**Phylogenetic analysis of Ebola virus secreted glycoprotein**

Mahsa Askary Hemmat

**Introduction:**

Ebola virus is a highly pathogenic virus that belongs to the family Filoviridae. The virus was first discovered in 1976 in Sudan and the Democratic Republic of Congo (formerly Zaire) and is named after the Ebola River, where the first outbreak occurred (Kuhn et al., 2010). The virus causes Ebola virus disease (EVD), which is a severe and often fatal illness in humans and other primates.

The Ebola virus genome is a single-stranded RNA molecule that is approximately 19 kilobases in length. The genome encodes for seven structural proteins, including the nucleoprotein, the glycoprotein, the matrix protein, the viral protein 35, the viral protein 24, the RNA-dependent RNA polymerase, and the viral protein 30.The glycoprotein is of particular interest as it is responsible for the virus's ability to infect host cells.

The Ebola virus glycoprotein is a type I transmembrane protein that is anchored to the viral envelope. The glycoprotein is composed of two subunits, GP1 and GP2, which are derived from the same precursor protein. The glycoprotein mediates viral entry into host cells by binding to specific receptors on the surface of the target cell. The glycoprotein is also responsible for the fusion of the viral and host cell membranes, which allows the virus to enter the host cell and initiate infection (Lee & Saphire, 2009).

Recent studies have shown that the soluble form of the Ebola virus glycoprotein, sGP, may play a role in the pathogenesis of EVD. sGP is a secreted protein that is produced during Ebola virus infection and is thought to interfere with the immune response of the host. sGP may also contribute to the vascular dysfunction that is observed in EVD patients.

Phylogenetic analysis of viruses is an important tool for understanding the evolution and spread of viral diseases. This analysis provides a way to identify and track the emergence and transmission of new viral strains, as well as to investigate the origins and evolutionary history of viruses.

One example is the use of phylogenetic analysis to investigate the origins of the 2014-2016 Ebola virus outbreak in West Africa. Phylogenetic analysis of viral sequences from patients in the outbreak revealed that the virus was most closely related to strains from Central Africa, indicating that it had likely been introduced into West Africa from that region (Baize et al., 2014; Dudas & Rambaut, 2014). This information was crucial for understanding the origins of the outbreak and informing efforts to control its spread.

Bayesian approach is a powerful statistical tool that has been increasingly used in phylogenetic analysis. Bayesian inference allows estimation of posterior probabilities of the different tree topologies, making the method more robust and informative than maximum likelihood (ML) methods. The Bayesian approach in phylogenetic analysis requires a prior probability distribution of the model parameters, which is then updated based on the likelihood of the data. The posterior probability distribution can be obtained using Markov chain Monte Carlo (MCMC) algorithms. The advantage of the Bayesian approach is that it provides a measure of the uncertainty in the estimate, which is reflected in the posterior probability distribution. In addition to the Bayesian approach, different protein substitution models can be used in phylogenetic analysis. These models describe the rates of amino acid substitutions in protein sequences and can affect the accuracy of the phylogenetic tree.

In our lab, we have an aptamer (single stranded DNA molecule) that is selected to bind to Zaire Ebola virus sGP. This aptamer also binds to Sudan Ebola virus sGP. I used Bayesian approach to construct phylogenetic trees for sGP sequences of different strains of Ebola virus to see how closely Zaire strain is related to other strains of the Ebola virus and ask the question if the aptamer is going to bind to other strains of the Ebola virus.

**Materials and Methods:**

Sequences of different strains of Ebola virus sGP were downloaded from the NCBI database. It is difficult to find an outgroup for sGP sequences because they just exist in Ebola virus. I did a Blast search to find a sequence of a protein that doesn’t belong to the Ebola virus genus, but close enough that could be aligned with other sequences. I used the Marburg virus GP sequence as the outgroup. I aligned the sequences using ClustalW and saved the alignment as a Nexus file. For building the phylogenetic trees, I used the MrBayes program. Number of substitution types was set as 6(GTR). I tried different substitution models to see the effect of different protein substitution models on the final phylogenetic tree. Rates variation across sites were set to “Invariable + gamma”. For the Markov Chain Monte Carlo parameters, number of generations was set to 10000 and sampling tree every 10 generation.

**Results:**

A total of 4 trees were generated using 4 different substitution models (Vt, Poisson, Dayhoff and Blossum62 substitution models).

Chart, box and whisker chart

Description automatically generated

Figure 1. Bayesian phylogenetic tree, Vt substitution model

Chart, box and whisker chart

Description automatically generated

Figure 2. Bayesian phylogenetic tree, Poisson substitution model

Chart, box and whisker chart

Description automatically generated

Figure 3. Figure 2. Bayesian phylogenetic tree, Dayhoff substitution model

Chart, box and whisker chart

Description automatically generated

Figure 4. Figure 2. Bayesian phylogenetic tree, Blossum62 substitution model

Discussion:

I have constructed 4 Bayesian phylogenetic trees using 4 different protein substitution models. The topology of trees are different which shows that choosing the right protein substitution model is very important when constructing the phylogenetic trees. From the previous analysis and taxonomic data, we know that all the Ebola virus strains belong to a genus and Marburg virus belongs to different virus genus. I expected that the Ebola virus strains to be in one clade and Marburg virus be the out group. Only in Poisson and Blossum62 substation models the Marburg virus is identifies as the outgroup correctly. Still in these two trees not all the Ebolavirus strains are grouped together. It seems like Zaire, Bombali, Thai, and Bundibugyo strains are grouped together While Sudan and Reston strains are more distant (Figure 2 and 4). This information could be very useful for our research, as we are looking for aptamers to be specific enough to just bind the Ebola virus not other viruses but also can bind to all the Ebola virus strains so it can detect the Ebola virus accurately. Our previous analysis has shown that the aptamer binds to the Sudan and Zaire Ebola virus. This phylogenetic analysis (Fig 2 and 4) shows that Zaire and Sudan are distantly related while the Bombali, Thai, and Bundibugyo strains are grouped together with Zaire. This data may suggest that the aptamer might bind to the sGP of Bombali, Thai, and Bundibugyo strains because they are closely related to the Zaire strain.

References:

Kuhn JH, Becker S, Ebihara H, et al. (2010). Proposal for a revised taxonomy of the family Filoviridae: classification, names of taxa and viruses, and virus abbreviations. Arch Virol, 155(12), 2083-103.

Lee JE, Saphire EO. (2009). Ebolavirus glycoprotein structure and mechanism of entry. Future Virol, 4(6), 621-35.