**Investigating the Transmission of Ebolavirus** **to Humans**

**Software Tools:**

* [MEGA](http://www.megasoftware.net/) – A software package for constructing phylogenetic trees using neighbor-joining, UPGMA, and maximum parsimony.
* [ClustalW](http://www.ebi.ac.uk/Tools/msa/clustalw2/) – A tool for constructing multiple sequence alignment. ClustalW alignment is integrated into MEGA.

**Databases:**

* [GenBank](http://www.ncbi.nlm.nih.gov/genbank/) – A database used for looking up DNA and protein sequences of organisms. GenBank can be queried directly from MEGA.
* [Ebola Genome Browser](http://genome-preview.cse.ucsc.edu/ebolaPortal/) - An information hub about the Ebola virus and interactive viewer of the Ebola genome.
* [NCBI Ebola Virus Variation](http://www.ncbi.nlm.nih.gov/genomes/VirusVariation/Database/nph-select2.cgi?cmd=database&taxid=186536) – A database containing various DNA and protein sequences for the Ebolavirus obtained from different host species, countries, and dates.

**Background:**

The 2014 Ebola outbreak was the deadliest Ebola epidemic in history, infecting over 26,000 people in West Africa and takin over 10,000 lives.

Fortunately, Ebola is not an airborne disease; it can only be spread through direct contact with body fluids of an infected individual. An infected individual becomes contagious after they begin to show symptoms of the disease, which can occur 2-21 days after exposure. Scientists are also investigating the possibility that the virus may be transmitted sexually in the semen of Ebola survivors, where the virus has been found 89 days after symptom onset, long after the virus can no longer be detected in the bloodstream.

Months of investigation resulted in the identification of patient zero for the 2014 outbreak as a 2-year old boy named Emile Ouamouno from Méliandou, Guinea. But how was he infected?

**Objective 1: Constructing a Multiple Alignment of Ebola Genome Sequences**

Smaller Ebola outbreaks have occurred in sub-Saharan Africa at different times since 1976. In fact, there are five different species in the *Ebolavirus* genus: Zaire (EBOV), Sudan (SUDV), Bundibugyo (BDBV), Tai Forest (TAFV), and Reston (RESTV). The first four of these species cause disease in humans. Our first biological objective is to place the strain causing the 2014 outbreak within the *Ebolavirus* phylogeny.

The following map shows the locations for which the Ebolavirus species are named in addition to the origin of the 2014 outbreak. The pins are, in order of appearance from left to right, Guinea (2014), Tai Forest, Zaire, Bundibugyo, and Sudan.

A map of africa with red points

AI-generated content may be incorrect.

**Based on the locations on the map above, to which species of Ebolavirus would you expect the 2014 outbreak in Guinea to be most closely related? Why?**

Since patient zero was found to be from Méliandou, Guinea, we would expect the origin species of Ebolavirus to be most closely related to the species from the closest geographical location, in this case Tai Forest.

In the main text, we began with the goal of constructing an evolutionary tree for a distance matrix holding the “distances” between every two pairs of present-day species under consideration. We saw that given a multiple alignment, we can construct a distance matrix for which the distance between two species is the number of differing symbols between their rows of the alignment.

Fortunately, MEGA includes the multiple alignment program ClustalW (which you may have encountered in the Application Challenge for [Comparing Genes, Genomes, and Proteins](http://coursera.org/course/comparinggenomes)). This allows us to create phylogenetic trees directly from a multiple alignment, removing the intermediate step of computing the distance matrix. Let’s see how MEGA does this using ten different Ebola genomes isolated from humans in different Ebola outbreaks.

First, download and install MEGA v. 6.0 from the [MEGA website](http://www.megasoftware.net/). Then, open the program, click the “Align” icon near the top of the application, choose “Edit/Build Alignment”, and create a new alignment. If asked, indicate that you are building a DNA alignment. At this point, an alignment explorer should open and you can start selecting sequences to align.

We are going to align the following ten DNA sequences. The length of these sequences fall in the range 18,875 – 18,959 base pairs.

|  |  |  |  |
| --- | --- | --- | --- |
| **Accession Number** | **Virus Species** | **Location** | **Date** |
| KJ660348 | ???? | Gueckedou, Guinea | 2014 |
| FJ217161 | BDBV | Bundibugyo, Uganda | 2007 |
| KC545393 | BDBV | Isiro, DRC | 2012 |
| AF272001 | EBOV | Yambuku, DRC | 1976 |
| KC242792 | EBOV | Mekouka, Gabon | 1994 |
| KC589025 | SUDV | Luwero, Uganda | 2012 |
| FJ968794 | SUDV | Sudan | 1976 |
| FJ217162 | TAFV | Tai Forest, Ivory Coast | 1994 |
| AF522874 | RESTV | Philippines | 1990 |
| FJ621583 | RESTV | Philippines | 2008 |

You can open these sequences in the alignment explorer by clicking on “Web” at the top of the application and selecting “Query GenBank”. Use the Accession Numbers on the table above to look up ten sequences that we are using to build a phylogenetic tree. To search GenBank, simply type an Accession Number into the search box at the top of the page, keep “Nucleotide” in the dropdown box, and click “Search”. You will be taken to the GenBank page for this Accession Number. Next to the Address Bar of the web browser window, you will see a button that says “Add To Alignment” with a red plus sign next to it. Click this button to add the sequence to your dataset.For “Sequence Label”, for the sake of simplicity and consistency for this assignment, delete what is automatically entered and instead enter the Accession Number you are searching for as well as the virus species it derives from.

Repeat for the other nine sequences until all ten sequences are populated. Your Alignment Explorer screen should resemble the following screenshot:

A screenshot of a computer

AI-generated content may be incorrect.

Now click on the “Alignment” menu and select “Align by ClustalW” (when prompted to select all, click "OK"). This will perform a multiple sequence alignment on your selected data. When prompted for parameters, use the default values.

The alignment takes a long time to complete (possibly up to an hour). Thus, while ClustalW is running in the background, we will review how to construct a phylogeny from a distance matrix using neighbor-joining.

**Objective 2: Applying the Neighbor-Joining Algorithm**

The neighbor-joining algorithm is one of the most popular methods for evolutionary tree reconstruction. We will first use the neighbor-joining algorithm to construct a small tree by hand.

Below is a distance matrix *D* containing distances between four different organisms labeled W, X, Y, and Z.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | **W** | **X** | **Y** | **Z** | | **W** | 0 | 11 | 2 | 16 | | **X** | 11 | 0 | 13 | 15 | | **Y** | 2 | 13 | 0 | 9 | | **Z** | 16 | 15 | 9 | 0 | |  |  |  |  |

Use the Neighbor-Joining algorithm to construct a phylogeny for the above distance matrix.

**Construct the neighbor-joining matrix *D\** from the distance matrix *D* given above.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **W** | **X** | **Y** | **Z** |
| **W** | 0 | -46 | -49 | -37 |
| **X** | -46 | 0 | -37 | -49 |
| **Y** | -49 | -37 | 0 | -46 |
| **Z** | -37 | -49 | -46 | 0 |

**Construct a 3x3 distance matrix *D*2 using the neighbor-joining matrix *D\**. “A” refers to joining W and Y as neighbors of the tree. You will need to fill in the two branches that were not joined in addition to all distances in the resulting 3x3 distance matrix.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **A** | X | Y |
| **A** | 0 | 11 | 11.5 |
| X | 11 | 0 | 15 |
| Y | 11.5 | 15 | 0 |

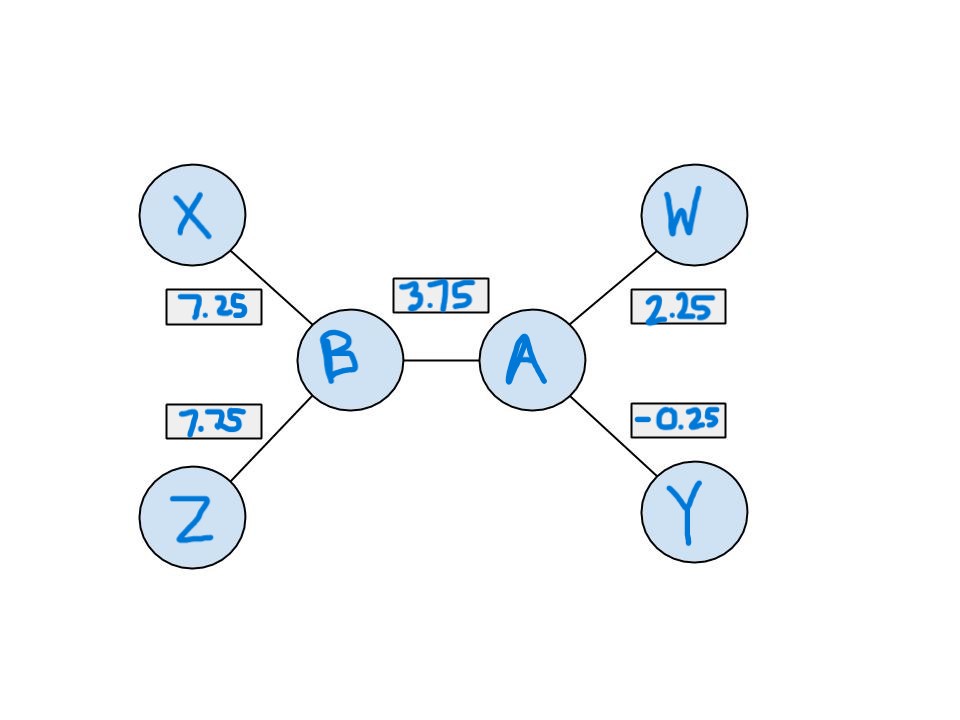
**Construct the neighbor-joining matrix *D*2*\** from the distance matrix *D*2 given above.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | A | X | Y |
| A | 0 | -37.5 | -37.5 |
| X | -37.5 | 0 | -37.5 |
| Y | -37.5 | -37.5 | 0 |

**Construct a 2x2 distance matrix *D*3 using the neighbor-joining matrix *D*2*\**. “B” refers to the new leaf formed by joining A and X in the neighbor-joining algorithm. You will need to fill in the branch that was not joined in addition to the distance in the resulting 2x2 distance matrix.**

|  |  |  |
| --- | --- | --- |
|  | **B** | Z |
| **B** | 0 | 7.75 |
| Z | 7.75 | 0 |

**Fill in the node labels and distances in the tree below. Specifically, fill in the circles with either the strain identifier (W, X, Y, Z), or labels A and B. Fill in the squares with the corresponding branch length. You may use any photo editor you wish (e.g. Microsoft Paint). Attach the filled-in tree to this question.**



**Objective 3: Using MEGA to Construct a Phylogenetic Tree**

Once ClustalW has finished constructing a multiple alignment of *Ebolavirus* sequences, you will need to load the alignment data to construct a tree. Click on the “Data” menu and select “Phylogenetic Analysis”. If the program asks you if the data is protein-coding nucleotide sequence data, select "No" (because we are aligning entire genomes, not just protein-coding regions). The data should now be loaded into the primary MEGA window. In the primary MEGA window, click on the “Phylogeny” icon and select “Construct/Test Neighbor-joining tree”. It should then ask if you want to use the currently active data. Select “Yes”. After clicking "Compute", you should now see an image of the evolutionary tree.

If you do not see branch lengths, you should enable them. To do so, click “View” at the top of the Tree Explorer. Hover over “Show/Hide”, then click on “Branch Lengths”. You should now be able to see all branch lengths of the tree. You can also go to “View” and then “Options” to adjust other properties of the tree. The “Tree” tab is especially of interest because you can adjust “Taxon Separation”, “Branch Length”, and “Tree Width”. (If you are having difficulty reading the branch lengths because they are too close together, you can click the “View” menu and select “Topology only”.) Click "View" --> "Options" --> "Branch" and select 5 decimal places of accuracy.

**Provide an image of the phylogenetic tree that you created.**

A graph of numbers and a number of numbers

AI-generated content may be incorrect.

**Based on the tree produced by Neighbor-Joining, which *Ebolavirus* species caused the 2014 outbreak? Why? Is it the same species that you predicted in question 1?**

The query sequence KJ660348 is most closely related to the EBOV species, which are from DRC and Gabon (Zaire). This was not the geographically closest (Tai Forest) as we predicted above.

**Objective 4: Constructing a phylogenetic tree with UPGMA.**

Previously, you generated a tree using the Neighbor-Joining algorithm. You will now use the alignment data generated in objective 3 to construct a phylogenetic tree using two other algorithms for phylogenetic tree reconstruction: the UPGMA and Maximum Parsimony algorithms.

Click on the “Phylogeny” button and click “Construct/Test UPGMA Tree”. When asked to use the currently active data, click "Yes". Use the default options and click “Compute”.

You should now see an image of the phylogenetic tree. If you do not see branch lengths, you should enable them. Click “View” at the top of the Tree Explorer. Hover over “Show/Hide” then click on “Branch Lengths”. You should now be able to see all branch lengths of the tree. You can also go to “View” and then “Options” to adjust other properties of the tree. The “Tree” tab is especially of interest because you can adjust Taxon Separation, Branch Length, and Tree Width.

**Include an image of the phylogenetic tree constructed by UPGMA (with branch lengths having five decimal places of accuracy).**

A graph of numbers and letters

AI-generated content may be incorrect.

**Where is the root of the tree produced by UPGMA?**

On the left, connecting the SUDV/SUV branch with the rest of the species.

**What is the total distance *D*maxof the tree produced by UPGMA (i.e., the distance from any leaf to the root)? Round your answer to three decimal places.**

The largest distance from leaf to root is 0.593 (SUV FJ968794).

**What is the distance in the UPGMA tree between the leaf corresponding to KJ660348 and the internal node at the beginning of the branch leading to this leaf? Round your answer to three decimal places.**

The distance from KJ660348 to the next internal node is 0.024.

**According to the UPGMA tree, how long did it take for the Ebola virus of 2014 to split from other Ebola viruses in the tree after their most recent common ancestor? How did you obtain this answer? Provide your answer in the distance units given in the tree (i.e. a decimal), not as actual time measurements (e.g. years).**

Subtracting the previous two answers, the amount of time for KJ660348 to split from other Ebola viruses in the tree after their most recent common ancestor is 0.569.

Export the tree (in Tree Explorer, go to File → Export Current Tree (Newick) → choose a name and click OK). In the MEGA home screen, go to “Clocks” → “Compute Timetree (RelTime-ML)” and open the file you just saved. Click “Yes” in the popup about specifying time calibration constraints. Click “New”. For “Taxon A” and “Taxon B”, choose the two BDBV entries. For “Calibration Name”, type “BDBV”. Set both “Min Divergence Time” and “Max Divergence Time” to 10 (we are estimating that the BDBV strains split apart from a common ancestor 10 years ago). Click “Save Changes” and click “Next Step”, then click “Compute”.

**According to the resulting tree, how many years ago did the 2014 Ebola strain (KJ660348) split apart from its closest common ancestor? In other words, what is the number next to the rectangle at the branching point that is KJ660348’s closest common ancestor with its nearest neighbors?**

The number next to the rectangle is 27.33, so it took KJ660348 27.33 years to split from its closest common ancestor.

**Objective 5: Constructing phylogenetic tree with Maximum Parsimony.**

We will now apply MEGA to construct a phylogenetic tree using the Maximum Parsimony algorithm. This time, click on the “Phylogeny” button and click “Construct/Test Maximum Parsimony Tree(s)”. Note that Maximum Parsimony requires multiple sequence alignment data that we have generated (it cannot be applied to a distance matrix). Use the default options and click “Compute”.

You should now see an image of the phylogenetic tree.

**Include an image of the phylogenetic tree that you created. (Note: Your tree does not need to have edge weights.)**

A diagram of a number tree

AI-generated content may be incorrect.

**Is the Maximum Parsimony tree rooted or unrooted?**

Unrooted

Under “Ancestors” (in the menu bar), click “Show All.” Notice that nucleotides appeared at every node of the tree.

**What does the nucleotide at a given node represent?**

The nucleotide at a given split represents what the nucleotide in the ancestral sequence was.

**What information can be inferred from this method that is unavailable in other tree-construction methods?**

From maximum parsimony we can see nucleotide predictions, so we get predictions of the entire genome for each ancestor.

This step requires Microsoft Excel or OpenOffice, so if you have neither of these, please install [OpenOffice](https://www.openoffice.org/) (which is free). Click on the “Ancestors” button in the toolbar, then “Export Changes List” (only shown if “Show All” is selected in the “Ancestors” menu), then click “OK”.

**How many sites have changes between KJ660348’s ancestor and KJ660348? How many entries are in the column that has an arrow pointing to KJ660348? (Note: you won't be able to do this exercise by hand. Consult the documentation for Excel/OpenOffice in order to count the sites. In Excel, the "CountA" function will be helpful.)**

The number of site changes is 10644.

**Conclusion**

We already observed that the virus that caused the epidemics found in West Africa in 2014 should be classified as *Zaire ebolavirus* (EBOV). However, the 2014 outbreak began not in Zaire (present-day Democratic Republic of the Congo), but over a thousand miles away in Guinea! What caused the virus to move so far without infecting a single patient along the way?

Animals are common reservoirs of disease, and viruses often live, multiply, and evolve in animal species before crossing over to humans. In the main text, we saw that the palm civet was the reservoir for SARS. Similarly, rats, chipmunks, and squirrels are reservoirs for bubonic plague; raccoons, skunks, and foxes are reservoirs for rabies; geese and ducks are reservoirs for bird flu; and ticks are the reservoir for Rocky Mountain spotted fever. In contrast, diseases like polio and smallpox have no animal reservoir. The lack of an animal reservoir makes it much easier to completely eradicate the disease, which is why smallpox was completely eradicated and polio has been limited down to just three countries (Nigeria, Pakistan, and Afghanistan).

**Look at the two figures below, which show the ranges of the Angolan free-tailed bat and little collared fruit bat, respectively. What can you conclude about these ranges?**

A map of africa with different countries/regions

AI-generated content may be incorrect.A map of africa with yellow areas

AI-generated content may be incorrect.

Both regions include Guinea and Zaire, so r transmission from Zaire to Guinea (and to the other regions identified) is possible via both the Angolan free-tailed bat the little collared fruit bat.

Emile Ouamouno, patient zero of the 2014 outbreak, frequently played near a hollow tree where bats nested. Researchers have therefore proposed that bats, as is the case for many other epidemics, may have caused the most recent SARS epidemic. In other words, the virus is still evolving within bats, and it crosses the species barrier to humans from time to time, causing an outbreak.

However, this theory is still just a hypothesis, and the fight against Ebola is an ongoing one, as scientists hold out hope for a vaccine that would save thousands of lives in the future. For more information about *Ebolavirus* genomes, you can visit the [Ebola Genome Browser](http://genome-preview.cse.ucsc.edu/ebolaPortal/) or the [NCBI Ebola Virus Variation](http://www.ncbi.nlm.nih.gov/genomes/VirusVariation/Database/nph-select2.cgi?cmd=database&taxid=186536) website. You may also find interesting background information on Ebola at the following links.

<http://www.who.int/mediacentre/factsheets/fs103/en/>

<http://apps.who.int/ebola/>

<http://www.who.int/reproductivehealth/topics/rtis/ebola-virus-semen/en/>

<http://www.npr.org/2014/10/23/358363535/why-do-ebola-mortality-rates-vary-so-widely>

<http://www.sciencealert.com/origin-of-2014-ebola-outbreak-traced-to-kids-favourite-hollowed-tree>