Using the Free Program MEGA to Build Phylogenetic Trees from Molecular Data

Molecular Evolutionary Genetics Analysis (MEGA) software is a free package that lets anyone build evolutionary trees in a user-friendly setup. There are several different options to choose from when building trees from molecular data in MEGA, but the most commonly used are Neighbor-Joining and Maximum – Likelihood, both of which give good estimates on the relationship between different molecular sequences. For this exercise, we will use data from GenBank and other sources, insert it into a text editor, import that data into MEGA, and create phylogenetic trees from the collected data.
In order to build evolutionary trees using the MEGA software package (www.megasoftware.net), we will use two of the most commonly used methods to infer evolutionary relationships among species by using their gene sequences. In this paper, we will introduce students to these two methods: Neighbor-Joining (NJ) and Maximum-Likelihood (ML) method.

For this exercise, we will create a phylogenetic tree of the primates, based on the molecular sequences of the surfactant protein A genes.
You will have to pull the **FASTA sequences** for the following viruses (GenBank accession numbers provided).

- *LC521925 (SARS-CoV-2, Japan, January 2020)*
- *MN908947 (SARS-CoV-2, China, December 2019)*
- *LR757996 (SARS-CoV-2, Wuhan, China, 1/1/2020)*
- *MT020880 (SARS-CoV-2, WA, USA, 1/25/2020)*
- *MT066176 (SARS-CoV-2, Belgium, 2/03/2020)*
- *MG772933 (bat-SL-CoVZC45, China, February 2017)*
- *GU553364 (SARS, China, 2003)*
- *KM015348 (MERS, UK, 2013)*
- *KY967354 (Human respirovirus 1, USA, 2015).*

Type each accession number to the search box of the Nucleotide database in NCBI. You may use Boolean operators to obtain multiple results. Hit Search. On the top right, click on the arrow next to the Send to: option. Select File under the *Choose Destination* option. Make sure you select FASTA as the *format*, and click on the Create File button. The Human respirovirus 1 will be used as the outgroup.

Phylogenetic tree construction

1. Open the MEGA-X software.
2. Click on Align then selected Edit/Build Alignment.
3. Create a new alignment. A secondary menu will appear that requires you to select an option. Choose Create a new alignment option and click OK.
4. A second submenu will appear asking you to select the type of sequence data that will be used to build the alignment. Select the DNA option. This will open the MEGA Alignment Explorer in a new window.
5. At the top of the MEGA Alignment Explorer Window select the Edit menu by clicking on it. From this menu, select the Insert Sequence From File option. This will open a new window.

6. Select the .fasta data files saved earlier.
7. Once you have selected all of the files that you wish to upload, select the Open button by clicking on it.
8. Once the sequences are loaded into MEGA, we want to align them. This is done by going Edit button and clicking on Select All.
9. Go to the Alignment menu at the top of the Alignment Explorer window. Click to open the dropdown menu and select Align by ClustalW by clicking on it.
10. This will open another window that is filled with ClustalW parameters. For our purposes, the default settings are adequate. Select the OK option at the bottom of this menu to proceed. This will set in motion the alignment algorithm. Aligning the sequences may take several minutes depending on the size and number of the sequences being examined.
11. Now, we need to save the Alignment Session so that the data is saved in a format that MEGA can use to build the phylogenetic trees. Save the Alignment Session by selecting the Data menu at the top right of the Alignment Explorer window. Click on the Save Alignment option. After this has been completed and the session saved, close the Alignment Explorer window.
12. Now, we need to open the saved alignment session by selecting the File menu in the general MEGA window and clicking on the Open a File/Session option.
13. This will bring up a second window where you have the choice to open the .MAS to analyze it or align it. Select Analyze by clicking on it.
14. To construct a phylogenetic tree, select the Phylogeny menu mid-way through the menu bar at the top of the MEGA window. Here we will continue our example with a Neighbor-Joining tree, but the process is the same for other types of phylogenetic trees. Select Construct/Test Neighbor-Joining Tree.
15. A menu will appear that asks if you want to use the currently active data sheet, select Yes.
16. This will open a dialogue box called Analysis Preferences. For the *Test of Phylogeny* select Bootstrap Method. In the field entitled *No. of Bootstrap Replications*, select 1000 to obtain stable estimates of reliability of the tree. For the *Substitution Type* start by selecting Nucleotide followed by selecting the Jukes-Cantor Model. Here all other fields are left at their default values. To generate the tree, click on OK.
17. After a few minutes, a tree will be generated. The length of time that this takes will depend in part on the length and number of sequences that are being used to create the tree.
18. The numbers on the branches of the tree represents the Bootstrap value, which is the statistical support that each branch receives by the Bootstrap analysis. Higher numbers mean that the branch has higher support and is most likely to be a real branch.
19. To simplify the tree, we now want to condense or cut out the branches that have less support and are less likely to be true branches. To do this, go to the Compute menu on the TreeExplorer menu and select Condensed Tree.
20. This opens a new menu, Tree Options. Select the Cutoff submenu and input 50 for the *Cut-off Value for Condensed Tree*, then click OK. Leave all other values at their defaults.
21. Now, the tree in the TreeExplorer will reflect the changes. All branches that have less than 50% support will have been removed.
22. The last thing that we need to do is to set our outgroup. In our example, it is *the human respirovirus 1.* To do this, right click on the branch that has *human respirovirus*. This brings up a submenu. In this submenu, select the Root option. This provides us with a rooted phylogenetic tree.

23. The tree can be saved as a PDF or printed out. To save the tree as a PDF, go to the Image menu and select Export as PDF. A window will pop up and you can save the file there. To print the tree, there is a Printer Icon that you can use just below the upper menu.
24. Repeat the process by creating a phylogeny with a Maximum-Likelihood Tree.

Additional Background Information:

The NJ and ML methods for building evolutionary trees rely on different statistical principles. In NJ, the least squares method is used along with pairwise evolutionary distances. In ML, the maximum likelihood is optimized such that the inferred tree is the most likely tree. Generally, they will produce very similar results, but NJ is much faster. Despite slight differences in the branching patterns between NJ and ML trees, they both are robust methods for building evolutionary trees. The Jukes-Cantor model is simply a mathematical model that describes the change of one the nucleotides in the DNA sequence to another one, over time. All of its parameters are automatically estimated by MEGA.

Both NJ and ML produce trees that are unrooted, even though they are frequently drawn from left to right. In this case, if one knows the outgroup then it can be used to properly root the tree. Choosing a proper outgroup can be a difficult task and may require some trial and error. A good outgroup should be similar to the sequences in question, but different enough so that the computer program can see the differences.

Homework:

For this exercise, you need to create a phylogenetic tree of the primates, based on the molecular sequences of the surfactant protein A genes.

You will have to pull the **FASTA sequences** for the SP-A genes (even if they are predicted or uncharacterized), of the following taxa:

- *Papio anubis* (olive baboon)
- *Theropithecus gelada* (gelada)
- *Macaca mulatta* (rhesus monkey)
- *Chlorocebus sabaeus* (green monkey)
- *Pan troglodytes* (chimpanzee)
- *Homo sapiens* (human)
- *Microcebus murinus* (gray mouse lemur)
- *Oryctolagus cuniculus* (rabbit, to be used as outgroup).

Prepare the phylogenetic trees made by both the NJ and ML methods. What conclusions can you make with regards to the evolutionary origin of the two SP-A genes?

References:

Newman, L, Duffus, ALJ, Lee, C, 2016, Using the Free Program MEGA to Build Phylogenetic Trees from Molecular Data, American Biology Teacher 78, 7: 608-612.