UGWU PASCHAL
+2348166207095 | ugwupaschal@gmail.com
https://www.linkedin.com/in/paschal-ugwu-52abb6229/
Omics Logic Genomics Final Project
Topic: Multiple Sequence Alignment and Phylogeny Analysis of selected strains of Ebola virus.
August, 2020

Disclaimer: All texts were copied directly from the project material available on the Omics Logic website, from which I partook in the Omics Logic Genomics workshop that led to working on the project, "Ebola Virus Epidemic: Deadly Mutations" with other students and researchers across the globe. Even though I worked personally on the topic, "Multiple Sequence Alignment and Phylogeny Analysis of selected strains of Ebola virus," I am not the originator/owner of most of the texts I used in this project.

Abstract

To understand whether genomes are different, we can align them to a reference (pairwise alignment) or compare multiple genomes to each other using multiple sequence alignment (MSA). Multiple Sequence Alignment is based on the assumption that genomes vary because of a natural process of mutation accumulation. The phylogenetic tree specifies a topology, length of branches, root positions, and their shapes. This project aims to determine which geological region's samples have the closest evolutionary relationship with the reference genome. In this project I studied the evolutionary distance between different Ebola virus strains with the help of a sequence repository called NCBI Virus. I search by virus the Ebolavirus, taxid: 186536. I used MEGA 11 software to perform Multiple Sequence Alignment of the selected samples against the consensus reference genome. The resulting aligned sequence was downloaded in MEGA format. The aligned sequences were then used to construct a phylogeny tree. From this project, I understand that samples with accession numbers: KC242795, KC242797, and KC242798 (all from Gabon) have the closest evolutionary relationship with the reference genome.

Background

When we compare multiple genomes from different samples of patients. There are multiple questions that arise in our mind. For example, we have genomes of various samples from different regions: Are these genomes different? How different

are they? Which portions of the genome are more variable? Which portions of the genomes are similar?

To understand whether genomes are different, we can align them to a reference (pairwise alignment) or compare multiple genomes to each other using multiple sequence alignment (MSA). This type of analysis allows us to calculate sequence similarity and find groups of similar genomes or regions in a genome that have more or less variation in a group of similar genomes. Multiple Sequence Alignment can be especially meaningful when we have genomes with good annotation about the origins, date of collection and potentially virulence or other types of phenotypes.

Once we are able to align multiple sequences from different samples using pairwise alignment or multiple sequence alignment. Next, how closely related are these genomes? Did multiple genomes or whole groups of genomes originate from the same "parent" genome? Thus, phylogenetic analysis can be performed to understand the relationship among them or their origin.

Multiple Sequence Alignment is based on the assumption that genomes vary because of a natural process of mutation accumulation. We assume that most mutations accumulate as a result of viruses adapting to an environment that includes host specificity (i.e. transfer between animal and human hosts), treatment, immune response as well as balance between lethal outcome for host and viral replication. This dynamic between virus and host is an evolutionary relationship. In an evolutionary relationship, an observed genomic representation has a link to its "ancestor" and a distance from current other genomes we come across frequently. To study these kinds of relationships, we can use a model of evolutionary processes to establish scale and distance between samples. This is called "phylogenetic analysis", it is typically performed by analyzing the genetic and molecular differences between DNA sequences to determine the evolutionary distance. There are different tree representations in phylogenetic analysis. Evolutionary distance is measured by studying the number of mutations that have occurred per site between homologous sequences, over the years from known samples.

As time passes, all biological species, including viruses and bacteria undergo significant changes. The process of change of a species over time is known as evolution. The evolution process of viruses is unpredictable because they change at irregular, high and varying rates. The viruses can evolve as a result of accumulation of

mutations(antigenic drift) or reunions with other species and generating a new species/strain(antigenic shift). All the changes are first engraved in the genome of the successful strains which later give rise to genetic lineages. In biology, we use the term phylogeny to describe the relationship between genetic lineages characterized by common ancestral origin. To deduce the phylogenetic relationship between different organisms, we build a phylogenetic tree diagram, in which the modern species and their direct and common ancestors occupy the terminal, internal, and the root, respectively.

The phylogenetic tree specifies a topology, length of branches, root positions, and their shapes. Fortunately, a multiplex mathematical apparatus can help us to understand phylogeny. This apparatus helps in assessing inter-species differences, the building of phylogenetic trees, and comparing them with each other. Additionally, it also helps to evaluate the suitability of data and trees through its thorough computational calculations. The rebuilt tree evaluates the real phylogeny that nearly remains unknown. The phylogenetic analysis helps in fundamental and applied virus research which encompasses epidemiology, diagnostics, forensic studies, phylogeography, evolutionary studies, and viral taxonomy.

Aim of Project

To determine which geological region's samples have the closest evolutionary relationship with the reference genome.

Method

In this project I studied the evolutionary distance between different Ebola virus strains with the help of a sequence repository called NCBI Virus. I search by virus the Ebolavirus, taxid: 186536 and it showed me tables of nucleotide sequences, protein sequences etc that are arranged by accession number, release date, geolocation, sequence length, host and collection date. I chose 9 different FASTA sequences on NCBI virus (full genomes, Ebola virus) from 3 different locations and years - 3 samples per site. Found a reference genome as well as references with known dates of collection. Then I downloaded the FASTA files and made sure my genome sequence title was in GeneBank format (GB).

I search on NCBI Virus for the following accession numbers: KC242795, KC242797, KC242798, KP178538, MH425138, KP240932, KP184503, KP120616, KR025228, NC_002549. I downloaded the selected nucleotide sequences in FASTA format and made sure the genome sequence title was in GeneBank format (GB). In

the next step of analysis, I used MEGA 11 software to perform Multiple Sequence Alignment of the selected samples against the consensus reference genome. The resulting aligned sequence was downloaded in MEGA format. The aligned sequences were then used to construct a phylogeny tree.

The evolutionary history was inferred by using the Maximum Likelihood method and Hasegawa-Kishino-Yano model (Hasegawa *et al*., 1985). The bootstrap consensus tree inferred from 1000 replicates (Felsenstein, 1985) is taken to represent the evolutionary history of the taxa analyzed. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test 1000 replicates are shown next to the branches (Felsenstein, 1985). Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. This analysis involved 9 nucleotide sequences. There were a total of 18959 positions in the final dataset. Evolutionary analyses were conducted in MEGA11 (Tamura *et al*., 2021).

Table 1: Reference and sample genomes used for multiple sequence alignment and phylogenetic analysis.

Accession	Isolate	Country	Collection_Date
NC_002549.1	Ebola virus/H.sapiens-tc/COD/1976/Yambuku-Mayinga	Democratic Republic of the Congo	1976
MH425138.1	Ebola virus/H.sapiens-wt/LBR/2014/Makona-LIBR140809	Liberia	8/9/2014
KP240932.1	Ebola virus/H.sapiens-wt/LBR/2014/Makona-201403261	Liberia	9/26/2014
KR025228.1	Ebola virus H.sapiens-wt/GBR/2015/Makona-UK3	United Kingdom	3/12/2015
KP178538.1	Ebolavirus/H.sapiens-wt/LIB/2014/Makona-201403007	Liberia	8/3/2014
KP184503.1	Ebola virus/H.sapiens-	United	8/25/2014

	tc/GBR/2014/Makona-UK1.1	Kingdom	
KP120616.1	Ebola virus/H.sapiens-wt/GBR/2014/Makona-UK1	United Kingdom	8/25/2014
KC242795.1	EBOV/H.sapiens-tc/GAB/1996/1Mbie	Gabon	1996
KC242797.1	EBOV/H.sapiens-tc/GAB/1996/1Oba	Gabon	1996
KC242798.1	EBOV/H.sapiens-tc/GAB/1996/1Ikot	Gabon	1996

Result

Figure 1: The phylogeny tree constructed from the the aligned sample and reference genomes.

From figure 1, samples with accession numbers: KC242795, KC242797, and KC242798 (all from Gabon) have the closest evolutionary relationship with the reference genome (as seen in the bootstrap values of 100 that connect them).

Conclusion

The main question that needs to be addressed during any viral outbreaks is about understanding its origin and identity. The answer of this question acts as a foundation to implement instant practical measures and potential planning. This helps

to specify the virus and detect it quickly so that the epidemic can be contained through the development of drugs and vaccines. This phylogenetic analysis is most helpful among all the techniques to determine the relationship between the existing and previously sequenced viruses. The use of phylogenetic analysis spans over a wide range of studies to help both fundamental and applied virus research, encompassing epidemiology, phylogeography, diagnostics, forensic studies, origin, evolution, and taxonomy of viruses.

References

Hasegawa M., Kishino H., and Yano T. (1985). Dating the human-ape split by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* 22:160-174.

Tamura K., Stecher G., and Kumar S. (2021). MEGA 11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution* https://doi.org/10.1093/molbev/msab120.

Felsenstein J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39:783-791.