sars-cov2_Wuhan_refseq.fasta

2.2.2. With the second "Choose file", upload betacoronavirus_sequences.fasta

Check "Allow unusual symbols"

Check "Same as input" for UPPERCASE / lowercase

Direction of nucleotide sequences: Same as input

Output order: Aligned

Sequence title: Same as input

Job name: use something obvious like e.g. betacoronavirus_aln_v1

Notify when finished: put your email address

2.3. Then click "submit". And wait a few seconds (click the link they give, if nothing happens initially). This algorithm should do the alignment fairly quickly. It is designed for very closely-related sequences (like the 2020 SARS-CoV-2 outbreak).

What we are doing here is aligning all of the sequences in "betacoronavirus_sequences.fasta" against the "sars-cov2_Wuhan_refseq.fasta" reference sequence. The program is MAFFT, and we are using settings that appropriate for closely-related virus sequences.

Technically we are including somewhat more distantly-related sequences (from SARS-1 etc.), so perhaps in a scientific publication we would use a slower/more thorough alignment algorithm, but this seems to work well on this dataset.

2.4. After a few seconds, click the link to see if the alignment job is finished.

You will see the MAFFT result. This shows the alignment, but we would like to get the alignment in FASTA format.

2.5. Click "Fasta format" at the top of the results page.

2.6. Download the fasta file to your working directory, as e.g. "betacoronavirus_aln_v1.fasta".

2.7. Open this alignment in AliView. Or, if you can't get AliView to work, click "View", then "Start MSAViewer in new window."

Answer these questions:

Q: Is this second sequence file aligned?

Q: Would this second sequence file be appropriate for phylogenetic analysis, as-is?

Part 3. Estimating the tree

3.1. Go back to the MAFFT webpage where you downloaded the aligned FASTA file from. To switch to the phylogeny subpage, go to the top of the webpage and click on "Tree".

Q: How many sequences are in this alignment?

Q: About how many total sites are in this alignment?

Q: What is a "site" in a multiple sequence alignment (MSA)?

Q: How many total nucleotide positions are there in the alignment? (equal #sequences times #sites).

Q: About how many gap-free sites are in this alignment?

Q: About how many conserved sites are in this alignment?

3.2. Let's estimate a phylogeny with neighbour-joining (NJ).

* Click "Conserved sites"
* Substitution model: Jukes-Cantor
* Bootstrap: On
* Number of resampling: 100
* Click: "Go!"

3.3. Watch the bootstrap replicates counting up.

Q: What is bootstrapping, in general statistical usage? (use Google)

Q: Why is bootstrapping a useful technique in general statistics? (use Google)

Q: What is a bootstrap replicate in phylogenetics? (use Google)

Q: Our analysis ran 100 bootstrap replicates. How many times is a phylogeny being estimated in our analysis?

3.4. Once the NJ/bootstrapping analysis completes, download the phylogeny:

* Go to "Tree file without terminal node number", click "Newick?"
* Save to your working directory as e.g. "betacoronavirus_aln_v1.nwk"
* Open the Newick tree file in your plain-text editor.

Q: What is the branch-length below the tip "MG772934_1_Bat_SARS-like_coronavirus_isolate_bat-SL-CoVZXC21__complete_genome"?

Q: What is the branch-length below the tip "MG772933_1_Bat_SARS-like_coronavirus_isolate_bat-SL-CoVZC45__complete_genome"?

Q: What is the branch-length below their common ancestor?

Q: What does branch-length mean for this non-dated, molecular phylogeny? (i.e., what are the units of branchlength here?)

Q: What is the bootstrap support percentage for the clade (MG772934_1_Bat_SARS-like_coronavirus_isolate_bat-SL-CoVZXC21__complete_genome, MG772933_1_Bat_SARS-like_coronavirus_isolate_bat-SL-CoVZC45__complete_genome)?

Q: What would a bootstrap support of 100 mean?

Q: What would a bootstrap support of 70 mean?

Q: What would a bootstrap support of 30 mean?

Part 4. Viewing & Interpreting the Tree

4.1. Open your newick tree file, "betacoronavirus_aln_v1.nwk", in FigTree.

4.2. When FigTree says "The node/branches of the tree are labelled...Please select a name for these values.", replace the word "label" with "bootstrap" or "bs".

4.3. Look carefully at the tree

Q: Is this a dated, or undated, phylogeny?

Q: Does the rooting of the tree make sense?

4.4. Let's re-root the tree with midpoint rooting.
* Click Tree -> Midpoint Root

4.5. Turn on viewing of bootstrap percentages:
* Check-mark on Branch Labels
* Click the triangle to see the options
* Display: bootstrap or bs (whichever you chose)

Q: Which branches tend have the highest bootstrap percentages, and why?

Summary Questions:

A helpful source that includes our sequences plus a few more that were not publicly available initially, is: https://nextstrain.org/groups/blab/sars-like-cov

Summary Question 1: SARS-1 (or SARS-CoV-1) was a dangerous outbreak that occurred in 2003, primarily in China, Hong Kong, Taiwan, and Canada. Over 8,000 cases were identified, with a Case Fatality Rate (CFR) of 11%, according to Wikipedia.

SARS-CoV-2 is a closely related virus. According to our phylogeny, did SARS-CoV-2 evolve from SARS-CoV-1, or not? How does the phylogeny support one position or another?

Summary Question 2: SARS-CoV-2 jumped into humans from some animal source, but there is a debate about which animal host specifically. What does our phylogeny allow us to say about this question?

Summary Question 3: Given unlimited funds, permissions, etc., what research would you recommend in order to answer the question of what animal SARS-CoV-2 came from?

Feel free to use additional sources (online, Wikipedia, etc.) to support your answer. (If you do, include a short references section; the exact format of the reference is not important, but please do include a link.)